A new REML (PX)EM algorithm for linear mixed models and factor analytic mixed models

Simon Malcolm Diffey

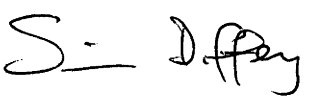
April 2012

A thesis submitted for the degree of Doctor of Philosophy
of the Australian National University

Australian
National
University

Declaration

The work in this thesis is my own except where otherwise stated.



Simon Malcolm Diffey

In memory of Andrew John Diffey (30/7/1976 to 30/12/2010) who died unexpectedly of an asthma attack. A hard worker who would have appreciated the amount of work this thesis represents.

Acknowledgements

This PhD would not have been possible without the financial support of the Grains Research Development Corporation and the New South Wales Department of Primary Industries. I would like to thank Brian Cullis for organising this support and his encouragement and advice throughout.

In his role as supervisor I would like to thank Alan Welsh for his guidance, patience, and seeing it through to the end. I would also like to thank Robin Thompson for his interest and thought provoking comments and Alison Smith for helping to clarify technical issues.

Finally, I would like to thank my family for their love, support, and understanding. In particular I wish to thank my wife Kristy. During the course of this PhD we have had two sons, bought a farm, and lost a brother. She has handled it all with grace, and where appropriate, good humour.

Abstract

Linear mixed models and factor analytic mixed models are routinely applied to biological data arising from designed experiments. The preferred method for estimating the parameters associated with these models is residual maximum likelihood (REML). Most statistical software packages available for the REML estimation of parameters associated with linear mixed models and factor analytic mixed models implement a Newton-Raphson type algorithm such as the expected information algorithm or the average information algorithm. There are two problems with these algorithms. Firstly, successive iterations of these algorithms are not guaranteed to increase the residual log-likelihood function. Secondly, parameter updates may not remain in their parameter space. Either may result in the algorithm failing to converge to a solution.

The REML expectation maximisation (REML EM) algorithm and the parameter expanded version of this algorithm (REML PX EM) are alternatives to Newton-Raphson type algorithms. Features of these two algorithms are that the residual log-likelihood may not decrease with successive iterations and parameter updates remain in their parameter space. Before the REML EM or REML PX EM algorithm can be considered practical alternatives to Newton-Raphson type algorithms two issues need to be addressed. Firstly, they can be notoriously slow to converge, particularly the REML EM algorithm. Secondly, compared to the average information algorithm, current implementations of these two algorithms are computationally more expensive at each iterate. This increased computational expense relates to calculating the trace of a matrix of the same order as the length of the observed data vector.

This thesis addresses the issues of speed and computational efficiency of the REML EM and REML PX EM algorithms for linear mixed models and factor analytic mixed models. The REML EM and REML PX EM algorithms require specification of the incomplete and missing data. A new specification of the incomplete data for linear mixed models and factor analytic mixed models is introduced which is shown be

computationally more efficient and we describe the conditions under which this new specification will have a faster rate of convergence. For factor analytic mixed models, model specification, a new parameter expansion, a new missing data specification, and the efficacy of using a less stringent stopping rule are considered. In an example plant breeding data set and in a simulation study it is shown that these innovations can drastically reduce the number of iterations to convergence.

The improvements to the REML EM and REML PX EM algorithms presented in this thesis make these two algorithms, particularly the latter, more likely to be implemented alongside Newton-Raphson type algorithms in statistical software packages for linear mixed models and factor analytic mixed models. In such a situation this would provide users of these models a viable alternative in the event of a Newton-Raphson type algorithm failing.

Contents

List of main notations

a	scalar
$\boldsymbol{a}$	vector
$\boldsymbol{A}$	matrix
$\boldsymbol{I}_n$	$n \times n$ identity matrix
$\boldsymbol{a}^T$	transpose of the vector $\boldsymbol{a}$
$\boldsymbol{A}^T$	transpose of the matrix $\boldsymbol{A}$
log	natural logarithm
det	determinant
tr	trace
E	expectation
V	variance
Cov	covariance
N	Gaussian distribution
L	likelihood function
ℓ	log-likelihood function
$\frac{\partial}{\partial \boldsymbol{\theta}}$	partial derivative with respect to the parameter vector $\boldsymbol{\theta}$
$\left\langle \cdot \right\rangle_{\boldsymbol{\theta} = \boldsymbol{\theta}_\star}$	evaluate at $\boldsymbol{\theta} = \boldsymbol{\theta}_\star$
$\lvert \cdot \rvert$	absolute value
$\lVert \cdot \rVert$	vector norm
$\lVert \cdot \rVert_F$	Frobenius matrix norm

Chapter 1

EM algorithm introduction

In this chapter we describe the EM algorithm in general terms, i.e., in terms which are not specific to linear mixed models or factor analytic mixed models. In particular we provide a derivation of the EM algorithm, introduce the missing information principle, describe a method for calculating the observed information matrix and for computing the rate of convergence of an EM algorithm. The rate of convergence will be used to compare two different REML EM algorithms for the linear mixed model in a later chapter.

1.1 A brief history of the EM algorithm

The term "EM algorithm" first appeared in the Dempster, Laird, and Rubin paper published in the Royal Statistical Society Series B journal in 1977 (Dempster et al., 1977). However, as noted by Dempster et al. (1977) a number of other authors had contributed to the underlying theory and method prior to their famous paper.

This history of papers with an EM flavour earlier than the Dempster et al. (1977) paper is hardly surprising as the EM algorithm was developed to compute maximum likelihood estimates from incomplete data - a problem of great practical interest. However, this previous work does not detract from the importance of the Dempster et al. (1977) paper. McLachlan and Krishnan (1997) note that it was in this paper that

> ideas were synthesized, a general formulation of the EM algorithm established, its properties investigated, and a host of traditional and nontraditional appli-

cations indicated.

The EM algorithm's popularity can be attributed to the number of problems that fall under the "incomplete data" umbrella. Besides the obvious cases where there are missing values in the data, a large number of problems can be formulated as incomplete data problems even though the incompleteness of the data is not immediately obvious. Linear mixed models, which are the subject of this thesis, are one such example.

1.2 EM algorithm derivation

The EM algorithm is an iterative optimisation method for computing maximum likelihood estimates (and residual or restricted maximum likelihood estimates) in cases where the observed data can be considered incomplete. The EM algorithm derives its name from two steps that are performed at each iteration. These are an expectation step (E-step) which is followed by a maximisation step (M-step).

There are a number of ways in which the EM algorithm can be derived. We present the traditional approach presented in the Dempster et al. (1977) paper. We will consider the observed data vector $\boldsymbol{y}_o = (y_{o_1}, ..., y_{o_n})^T$ as incomplete and a random sample from a distribution having a probability density function that is parameterised by $\boldsymbol{\theta} = (\theta_1, ..., \theta_d)^T$, a vector of unknown parameters with parameter space $\boldsymbol{\Theta}$. It is assumed that there exists a complete data vector $\boldsymbol{y}_c$ which consists of the observed data vector and the missing data vector $\boldsymbol{y}_m$, i.e., $\boldsymbol{y}_c = (\boldsymbol{y}_o^T, \boldsymbol{y}_m^T)^T$.

Define the density function for the complete data vector as

$$f(\boldsymbol{y}_c; \boldsymbol{\theta}) = f(\boldsymbol{y}_o, \boldsymbol{y}_m; \boldsymbol{\theta}).$$

If we let $S(\boldsymbol{y}_m)$ denote the sample space of $\boldsymbol{y}_m$, the density function for the observed data vector can be expressed as

$$g(\boldsymbol{y}_o; \boldsymbol{\theta}) = \int_{S(\boldsymbol{y}_m)} f(\boldsymbol{y}_o, \boldsymbol{y}_m; \boldsymbol{\theta}) d\boldsymbol{y}_m.$$

The density function for the observed data vector can also be defined in terms of the complete data vector and a conditional density function, i.e.,

$$g(\boldsymbol{y}_o;\boldsymbol{\theta}) = \frac{f(\boldsymbol{y}_o,\boldsymbol{y}_m;\boldsymbol{\theta})}{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta})}.$$

Taking natural logarithms we have

$$\log\{g(\boldsymbol{y}_o;\boldsymbol{\theta})\} = \log\{f(\boldsymbol{y}_o,\boldsymbol{y}_m;\boldsymbol{\theta})\} - \log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta})\} \tag{1.1}$$

$$\ell(\boldsymbol{\theta};\boldsymbol{y}_o) = \log\{f(\boldsymbol{y}_c;\boldsymbol{\theta})\} - \log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta})\}, \tag{1.2}$$

where $\ell(\boldsymbol{\theta};\boldsymbol{y}_o) = \log\{g(\boldsymbol{y}_o;\boldsymbol{\theta})\}$ is a log-likelihood function based on the observed data and which will be referred to as the observed data log-likelihood function. Note that the log-density function $\log\{f(\boldsymbol{y}_c;\boldsymbol{\theta})\}$ cannot be referred to as a log-likelihood function as $\boldsymbol{y}_m$ is not observed. Maximum likelihood estimates for $\boldsymbol{\theta}$ are obtained by finding values of $\boldsymbol{\theta}$ that maximize $\ell(\boldsymbol{\theta};\boldsymbol{y}_o)$, i.e.,

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta};\boldsymbol{y}_o).$$

The problem is that $\hat{\boldsymbol{\theta}}$ is often difficult to compute. The EM algorithm approaches this problem by finding the maximum likelihood estimate of $\boldsymbol{\theta}$ indirectly by considering the complete data log-density function, i.e., $\log\{f(\boldsymbol{y}_c;\boldsymbol{\theta})\}$. As this function is unobservable it is replaced with the current conditional expectation of $\log\{f(\boldsymbol{y}_c;\boldsymbol{\theta})\}$ given the observed data $\boldsymbol{y}_o$ and the current estimate of $\boldsymbol{\theta}$.

Let the w-th iterate of $\boldsymbol{\theta}$ be denoted $\boldsymbol{\theta}^{(w)}$. Multiplying both sides of (1.2) by $k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})$ and integrating with respect to $\boldsymbol{y}_m$ we have

$$\ell(\boldsymbol{\theta};\boldsymbol{y}_o) = Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(w)}) - H(\boldsymbol{\theta};\boldsymbol{\theta}^{(w)}), \tag{1.3}$$

where

$$\begin{aligned} Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(w)}) &= \int_{S(\boldsymbol{y}_m)} \log\{f(\boldsymbol{y}_o,\boldsymbol{y}_m;\boldsymbol{\theta})\}k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})d\boldsymbol{y}_m, \\ H(\boldsymbol{\theta},\boldsymbol{\theta}^{(w)}) &= \int_{S(\boldsymbol{y}_m)} \log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta})\}k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})d\boldsymbol{y}_m. \end{aligned}$$

The E and M-steps of the EM algorithm can be written

$$\begin{aligned}
\text{E-step: Calculate } Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(w)}) &= \int_{S(\boldsymbol{y}_m)} \log\{f(\boldsymbol{y}_o,\boldsymbol{y}_m;\boldsymbol{\theta})\}k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})d\boldsymbol{y}_m \\
&= E[\log\{f(\boldsymbol{y}_c;\boldsymbol{\theta})\}|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)}] \\
\text{M-step: } \boldsymbol{\theta}^{(w+1)} &= \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(w)}).
\end{aligned}$$

Although we have ignored the $H(.)$ function we can form a bound on the difference $H(\boldsymbol{\theta}^{(w+1)},\boldsymbol{\theta}^{(w)}) - H(\boldsymbol{\theta}^{(w)},\boldsymbol{\theta}^{(w)})$ through the use of Jensen's inequality. Forming this bound enables us to show that each successive iteration of the EM algorithm does not decrease the observed data log-likelihood function and that iteratively maximising $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(w)})$ will maximise $\ell(\boldsymbol{\theta};\boldsymbol{y}_o)$. This attribute of not decreasing the observed data log-likelihood with successive iterations is known as monotonic convergence. For the EM algorithm

$$\begin{aligned}
H(\boldsymbol{\theta}^{(w+1)},\boldsymbol{\theta}^{(w)}) - H(\boldsymbol{\theta}^{(w)},\boldsymbol{\theta}^{(w)}) &= \int_{S(\boldsymbol{y}_m)} \log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w+1)})\}k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})d\boldsymbol{y}_m \\
&\quad - \int_{S(\boldsymbol{y}_m)} \log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})d\boldsymbol{y}_m \\
&= \int_{S(\boldsymbol{y}_m)} \log\left\{\frac{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w+1)})}{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})}\right\} k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})d\boldsymbol{y}_m \\
&\leq \log\left\{\int_{S(\boldsymbol{y}_m)} \frac{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w+1)})}{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})}.k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})d\boldsymbol{y}_m\right\} \\
&= \log\left\{\int_{S(\boldsymbol{y}_m)} k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w+1)})d\boldsymbol{y}_m\right\} \\
&= 0, \qquad (1.4)
\end{aligned}$$

where the inequality in (1.4) is a result of Jensen's inequality. It follows that

$$-\{H(\boldsymbol{\theta}^{(w+1)};\boldsymbol{\theta}^{(w)}) - H(\boldsymbol{\theta}^{(w)};\boldsymbol{\theta}^{(w)})\} \geq 0. \qquad (1.5)$$

The inequality (1.5) becomes relevant if we consider finding $\boldsymbol{\theta}^{(w+1)}$ such that

$$\begin{aligned}
&\ell(\boldsymbol{\theta}^{(w+1)}; \boldsymbol{y}_o) \geq \ell(\boldsymbol{\theta}^{(w)}; \boldsymbol{y}_o) \\
&Q(\boldsymbol{\theta}^{(w+1)}; \boldsymbol{\theta}^{(w)}) - H(\boldsymbol{\theta}^{(w+1)}; \boldsymbol{\theta}^{(w)}) \geq Q(\boldsymbol{\theta}^{(w)}; \boldsymbol{\theta}^{(w)}) - H(\boldsymbol{\theta}^{(w)}; \boldsymbol{\theta}^{(w)}) \\
&Q(\boldsymbol{\theta}^{(w+1)}; \boldsymbol{\theta}^{(w)}) - Q(\boldsymbol{\theta}^{(w)}; \boldsymbol{\theta}^{(w)}) - \{H(\boldsymbol{\theta}^{(w+1)}; \boldsymbol{\theta}^{(w)}) - H(\boldsymbol{\theta}^{(w)}; \boldsymbol{\theta}^{(w)})\} \geq 0.
\end{aligned} \tag{1.6}$$

The inequality presented in (1.6) is positive for any $\boldsymbol{\theta}^{(w+1)}$ if

$$Q(\boldsymbol{\theta}^{(w+1)}; \boldsymbol{\theta}^{(w)}) - Q(\boldsymbol{\theta}^{(w)}; \boldsymbol{\theta}^{(w)}) \geq 0.$$

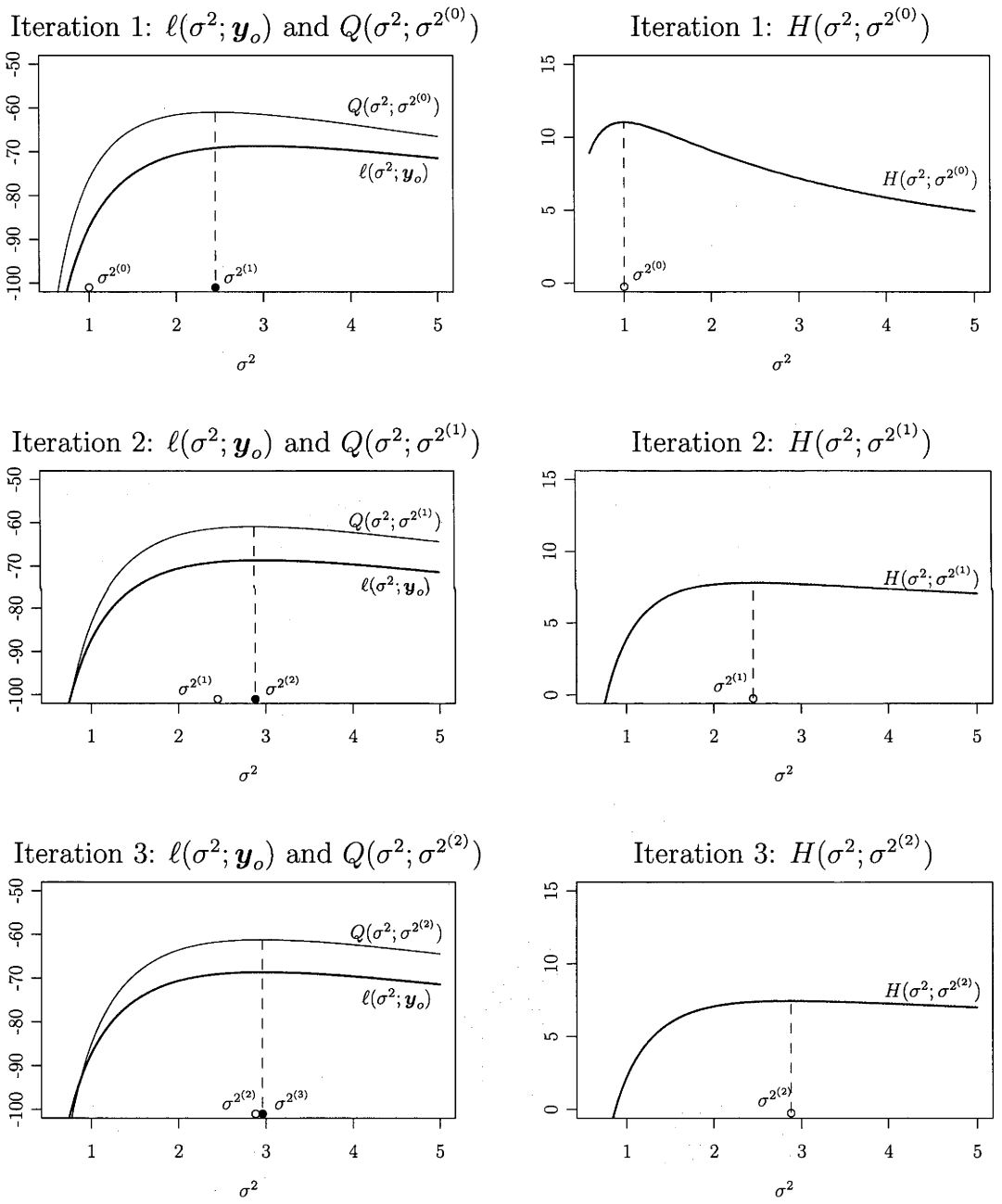
Choosing $\boldsymbol{\theta}^{(w+1)}$ such that $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(w)})$ is maximised for fixed $\boldsymbol{\theta}^{(w)}$ is the M-step of the EM algorithm and it follows that

$$\ell(\boldsymbol{\theta}^{(w+1)}; \boldsymbol{y}_o) \geq \ell(\boldsymbol{\theta}^{(w)}; \boldsymbol{y}_o),$$

establishing the result that each iteration of the EM algorithm increases the observed data log-likelihood and iteratively maximising $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(w)})$ will maximise $\ell(\boldsymbol{\theta}; \boldsymbol{y}_o)$. Jensen's inequality also appears in Wald's proof of the consistency of maximum likelihood estimators (Wald, 1949) and is therefore fundamental to maximum likelihood estimation.

A graphical illustration of iteratively maximising $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(w)})$ is provided in Figure 1.1. The main purpose of Figure 1.1 is to graphically illustrate that $H(\boldsymbol{\theta}; \boldsymbol{\theta}^{(w)})$ is maximised at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(w)}$. This result is fundamental to the EM algorithm having monotonic convergence in $\ell(\boldsymbol{\theta}; \boldsymbol{y}_o)$.

Figure 1.1: Graphical illustration of the EM algorithm being used to compute a maximum likelihood estimate of $\boldsymbol{\theta} = \sigma^2$ by iteratively maximising $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(w)})$ (for 3 iterations) and of $H(\boldsymbol{\theta}; \boldsymbol{\theta}^{(w)})$ being maximised at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(w)}$.



1.3 Score statistic

The gradient vector of the observed data log-likelihood function is

$$\boldsymbol{\mathcal{S}}_o(\boldsymbol{y}_o;\boldsymbol{\theta}) = \frac{\partial \log\{g(\boldsymbol{y}_o;\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}}$$

and the gradient vector of the complete data log-density function is

$$\boldsymbol{\mathcal{S}}_c(\boldsymbol{y}_c;\boldsymbol{\theta}) = \frac{\partial \log\{f(\boldsymbol{y}_o,\boldsymbol{y}_m;\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}}.$$

The observed data score statistic can be expressed as the conditional expectation of the complete data score statistic (Orchard and Woodbury, 1972), i.e.,

$$\boldsymbol{\mathcal{S}}_o(\boldsymbol{y}_o;\boldsymbol{\theta}) = E\{\boldsymbol{\mathcal{S}}_c(\boldsymbol{y}_c;\boldsymbol{\theta})|\boldsymbol{y}_o;\boldsymbol{\theta}\}.$$

This result is developed below, assuming regularity conditions that permit the interchange of differentiation and integration hold. We can write

$$\begin{aligned}
\boldsymbol{\mathcal{S}}_o(\boldsymbol{y}_o;\boldsymbol{\theta}) &= \frac{\partial \log\{g(\boldsymbol{y}_o;\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}} \\
&= \frac{\frac{\partial}{\partial \boldsymbol{\theta}}\{g(\boldsymbol{y}_o;\boldsymbol{\theta})\}}{g(\boldsymbol{y}_o;\boldsymbol{\theta})} \\
&= \frac{1}{g(\boldsymbol{y}_o;\boldsymbol{\theta})}\frac{\partial}{\partial \boldsymbol{\theta}}\left\{\int_{S(\boldsymbol{y}_m)} f(\boldsymbol{y}_o,\boldsymbol{y}_m;\boldsymbol{\theta})d\boldsymbol{y}_m\right\} \\
&= \frac{1}{g(\boldsymbol{y}_o;\boldsymbol{\theta})}\int_{S(\boldsymbol{y}_m)} \frac{\partial}{\partial \boldsymbol{\theta}}\left\{f(\boldsymbol{y}_o,\boldsymbol{y}_m;\boldsymbol{\theta})\right\}d\boldsymbol{y}_m \\
&= \int_{S(\boldsymbol{y}_m)} \frac{\frac{\partial}{\partial \boldsymbol{\theta}}\{f(\boldsymbol{y}_o,\boldsymbol{y}_m;\boldsymbol{\theta})\}}{f(\boldsymbol{y}_o,\boldsymbol{y}_m;\boldsymbol{\theta})}\frac{f(\boldsymbol{y}_o,\boldsymbol{y}_m;\boldsymbol{\theta})}{g(\boldsymbol{y}_o;\boldsymbol{\theta})}d\boldsymbol{y}_m \\
&= \int_{S(\boldsymbol{y}_m)} \frac{\partial}{\partial \boldsymbol{\theta}}\left[\log\{f(\boldsymbol{y}_o,\boldsymbol{y}_m;\boldsymbol{\theta})\}\right]k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta})d\boldsymbol{y}_m \\
&= E\{\boldsymbol{\mathcal{S}}_c(\boldsymbol{y}_c;\boldsymbol{\theta})|\boldsymbol{y}_o;\boldsymbol{\theta}\}.
\end{aligned}$$

We can also express the observed data score statistic as

$$
\begin{aligned}
\mathcal{S}_o(\boldsymbol{y}_o;\boldsymbol{\theta}) &= \left\langle \frac{\partial Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(w)})}{\partial \boldsymbol{\theta}} \right\rangle_{\boldsymbol{\theta}^{(w)}=\boldsymbol{\theta}} \qquad (1.7)\\
&= \left\langle \int_{S(\boldsymbol{y}_m)} \frac{\partial}{\partial \boldsymbol{\theta}} \left[\log\{f(\boldsymbol{y}_o,\boldsymbol{y}_m;\boldsymbol{\theta})\} k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)}) \right] d\boldsymbol{y}_m \right\rangle_{\boldsymbol{\theta}^{(w)}=\boldsymbol{\theta}}\\
&= \left\langle \int_{S(\boldsymbol{y}_m)} \frac{\frac{\partial}{\partial \boldsymbol{\theta}}\{f(\boldsymbol{y}_o,\boldsymbol{y}_m;\boldsymbol{\theta})\}}{f(\boldsymbol{y}_o,\boldsymbol{y}_m;\boldsymbol{\theta})} k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)}) d\boldsymbol{y}_m \right\rangle_{\boldsymbol{\theta}^{(w)}=\boldsymbol{\theta}}\\
&= \left\langle \frac{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})}{f(\boldsymbol{y}_o,\boldsymbol{y}_m;\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \int_{S(\boldsymbol{y}_m)} f(\boldsymbol{y}_o,\boldsymbol{y}_m;\boldsymbol{\theta}) d\boldsymbol{y}_m \right\} \right\rangle_{\boldsymbol{\theta}^{(w)}=\boldsymbol{\theta}}\\
&= \left\langle \frac{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})}{f(\boldsymbol{y}_o,\boldsymbol{y}_m;\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ g(\boldsymbol{y}_o;\boldsymbol{\theta}) d\boldsymbol{y}_m \right\} \right\rangle_{\boldsymbol{\theta}^{(w)}=\boldsymbol{\theta}}\\
&= \frac{\frac{\partial}{\partial \boldsymbol{\theta}}\{g(\boldsymbol{y}_o;\boldsymbol{\theta})\}}{g(\boldsymbol{y}_o;\boldsymbol{\theta})}\\
&= \frac{\partial \log\{g(\boldsymbol{y}_o;\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}},
\end{aligned}
$$

as required. This result is referred to as the self-consistency property of the EM algorithm.

1.4 Missing information principle and observed information matrix

Often the justification for using the EM algorithm instead of another algorithm is that the complete data log-likelihood function is easier to consider than the observed data log-likelihood function. However, an early criticism of the EM algorithm was that it didn't provide an estimate of the observed information matrix and, to estimate this matrix, derivatives of the observed data log-likelihood function had to be considered anyway. Two notable papers that address the issue of finding the observed information matrix when using the EM algorithm are Louis (1982) and Oakes (1999). We will consider the derivation of the observed information matrix presented by Oakes (1999) where the observed information matrix can be estimated using only derivatives of the $Q(.)$ function. Before proceeding with the Oakes (1999) derivation of the observed information matrix it is useful to consider the so-called missing information principle. Beginning with the relation in (1.1) we can write

$$-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big[\log\{g(\boldsymbol{y}_o;\boldsymbol{\theta})\}\Big] = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big[\log\{f(\boldsymbol{y}_o,\boldsymbol{y}_m;\boldsymbol{\theta})\}\Big] + \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big[\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta})\}\Big].$$

Taking expectations on both sides over $k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta})$ we can write

$$E\{\mathrm{I}_o(\boldsymbol{\theta};\boldsymbol{y}_o)|\boldsymbol{y}_o;\boldsymbol{\theta}\} = E\{\mathrm{I}_c(\boldsymbol{\theta};\boldsymbol{y}_c)|\boldsymbol{y}_o;\boldsymbol{\theta}\} - E\{\mathrm{I}_m(\boldsymbol{\theta};\boldsymbol{y}_m)|\boldsymbol{y}_o;\boldsymbol{\theta}\}$$

which can be re-expressed as

$$\mathrm{I}_o(\boldsymbol{\theta};\boldsymbol{y}_o) = \boldsymbol{\mathcal{I}}_c(\boldsymbol{\theta};\boldsymbol{y}_c) - \boldsymbol{\mathcal{I}}_m(\boldsymbol{\theta};\boldsymbol{y}_m),$$

where the symbol $\boldsymbol{\mathcal{I}}$ indicates that it is a conditional expectation of an information matrix that is being considered. The above identity has been interpreted as

observed information = complete information − missing information

which has been called the missing information principle and can be attributed to Orchard and Woodbury (1972). We now present the Oakes (1999) derivation of the observed information matrix using only derivatives of the $Q(.)$ function. In the following development the regularity conditions required for the interchange of differentiation and integration are assumed. Consider the second order partial derivatives,

$$\begin{aligned}\frac{\partial^2\ell(\boldsymbol{\theta};\boldsymbol{y}_o)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} &= \frac{\partial^2 Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(w)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} - \frac{\partial^2 H(\boldsymbol{\theta};\boldsymbol{\theta}^{(w)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} \\ &= \frac{\partial^2 Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(w)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} - E\left[\frac{\partial^2\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta})\}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\middle|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)}\right] \qquad (1.8)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2\ell(\boldsymbol{\theta};\boldsymbol{y}_o)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{(w)^T}} &= \frac{\partial^2 Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(w)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{(w)^T}} - \frac{\partial^2 H(\boldsymbol{\theta};\boldsymbol{\theta}^{(w)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{(w)^T}} \\ \mathbf{0} &= \frac{\partial^2 Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(w)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{(w)^T}} - E\left[\frac{\partial^2\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta})\}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{(w)^T}}\middle|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)}\right]. \qquad (1.9)\end{aligned}$$

The second term on the right hand side of (1.9) can be written

$$
\begin{aligned}
& E\left[\frac{\partial^2 \log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta})\}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{(w)^T}}\middle|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)}\right] \\
&= \frac{\partial}{\partial\boldsymbol{\theta}}\left(\int_{S(\boldsymbol{y}_m)} \frac{\partial}{\partial\boldsymbol{\theta}^{(w)^T}}\left[\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta})\}k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\right]d\boldsymbol{y}_m\right) \\
&= \int_{S(\boldsymbol{y}_m)} \frac{\partial}{\partial\boldsymbol{\theta}}\left[\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta})\}\frac{\partial k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})}{\partial\boldsymbol{\theta}^{(w)^T}}\right]d\boldsymbol{y}_m \\
&= \int_{S(\boldsymbol{y}_m)} \frac{\frac{\partial}{\partial\boldsymbol{\theta}}\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta})\}}{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta})}\frac{\frac{\partial}{\partial\boldsymbol{\theta}^{(w)^T}}\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}}{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})}k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})d\boldsymbol{y}_m \\
&= \int_{S(\boldsymbol{y}_m)} \frac{\partial}{\partial\boldsymbol{\theta}}\left[\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta})\}\right]\frac{\partial}{\partial\boldsymbol{\theta}^{(w)^T}}\left[\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}\right]k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})d\boldsymbol{y}_m \\
&= E\left(\frac{\partial}{\partial\boldsymbol{\theta}}\left[\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta})\}\right]\frac{\partial}{\partial\boldsymbol{\theta}^{(w)^T}}\left[\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}\right]\middle|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)}\right). \qquad (1.10)
\end{aligned}
$$

Substituting (1.10) into (1.9) and adding (1.8) and (1.9) we have

$$
\begin{aligned}
\frac{\partial^2\ell(\boldsymbol{\theta};\boldsymbol{y}_o)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} &= \frac{\partial^2 Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(w)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} + \frac{\partial^2 Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(w)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{(w)^T}} \\
&\quad -E\left[\frac{\partial^2 \log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta})\}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\middle|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)}\right] \\
&\quad -E\left(\frac{\partial}{\partial\boldsymbol{\theta}}\left[\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta})\}\right]\frac{\partial}{\partial\boldsymbol{\theta}^{(w)^T}}\left[\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}\right]\middle|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)}\right). \qquad (1.11)
\end{aligned}
$$

A key result in Oakes (1999) is the identity

$$
\begin{aligned}
& E\left[\frac{\partial^2 \log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}}{\partial\boldsymbol{\theta}^{(w)}\partial\boldsymbol{\theta}^{(w)^T}}\middle|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)}\right] \\
&= -E\left(\frac{\partial}{\partial\boldsymbol{\theta}^{(w)}}\left[\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}\right]\frac{\partial}{\partial\boldsymbol{\theta}^{(w)^T}}\left[\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}\right]\middle|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)}\right) \qquad (1.12)
\end{aligned}
$$

which is derived in Appendix B. Evaluating (1.11) at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(w)}$ and using the identity in (1.12) we have

$$\frac{\partial^2 \ell(\boldsymbol{\theta}^{(w)}; \boldsymbol{y}_o)}{\partial \boldsymbol{\theta}^{(w)} \partial \boldsymbol{\theta}^{(w)^T}} = \left\langle \frac{\partial^2 Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(w)})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} + \frac{\partial^2 Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(w)})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{(w)^T}} \right\rangle_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(w)}} \tag{1.13}$$

which is the result presented by Oakes (1999). It follows that the complete and missing information matrices are given by

$$\begin{aligned} \mathcal{I}_c(\boldsymbol{\theta}^{(w)}; \boldsymbol{y}_c) &= -\left\langle \frac{\partial^2 Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(w)})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right\rangle_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(w)}}, \\ \mathcal{I}_m(\boldsymbol{\theta}^{(w)}; \boldsymbol{y}_m) &= \left\langle \frac{\partial^2 Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(w)})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{(w)^T}} \right\rangle_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(w)}}. \end{aligned}$$

As noted by Oakes (1999) some noteworthy features of (1.13) are that the relation is valid for all $\boldsymbol{\theta}^{(w)}$, the second term on the right hand side can be interpreted as the missing information, and the partitioning of the observed information matrix into two parts may be of some benefit when the inverse of the observed information matrix is required.

1.5 Rate of convergence

In this thesis we will be comparing EM algorithms using different complete data specifications. One measure of algorithm performance is the number of iterations the algorithm takes to converge. A more rigorous approach to measuring algorithm performance is to measure the rate of convergence. For θ_i, $i = 1, \ldots, d$ the rate of convergence can be measured empirically using

$$r_i = \lim_{w \to \infty} \frac{|\theta_i^{(w+1)} - \theta_i^{(w)}|}{|\theta_i^{(w)} - \theta_i^{(w-1)}|}.$$

The global rate of convergence can be measured empirically using

$$r = \lim_{w \to \infty} \frac{\|\boldsymbol{\theta}^{(w+1)} - \boldsymbol{\theta}^{(w)}\|}{\|\boldsymbol{\theta}^{(w)} - \boldsymbol{\theta}^{(w-1)}\|}$$

or by

$$r = \max r_i \quad i = 1, \ldots, d.$$

Empirical rates of convergence are only useful when considering an EM algorithm applied to specific examples. A more general formulation of the global rate of convergence can be derived by considering each iteration of the EM algorithm to involve a mapping $\mathcal{M}$ from a set of values $\boldsymbol{\theta}^{(w)}$ in Θ to another set of values $\boldsymbol{\theta}^{(w+1)}$ also in Θ, i.e.,

$$\boldsymbol{\theta}^{(w+1)} = \mathcal{M}(\boldsymbol{\theta}^{(w)}) \quad w = 0, 1, 2, \ldots \ .$$

Following the development in McLachlan and Krishnan (1997) we can say that if $\boldsymbol{\theta}^{(w)}$ converges to some point $\boldsymbol{\theta}^{\star}$ then $\boldsymbol{\theta}^{\star}$ must satisfy

$$\boldsymbol{\theta}^{\star} = \mathcal{M}(\boldsymbol{\theta}^{\star}).$$

A Taylor series expansion of $\boldsymbol{\theta}^{(w+1)} = \mathcal{M}(\boldsymbol{\theta}^{(w)})$ about $\boldsymbol{\theta}^{\star}$ gives

$$\boldsymbol{\theta}^{(w+1)} - \boldsymbol{\theta}^{\star} \approx \mathcal{J}(\boldsymbol{\theta}^{\star})(\boldsymbol{\theta}^{(w)} - \boldsymbol{\theta}^{\star}),$$

where $\mathcal{J}(\boldsymbol{\theta}^{\star})$ is the $d \times d$ Jacobian matrix for $\mathcal{M}(\boldsymbol{\theta}) = \{\mathcal{M}_1(\boldsymbol{\theta}), \ldots, \mathcal{M}_d(\boldsymbol{\theta})\}^T$ evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^{\star}$, i.e.,

$$\mathcal{J}(\boldsymbol{\theta}^{\star}) = \left\langle \begin{pmatrix} \frac{\partial \mathcal{M}_1(\boldsymbol{\theta})}{\partial \theta_1} & \cdots & \frac{\partial \mathcal{M}_1(\boldsymbol{\theta})}{\partial \theta_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{M}_d(\boldsymbol{\theta})}{\partial \theta_1} & \cdots & \frac{\partial \mathcal{M}_d(\boldsymbol{\theta})}{\partial \theta_d} \end{pmatrix} \right\rangle_{\boldsymbol{\theta} = \boldsymbol{\theta}^{\star}}.$$

The EM algorithm is a first order algorithm having linear convergence with rate matrix $\mathcal{J}(\boldsymbol{\theta}^{\star})$. The global rate of convergence is the largest eigenvalue of $\mathcal{J}(\boldsymbol{\theta}^{\star})$, denoted $\lambda_{\max}$. A small value of $\lambda_{\max}$ implies fast convergence. Dempster et al. (1977) showed that $\mathcal{J}(\boldsymbol{\theta}^{\star})$ can be expressed as

$$\begin{aligned} \mathcal{J}(\boldsymbol{\theta}^{\star}) &= \boldsymbol{I}_d - \boldsymbol{\mathcal{I}}_c^{-1}(\boldsymbol{\theta}^{\star}; \boldsymbol{y}_o)\mathrm{I}_o(\boldsymbol{\theta}^{\star}; \boldsymbol{y}_o) \\ &= \boldsymbol{\mathcal{I}}_c^{-1}(\boldsymbol{\theta}^{\star}; \boldsymbol{y}_o)\boldsymbol{\mathcal{I}}_m(\boldsymbol{\theta}^{\star}; \boldsymbol{y}_o) \end{aligned}$$

i.e., the rate of convergence is determined by the amount of missing information and the greater the proportion of missing information the slower the rate of convergence. Using $\lambda_{\max}$ as a measure of the global rate of convergence allows us to make general

comparisons regarding EM algorithms using different complete data specifications. If we can show that $\lambda_{\max}$ for one specification is always less than that for another specification, then all else being equal we will prefer the specification with smaller $\lambda_{\max}$.

Chapter 2

The linear mixed model as an incomplete data problem and a conditional derivation of REML

Many authors, including Dempster et al. (1977), note that it can be unclear how estimating the parameters of a linear mixed model can be viewed as an incomplete data problem. We present Henderson's mixed model equations as we believe it provides a natural starting point from which to consider an EM algorithm for the linear mixed model. Furthermore, we use the coefficient matrix associated with Henderson's mixed model equations extensively in later chapters. This thesis is concerned with the development of a REML EM algorithm for the linear mixed model. We present the conditional derivation of REML derived by Verbyla (1990). This derivation provides the basis for a new specification of a REML EM algorithm for the linear mixed model considered in a later chapter and presents terminology that will used throughout the rest of this thesis. We begin this chapter by considering a linear mixed model specification.

2.1 Linear mixed model specification

The linear mixed model can be written as

$$\boldsymbol{y}_o = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}, \tag{2.1}$$

where $\boldsymbol{y}_o$ is a $n \times 1$ vector of observed data, $\boldsymbol{\beta}$ is a $t \times 1$ vector of fixed effects, $\boldsymbol{X}$ is a $n \times t$ design matrix of full column rank, $\boldsymbol{u}$ is a $b \times 1$ vector of random effects, $\boldsymbol{Z}$ is a $n \times b$ design matrix, and $\boldsymbol{e}$ is a $n \times 1$ vector of residuals. It is assumed that $\boldsymbol{u}$ and $\boldsymbol{e}$ are realisations of a random variable with the joint distribution

$$\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{e} \end{pmatrix} \sim N\left\{ \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \sigma_H^2 \begin{pmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{pmatrix} \right\},$$

where $\boldsymbol{G} = \boldsymbol{G}(\gamma)$, $\boldsymbol{R} = \sigma^2\Sigma$, and $\Sigma = \Sigma(\phi)$. The vectors γ and ϕ are vectors of variance parameters associated with random effects and residuals respectively. The variance matrices $\boldsymbol{G}$, $\boldsymbol{R}$, and Σ are assumed to be symmetric positive definite. The distribution of the data vector $\boldsymbol{y}_o$ is therefore Gaussian with mean $\boldsymbol{X\beta}$ and variance $\sigma_H^2\boldsymbol{H}$ where $\boldsymbol{H} = \boldsymbol{ZGZ}^T + \boldsymbol{R}$. The variance matrix $\boldsymbol{H}$ is also assumed to be symmetric positive definite.

This specification of the linear mixed model is completely general. However, when σ^2 and σ_H^2 are both included in the model these two parameters are unidentifiable, i.e., they cannot both be estimated. They are both included in the mixed model specification to allow for different parameterisations, namely the parameter vectors γ and ϕ can be considered as either variance components or variance ratios. The problem of identifiability is overcome by setting at least one of them to 1. In this thesis we will consider $\sigma_H^2 = 1$ so that the elements of γ are variance components and when Σ is a variance-covariance matrix the elements of ϕ are ratios relative to σ^2. The joint distribution of $\boldsymbol{y}_o$, $\boldsymbol{u}$, and $\boldsymbol{e}$ is

$$\begin{pmatrix} \boldsymbol{y}_o \\ \boldsymbol{u} \\ \boldsymbol{e} \end{pmatrix} \sim N\left\{ \begin{pmatrix} \boldsymbol{X\beta} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{H} & \boldsymbol{ZG} & \boldsymbol{R} \\ \boldsymbol{GZ}^T & \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{R} & \boldsymbol{0} & \boldsymbol{R} \end{pmatrix} \right\}. \tag{2.2}$$

Fitting the model given in (2.1) involves estimating a vector of variance parameters $\boldsymbol{\kappa} = (\sigma^2, \gamma^T, \phi^T)^T$ and the vector of fixed effects $\boldsymbol{\beta}$.

2.2 Henderson's mixed model equations

Charles Roy Henderson revolutionised the application of linear mixed models in the biological sciences in two important papers in the 1950's. In the first (Henderson, 1953) three methods, known as Method I, II, and III were presented for finding

maximum likelihood estimates of the parameters associated with a linear mixed model. The second (Henderson et al., 1959) presents the now famous mixed model equations. Henderson's mixed model equations are relevant to this thesis in two ways. Firstly, both Henderson's mixed model equations and a ML EM algorithm for linear mixed models are derived by considering the same log-density function. Secondly, the coefficient matrix associated with Henderson's mixed model equations and its inverse are used to compare the computational requirements of alternative REML EM algorithms.

Henderson's mixed model equations are derived by considering the log joint density function of $\boldsymbol{y}_o$ and $\boldsymbol{u}$, i.e.

$$\log\{f(\boldsymbol{y}_o, \boldsymbol{u})\} = \log\{f(\boldsymbol{y}_o|\boldsymbol{u})\} + \log\{f(\boldsymbol{u})\}.$$

Obtaining the conditional distribution $f(\boldsymbol{y}_o|\boldsymbol{u})$ requires a well known result in multivariate normal statistics (Result A.1) and using the joint distribution given in (2.2). The conditional distribution $f(\boldsymbol{y}_o|\boldsymbol{u})$ is

$$\boldsymbol{y}_o|\boldsymbol{u} \sim N\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \boldsymbol{R}\right).$$

We define the vector of unknown scale and location parameters as $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\kappa}^T)^T$ where $\boldsymbol{\kappa} = (\sigma^2, \boldsymbol{\gamma}^T, \boldsymbol{\phi}^T)^T$. The log joint density function of $\boldsymbol{y}_o$ and $\boldsymbol{u}$ is therefore (excluding constants)

$$\begin{aligned}\log\{f(\boldsymbol{y}_o, \boldsymbol{u}; \boldsymbol{\theta})\} &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{R})\} + \log\{\det(\boldsymbol{G})\} \\ &\quad + (\boldsymbol{y}_o - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\dot{\boldsymbol{u}})^T\boldsymbol{R}^{-1}(\boldsymbol{y}_o - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}) + \boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u}\Big].\end{aligned} \tag{2.3}$$

Maximising (2.3) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{u}$ and equating to zero gives

$$\begin{aligned}\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\tilde{\boldsymbol{u}} &= \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{y}_o \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}} - (\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}) &= \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y}_o,\end{aligned}$$

which are known as Henderson's mixed model equations. They can be written in matrix form as

$$\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{G}^{-1}+\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{y}_o \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y}_o \end{pmatrix},$$

where $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{u}}$ is the best linear unbiased predictor (BLUP) of $\boldsymbol{u}$. It is convenient to define the coefficient matrix of Hendersons mixed model equations as

$$\begin{aligned} \boldsymbol{C} &= \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{C}_{XX} & \boldsymbol{C}_{XZ} \\ \boldsymbol{C}_{ZX} & \boldsymbol{C}_{ZZ} \end{pmatrix} \end{aligned}$$

and the inverse of the coefficient matrix as

$$\boldsymbol{C}^{-1} = \begin{pmatrix} \boldsymbol{C}^{XX} & \boldsymbol{C}^{XZ} \\ \boldsymbol{C}^{ZX} & \boldsymbol{C}^{ZZ} \end{pmatrix}.$$

Using Result A.3.6 it can be shown that

$$\begin{aligned} \boldsymbol{C}^{XX} &= (\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}, && \text{using Result A.3.7,} \\ \boldsymbol{C}^{XZ} &= \boldsymbol{C}^{ZX^T} = -(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G}, && \text{using Result A.3.7,} \\ \boldsymbol{C}^{ZZ} &= (\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z}+\boldsymbol{G}^{-1})^{-1}, && \text{using Result A.3.5,} \end{aligned}$$

where $\boldsymbol{S} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{R}^{-1}$. Throughout this thesis we will define the conditional variances required for computation of the E-step in the EM algorithm for linear mixed models in terms related to the coefficient matrix of Henderson's mixed model equations.

2.3 Missing data specification for the linear mixed model

Dempster et al. (1977) in Section 4.4 of their paper provide an example of an EM algorithm for a linear mixed model. They note that

> in the context of linear models...the relevance of incomplete-data concepts is at first sight remote.

However, since Henderson's mixed model equations are based on the joint density function of $\boldsymbol{y}_o$ and $\boldsymbol{u}$, a natural choice for the complete data when formulating an EM algorithm for a linear mixed model is $\boldsymbol{y}_c = (\boldsymbol{y}_o^T, \boldsymbol{u}^T)^T$ where $\boldsymbol{y}_o$ and $\boldsymbol{u}$ are the incomplete and missing data vectors respectively. Using EM algorithm terminology, equation (2.3) is referred to as the complete data log-density function.

It is well known that maximum likelihood estimates of variance components are biased in small samples. Therefore, we do not present an EM algorithm for a linear mixed model in this thesis. However, it is worth noting that considering the vector of random effects $\boldsymbol{u}$ as the missing data is the approach taken later when we consider two REML EM algorithms for the linear mixed model. The second of these is a new complete data specification based on the Verbyla (1990) conditional derivation of REML.

2.4 Conditional derivation of REML

The preferred method for estimating the parameters associated with a linear mixed model is residual or restricted maximum likelihood (REML). The original reference for REML is the paper by Patterson and Thompson (1971), however the new complete data specification we consider later is based on the conditional derivation of REML presented in Verbyla (1990). The conditional derivation of REML begins by considering the transformation

$$\boldsymbol{L}^T\boldsymbol{y}_o = \begin{pmatrix} \boldsymbol{L}_1^T\boldsymbol{y}_o \\ \boldsymbol{L}_2^T\boldsymbol{y}_o \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix},$$

where $\boldsymbol{L} = (\boldsymbol{L}_1, \boldsymbol{L}_2)$ is a non-singular matrix, with $\boldsymbol{L}_1$ and $\boldsymbol{L}_2$ $n \times t$ and $n \times (n-t)$ matrices respectively, both of full-column rank and chosen to satisfy $\boldsymbol{L}_1^T\boldsymbol{X} = \boldsymbol{I}_t$ and $\boldsymbol{L}_2^T\boldsymbol{X} = \boldsymbol{0}$. It is worth noting that the $n-t \times 1$ vector $\boldsymbol{y}_2 = \boldsymbol{L}_2^T\boldsymbol{y}_o$ will play a key role in the development of a new complete data specification for linear mixed models and factor analytic mixed models. The distribution of the transformed data is

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} \sim N\left\{ \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{L}_1^T\boldsymbol{H}\boldsymbol{L}_1 & \boldsymbol{L}_1^T\boldsymbol{H}\boldsymbol{L}_2 \\ \boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_1 & \boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2 \end{pmatrix} \right\}.$$

Following the example of Verbyla (1990) it can be shown that

$$\begin{aligned}
\boldsymbol{L}^T\boldsymbol{y}_o &\sim N\left(\boldsymbol{L}^T\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{L}^T\boldsymbol{H}\boldsymbol{L}\right), \\
\boldsymbol{y}_1|\boldsymbol{y}_2 &\sim N\left\{\boldsymbol{\beta}, (\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\right\}, \\
\boldsymbol{y}_2 &\sim N\left(\boldsymbol{0}, \boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2\right).
\end{aligned}$$

Based on the relation

$$f(\boldsymbol{L}^T\boldsymbol{y}_o;\boldsymbol{\theta}) = f(\boldsymbol{y}_2;\boldsymbol{\kappa})f(\boldsymbol{y}_1|\boldsymbol{y}_2;\boldsymbol{\theta}),$$

we can equate the determinants, i.e.,

$$\begin{aligned}
\det(\boldsymbol{L}^T\boldsymbol{H}\boldsymbol{L}) &= \det\{(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\}\det(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2), \\
\det(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2) &= \det(\boldsymbol{L}^T\boldsymbol{L})\det(\boldsymbol{H})\det(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}). \qquad (2.4)
\end{aligned}$$

The identity (2.4) will be useful later in the development of the residual log-likelihood function that is most commonly presented. The log-likelihood function of $\boldsymbol{L}^T\boldsymbol{y}_o$ can be expressed as

$$\ell(\boldsymbol{\theta};\boldsymbol{L}^T\boldsymbol{y}_o) = \ell(\boldsymbol{\kappa};\boldsymbol{y}_2) + \ell(\boldsymbol{\theta};\boldsymbol{y}_1|\boldsymbol{y}_2). \qquad (2.5)$$

The conditional log-likelihood function $\ell(\boldsymbol{\theta};\boldsymbol{y}_1|\boldsymbol{y}_2)$ is used to estimate the fixed effects, after which there is no information left for the estimation of variance parameters. Therefore, the log-likelihood function associated with the marginal density of $\boldsymbol{y}_2$, i.e., $\ell(\boldsymbol{\kappa};\boldsymbol{y}_2)$ is used for the REML estimation of variance parameters and can be expressed (excluding constants) as

$$\begin{aligned}
\ell(\boldsymbol{\kappa};\boldsymbol{y}_2) &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)\} + \boldsymbol{y}_2^T(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{y}_2\Big] \qquad (2.6) \\
&= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{H})\} + \log\{\det(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})\} + \boldsymbol{y}_o^T\boldsymbol{P}\boldsymbol{y}_o\Big], \qquad (2.7)
\end{aligned}$$

where going from (2.6) to (2.7) makes use of (2.4) and Result A.3.8, i.e.,

$$\begin{aligned} \boldsymbol{P} &= \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T \\ &= \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}. \end{aligned}$$

The expression given in (2.7) is the form of the residual log-likelihood function most commonly presented.

Although we have concentrated on the conditional derivation of REML presented in Verbyla (1990) it is worth noting that much earlier Kalbfleisch and Sprott (1970) considered factorising a distribution, parameterised by two parameters of interest, into a marginal and conditional distribution in much the same way that the likelihood is factorised in Verbyla (1990).

Chapter 3

Current REML EM algorithms for the linear mixed model

There are two specifications of a REML EM algorithm for the linear mixed model currently in the literature. The first and most widely used is the random effect specification which assumes $\boldsymbol{\beta}$ is a vector of random effects with variance tending to infinity. The random effect approach was suggested by Dempster et al. (1977) and followed by others including Laird and Ware (1982) and Foulley et al. (2000). The other approach is based on the expectation computed in the E-step being conditional on the transformed observed data vector that is commonly associated with the residual log-likelihood function. This fixed effect approach was taken by Cullis et al. (2004) and Knight (2008). It will be shown that from a practical implementation point of view the two approaches are equivalent.

This chapter proceeds by considering the linear mixed model specification, the complete data specification, the complete data log-density function, and the conditional distributions required in the E-step for both the random effect and fixed effect approaches. We show that the complete data log-density function and the conditional distributions required are exactly the same for both approaches. This has not been shown in the literature previously. This chapter ends by considering the E and M-steps for the fixed effect approach in more detail and the complete and missing data information matrices associated with this approach.

3.1 Linear mixed model specification - random effect approach

The linear mixed model is specified in the same way as in Section 2.1 except that $\boldsymbol{\beta}$ is considered to be a $t \times 1$ vector of random effects. We define the joint distribution of $\boldsymbol{y}_o$, $\boldsymbol{\beta}$, $\boldsymbol{u}$, and $\boldsymbol{e}$ as

$$\begin{pmatrix} \boldsymbol{y}_o \\ \boldsymbol{\beta} \\ \boldsymbol{u} \\ \boldsymbol{e} \end{pmatrix} \sim N \left\{ \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{F} & \boldsymbol{XB} & \boldsymbol{ZG} & \boldsymbol{R} \\ \boldsymbol{BX}^T & \boldsymbol{B} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{GZ}^T & \mathbf{0} & \boldsymbol{G} & \mathbf{0} \\ \boldsymbol{R} & \mathbf{0} & \mathbf{0} & \boldsymbol{R} \end{pmatrix} \right\}, \tag{3.1}$$

where

$$\begin{aligned} \boldsymbol{F} &= \boldsymbol{XBX}^T + \boldsymbol{ZGZ}^T + \boldsymbol{R} \\ &= \boldsymbol{XBX}^T + \boldsymbol{H} \end{aligned}$$

and

$$\boldsymbol{F}^{-1} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{B}^{-1} + \boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}, \qquad \text{using Result A.3.5.}$$

The random effect approach assumes that the variance of $\boldsymbol{\beta}$ tends to infinity, i.e., $\boldsymbol{B}^{-1} \to \mathbf{0}$. Therefore

$$\begin{aligned} \boldsymbol{F}^{-1} &\to \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}, \\ \boldsymbol{F}^{-1} &\to \boldsymbol{P}. \end{aligned} \tag{3.2}$$

From here on we will assume that if $\boldsymbol{B}^{-1} \to \mathbf{0}$ then $\boldsymbol{F}^{-1} = \boldsymbol{P}$.

3.2 Complete data specification - random effect approach

Under the random effect approach the complete data is defined as $\boldsymbol{y}_c = (\boldsymbol{y}_o^T, \boldsymbol{\beta}^T, \boldsymbol{u}^T)^T$. We define $\boldsymbol{\kappa} = (\sigma^2, \boldsymbol{\gamma}^T, \boldsymbol{\phi}^T)^T$ and note that the complete data density function can be expressed as

$$f(\boldsymbol{y}_c; \boldsymbol{\kappa}) = f(\boldsymbol{y}_o|\boldsymbol{\beta}, \boldsymbol{u}; \sigma^2, \boldsymbol{\phi}) f(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\gamma}).$$

Since it is assumed $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are independent and that $\boldsymbol{B}^{-1} \to \boldsymbol{0}$, then

$$f(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\gamma}) \propto f(\boldsymbol{u}; \boldsymbol{\gamma}).$$

Using (3.1) and Result A.1 it can be shown that

$$\boldsymbol{y}_o|\boldsymbol{\beta}, \boldsymbol{u} \sim N\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \boldsymbol{R}\right).$$

The complete data log-density function is therefore

$$\begin{aligned}
\log\{f(\boldsymbol{y}_c; \boldsymbol{\kappa})\} &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{R})\} + \log\{\det(\boldsymbol{G})\} + \boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u} \\
&\quad + (\boldsymbol{y}_o - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})^T\boldsymbol{R}^{-1}(\boldsymbol{y}_o - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})\Big] \\
&= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{R})\} + \log\{\det(\boldsymbol{G})\} \\
&\quad + \boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u} + \boldsymbol{e}^T\boldsymbol{R}^{-1}\boldsymbol{e}\Big], \qquad (3.3)
\end{aligned}$$

which is equivalent to the log joint density function of $\boldsymbol{y}_o$ and $\boldsymbol{u}$ used in deriving Henderson's mixed model equations. The E-step of a REML EM algorithm using the complete data specification $\boldsymbol{y}_c = (\boldsymbol{y}_o^T, \boldsymbol{\beta}^T, \boldsymbol{u}^T)^T$ requires the conditional distributions $\boldsymbol{u}|\boldsymbol{y}_o$ and $\boldsymbol{e}|\boldsymbol{y}_o$. Using (3.1) and Result A.1 we have

$$\begin{aligned}
\boldsymbol{u}|\boldsymbol{y}_o &\sim N\left(\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{F}^{-1}\boldsymbol{y}_o, \boldsymbol{G}-\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{F}^{-1}\boldsymbol{Z}\boldsymbol{G}\right), \\
&\sim N\left(\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{y}_o, \boldsymbol{G}-\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G}\right) \qquad \text{using (3.2)}, \\
&\sim N\left(\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{y}_o, \boldsymbol{C}^{ZZ}\right) \qquad \text{using Result A.3.10}
\end{aligned} \tag{3.4}$$

and

$$\begin{aligned}
\boldsymbol{e}|\boldsymbol{y}_o &\sim N\left(\boldsymbol{R}\boldsymbol{F}^{-1}\boldsymbol{y}_o, \boldsymbol{R}-\boldsymbol{R}\boldsymbol{F}^{-1}\boldsymbol{R}\right), \\
&\sim N\left(\boldsymbol{R}\boldsymbol{P}\boldsymbol{y}_o, \boldsymbol{R}-\boldsymbol{R}\boldsymbol{P}\boldsymbol{R}\right) \qquad \text{using (3.2)}, \\
&\sim N\left(\boldsymbol{R}\boldsymbol{P}\boldsymbol{y}_o, \boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}^T\right) \qquad \text{using Result A.3.9},
\end{aligned} \tag{3.5}$$

where $\boldsymbol{W} = (\boldsymbol{X} \;\; \boldsymbol{Z})$. An important feature of (3.4) and (3.5) is that the conditional variances associated with $\boldsymbol{u}|\boldsymbol{y}_o$ and $\boldsymbol{e}|\boldsymbol{y}_o$ are $b \times b$ and $n \times n$ matrices respectively. Therefore, the conditional variance of $\boldsymbol{e}|\boldsymbol{y}_o$ can require significantly more computation time than the conditional variance of $\boldsymbol{u}|\boldsymbol{y}_o$.

3.3 Linear mixed model specification - fixed effect approach

The linear mixed model is specified in the same way as in Section 2.1. Under the fixed effect approach we require the joint distribution of $\boldsymbol{y}_o$, $\boldsymbol{u}$, and $\boldsymbol{e}$ given in (2.2) to form the complete data log-density function. To form the conditional distributions necessary for computation of the E-step we require the joint distribution of $\boldsymbol{y}_2$, $\boldsymbol{u}$, and $\boldsymbol{e}$ which can be shown to be

$$\begin{pmatrix} \boldsymbol{y}_2 \\ \boldsymbol{u} \\ \boldsymbol{e} \end{pmatrix} \sim N\left\{ \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2 & \boldsymbol{L}_2^T\boldsymbol{Z}\boldsymbol{G} & \boldsymbol{L}_2^T\boldsymbol{R} \\ \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{L}_2 & \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{R}\boldsymbol{L}_2 & \boldsymbol{0} & \boldsymbol{R} \end{pmatrix} \right\} \tag{3.6}$$

3.4 Complete data specification - fixed effect approach

Under the fixed effect approach the complete data is defined as $\boldsymbol{y}_c = (\boldsymbol{y}_o^T, \boldsymbol{u}^T)^T$. We define $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\kappa}^T)^T$ and $\boldsymbol{\kappa} = (\sigma^2, \boldsymbol{\gamma}^T, \boldsymbol{\phi}^T)^T$ and note that the complete data density function can be expressed as

$$f(\boldsymbol{y}_c; \boldsymbol{\theta}) = f(\boldsymbol{y}_o|\boldsymbol{u}; \boldsymbol{\theta}) f(\boldsymbol{u}; \boldsymbol{\gamma}).$$

Using (2.2) and Result A.1 it can be shown that

$$\boldsymbol{y}_o|\boldsymbol{u} \sim N\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \boldsymbol{R}\right). \tag{3.7}$$

The complete data log-density function is therefore

$$\begin{aligned}
\log\{f(\boldsymbol{y}_c; \boldsymbol{\theta})\} &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{R})\} + \log\{\det(\boldsymbol{G})\} + \boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u} \\
&\quad + (\boldsymbol{y}_o - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})^T\boldsymbol{R}^{-1}(\boldsymbol{y}_o - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})\Big] \\
&= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{R})\} + \log\{\det(\boldsymbol{G})\} \\
&\quad + \boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u} + \boldsymbol{e}^T\boldsymbol{R}^{-1}\boldsymbol{e}\Big],
\end{aligned} \tag{3.8}$$

which is exactly the same as the random effect complete data specification in (3.3). Note that the difference between the parameterisation of the complete data log-density under the two approaches is due to the different treatment of $\boldsymbol{\beta}$. Under the random effect approach, the complete data log-likelihood function is parameterised by $\boldsymbol{\kappa}$ and under the fixed effect approach by $\boldsymbol{\theta}$, including when we make the substitution $\boldsymbol{e} = \boldsymbol{y}_o - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}$.

The E-step of the REML EM algorithm using the fixed effect approach requires the conditional distributions $\boldsymbol{u}|\boldsymbol{y}_2$ and $\boldsymbol{e}|\boldsymbol{y}_2$. Using (3.6) and Result A.1 we have

$$\begin{aligned}
\boldsymbol{u}|\boldsymbol{y}_2 &\sim N\left\{\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T\boldsymbol{y}_o, \boldsymbol{G}-\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T\boldsymbol{Z}\boldsymbol{G}\right\}, \\
&\sim N\left(\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{y}_o, \boldsymbol{G}-\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G}\right) && \text{using Result A.3.8,} \\
&\sim N\left(\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{y}_o, \boldsymbol{C}^{ZZ}\right) && \text{using Result A.3.10}
\end{aligned} \tag{3.9}$$

and

$$\begin{aligned}
\boldsymbol{e}|\boldsymbol{y}_2 &\sim N\left\{\boldsymbol{R}\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T\boldsymbol{y}_o, \boldsymbol{R}-\boldsymbol{R}\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T\boldsymbol{R}\right\}, \\
&\sim N\left(\boldsymbol{R}\boldsymbol{P}\boldsymbol{y}_o, \boldsymbol{R}-\boldsymbol{R}\boldsymbol{P}\boldsymbol{R}\right) && \text{using Result A.3.8,} \\
&\sim N\left(\boldsymbol{R}\boldsymbol{P}\boldsymbol{y}_o, \boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}^T\right) && \text{using Result A.3.9.}
\end{aligned} \tag{3.10}$$

Note that (3.9) and (3.10) are exactly the same as (3.4) and (3.5) respectively. Since the complete data specification and the conditional distributions necessary for computation of the E-step are exactly the same for the two approaches the random effect and fixed effect approach are equivalent from a practical implementation point of view. From here on we will only consider the fixed effect approach.

3.5 REML EM algorithm E-step

We use the superscript (P) to define $\boldsymbol{y}_c^{(P)} = (\boldsymbol{y}_o^T, \boldsymbol{u}^T)^T$ to distinguish this complete data specification and its associated $Q(.)$ function from the new complete data specification considered later. The log-density function associated with $\boldsymbol{y}_c^{(P)}$ can be partitioned as

$$\log\{f(\boldsymbol{y}_c^{(P)};\boldsymbol{\theta})\} = \log\{f(\boldsymbol{e};\boldsymbol{\theta})\} + \log\{f(\boldsymbol{u};\boldsymbol{\gamma})\}, \tag{3.11}$$

where

$$\begin{aligned}
\log\{f(\boldsymbol{e};\boldsymbol{\theta})\} &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{R})\} + \boldsymbol{e}^T\boldsymbol{R}^{-1}\boldsymbol{e}\Big], \\
\log\{f(\boldsymbol{u};\boldsymbol{\gamma})\} &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{G})\} + \boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u}\Big].
\end{aligned}$$

The E-step involves taking the expectation of (3.11) over $k(\boldsymbol{u}|\boldsymbol{y}_2;\boldsymbol{\kappa}^{(w)})$ where $\boldsymbol{\kappa}^{(w)}$ is the w-th iterate of $\boldsymbol{\kappa}$. We can write

$$\begin{aligned} Q^{(P)}(\boldsymbol{\kappa};\boldsymbol{\kappa}^{(w)}) &= E\big[\log\{f(\boldsymbol{y}_c^{(P)};\boldsymbol{\theta})\}|\boldsymbol{y}_2;\boldsymbol{\kappa}^{(w)}\big] \\ &= E\big[\log\{f(\boldsymbol{e};\boldsymbol{\theta})\}|\boldsymbol{y}_2;\boldsymbol{\kappa}^{(w)}\big] + E\big[\log\{f(\boldsymbol{u};\boldsymbol{\gamma})\}|\boldsymbol{y}_2;\boldsymbol{\kappa}^{(w)}\big] \\ &= Q^{(P)}(\sigma^2,\boldsymbol{\phi};\boldsymbol{\kappa}^{(w)}) + Q^{(P)}(\boldsymbol{\gamma};\boldsymbol{\kappa}^{(w)}). \end{aligned}$$

Upon taking the expectation of $\log\{f(\boldsymbol{y}_c^{(P)};\boldsymbol{\theta})$ over $k(\boldsymbol{u}|\boldsymbol{y}_2;\boldsymbol{\kappa}^{(w)})$ the associated $Q(.)$ function is now parameterised by $\boldsymbol{\kappa}$ (and $\boldsymbol{\kappa}^{(w)}$). Using Result A.3.2 we can write

$$\begin{aligned} Q^{(P)}(\sigma^2,\boldsymbol{\phi};\boldsymbol{\kappa}^{(w)}) &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{R})\} + \tilde{\boldsymbol{e}}^{(w)^T}\boldsymbol{R}^{-1}\tilde{\boldsymbol{e}}^{(w)} + \mathrm{tr}\big(\boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1^{(w)}}\boldsymbol{W}^T\big)\Big], \\ Q^{(P)}(\boldsymbol{\gamma};\boldsymbol{\kappa}^{(w)}) &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{G})\} + \tilde{\boldsymbol{u}}^{(w)^T}\boldsymbol{G}^{-1}\tilde{\boldsymbol{u}}^{(w)} + \mathrm{tr}\big(\boldsymbol{G}^{-1}\boldsymbol{C}^{ZZ^{(w)}}\big)\Big], \end{aligned}$$

where

$$\tilde{\boldsymbol{e}}^{(w)} = E\big(\boldsymbol{e}|\boldsymbol{y}_2;\boldsymbol{\kappa}^{(w)}\big) = \boldsymbol{R}^{(w)}\boldsymbol{P}^{(w)}\boldsymbol{y}_o, \tag{3.12}$$

$$\tilde{\boldsymbol{u}}^{(w)} = E\big(\boldsymbol{u}|\boldsymbol{y}_2;\boldsymbol{\kappa}^{(w)}\big) = \boldsymbol{G}^{(w)}\boldsymbol{Z}^T\boldsymbol{P}^{(w)}\boldsymbol{y}_o. \tag{3.13}$$

This completes the E-step.

3.6 REML EM algorithm M-step

The M-step involves maximising $Q^{(P)}(\boldsymbol{\kappa};\boldsymbol{\kappa}^{(w)})$ with respect to $\boldsymbol{\kappa}$ to find an update for $\boldsymbol{\kappa} = (\sigma^2,\boldsymbol{\gamma}^T,\boldsymbol{\phi}^T)^T$. The partitioning of $Q^{(P)}(\boldsymbol{\kappa};\boldsymbol{\kappa}^{(w)})$ into $Q^{(P)}(\sigma^2,\boldsymbol{\phi};\boldsymbol{\kappa}^{(w)})$ and $Q^{(P)}(\boldsymbol{\gamma};\boldsymbol{\kappa}^{(w)})$ simplifies the process. We have not provided an updating equation for $\boldsymbol{\beta}$ as the maximum likelihood estimate for $\boldsymbol{\beta}$ can be obtained at the end of the REML EM algorithm by using the generalised least squares estimate of $\boldsymbol{\beta}$. We use the notation in (3.12) and (3.13) when considering the updating equations for σ^2, γ_{ij} and ϕ_{gh}.

3.6.1 Update for σ^2

Noting that $\boldsymbol{R} = \sigma^2\boldsymbol{\Sigma}$ and differentiating $Q^{(P)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})$ with respect to σ^2 we have

$$\frac{\partial Q^{(P)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})}{\partial \sigma^2} = -\frac{1}{2}\left\{\frac{n}{\sigma^2} - \frac{1}{(\sigma^2)^2}\tilde{\boldsymbol{e}}^{(w)^T}\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{e}}^{(w)} - \frac{1}{(\sigma^2)^2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1^{(w)}}\boldsymbol{W}^T\right)\right\}.$$

Equating the derivative to zero we have

$$\sigma^{2^{(w+1)}} = \frac{1}{n}\left\{\tilde{\boldsymbol{e}}^{(w)^T}\boldsymbol{\Sigma}^{-1^{(w)}}\tilde{\boldsymbol{e}}^{(w)} + \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1^{(w)}}\boldsymbol{W}\boldsymbol{C}^{-1^{(w)}}\boldsymbol{W}^T\right)\right\}. \tag{3.14}$$

3.6.2 Update for γ_{gh}

To find an update for γ_{gh}, $g = 1, \ldots, b$, $h = 1, \ldots, b$, $i \geq j$ involves differentiating $Q^{(P)}(\boldsymbol{\gamma}; \boldsymbol{\kappa}^{(w)})$ with respect to $\boldsymbol{\gamma}_{gh}$. Using Result A.4.1 and Result A.4.2 we have

$$\frac{\partial Q^{(P)}(\boldsymbol{\gamma}; \boldsymbol{\kappa}^{(w)})}{\partial \gamma_{gh}} = -\frac{1}{2}\left\{\mathrm{tr}\left(\boldsymbol{G}^{-1}\frac{\partial \boldsymbol{G}}{\partial \gamma_{gh}}\right) - \tilde{\boldsymbol{u}}^{(w)^T}\boldsymbol{G}^{-1}\frac{\partial \boldsymbol{G}}{\partial \gamma_{gh}}\boldsymbol{G}^{-1}\tilde{\boldsymbol{u}}^{(w)} - \mathrm{tr}\left(\boldsymbol{G}^{-1}\frac{\partial \boldsymbol{G}}{\partial \gamma_{gh}}\boldsymbol{G}^{-1}\boldsymbol{C}^{ZZ^{(w)}}\right)\right\}.$$

The updating formula for γ_{gh} depends on the specific form of $\boldsymbol{G}$. For a one-way random effects model $\boldsymbol{G}$ has the simple form $\boldsymbol{G} = \sigma_u^2\boldsymbol{I}_b$. In this case the updating equation is

$$\sigma_u^{2^{(w+1)}} = \frac{1}{b}\left\{\tilde{\boldsymbol{u}}^{(w)^T}\tilde{\boldsymbol{u}}^{(w)} + \mathrm{tr}\left(\boldsymbol{C}^{ZZ^{(w)}}\right)\right\}.$$

3.6.3 Update for ϕ_{ij}

To find an update for ϕ_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, n$, $i \geq j$ involves differentiating $Q^{(P)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})$ with respect to ϕ_{ij}. Noting that $\boldsymbol{R} = \sigma^2\boldsymbol{\Sigma}$ and using Result A.4.1 and Result A.4.2 we have

$$\frac{\partial Q^{(P)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})}{\partial \phi_{ij}} = -\frac{1}{2}\left\{ \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_{ij}} \right) - \frac{1}{\sigma^2} \tilde{\boldsymbol{e}}^{(w)^T} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_{ij}} \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{e}}^{(w)} - \frac{1}{\sigma^2} \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_{ij}} \boldsymbol{\Sigma}^{-1} \boldsymbol{W} \boldsymbol{C}^{-1^{(w)}} \boldsymbol{W}^T \right) \right\}.$$

The updating equation for ϕ_{ij} depends on the specific form of $\boldsymbol{\Sigma}$. In this thesis we only consider models where $\boldsymbol{\Sigma} = \boldsymbol{I}_n$.

This completes the M-step.

3.7 Complete data information matrix

We use the superscript (P) when considering the complete and missing data information matrices associated with $\boldsymbol{y}_c^{(P)}$. This is done to distinguish these information matrices from the complete and missing data information matrices associated with the new complete data specification we consider later. The complete and missing data information matrices are symmetric and in the interests of parsimony we only present the upper right triangle for each.

The complete data information matrix based on $\boldsymbol{y}_c^{(P)}$ is

$$\boldsymbol{\mathcal{I}}_c^{(P)}(\boldsymbol{\kappa}^{(w)}; \boldsymbol{y}_c^{(P)}) = \begin{pmatrix} \mathcal{I}^{(P)}_{c_{(\sigma^2)^2}} & \boldsymbol{0} & \boldsymbol{\mathcal{I}}^{(P)}_{c_{\sigma^2\boldsymbol{\phi}}} \\ & \boldsymbol{\mathcal{I}}^{(P)}_{c_{\boldsymbol{\gamma},\boldsymbol{\gamma}}} & \boldsymbol{0} \\ & & \boldsymbol{\mathcal{I}}^{(P)}_{c_{\boldsymbol{\phi},\boldsymbol{\phi}}} \end{pmatrix},$$

where the zeroes in the complete data information matrix are due to the partitioning of $Q^{(P)}(\boldsymbol{\kappa}; \boldsymbol{\kappa}^{(w)})$ into $Q^{(P)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})$ and $Q^{(P)}(\boldsymbol{\gamma}; \boldsymbol{\kappa}^{(w)})$. Using the Oakes (1999) result presented in (1.13) we have

$$\mathcal{I}^{(P)}_{c_{(\sigma^2)^2}} = -\left\langle \frac{\partial^2 Q^{(P)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})}{\partial(\sigma^2)^2} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}},$$
$$\mathcal{I}^{(P)}_{c_{\boldsymbol{\gamma},\boldsymbol{\gamma}}} = -\left\langle \frac{\partial^2 Q^{(P)}(\boldsymbol{\gamma}; \boldsymbol{\kappa}^{(w)})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}},$$
$$\mathcal{I}^{(P)}_{c_{\boldsymbol{\phi},\boldsymbol{\phi}}} = -\left\langle \frac{\partial^2 Q^{(P)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}},$$
$$\mathcal{I}^{(P)}_{c_{\sigma^2,\boldsymbol{\phi}}} = -\left\langle \frac{\partial^2 Q^{(P)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})}{\partial\sigma^2\partial\boldsymbol{\phi}^T} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}}.$$

We present each in turn. The linear algebra involved is straightforward but tedious and therefore is omitted.

3.7.1 Element $\mathcal{I}^{(P)}_{c_{(\sigma^2)^2}}$

It can be shown that

$$\mathcal{I}^{(P)}_{c_{(\sigma^2)^2}} = \frac{n}{2\left(\sigma^{2(w)}\right)^2} + \frac{1}{\left(\sigma^{2(w)}\right)^2}\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\boldsymbol{P}^{(w)}\boldsymbol{y}_o - \frac{1}{\left(\sigma^{2(w)}\right)^2}\mathrm{tr}\left(\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\right). \quad (3.15)$$

3.7.2 Elements of $\mathcal{I}^{(P)}_{c_{\boldsymbol{\gamma},\boldsymbol{\gamma}}}$

For $g = 1, \ldots, b$, $h = 1, \ldots, b$, $g \geq h$ and $q = 1, \ldots, b$, $r = 1, \ldots, b$, $q \geq r$ it can be shown that

$$
\begin{aligned}
\mathcal{I}^{(P)}_{c_{\gamma_{gh},\gamma_{qr}}} = \frac{1}{2}\Bigg\{ & \operatorname{tr}\Big(\boldsymbol{G}^{-1^{(w)}}\frac{\partial^2\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{gh}\partial\gamma^{(w)}_{qr}}\Big) - \operatorname{tr}\Big(\boldsymbol{G}^{-1^{(w)}}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{qr}}\boldsymbol{G}^{-1^{(w)}}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{gh}}\Big) \\
& + \tilde{\boldsymbol{u}}^{(w)^T}\boldsymbol{G}^{-1^{(w)}}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{gh}}\boldsymbol{G}^{-1^{(w)}}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{qr}}\boldsymbol{G}^{-1^{(w)}}\tilde{\boldsymbol{u}}^{(w)} \\
& + \tilde{\boldsymbol{u}}^{(w)^T}\boldsymbol{G}^{-1^{(w)}}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{qr}}\boldsymbol{G}^{-1^{(w)}}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{gh}}\boldsymbol{G}^{-1^{(w)}}\tilde{\boldsymbol{u}}^{(w)} \\
& - \tilde{\boldsymbol{u}}^{(w)^T}\boldsymbol{G}^{-1^{(w)}}\frac{\partial^2\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{gh}\partial\gamma^{(w)}_{qr}}\boldsymbol{G}^{-1^{(w)}}\tilde{\boldsymbol{u}}^{(w)} - \operatorname{tr}\Big(\boldsymbol{G}^{-1^{(w)}}\frac{\partial^2\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{gh}\partial\gamma^{(w)}_{qr}}\boldsymbol{G}^{-1^{(w)}}\boldsymbol{C}^{ZZ^{(w)}}\Big) \\
& + \operatorname{tr}\Big(\boldsymbol{G}^{-1^{(w)}}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{gh}}\boldsymbol{G}^{-1^{(w)}}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{qr}}\boldsymbol{G}^{-1^{(w)}}\boldsymbol{C}^{ZZ^{(w)}}\Big) \\
& + \operatorname{tr}\Big(\boldsymbol{G}^{-1^{(w)}}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{qr}}\boldsymbol{G}^{-1^{(w)}}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{gh}}\boldsymbol{G}^{-1^{(w)}}\boldsymbol{C}^{ZZ^{(w)}}\Big)\Bigg\}. \qquad (3.16)
\end{aligned}
$$

3.7.3 Elements of $\mathcal{I}^{(P)}_{c_{\phi,\phi}}$

For $i = 1,\ldots,n$, $j = 1,\ldots,n$, $i \geq j$ and $k = 1,\ldots,n$, $l = 1,\ldots,n$, $k \geq l$ it can be shown that

$$
\begin{aligned}
\mathcal{I}^{(P)}_{c_{\phi_{ij},\phi_{kl}}} = \frac{1}{2}\Bigg\{ & \big(\sigma^{2^{(w)}}\big)^2\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{ij}}\boldsymbol{R}^{-1^{(w)}}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{kl}}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\
& + \big(\sigma^{2^{(w)}}\big)^2\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{kl}}\boldsymbol{R}^{-1^{(w)}}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{ij}}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\
& - \sigma^{2^{(w)}}\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\frac{\partial^2\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{ij}\partial\phi^{(w)}_{kl}}\boldsymbol{P}^{(w)}\boldsymbol{y}_o + \sigma^{2^{(w)}}\operatorname{tr}\Big(\frac{\partial^2\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{ij}\partial\phi^{(w)}_{kl}}\boldsymbol{P}^{(w)}\Big) \\
& + \big(\sigma^{2^{(w)}}\big)^2\operatorname{tr}\Big\{\Big(\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{kl}}\boldsymbol{R}^{-1^{(w)}} - \frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{kl}}\boldsymbol{P}^{(w)}\Big)^T\Big(\boldsymbol{R}^{-1^{(w)}}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{ij}} - \boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{ij}}\Big)\Big\} \\
& - \big(\sigma^{2^{(w)}}\big)^2\operatorname{tr}\Big(\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{kl}}\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{ij}}\Big)\Bigg\}. \qquad (3.17)
\end{aligned}
$$

3.7.4 Elements of $\mathcal{I}^{(P)}_{c_{\sigma^2,\phi}}$

For $i = 1,\ldots,n$, $j = 1,\ldots,n$, $i \geq j$ it can be shown that

$$\mathcal{I}^{(P)}_{c_{\sigma^{2(w)}\phi^{(w)}_{ij}}} = \frac{1}{2}\Big\{ \boldsymbol{y}_o^T \boldsymbol{P}^{(w)} \frac{\partial \boldsymbol{\Sigma}^{(w)}}{\partial \phi^{(w)}_{ij}} \boldsymbol{P}^{(w)} \boldsymbol{y}_o + \mathrm{tr}\Big(\frac{\partial \boldsymbol{\Sigma}^{(w)}}{\partial \phi^{(w)}_{ij}} \boldsymbol{R}^{-1^{(w)}}\Big) - \mathrm{tr}\Big(\frac{\partial \boldsymbol{\Sigma}^{(w)}}{\partial \phi^{(w)}_{ij}} \boldsymbol{P}^{(w)}\Big)\Big\}. \tag{3.18}$$

This concludes the section on the complete data information matrix.

3.8 Missing data information matrix

The missing data information matrix based on $\boldsymbol{y}_c^{(P)}$ and using the Oakes (1999) result presented in (1.13) is

$$\mathcal{I}^{(P)}_m(\boldsymbol{\kappa}^{(w)}; \boldsymbol{y}_c^{(P)}) = \begin{pmatrix} \mathcal{I}^{(P)}_{m_{(\sigma^2)^2}} & \mathcal{I}^{(P)}_{m_{\sigma^2,\gamma}} & \mathcal{I}^{(P)}_{m_{\sigma^2,\phi}} \\ & \mathcal{I}^{(P)}_{m_{\gamma,\gamma}} & \mathcal{I}^{(P)}_{m_{\sigma^2,\gamma}} \\ & & \mathcal{I}^{(P)}_{m_{\phi,\phi}} \end{pmatrix},$$

where

$$\begin{aligned}
\mathcal{I}^{(P)}_{m_{(\sigma^2)^2}} &= \left\langle \frac{\partial^2 Q^{(P)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})}{\partial \sigma^2 \partial \sigma^{2(w)}} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}}, \\
\mathcal{I}^{(P)}_{m_{\gamma,\gamma}} &= \left\langle \frac{\partial^2 Q^{(P)}(\boldsymbol{\gamma}; \boldsymbol{\kappa}^{(w)})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^{(w)^T}} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}}, \\
\mathcal{I}^{(P)}_{m_{\boldsymbol{\phi},\boldsymbol{\phi}}} &= \left\langle \frac{\partial^2 Q^{(P)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^{(w)^T}} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}}, \\
\mathcal{I}^{(P)}_{m_{\sigma^2,\gamma}} &= \left\langle \frac{\partial^2 Q^{(P)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})}{\partial \sigma^2 \partial \boldsymbol{\gamma}^{(w)^T}} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}}, \\
\mathcal{I}^{(P)}_{m_{\sigma^2,\boldsymbol{\phi}}} &= \left\langle \frac{\partial^2 Q^{(P)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})}{\partial \sigma^2 \partial \boldsymbol{\phi}^{(w)^T}} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}}, \\
\mathcal{I}^{(P)}_{m_{\boldsymbol{\gamma},\boldsymbol{\phi}}} &= \left\langle \frac{\partial^2 Q^{(P)}(\boldsymbol{\gamma}; \boldsymbol{\kappa}^{(w)})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\phi}^{(w)^T}} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}}.
\end{aligned}$$

We make use of Result A.4.3 and present each in turn.

3.8.1 Element $\mathcal{I}^{(P)}_{m_{(\sigma^2)^2}}$

It can be shown that

$$\begin{aligned}\mathcal{I}^{(P)}_{m_{(\sigma^2)^2}} &= -\frac{1}{2\left(\sigma^{2(w)}\right)^2}\Big\{-n - 2\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ &\quad +2\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ &\quad +2\mathrm{tr}\Big(\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\Big) - \mathrm{tr}\Big(\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\Big)\Big\}. \qquad (3.19)\end{aligned}$$

3.8.2 Elements of $\mathcal{I}^{(P)}_{m_{\gamma,\gamma}}$

For $g = 1,\ldots,b$, $h = 1,\ldots,b$, $g \geq h$ and $q = 1,\ldots,b$, $r = 1,\ldots,b$, $q \geq r$ it can be shown that

$$\begin{aligned}\mathcal{I}^{(P)}_{m_{\gamma gh,\gamma qr}} &= -\frac{1}{2}\Bigg[\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\boldsymbol{Z}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{qr}}\boldsymbol{Z}^T\boldsymbol{P}^{(w)}\boldsymbol{Z}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{gh}}\boldsymbol{Z}^T\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ &+\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\boldsymbol{Z}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{gh}}\boldsymbol{Z}^T\boldsymbol{P}^{(w)}\boldsymbol{Z}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{qr}}\boldsymbol{Z}^T\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ &-\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\boldsymbol{Z}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{qr}}\boldsymbol{G}^{-1^{(w)}}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{gh}}\boldsymbol{Z}^T\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ &-\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\boldsymbol{Z}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{gh}}\boldsymbol{G}^{-1^{(w)}}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{qr}}\boldsymbol{Z}^T\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ &-\mathrm{tr}\Big\{\Big(\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{gh}}\boldsymbol{G}^{-1^{(w)}} - \frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{gh}}\boldsymbol{Z}^T\boldsymbol{P}^{(w)}\boldsymbol{Z}\Big)^T\Big(\boldsymbol{G}^{-1^{(w)}}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{qr}} - \boldsymbol{Z}^T\boldsymbol{P}^{(w)}\boldsymbol{Z}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma^{(w)}_{qr}}\Big)\Big\}\Bigg]. \qquad (3.20)\end{aligned}$$

3.8.3 Elements of $\mathcal{I}^{(P)}_{m_{\phi,\phi}}$

For $i = 1,\ldots,n$, $j = 1,\ldots,n$, $i \geq j$ and $k = 1,\ldots,n$, $l = 1,\ldots,n$, $k \geq l$ it can be shown that

$$\begin{aligned}\mathcal{I}^{(P)}_{m_{\phi_{ij},\phi_{kl}}} = & -\frac{\left(\sigma^{2(w)}\right)^2}{2}\Bigg[\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{kl}^{(w)}}\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ & +\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{kl}^{(w)}}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ & -\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{kl}^{(w)}}\boldsymbol{R}^{-1^{(w)}}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ & -\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\boldsymbol{R}^{-1^{(w)}}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{kl}^{(w)}}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ & -\mathrm{tr}\Big\{\Big(\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{kl}^{(w)}}\boldsymbol{R}^{-1^{(w)}}-\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{kl}^{(w)}}\boldsymbol{P}^{(w)}\Big)^T\Big(\boldsymbol{R}^{-1^{(w)}}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}-\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\Big)\Big\}\Bigg].\end{aligned} \tag{3.21}$$

3.8.4 Elements of $\mathcal{I}^{(P)}_{m_{\sigma^2,\gamma}}$

For $g = 1,\ldots,b$, $h = 1,\ldots,b$, $g \geq h$ it can be shown that

$$\begin{aligned}\mathcal{I}^{(P)}_{m_{\sigma^2,\gamma_{gh}}} = & -\frac{1}{2\sigma^{2(w)}}\Bigg\{\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\boldsymbol{Z}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma_{gh}^{(w)}}\boldsymbol{Z}^T\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ & +\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\boldsymbol{P}^{(w)}\boldsymbol{Z}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma_{gh}^{(w)}}\boldsymbol{Z}^T\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ & -\mathrm{tr}\Big(\boldsymbol{P}^{(w)}\boldsymbol{Z}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma_{gh}^{(w)}}\boldsymbol{Z}^T\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\Big)\Bigg\}.\end{aligned} \tag{3.22}$$

3.8.5 Elements of $\mathcal{I}^{(P)}_{m_{\sigma^2,\phi}}$

For $i = 1,\ldots,n$, $j = 1,\ldots,n$, $i \geq j$ it can be shown that

$$\begin{aligned}\mathcal{I}^{(P)}_{m_{\sigma^2,\phi_{ij}}} &= -\frac{1}{2}\Bigg\{\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ &\quad +\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\boldsymbol{P}^{(w)}\boldsymbol{y}_o - 2\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ &\quad -\mathrm{tr}\Big(\boldsymbol{R}^{-1^{(w)}}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\Big) + 2\mathrm{tr}\Big(\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\boldsymbol{P}^{(w)}\Big) \\ &\quad -\mathrm{tr}\Big(\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\Big)\Bigg\}. \end{aligned} \tag{3.23}$$

3.8.6 Elements of $\mathcal{I}^{(P)}_{m_{\gamma,\phi}}$

For $g = 1,\ldots,b$, $h = 1,\ldots,b$, $g \geq h$ and $i = 1,\ldots,n$, $j = 1,\ldots,n$, $i \geq j$ it can be shown that

$$\begin{aligned}\mathcal{I}^{(P)}_{m_{\gamma_{gh},\phi_{ij}}} &= -\frac{\sigma^{2^{(w)}}}{2}\Bigg\{\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\boldsymbol{P}^{(w)}\boldsymbol{Z}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma_{gh}^{(w)}}\boldsymbol{Z}^T\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ &\quad +\boldsymbol{P}^{(w)}\boldsymbol{Z}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma_{gh}^{(w)}}\boldsymbol{Z}^T\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ &\quad +\mathrm{tr}\Big(\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\boldsymbol{P}^{(w)}\boldsymbol{Z}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma_{gh}^{(w)}}\boldsymbol{Z}^T\Big)\Bigg\}. \end{aligned} \tag{3.24}$$

This concludes the section on the missing data information matrix.

An example data set where a REML EM algorithm based on the complete data specification $\boldsymbol{y}_c^{(P)} = (\boldsymbol{y}_o^T, \boldsymbol{u}^T)^T$ is considered in the next chapter where the complete data specification $\boldsymbol{y}_c^{(P)}$ is compared to a new complete data specification.

Chapter 4

A new REML EM algorithm for linear mixed models

In this chapter we present a new REML EM algorithm for linear mixed models based on a new specification of the complete data. This new complete data specification is based upon the conditional derivation of REML presented by Verbyla (1990). This specification is computationally more efficient than the complete data specification considered in the previous chapter as the E-step under the new specification involves computing the trace of two matrices each of order b compared to a matrix of order b and a matrix of order n under the previously published complete data specification. Furthermore, we provide the conditions under which the new specification has a faster rate of convergence than the previously published specification. This chapter concludes by comparing two REML EM algorithms applied to an example data set. One algorithm is based on the new complete data specification and the other on the previously published specification.

4.1 New complete data specification

We consider a linear mixed model specified in the same way as in Section 2.1. We define a new complete data specification $\boldsymbol{y}_c^{(U)} = (\boldsymbol{y}_2^T, \boldsymbol{u}^T)^T$. This differs from the previously published specification $\boldsymbol{y}_c^{(P)} = (\boldsymbol{y}_o^T, \boldsymbol{u}^T)^T$ by considering the incomplete data to be the transformed observed data vector associated with the marginal log-density function in (2.5) which is used for the REML estimation of variance parameters. There are a number of advantages in using the complete data specification

$\boldsymbol{y}_c^{(U)}$. Firstly, to form the complete data log-density function only requires the joint distribution of $\boldsymbol{y}_2$ and $\boldsymbol{u}$ which can be ascertained from (3.6) but is reproduced here for convenience

$$\begin{pmatrix} \boldsymbol{y}_2 \\ \boldsymbol{u} \end{pmatrix} \sim N\left\{ \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2 & \boldsymbol{L}_2^T\boldsymbol{Z}\boldsymbol{G} \\ \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{L}_2 & \boldsymbol{G} \end{pmatrix} \right\}. \tag{4.1}$$

Using (2.2) and Result A.1 or by simple modification of the expectation and variance associated with the distribution of $\boldsymbol{y}_o|\boldsymbol{u}$ presented in (3.7) it can be shown that

$$\boldsymbol{y}_2|\boldsymbol{u} \sim N\left(\boldsymbol{L}_2^T\boldsymbol{Z}\boldsymbol{u}, \boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2\right).$$

The complete data log density function is therefore (excluding constants)

$$\begin{aligned} \log\{f(\boldsymbol{y}_c^{(U)};\boldsymbol{\kappa})\} &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2)\} + \log\{\det(\boldsymbol{G})\} \\ &\quad +(\boldsymbol{y}_o - \boldsymbol{Z}\boldsymbol{u})^T\boldsymbol{S}(\boldsymbol{y}_o - \boldsymbol{Z}\boldsymbol{u}) + \boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u}\Big], \end{aligned} \tag{4.2}$$

where

$$\begin{aligned} \boldsymbol{S} &= \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T \\ &= \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{R}^{-1}. \qquad \text{using Result A.3.8} \end{aligned}$$

The relationship between the complete data log density functions based on $\boldsymbol{y}_c^{(U)}$ and $\boldsymbol{y}_c^{(P)}$ can be shown to be

$$\log\{f(\boldsymbol{y}_c^{(P)};\boldsymbol{\theta})\} = \log\{f(\boldsymbol{y}_c^{(U)};\boldsymbol{\kappa})\} + \log\{f(\boldsymbol{y}_1|\boldsymbol{y}_2;\boldsymbol{\theta})\},$$

which is similar to the relation presented in (2.5) based on the Verbyla (1990) conditional derivation of REML, i.e.,

$$\ell(\boldsymbol{\theta};\boldsymbol{L}^T\boldsymbol{y}_o) = \ell(\boldsymbol{\kappa};\boldsymbol{y}_2) + \ell(\boldsymbol{\theta};\boldsymbol{y}_1|\boldsymbol{y}_2).$$

This suggests that using the complete data specification $\boldsymbol{y}_c^{(P)}$ would be similar to using $\ell(\boldsymbol{\theta};\boldsymbol{L}^T\boldsymbol{y}_o)$ for the REML estimation of variance parameters. While it is

possible to use both for the REML estimation of variance parameters the task is simplified when considering the complete data specification $\boldsymbol{y}_c^{(U)}$ or using the residual log-likelihood function $\ell(\boldsymbol{\kappa}; \boldsymbol{y}_2)$. The advantages of using $\boldsymbol{y}_c^{(U)}$ compared to $\boldsymbol{y}_c^{(P)}$ are more apparent when considering the E-step of a REML EM algorithm based on the complete data specification $\boldsymbol{y}_c^{(U)}$.

4.2 REML EM algorithm E-step

The complete data log-density function for $\boldsymbol{y}_c^{(U)}$ can be partitioned as

$$\log\{f(\boldsymbol{y}_c^{(U)}; \boldsymbol{\kappa})\} = \log\{f(\boldsymbol{L}_2^T \boldsymbol{e}; \sigma^2, \boldsymbol{\phi})\} + \log\{f(\boldsymbol{u}; \boldsymbol{\gamma})\},$$

where

$$\log\{f(\boldsymbol{L}_2^T \boldsymbol{e}; \sigma^2, \boldsymbol{\phi})\} = -\frac{1}{2}\Big[\log\{\det(\boldsymbol{L}_2^T \boldsymbol{R} \boldsymbol{L}_2)\} + (\boldsymbol{y}_o - \boldsymbol{Z}\boldsymbol{u})^T \boldsymbol{S} (\boldsymbol{y}_o - \boldsymbol{Z}\boldsymbol{u})\Big],$$
$$\log\{f(\boldsymbol{u}; \boldsymbol{\gamma})\} = -\frac{1}{2}\Big[\log\{\det(\boldsymbol{G})\} + \boldsymbol{u}^T \boldsymbol{G}^{-1} \boldsymbol{u}\Big].$$

The E-step involves taking the conditional expectation of (4.2) over $k(\boldsymbol{u}|\boldsymbol{y}_2; \boldsymbol{\kappa}^{(w)})$ where $\boldsymbol{\kappa}^{(w)}$ is the w-th iterate of $\boldsymbol{\kappa}$. We can write

$$\begin{aligned} Q^{(U)}(\boldsymbol{\kappa}; \boldsymbol{\kappa}^{(w)}) &= E\big[\log\{f(\boldsymbol{y}_c^{(U)}; \boldsymbol{\kappa})\} | \boldsymbol{y}_2; \boldsymbol{\kappa}^{(w)}\big] \\ &= E\big[\log\{f(\boldsymbol{L}_2^T \boldsymbol{e}; \sigma^2, \boldsymbol{\phi})\} | \boldsymbol{y}_2; \boldsymbol{\kappa}^{(w)}\big] + E\big[\log\{f(\boldsymbol{u}; \boldsymbol{\gamma})\} | \boldsymbol{y}_2; \boldsymbol{\kappa}^{(w)}\big] \\ &= Q^{(U)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)}) + Q^{(U)}(\boldsymbol{\gamma}; \boldsymbol{\kappa}^{(w)}). \end{aligned}$$

Using Result A.3.2 we have

$$\begin{aligned} Q^{(U)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)}) &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{L}_2^T \boldsymbol{R} \boldsymbol{L}_2)\} + (\boldsymbol{y}_o - \boldsymbol{Z}\tilde{\boldsymbol{u}}^{(w)})^T \boldsymbol{S} (\boldsymbol{y}_o - \boldsymbol{Z}\tilde{\boldsymbol{u}}^{(w)}) \\ &\quad + \mathrm{tr}\Big(\boldsymbol{Z}^T \boldsymbol{S} \boldsymbol{Z} \boldsymbol{C}^{ZZ^{(w)}}\Big)\Big], \qquad (4.3) \\ Q^{(U)}(\boldsymbol{\gamma}; \boldsymbol{\kappa}^{(w)}) &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{G})\} + \tilde{\boldsymbol{u}}^{(w)^T} \boldsymbol{G}^{-1} \tilde{\boldsymbol{u}}^{(w)} + \mathrm{tr}\Big(\boldsymbol{G}^{-1} \boldsymbol{C}^{ZZ^{(w)}}\Big)\Big], \qquad (4.4) \end{aligned}$$

where

$$\tilde{\boldsymbol{u}}^{(w)} = E\big(\boldsymbol{u}|\boldsymbol{y}_2;\boldsymbol{\kappa}^{(w)}\big) = \boldsymbol{G}^{(w)}\boldsymbol{Z}^T\boldsymbol{P}^{(w)}\boldsymbol{y}_o.$$

It is worth noting that $Q^{(U)}(\boldsymbol{\gamma};\boldsymbol{\kappa}^{(w)})$ is equivalent to $Q^{(P)}(\boldsymbol{\gamma};\boldsymbol{\kappa}^{(w)})$.

A feature of $Q^{(U)}(\boldsymbol{\kappa};\boldsymbol{\kappa}^{(w)})$ is that it requires knowledge of only one conditional distribution, namely $\boldsymbol{u}|\boldsymbol{y}_2$. This differs from the $Q(.)$ function associated with the complete data specification $\boldsymbol{y}_c^{(P)}$ which requires knowledge of both $\boldsymbol{u}|\boldsymbol{y}_2$ and $\boldsymbol{e}|\boldsymbol{y}_2$. This difference is notable because the $Q(.)$ function associated with $\boldsymbol{y}_c^{(U)}$ involves computing the traces of two matrices both of order b whereas the $Q(.)$ function associated with $\boldsymbol{y}_c^{(P)}$ also involves computing the trace of two matrices, one of order b but the other of order n. Depending on the linear mixed model being fitted the computing time associated with $\boldsymbol{y}_c^{(U)}$ can be significantly less than that associated with $\boldsymbol{y}_c^{(P)}$.

Although we do not consider in this thesis models where the E-step is intractable and where a stochastic variant of a REML EM algorithm is used to approximate the $Q(.)$ function, the use of the complete data specification $\boldsymbol{y}_c^{(U)}$ in these cases might be of some interest as approximating the $Q(.)$ function would only require drawing samples from one conditional distribution rather than the two required when using $\boldsymbol{y}_c^{(P)}$.

4.3 REML EM algorithm M-step

We consider the updating equations for σ^2, γ_{kl}, and ϕ_{ij} in turn.

4.3.1 Update for σ^2

Noting that $\boldsymbol{R} = \sigma^2\boldsymbol{\Sigma}$ and differentiating $Q^{(U)}(\sigma^2,\boldsymbol{\phi};\boldsymbol{\kappa}^{(w)})$ with respect to σ^2 we have

$$\begin{aligned}\frac{\partial Q^{(U)}(\sigma^2,\boldsymbol{\phi};\boldsymbol{\kappa}^{(w)})}{\partial\sigma^2} &= -\frac{1}{2}\Big\{\frac{n-t}{\sigma^2} - \frac{1}{(\sigma^2)^2}(\boldsymbol{y}_o - \boldsymbol{Z}\tilde{\boldsymbol{u}}^{(w)})^T\boldsymbol{U}(\boldsymbol{y}_o - \boldsymbol{Z}\tilde{\boldsymbol{u}}^{(w)}) \\ &\quad -\frac{1}{(\sigma^2)^2}\mathrm{tr}\Big(\boldsymbol{Z}^T\boldsymbol{U}\boldsymbol{Z}\boldsymbol{C}^{ZZ^{(w)}}\Big)\Big\},\end{aligned}$$

where $\boldsymbol{U} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}$. Equating to zero we have

$$\sigma^{2^{(w+1)}} = \frac{1}{n-t}\left\{(\boldsymbol{y}_o - \boldsymbol{Z}\tilde{\boldsymbol{u}}^{(w)})^T\boldsymbol{U}(\boldsymbol{y}_o - \boldsymbol{Z}\tilde{\boldsymbol{u}}^{(w)}) + \mathrm{tr}\Big(\boldsymbol{Z}^T\boldsymbol{U}\boldsymbol{Z}\boldsymbol{C}^{ZZ^{(w)}}\Big)\right\}. \tag{4.5}$$

4.3.2 Update for γ_{gh}

Since $Q^{(U)}(\boldsymbol{\gamma};\boldsymbol{\kappa}^{(w)})$ is equivalent to $Q^{(P)}(\boldsymbol{\gamma};\boldsymbol{\kappa}^{(w)})$ the update for γ_{gh}, $g = 1,\ldots,b$, $h = 1,\ldots,b$, $g \geq h$ is the same as that presented in Section 3.6.2. It is reproduced here for convenience. Differentiating $Q^{(P)}(\boldsymbol{\gamma};\boldsymbol{\kappa}^{(w)})$ with respect to $\boldsymbol{\gamma}_{gh}$ we have

$$\begin{aligned}\frac{\partial Q^{(P)}(\boldsymbol{\gamma};\boldsymbol{\kappa}^{(w)})}{\partial\gamma_{gh}} &= -\frac{1}{2}\left\{\mathrm{tr}\left(\boldsymbol{G}^{-1}\frac{\partial\boldsymbol{G}}{\partial\gamma_{gh}}\right) - \tilde{\boldsymbol{u}}^{(w)^T}\boldsymbol{G}^{-1}\frac{\partial\boldsymbol{G}}{\partial\gamma_{gh}}\boldsymbol{G}^{-1}\tilde{\boldsymbol{u}}^{(w)}\right.\\ &\qquad \left.-\mathrm{tr}\left(\boldsymbol{G}^{-1}\frac{\partial\boldsymbol{G}}{\partial\gamma_i}\boldsymbol{G}^{-1}\boldsymbol{C}^{ZZ^{(w)}}\right)\right\}.\end{aligned}$$

The updating formula for γ_{gh} depends on the specific form of $\boldsymbol{G}$. For a one-way random effects model $\boldsymbol{G}$ has the simple form $\boldsymbol{G} = \sigma_u^2\boldsymbol{I}_b$. In this case the updating equation is

$$\sigma_u^{2^{(w+1)}} = \frac{1}{b}\left\{\tilde{\boldsymbol{u}}^{(w)^T}\tilde{\boldsymbol{u}}^{(w)} + \mathrm{tr}\Big(\boldsymbol{C}^{ZZ^{(w)}}\Big)\right\}. \tag{4.6}$$

4.3.3 Update for ϕ_{ij}

To find an update for ϕ_{ij}, $i = 1,\ldots,n$, $j = 1,\ldots,n$, $i \geq j$ involves differentiating $Q^{(U)}(\sigma^2,\boldsymbol{\phi};\boldsymbol{\kappa}^{(w)})$ with respect to ϕ_{ij}. Noting that $\boldsymbol{R} = \sigma^2\boldsymbol{\Sigma}$ and using Result A.4.1 and Result A.4.3 we have

$$\begin{aligned}\frac{\partial Q^{(U)}(\sigma^2,\boldsymbol{\phi};\boldsymbol{\kappa}^{(w)})}{\partial\phi_{ij}} &= -\frac{1}{2}\left\{\mathrm{tr}\Big(\boldsymbol{U}\frac{\partial\boldsymbol{\Sigma}}{\partial\phi_{ij}}\Big) - \frac{1}{\sigma^2}(\boldsymbol{y}_o - \boldsymbol{Z}\tilde{\boldsymbol{u}}^{(w)})^T\boldsymbol{U}\frac{\partial\boldsymbol{\Sigma}}{\partial\phi_{ij}}\boldsymbol{U}(\boldsymbol{y}_o - \boldsymbol{Z}\tilde{\boldsymbol{u}}^{(w)})\right.\\ &\qquad \left.-\frac{1}{\sigma^2}\mathrm{tr}\Big(\boldsymbol{Z}^T\boldsymbol{U}\frac{\partial\boldsymbol{\Sigma}}{\partial\phi_{ij}}\boldsymbol{U}\boldsymbol{Z}\boldsymbol{C}^{ZZ^{(w)}}\Big)\right\},\end{aligned}$$

where $\boldsymbol{U} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}$.The updating formula for ϕ_{ij} depends on the specific form of $\boldsymbol{\Sigma}$. In this thesis we only consider models where

$\boldsymbol{\Sigma} = \boldsymbol{I}_n$.

4.4 Complete data information matrix

We use the superscript (U) when considering the complete and missing data information matrices associated with $\boldsymbol{y}_c^{(U)}$. The complete and missing data information matrices are symmetric and in the interests of parsimony we only present the upper right triangle for each.

The complete data information matrix based on $\boldsymbol{y}_c^{(U)}$ is

$$\boldsymbol{\mathcal{I}}_c^{(U)}(\boldsymbol{\kappa}^{(w)}; \boldsymbol{y}_c^{(U)}) = \begin{pmatrix} \mathcal{I}^{(U)}_{c_{(\sigma^2)^2}} & \mathbf{0} & \boldsymbol{\mathcal{I}}^{(U)}_{c_{\sigma^2\boldsymbol{\phi}}} \\ & \boldsymbol{\mathcal{I}}^{(U)}_{c_{\boldsymbol{\gamma},\boldsymbol{\gamma}}} & \mathbf{0} \\ & & \boldsymbol{\mathcal{I}}^{(U)}_{c_{\boldsymbol{\phi},\boldsymbol{\phi}}} \end{pmatrix},$$

where the zeroes in the complete data information matrix are due to the partitioning of $Q^{(U)}(\boldsymbol{\kappa}; \boldsymbol{\kappa}^{(w)})$ into $Q^{(U)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})$ and $Q^{(U)}(\boldsymbol{\gamma}; \boldsymbol{\kappa}^{(w)})$. Using the Oakes (1999) result presented in (1.13) we have

$$\begin{aligned}
\mathcal{I}^{(U)}_{c_{(\sigma^2)^2}} &= -\left\langle \frac{\partial^2 Q^{(U)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})}{\partial(\sigma^2)^2} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}} \\
\boldsymbol{\mathcal{I}}^{(U)}_{c_{\boldsymbol{\gamma},\boldsymbol{\gamma}}} &= -\left\langle \frac{\partial^2 Q^{(U)}(\boldsymbol{\gamma}; \boldsymbol{\kappa}^{(w)})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}} \\
\boldsymbol{\mathcal{I}}^{(U)}_{c_{\boldsymbol{\phi},\boldsymbol{\phi}}} &= -\left\langle \frac{\partial^2 Q^{(U)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}} \\
\boldsymbol{\mathcal{I}}^{(U)}_{c_{\sigma^2,\boldsymbol{\phi}}} &= -\left\langle \frac{\partial^2 Q^{(U)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})}{\partial\sigma^2\partial\boldsymbol{\phi}^T} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}}.
\end{aligned}$$

We present each in turn and note any difference with their counterpart associated with $\boldsymbol{y}_c^{(P)}$.

4.4.1 Element $\mathcal{I}^{(U)}_{c_{(\sigma^2)^2}}$

It can be shown that

$$\mathcal{I}^{(U)}_{c_{(\sigma^2)^2}} = \frac{n-t}{2\left(\sigma^{2(w)}\right)^2} + \frac{1}{\left(\sigma^{2(w)}\right)^2}\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\boldsymbol{P}^{(w)}\boldsymbol{y}_o - \frac{1}{\left(\sigma^{2(w)}\right)^2}\mathrm{tr}\Big(\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\Big). \quad (4.7)$$

Comparing $\mathcal{I}^{(U)}_{c_{(\sigma^2)^2}}$ and $\mathcal{I}^{(P)}_{c_{(\sigma^2)^2}}$ we have

$$\mathcal{I}^{(U)}_{c_{(\sigma^2)^2}} - \mathcal{I}^{(P)}_{c_{(\sigma^2)^2}} = -\frac{t}{2\left(\sigma^{2(w)}\right)^2}.$$

For $t \geq 1$ and $\sigma^{2(w)} > 0$ we have the following inequality

$$\mathcal{I}^{(U)}_{c_{(\sigma^2)^2}} < \mathcal{I}^{(P)}_{c_{(\sigma^2)^2}}.$$

4.4.2 Elements of $\mathcal{I}^{(U)}_{c_{\gamma,\gamma}}$

Since $\mathcal{I}^{(U)}_{c_{\gamma,\gamma}}$ involves the second partial differentials of $Q^{(U)}(\boldsymbol{\gamma};\boldsymbol{\kappa}^{(w)})$ and $Q^{(U)}(\boldsymbol{\gamma};\boldsymbol{\kappa}^{(w)})$ is equivalent to $Q^{(P)}(\boldsymbol{\gamma};\boldsymbol{\kappa}^{(w)})$ then the elements for $\mathcal{I}^{(U)}_{c_{\gamma,\gamma}}$ are equivalent to those for $\mathcal{I}^{(P)}_{c_{\gamma,\gamma}}$ which are presented in (3.16).

4.4.3 Elements of $\mathcal{I}^{(U)}_{c_{\phi,\phi}}$

For $i = 1,\ldots,n$, $j = 1,\ldots,n$, $i \geq j$ and $k = 1,\ldots,n$, $l = 1,\ldots,n$, $k \geq l$ it can be shown that

$$\begin{aligned}
\mathcal{I}^{(U)}_{c_{\phi_{ij},\phi_{kl}}} = \frac{1}{2}\Bigg\{ & \left(\sigma^{2(w)}\right)^2\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\boldsymbol{S}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{kl}^{(w)}}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\
& + \left(\sigma^{2(w)}\right)^2\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{kl}^{(w)}}\boldsymbol{S}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\
& - \sigma^{2(w)}\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\frac{\partial^2\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}\partial\phi_{kl}^{(w)}}\boldsymbol{P}^{(w)}\boldsymbol{y}_o + \sigma^{2(w)}\mathrm{tr}\Big(\frac{\partial^2\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}\partial\phi_{kl}^{(w)}}\boldsymbol{P}^{(w)}\Big) \\
& + \left(\sigma^{2(w)}\right)^2\mathrm{tr}\Bigg\{\Big(\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{kl}^{(w)}}\boldsymbol{S}^{(w)} - \frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{kl}^{(w)}}\boldsymbol{P}^{(w)}\Big)^T\Big(\boldsymbol{S}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}} - \boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\Big)\Bigg\} \\
& - \left(\sigma^{2(w)}\right)^2\mathrm{tr}\Big(\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{kl}^{(w)}}\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\Big)\Bigg\}. \qquad (4.8)
\end{aligned}$$

The difference between $\mathcal{I}^{(U)}_{c_{\phi_{ij},\phi_{kl}}}$ and $\mathcal{I}^{(P)}_{c_{\phi_{ij},\phi_{kl}}}$ is that $\boldsymbol{R}^{-1^{(w)}}$ replaces $\boldsymbol{S}^{(w)}$ in three terms in the later.

4.4.4 Elements of $\mathcal{I}^{(U)}_{c_{\sigma^2,\phi}}$

For $i = 1, \ldots, n$, $j = 1, \ldots, n$, $i \geq j$ it can be shown that

$$\begin{aligned} \mathcal{I}^{(U)}_{c_{\sigma^{2(w)}\phi^{(w)}_{ij}}} &= \frac{1}{2}\Bigg\{\boldsymbol{y}_o^T \boldsymbol{P}^{(w)} \frac{\partial \boldsymbol{\Sigma}^{(w)}}{\partial \phi^{(w)}_{ij}} \boldsymbol{P}^{(w)} \boldsymbol{y}_o \\ &\quad + \mathrm{tr}\Big(\frac{\partial \boldsymbol{\Sigma}^{(w)}}{\partial \phi^{(w)}_{ij}} \boldsymbol{S}^{(w)}\Big) - \mathrm{tr}\Big(\frac{\partial \boldsymbol{\Sigma}^{(w)}}{\partial \phi^{(w)}_{ij}} \boldsymbol{P}^{(w)}\Big)\Bigg\}. \end{aligned} \tag{4.9}$$

The difference between $\mathcal{I}^{(U)}_{c_{\sigma^{2(w)}\phi^{(w)}_{ij}}}$ and $\mathcal{I}^{(P)}_{c_{\sigma^{2(w)}\phi^{(w)}_{ij}}}$ is that $\boldsymbol{R}^{-1^{(w)}}$ replaces $\boldsymbol{S}^{(w)}$ in one trace term in the latter.

This concludes the section on the complete data information matrix.

4.5 Missing data information matrix

The missing data information matrix based on $\boldsymbol{y}_c^{(U)}$ and using the Oakes (1999) result presented in (1.13) is

$$\mathcal{I}^{(U)}_m(\boldsymbol{\kappa}^{(w)}; \boldsymbol{y}_c^{(U)}) = \begin{pmatrix} \mathcal{I}^{(U)}_{m_{(\sigma^2)^2}} & \mathcal{I}^{(U)}_{m_{\sigma^2,\boldsymbol{\gamma}}} & \mathcal{I}^{(U)}_{m_{\sigma^2,\boldsymbol{\phi}}} \\ & \mathcal{I}^{(U)}_{m_{\boldsymbol{\gamma},\boldsymbol{\gamma}}} & \mathcal{I}^{(U)}_{m_{\sigma^2,\boldsymbol{\gamma}}} \\ & & \mathcal{I}^{(U)}_{m_{\boldsymbol{\phi},\boldsymbol{\phi}}} \end{pmatrix},$$

where

$$
\begin{aligned}
\mathcal{I}^{(U)}_{m_{(\sigma^2)^2}} &= \left\langle \frac{\partial^2 Q^{(U)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})}{\partial \sigma^2 \partial \sigma^{2(w)}} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}} \\
\mathcal{I}^{(U)}_{m_{\boldsymbol{\gamma},\boldsymbol{\gamma}}} &= \left\langle \frac{\partial^2 Q^{(U)}(\boldsymbol{\gamma}; \boldsymbol{\kappa}^{(w)})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^{(w)^T}} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}} \\
\mathcal{I}^{(U)}_{m_{\boldsymbol{\phi},\boldsymbol{\phi}}} &= \left\langle \frac{\partial^2 Q^{(U)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^{(w)^T}} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}} \\
\mathcal{I}^{(U)}_{m_{\sigma^2,\boldsymbol{\gamma}}} &= \left\langle \frac{\partial^2 Q^{(U)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})}{\partial \sigma^2 \partial \boldsymbol{\gamma}^{(w)^T}} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}} \\
\mathcal{I}^{(U)}_{m_{\sigma^2,\boldsymbol{\phi}}} &= \left\langle \frac{\partial^2 Q^{(U)}(\sigma^2, \boldsymbol{\phi}; \boldsymbol{\kappa}^{(w)})}{\partial \sigma^2 \partial \boldsymbol{\phi}^{(w)^T}} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}} \\
\mathcal{I}^{(U)}_{m_{\boldsymbol{\gamma},\boldsymbol{\phi}}} &= \left\langle \frac{\partial^2 Q^{(U)}(\boldsymbol{\gamma}; \boldsymbol{\kappa}^{(w)})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\phi}^{(w)^T}} \right\rangle_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^{(w)}}.
\end{aligned}
$$

We make use of Result A.4.3, Result A.3.11, Result A.3.12, Result A.3.13, Result A.3.14 and present each in turn and note any differences with their counterpart associated with $\boldsymbol{y}_c^{(P)}$.

4.5.1 Element $\mathcal{I}^{(U)}_{m_{(\sigma^2)^2}}$

It can be shown that

$$
\begin{aligned}
\mathcal{I}^{(P)}_{m_{(\sigma^2)^2}} &= -\frac{1}{2\left(\sigma^{2(w)}\right)^2}\Big\{-(n-t) - 2\boldsymbol{y}_o^T \boldsymbol{P}^{(w)} \boldsymbol{R}^{(w)} \boldsymbol{P}^{(w)} \boldsymbol{y}_o \\
&\quad + 2\boldsymbol{y}_o^T \boldsymbol{P}^{(w)} \boldsymbol{R}^{(w)} \boldsymbol{P}^{(w)} \boldsymbol{R}^{(w)} \boldsymbol{P}^{(w)} \boldsymbol{y}_o \\
&\quad + 2\mathrm{tr}\Big(\boldsymbol{P}^{(w)} \boldsymbol{R}^{(w)}\Big) - \mathrm{tr}\Big(\boldsymbol{P}^{(w)} \boldsymbol{R}^{(w)} \boldsymbol{P}^{(w)} \boldsymbol{R}^{(w)}\Big)\Big\}. \qquad (4.10)
\end{aligned}
$$

Comparing $\mathcal{I}^{(U)}_{m_{(\sigma^2)^2}}$ and $\mathcal{I}^{(P)}_{m_{(\sigma^2)^2}}$ we have

$$
\mathcal{I}^{(U)}_{m_{(\sigma^2)^2}} - \mathcal{I}^{(P)}_{m_{(\sigma^2)^2}} = -\frac{t}{2\left(\sigma^{2(w)}\right)^2}.
$$

For $t \geq 1$ and $\sigma^{2(w)} > 0$ we have the following inequality

$$
\mathcal{I}^{(U)}_{m_{(\sigma^2)^2}} < \mathcal{I}^{(P)}_{m_{(\sigma^2)^2}}.
$$

4.5.2 Elements of $\mathcal{I}^{(U)}_{m_{\gamma,\gamma}}$

Since $\mathcal{I}^{(U)}_{m_{\gamma,\gamma}}$ involves the second partial differentials of $Q^{(U)}(\boldsymbol{\gamma};\boldsymbol{\kappa}^{(w)})$ and $Q^{(U)}(\boldsymbol{\gamma};\boldsymbol{\kappa}^{(w)})$ is equivalent to $Q^{(P)}(\boldsymbol{\gamma};\boldsymbol{\kappa}^{(w)})$ then the elements for $\mathcal{I}^{(U)}_{m_{\gamma,\gamma}}$ are equivalent to those for $\mathcal{I}^{(P)}_{m_{\gamma,\gamma}}$ which are presented in (3.20).

4.5.3 Elements of $\mathcal{I}^{(U)}_{m_{\phi,\phi}}$

For $i = 1,\ldots,n$, $j = 1,\ldots,n$, $i \geq j$ and $k = 1,\ldots,n$, $l = 1,\ldots,n$, $k \geq l$ it can be shown that

$$\begin{aligned}
\mathcal{I}^{(U)}_{m_{\phi_{ij},\phi_{kl}}} = & -\frac{\left(\sigma^{2(w)}\right)^2}{2}\Bigg[\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{kl}}\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{ij}}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\
& +\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{ij}}\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{kl}}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\
& -\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{kl}}\boldsymbol{S}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{ij}}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\
& -\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{ij}}\boldsymbol{S}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{kl}}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\
& -\mathrm{tr}\Big\{\Big(\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{kl}}\boldsymbol{S}^{(w)}-\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{kl}}\boldsymbol{P}^{(w)}\Big)^T\Big(\boldsymbol{S}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{ij}}-\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi^{(w)}_{ij}}\Big)\Big\}\Bigg].
\end{aligned} \tag{4.11}$$

The difference between $\mathcal{I}^{(U)}_{m_{\phi_{ij},\phi_{kl}}}$ and $\mathcal{I}^{(P)}_{m_{\phi_{ij},\phi_{kl}}}$ is that $\boldsymbol{R}^{-1(w)}$ replaces $\boldsymbol{S}^{(w)}$ in three terms in the latter.

4.5.4 Elements of $\mathcal{I}^{(U)}_{m_{\sigma^2,\gamma}}$

For $g = 1,\ldots,b$, $h = 1,\ldots,b$, $g \geq h$ it can be shown that

$$\begin{aligned}\mathcal{I}^{(U)}_{m_{\sigma^2},\gamma_{gh}} &= -\frac{1}{2\sigma^{2(w)}}\Bigg\{\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\boldsymbol{Z}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma_{gh}^{(w)}}\boldsymbol{Z}^T\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ &\quad +\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\boldsymbol{P}^{(w)}\boldsymbol{Z}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma_{gh}^{(w)}}\boldsymbol{Z}^T\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ &\quad -\mathrm{tr}\Big(\boldsymbol{P}^{(w)}\boldsymbol{Z}\frac{\partial\boldsymbol{G}^{(w)}}{\partial\gamma_{gh}^{(w)}}\boldsymbol{Z}^T\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\Big)\Bigg\}. \qquad (4.12)\end{aligned}$$

This is equivalent to $\mathcal{I}^{(P)}_{m_{\sigma^2},\gamma_{gh}}$ as presented in (3.22).

4.5.5 Elements of $\mathcal{I}^{(U)}_{m_{\sigma^2},\phi}$

For $i = 1,\ldots,n$, $j = 1,\ldots,n$, $i \geq j$ it can be shown that

$$\begin{aligned}\mathcal{I}^{(U)}_{m_{\sigma^2},\phi_{ij}} &= -\frac{1}{2}\Bigg\{\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ &\quad +\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\boldsymbol{P}^{(w)}\boldsymbol{y}_o - 2\boldsymbol{y}_o^T\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\boldsymbol{P}^{(w)}\boldsymbol{y}_o \\ &\quad -\mathrm{tr}\Big(\boldsymbol{S}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\Big) + 2\mathrm{tr}\Big(\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\boldsymbol{P}^{(w)}\Big) \\ &\quad -\mathrm{tr}\Big(\boldsymbol{P}^{(w)}\frac{\partial\boldsymbol{\Sigma}^{(w)}}{\partial\phi_{ij}^{(w)}}\boldsymbol{P}^{(w)}\boldsymbol{R}^{(w)}\Big)\Bigg\}. \qquad (4.13)\end{aligned}$$

The difference between $\mathcal{I}^{(U)}_{m_{\sigma^2},\phi_{ij}}$ and $\mathcal{I}^{(U)}_{m_{\sigma^2},\phi_{ij}}$ is that $\boldsymbol{R}^{-1(w)}$ replaces $\boldsymbol{S}^{(w)}$ in a single trace term in the latter.

4.5.6 Elements of $\mathcal{I}^{(U)}_{m_{\gamma},\phi}$

Since $\mathcal{I}^{(U)}_{m_{\gamma},\phi}$ involves the second partial differentials of $Q^{(U)}(\boldsymbol{\gamma};\boldsymbol{\kappa}^{(w)})$ and $Q^{(U)}(\boldsymbol{\gamma};\boldsymbol{\kappa}^{(w)})$ is equivalent to $Q^{(P)}(\boldsymbol{\gamma};\boldsymbol{\kappa}^{(w)})$ then the elements for $\mathcal{I}^{(U)}_{m_{\gamma},\phi}$ are equivalent to those for $\mathcal{I}^{(P)}_{m_{\gamma},\phi}$ which are presented in (3.24).

4.6 Comparing rates of convergence between $\boldsymbol{y}_c^{(U)}$ and $\boldsymbol{y}_c^{(P)}$

For ease of notation we define

$$\begin{aligned}\boldsymbol{\mathcal{I}}_c^{(U)} &= \boldsymbol{\mathcal{I}}_c^{(U)}(\boldsymbol{\kappa}^\star; \boldsymbol{y}_c^{(U)}),\\ \boldsymbol{\mathcal{I}}_c^{(P)} &= \boldsymbol{\mathcal{I}}_c^{(P)}(\boldsymbol{\kappa}^\star; \boldsymbol{y}_c^{(P)}),\\ \mathrm{I}_o &= \mathrm{I}_o(\boldsymbol{\kappa}^\star; \boldsymbol{y}_o),\end{aligned}$$

where $\boldsymbol{\kappa}^{(w)}$ has converged to $\boldsymbol{\kappa}^\star$. We define the rate matrices

$$\begin{aligned}\boldsymbol{\mathcal{J}}^{(U)}(\boldsymbol{\kappa}^\star) &= \boldsymbol{I}_d - \boldsymbol{\mathcal{I}}_c^{(U)^{-1}}\mathrm{I}_o,\\ \boldsymbol{\mathcal{J}}^{(P)}(\boldsymbol{\kappa}^\star) &= \boldsymbol{I}_d - \boldsymbol{\mathcal{I}}_c^{(P)^{-1}}\mathrm{I}_o,\end{aligned}$$

which can be written as

$$\begin{aligned}\boldsymbol{I}_d - \boldsymbol{\mathcal{J}}^{(U)}(\boldsymbol{\kappa}^\star) &= \boldsymbol{\mathcal{I}}_c^{(U)^{-1}}\mathrm{I}_o,\\ \boldsymbol{I}_d - \boldsymbol{\mathcal{J}}^{(P)}(\boldsymbol{\kappa}^\star) &= \boldsymbol{\mathcal{I}}_c^{(P)^{-1}}\mathrm{I}_o.\end{aligned}$$

The rate of convergence associated with the complete data specification $\boldsymbol{y}_c^{(U)}$ and $\boldsymbol{y}_c^{(P)}$ is $\lambda_{\min}\{\boldsymbol{I}_d - \boldsymbol{\mathcal{J}}^{(U)}(\boldsymbol{\kappa}^\star)\}$ and $\lambda_{\min}\{\boldsymbol{I}_d - \boldsymbol{\mathcal{J}}^{(P)}(\boldsymbol{\kappa}^\star)\}$ respectively (McLachlan and Krishnan, 1997, Section 3.9.1). The term $\lambda_{\min}(.)$ refers to the minimum eigenvalue. We aim to show

$$\lambda_{\min}\{\boldsymbol{I}_d - \boldsymbol{\mathcal{J}}^{(U)}(\boldsymbol{\kappa}^\star)\} \geq \lambda_{\min}\{\boldsymbol{I}_d - \boldsymbol{\mathcal{J}}^{(P)}(\boldsymbol{\kappa}^\star)\}, \tag{4.14}$$

i.e., the rate of convergence of a REML EM algorithm using the complete data specification $\boldsymbol{y}_c^{(U)}$ is faster than the rate of convergence of a REML EM algorithm using the complete data specification $\boldsymbol{y}_c^{(P)}$. We have not been able to prove that this inequality always holds. However, in our experience the rate of convergence when using $\boldsymbol{y}_c^{(U)}$ has always been faster than that when using $\boldsymbol{y}_c^{(P)}$. By considering a Loewner partial ordering (see Appendix C) of the complete information matrices

associated with $\boldsymbol{y}_c^{(U)}$ and $\boldsymbol{y}_c^{(P)}$ we are able to present the conditions under which the inequality in (4.14) will hold.

If we consider the difference between $\boldsymbol{\mathcal{I}}_c^{(P)}$ and $\boldsymbol{\mathcal{I}}_c^{(U)}$ to be positive definite then we can define the following Loewner partial ordering

$$\boldsymbol{\mathcal{I}}_c^{(U)} \preceq \boldsymbol{\mathcal{I}}_c^{(P)}.$$

Using Corollary C.1 (a) we can write

$$\boldsymbol{\mathcal{I}}_c^{(U)^{-1}} \succeq \boldsymbol{\mathcal{I}}_c^{(P)^{-1}}.$$

As a consequence of Corollary C.1 (b) we have the following inequalities between the complete data information matrices associated with $\boldsymbol{y}_c^{(U)}$ and $\boldsymbol{y}_c^{(P)}$

$$\det(\boldsymbol{\mathcal{I}}_c^{(U)^{-1}}) \geq \det(\boldsymbol{\mathcal{I}}_c^{(P)^{-1}}), \tag{4.15}$$

$$\frac{1}{\det(\boldsymbol{\mathcal{I}}_c^{(U)})} \geq \frac{1}{\det(\boldsymbol{\mathcal{I}}_c^{(P)})}, \tag{4.16}$$

$$\det(\boldsymbol{\mathcal{I}}_c^{(U)}) \leq \det(\boldsymbol{\mathcal{I}}_c^{(P)}), \tag{4.17}$$

where going from (4.15) to (4.16) uses the second relation in Result A.3.3. Multiplying both sides of (4.17) by $\det(\mathrm{I}_o)$ and using the third relation in Result A.3.3 we have

$$\begin{aligned}
\det(\boldsymbol{\mathcal{I}}_c^{(U)^{-1}})\det(\mathrm{I}_o) &\geq \det(\boldsymbol{\mathcal{I}}_c^{(P)^{-1}})\det(\mathrm{I}_o),\\
\det(\boldsymbol{\mathcal{I}}_c^{(U)^{-1}}\mathrm{I}_o) &\geq \det(\boldsymbol{\mathcal{I}}_c^{(P)^{-1}}\mathrm{I}_o),\\
\det\{\boldsymbol{I}_d - \boldsymbol{\mathcal{J}}^{(U)}(\boldsymbol{\kappa}^\star)\} &\geq \det\{\boldsymbol{I}_d - \boldsymbol{\mathcal{J}}^{(P)}(\boldsymbol{\kappa}^\star)\}.
\end{aligned}$$

As a consequence of Corollary C.1 (b) and (c) it follows that

$$\boldsymbol{\lambda}_k\{\boldsymbol{I}_d - \boldsymbol{\mathcal{J}}^{(U)}(\boldsymbol{\kappa}^\star)\} \geq \boldsymbol{\lambda}_k\{\boldsymbol{I}_d - \boldsymbol{\mathcal{J}}^{(P)}(\boldsymbol{\kappa}^\star)\},$$

where $\boldsymbol{\lambda}_k(.)$ is a vector of k eigenvalues. If the k eigenvalues of $\boldsymbol{I}_d - \boldsymbol{\mathcal{J}}^{(U)}(\boldsymbol{\kappa}^\star)$ and $\boldsymbol{I}_d - \boldsymbol{\mathcal{J}}^{(P)}(\boldsymbol{\kappa}^\star)$ are arranged in descending order then

$$\lambda_{\min}\{\boldsymbol{I}_d - \boldsymbol{\mathcal{J}}^{(U)}(\boldsymbol{\kappa}^\star)\} \geq \lambda_{\min}\{\boldsymbol{I}_d - \boldsymbol{\mathcal{J}}^{(P)}(\boldsymbol{\kappa}^\star)\},$$

which concludes the proof. The key part of the proof is the inequality in (4.17). Elements of the complete data information matrices for both $\boldsymbol{y}_c^{(U)}$ and $\boldsymbol{y}_c^{(P)}$ can be used to check if the difference between $\boldsymbol{\mathcal{I}}_c^{(P)}$ and $\boldsymbol{\mathcal{I}}_c^{(U)}$ is positive definite for particular cases.

4.7 Example: lamb weight data

We apply a REML EM algorithm using the complete data specifications $\boldsymbol{y}_c^{(U)}$ and $\boldsymbol{y}_c^{(P)}$ to the lamb weight data presented in Callanan and Harville (1991) and provided in Table 4.1.

The data consists of birth weights (in pounds) of single-birth male lambs that are the progeny of 62 ewes. The age of the ewes are categorised into 3 groups ($1 = 1-2$ years; $2 = 2-3$ years; $3 =$ over 3 years). The lambs were sired by one of 23 rams which belonged to 5 different population lines. A possible model for lamb birth weight involves fitting sire as a random effect and age and line as fixed effects (along with the overall mean). Using symbolic notation this model can be written as

$$\text{weight} \sim \text{mean} + \text{age} + \text{line} + \textbf{sire},$$

where bold type is used to indicate that a term is being fitted as a random effect. Writing this model in matrix notation we have

$$\boldsymbol{y}_o = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e},$$

where $\boldsymbol{y}_o$ is a 62×1 vector of observed birth weights, $\boldsymbol{\beta}$ is a 7×1 vector of fixed effects comprising the overall mean and the main effects of age and line. The 62×7 design matrix $\boldsymbol{X}$ is of full-column rank. The vector $\boldsymbol{u}$ is a 23×1 vector of random sire effects and $\boldsymbol{Z}$ is its 62×23 design matrix. The vector $\boldsymbol{e}$ is a 62×1 vector of residuals. We assume that $\boldsymbol{u}$ and $\boldsymbol{e}$ are independent of each other and have the following distributions

Table 4.1: Lamb weight data.

Sire	DAMage	Line	weight	Sire	DAMage	Line	weight
11	1	1	6.2	33	3	3	13.0
12	1	1	13.0	34	2	3	12.0
13	1	1	9.5	41	1	4	9.2
13	1	1	10.1	41	1	4	10.6
13	1	1	11.4	41	1	4	10.6
13	2	1	11.8	41	3	4	7.7
13	3	1	12.9	41	3	4	10.0
13	3	1	13.1	41	3	4	11.2
14	1	1	10.4	42	1	4	10.2
14	2	1	8.5	42	1	4	10.9
21	3	2	13.5	43	1	4	11.7
22	2	2	10.1	43	3	4	9.9
22	3	2	11.0	51	1	5	11.7
22	3	2	14.0	51	1	5	12.6
22	3	2	15.5	52	1	5	9.0
23	1	2	12.0	52	3	5	11.0
24	1	2	11.5	53	3	5	9.0
24	3	2	10.8	53	3	5	12.0
31	2	3	9.0	54	3	5	9.9
31	3	3	9.5	55	2	5	13.5
31	3	3	12.6	56	2	5	10.9
32	1	3	11.0	56	3	5	5.9
32	2	3	10.1	57	2	5	10.0
32	2	3	11.7	57	2	5	12.7
32	3	3	8.5	57	3	5	13.2
32	3	3	8.8	57	3	5	13.3
32	3	3	9.9	58	1	5	10.7
32	3	3	10.9	58	1	5	11.0
32	3	3	11.0	58	1	5	12.5
32	3	3	13.9	58	3	5	9.0
33	1	3	11.6	58	3	5	10.2

$$\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma_s^2 \boldsymbol{I}_{23}),$$
$$\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{62}),$$

so that $\boldsymbol{R} = \sigma^2 \boldsymbol{I}_{62}$ and $\boldsymbol{G} = \sigma_s^2 \boldsymbol{I}_{23}$. The updating equation for σ^2 is provided in (3.14) and (4.5) for $\boldsymbol{y}_c^{(P)}$ and $\boldsymbol{y}_c^{(U)}$ respectively. The updating equation for σ_s^2 is provided in (4.6) and applies to both complete data specifications. We define $\boldsymbol{\kappa} = (\sigma^2, \sigma_s^2)^T$ to be the vector of variance parameters to be estimated using a REML EM algorithm. The convergence criterion of Foulley and Van Dyk (2000) was used so convergence is defined as being achieved on the w-th iteration if

$$\sqrt{\frac{(\boldsymbol{\kappa}^{(w+1)} - \boldsymbol{\kappa}^{(w)})^T (\boldsymbol{\kappa}^{(w+1)} - \boldsymbol{\kappa}^{(w)})}{\boldsymbol{\kappa}^{T(w)} \boldsymbol{\kappa}^{(w)}}} < \varepsilon, \tag{4.18}$$

where $\varepsilon = 10^{-8}$. Two sets of starting values were considered for both complete data specifications. The first set of starting values were $\boldsymbol{\kappa}^{(0)} = (\sigma^{2(0)}, \sigma_s^{2(0)})^T = (1, 0.01)^T$ and the second set of starting values were $\boldsymbol{\kappa}^{(0)} = (\sigma^{2(0)}, \sigma_s^{2(0)})^T = (1, 5)^T$. The statistical software package R (R Development Core Team, 2011) was used to implement both REML EM algorithms. REML estimates of σ^2 and σ_s^2, denoted $\sigma^{2\star}$ and $\sigma_s^{2\star}$, the number of iterations until convergence and the rate of convergence for both complete data specifications and both sets of starting values are provided in Table 4.2.

Table 4.2: Results for two REML EM algorithms applied to the lamb weight data.

Complete data	$\sigma^{2\star}$	$\sigma_s^{2\star}$	Iterations	Rate of convergence
	$\sigma^{2(0)} = 1, \sigma_s^{2(0)} = 0.01$			
$\boldsymbol{y}_c^{(P)}$	2.9615972	0.5170759	1296	0.9630649
$\boldsymbol{y}_c^{(U)}$	2.9615972	0.5170759	1296	0.9629925
	$\sigma^{2(0)} = 1, \sigma_s^{2(0)} = 5$			
$\boldsymbol{y}_c^{(P)}$	2.9615966	0.5170773	342	0.9630648
$\boldsymbol{y}_c^{(U)}$	2.9615966	0.5170773	341	0.9629924

For both sets of starting values and for both complete data specifications the observed information matrix was found to be (to 4 decimal places)

$$\mathrm{I}_o(\boldsymbol{\kappa}^\star; \boldsymbol{y}_o) = \begin{pmatrix} 2.6598 & 1.1753 \\ 1.1753 & 2.13448 \end{pmatrix},$$

which can be used to calculate the standard errors associated with the REML estimates of σ^2 and σ_s^2. It is worth noting that the empirical rate of convergence can be used to check the rate of convergence and the observed information matrix.

Based on the rates of convergence in Table 4.2 there appears to be little to gain in using $\boldsymbol{y}_c^{(U)}$ instead of $\boldsymbol{y}_c^{(P)}$. Our experience with a number of data sets is that the difference in the rate of convergence between $\boldsymbol{y}_c^{(U)}$ and $\boldsymbol{y}_c^{(P)}$ when using a REML EM algorithm is only minor. However, this difference can be significant when a parameter expanded version of the REML EM algorithm, referred to as a REML PX EM algorithm, is implemented. We consider a REML PX EM algorithm for both complete data specifications in the next chapter.

Chapter 5

Two REML PX EM algorithms for the linear mixed model

A common criticism of the EM algorithm is that it can be slow to converge. Since the Dempster et al. (1977) paper there have been a number of schemes suggested for speeding up the convergence of the EM algorithm. In this chapter we will consider one such scheme, known as the parameter expanded (PX) EM algorithm (Liu et al., 1998). From a practical implementation point of view a PX EM algorithm is preferred to an EM algorithm as there is nothing to lose and often a lot to gain. Nothing to lose in the sense that the PX EM algorithm, like the EM algorithm, has monotonic convergence. Furthermore, the E-step of the PX EM algorithm is exactly the same as the E-step of the EM algorithm. The gains in using the PX EM algorithm compared to the EM algorithm are in rates of convergence. Liu et al. (1998) show that the PX EM algorithm has a rate of convergence that is no slower than the EM algorithm but is often much faster. In this chapter we will consider REML PX EM algorithms for the two complete data specifications considered in the previous chapter.

5.1 Linear mixed model specification

For the PX EM algorithm the linear mixed model is formulated by introducing the auxillary parameter $\boldsymbol{\lambda}$ (say) so that the linear mixed model in (2.1) is expanded to

$$\boldsymbol{y}_o = \boldsymbol{X}\boldsymbol{\beta}_* + \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{\epsilon}, \tag{5.1}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are $n \times t$ and $n \times b$ design matrices respectively, and $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}(\boldsymbol{\lambda})$ is a $b \times b$ real invertible matrix which is a function of the $v \times 1$ auxillary parameter vector $\boldsymbol{\lambda}$. It is assumed that the joint distribution of $\boldsymbol{f}$ and $\boldsymbol{\epsilon}$ is

$$\begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N\left\{ \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_* \end{pmatrix} \right\},$$

where $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{d})$ and $\boldsymbol{R}_* = \sigma_*^2 \boldsymbol{\Sigma}_*(\boldsymbol{\phi}_*)$ are symmetric positive definite matrices. The marginal distribution of $\boldsymbol{y}_o$ under the parameter expanded model is

$$\boldsymbol{y}_o \sim N(\boldsymbol{X}\boldsymbol{\beta}_*, \boldsymbol{H}_*)$$

where $\boldsymbol{H}_* = \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{D}\boldsymbol{\Lambda}^T\boldsymbol{Z}^T + \boldsymbol{R}_*$. The expanded parameter set for the model in (5.1) is $\boldsymbol{\mathcal{K}} = (\boldsymbol{\kappa}_*, \boldsymbol{\lambda}^T)^T$ where $\boldsymbol{\kappa}_* = (\sigma_*^2, \boldsymbol{d}^T, \boldsymbol{\phi}_*^T)^T$. The subscript $*$ is used to distinguish parameters associated with the linear mixed model specified in (5.1) and the linear mixed model specified in (2.1). The expanded parameter vector $\boldsymbol{\mathcal{K}}$ must satisfy two conditions:

1. $\boldsymbol{\mathcal{K}}$ can be reduced to the original parameter vector $\boldsymbol{\kappa} = (\sigma^2, \boldsymbol{\gamma}^T, \boldsymbol{\phi}^T)^T$ by a many-to-one reduction function. For the model specified in (5.1) this reduction function, labelled $\mathcal{R}$ is $\boldsymbol{\kappa} = (\sigma^2, \boldsymbol{\gamma}^T, \boldsymbol{\phi}^T)^T = \mathcal{R}(\boldsymbol{\mathcal{K}}) = (\sigma_*^2, \text{vech}(\boldsymbol{\Lambda}\boldsymbol{D}\boldsymbol{\Lambda}^T), \boldsymbol{\phi}_*^T)^T$.

2. When the auxillary parameter $\boldsymbol{\lambda}$ is set to its null value $\boldsymbol{\lambda}_0$ (say) the expanded complete data model is reduced to the original complete data model. If we define $\boldsymbol{u} = \boldsymbol{\Lambda}\boldsymbol{f}$ then $\boldsymbol{G} = \boldsymbol{\Lambda}\boldsymbol{D}\boldsymbol{\Lambda}^T$ and when $\boldsymbol{\Lambda}(\boldsymbol{\lambda}_0) = \boldsymbol{I}_b$ the expanded model in (5.1) is reduced to the original model in (2.1) with $\boldsymbol{u} = \boldsymbol{f}$ and $\boldsymbol{G} = \boldsymbol{D}$.

Liu et al. (1998) state that

> The underlying statistical principle of PX-EM is to perform a 'covariance adjustment' to correct the M step, capitalising on extra information captured in the imputed complete data...

An intuitively appealing way to consider this 'covariance adjustment' is to compare the joint distribution of the incomplete and missing data under the parameter expanded model, i.e.,

$$\begin{pmatrix} \boldsymbol{y}_o \\ \boldsymbol{f} \end{pmatrix} \sim N\left\{ \begin{pmatrix} \boldsymbol{X}\boldsymbol{\beta}_* \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{H}_* & \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{D} \\ \boldsymbol{D}\boldsymbol{\Lambda}^T\boldsymbol{Z}^T & \boldsymbol{D} \end{pmatrix} \right\}$$

with the joint distribution of the incomplete and missing data under the original model, i.e.,

$$\begin{pmatrix} \boldsymbol{y}_o \\ \boldsymbol{u} \end{pmatrix} \sim N\left\{ \begin{pmatrix} \boldsymbol{X}\boldsymbol{\beta} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{H} & \boldsymbol{Z}\boldsymbol{G} \\ \boldsymbol{G}\boldsymbol{Z}^T & \boldsymbol{D} \end{pmatrix} \right\}.$$

The idea of parameter expansion as a 'covariance adjustment' is easily seen by the introduction of $\boldsymbol{\Lambda}$ into the covariance of the former. As Liu et al. (1998) note, the covariance of the later does not generally reflect the "true" relationship between the imputed missing data $\tilde{\boldsymbol{u}}^{(w+1)}$ and the known incomplete data vector $\boldsymbol{y}_o$, particularly in the initial iterations of an EM algorithm. The auxillary parameter $\boldsymbol{\lambda}$ is used to "correct" the estimate of $\boldsymbol{G}$ to produce an adjusted estimate. At convergence there is no longer any need to adjust the estimate of $\boldsymbol{G}$ and $\boldsymbol{\Lambda}(\boldsymbol{\lambda}) = \boldsymbol{I}_b$.

5.2 REML PX EM algorithm for $\boldsymbol{y}_c^{(P)}$

A REML PX EM algorithm for the linear mixed model using the complete data specification $\boldsymbol{y}_c^{(P)} = (\boldsymbol{y}_o^T, \boldsymbol{f}^T)^T$ was considered by Foulley and Van Dyk (2000) who used the random effects approach, and Cullis et al. (2004) and Knight (2008) who used the fixed effects approach. Our exposition of a REML PX EM algorithm using $\boldsymbol{y}_c^{(P)}$ is brief as the detail can be found in the aforementioned publications.

5.2.1 Complete data log density function

The complete data log-density function for $\boldsymbol{y}_c^{(P)}$ is

$$\begin{aligned} \log\{f_X(\boldsymbol{y}_c^{(P)}; \boldsymbol{\Theta})\} &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{R}_*)\} + \log\{\det(\boldsymbol{D})\} \\ &\quad + \boldsymbol{f}^T\boldsymbol{D}^{-1}\boldsymbol{f} + \boldsymbol{\epsilon}^T\boldsymbol{R}_*^{-1}\boldsymbol{\epsilon}\Big], \end{aligned} \tag{5.2}$$

where $\boldsymbol{\epsilon} = \boldsymbol{y}_o - \boldsymbol{X}\boldsymbol{\beta}_* - \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{f}$ and $\boldsymbol{\Theta} = (\boldsymbol{\beta}_*^T, \boldsymbol{\kappa}_*^T, \boldsymbol{\lambda}^T)^T$. The subscript X will be used to distinguish between the complete data log-density function and the $Q(.)$ function associated with the model in (5.1) and the model in (2.1).

5.2.2 REML PX EM algorithm E-step

The complete data log-density function in (5.2) can be partitioned

$$\log\{f_X(\boldsymbol{y}_c^{(P)}; \boldsymbol{\Theta})\} = \log\{f_X(\boldsymbol{\epsilon}; \boldsymbol{\Theta})\} + \log\{f_X(\boldsymbol{f}; \boldsymbol{d})\},$$

where

$$\log\{f_X(\boldsymbol{\epsilon}; \boldsymbol{\Theta})\} = -\frac{1}{2}\Big[\log\{\det(\boldsymbol{R}_*)\} + \boldsymbol{\epsilon}^T\boldsymbol{R}_*^{-1}\boldsymbol{\epsilon}\Big],$$
$$\log\{f_X(\boldsymbol{f}; \boldsymbol{d})\} = -\frac{1}{2}\Big[\log\{\det(\boldsymbol{D})\} + \boldsymbol{f}^T\boldsymbol{D}^{-1}\boldsymbol{f}\Big].$$

The E-step of the REML PX EM algorithm involves taking the expectation of (5.2) over $k(\boldsymbol{f}|\boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)})$ where $\boldsymbol{\mathcal{K}}^{(w)} = \{\boldsymbol{\kappa}_*^{(w)^T}, (\boldsymbol{\lambda} = \boldsymbol{\lambda}_0)^T\}^T$. We can write

$$\begin{aligned} Q_X^{(P)}(\boldsymbol{\mathcal{K}}; \boldsymbol{\mathcal{K}}^{(w)}) &= E\big[\log\{f_X(\boldsymbol{y}_c^{(P)}; \boldsymbol{\Theta})\}|\boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\big] \\ &= E\big[\log\{f_X(\boldsymbol{\epsilon}; \boldsymbol{\Theta})\}|\boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\big] + E\big[\log\{f_X(\boldsymbol{f}; \boldsymbol{d})\}|\boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\big] \\ &= Q_X^{(P)}(\sigma_*^2, \boldsymbol{\phi}_*, \boldsymbol{\lambda}; \boldsymbol{\mathcal{K}}^{(w)}) + Q_X^{(P)}(\boldsymbol{d}; \boldsymbol{\mathcal{K}}^{(w)}), \end{aligned}$$

where

$$Q_X^{(P)}(\sigma_*^2, \boldsymbol{\phi}_*, \boldsymbol{\lambda}; \boldsymbol{\mathcal{K}}^{(w)}) = -\frac{1}{2}\Big[\log\{\det(\boldsymbol{R}_*)\} + E\big(\boldsymbol{\epsilon}^T\boldsymbol{R}_*^{-1}\boldsymbol{\epsilon}|\boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\big)\Big],$$
$$Q_X^{(P)}(\boldsymbol{d}; \boldsymbol{\mathcal{K}}^{(w)}) = -\frac{1}{2}\Big[\log\{\det(\boldsymbol{D})\} + E\big(\boldsymbol{f}^T\boldsymbol{D}^{-1}\boldsymbol{f}|\boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\big)\Big].$$

When $\boldsymbol{\Lambda}(\boldsymbol{\lambda}_0) = \boldsymbol{I}_b$ the REML PX EM algorithm E-step is the same as the REML EM algorithm E-step.

5.2.3 REML PX EM algorithm M-step

The M-step of the REML PX EM algorithm involves maximising $Q_X^{(P)}(\mathcal{K};\mathcal{K}^{(w)})$ with respect to $\mathcal{K} = (\sigma_*^2, \boldsymbol{d}^T, \boldsymbol{\phi}_*^T, \boldsymbol{\lambda}^T)^T$ and then applying the reduction function $\mathcal{R}(\mathcal{K}) = (\sigma_*^2, \text{vech}(\boldsymbol{\Lambda}\boldsymbol{D}\boldsymbol{\Lambda}^T), \boldsymbol{\phi}_*^T)^T$ to obtain estimates of $\boldsymbol{\kappa} = (\sigma^2, \boldsymbol{\gamma}^T, \boldsymbol{\phi}^T)^T$.

Update for σ_*^2

The update for σ_*^2 is the same as that presented in (3.14), i.e.,

$$\sigma_*^{2^{(w+1)}} = \frac{1}{n}\left\{\tilde{\boldsymbol{e}}^{(w)^T}\boldsymbol{\Sigma}^{-1^{(w)}}\tilde{\boldsymbol{e}}^{(w)} + \text{tr}\left(\boldsymbol{\Sigma}^{-1^{(w)}}\boldsymbol{W}\boldsymbol{C}^{-1^{(w)}}\boldsymbol{W}^T\right)\right\}. \tag{5.3}$$

Applying the reduction function we have $\sigma^{2^{(w+1)}} = \sigma_*^{2^{(w+1)}}$.

Update for d_{gh}

The update for d_{gh} is the same as that for γ_{gh} when using the REML EM algorithm except we substitute $\boldsymbol{D}$ for $\boldsymbol{G}$, i.e., differentiating $Q_X^{(P)}(\boldsymbol{d};\mathcal{K}^{(w)})$ with respect to d_{gh} we have

$$\frac{\partial Q_X^{(P)}(\boldsymbol{d};\mathcal{K}^{(w)})}{\partial d_{gh}} = -\frac{1}{2}\left\{\text{tr}\left(\boldsymbol{D}^{-1}\frac{\partial \boldsymbol{D}}{\partial d_{gh}}\right) - E\left(\boldsymbol{f}^T\boldsymbol{D}^{-1}\frac{\partial \boldsymbol{D}}{\partial d_{gh}}\boldsymbol{D}^{-1}\boldsymbol{f}\middle|\boldsymbol{y}_2;\mathcal{K}^{(w)}\right)\right\}.$$

The updating formula for d_{gh} depends on the specific form of $\boldsymbol{D}$. If we define the update for $\boldsymbol{D}$ as $\boldsymbol{D}^{(w+1)}$ then applying the reduction function we have $\boldsymbol{\gamma}^{(w+1)} = \text{vech}(\boldsymbol{\Lambda}^{(w+1)}\boldsymbol{D}^{(w+1)}\boldsymbol{\Lambda}^{(w+1)^T})$. For a one-way random effects model $\boldsymbol{D}$ has the form $\boldsymbol{D} = d\boldsymbol{I}_b$. In this case the updating equation is

$$d^{(w+1)} = \frac{1}{b}\left\{\tilde{\boldsymbol{u}}^{(w)^T}\tilde{\boldsymbol{u}}^{(w)} + \text{tr}\left(\boldsymbol{C}^{ZZ^{(w)}}\right)\right\}, \tag{5.4}$$

and the reduction function is $\sigma_u^{2^{(w+1)}} = \lambda^{2^{(w+1)}}d^{(w+1)}$.

Update for $\phi_{*_{ij}}$

The update for $\phi_{*_{ij}}$ is the same as that for the REML EM algorithm. Differentiating $Q_X^{(P)}(\sigma_*^2, \boldsymbol{\phi}_*, \boldsymbol{\lambda}; \boldsymbol{\mathcal{K}}_*^{(w)})$ with respect to $\phi_{*_{ij}}$ we have

$$\frac{\partial Q_X^{(P)}(\sigma_*^2, \boldsymbol{\phi}_*, \boldsymbol{\lambda}; \boldsymbol{\mathcal{K}}^{(w)})}{\partial \phi_{*_{ij}}} = -\frac{1}{2}\left\{ \mathrm{tr}\left(\boldsymbol{\Sigma}_*^{-1} \frac{\partial \boldsymbol{\Sigma}_*}{\partial \phi_{*_{ij}}} \right) - \frac{1}{\sigma_*^2} E\left(\boldsymbol{\epsilon}^T \boldsymbol{\Sigma}_*^{-1} \frac{\partial \boldsymbol{\Sigma}_*}{\partial \phi_{*_{ij}}} \boldsymbol{\Sigma}_*^{-1} \boldsymbol{\epsilon} \middle| \boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)} \right) \right\}.$$

The updating equation for $\phi_{*_{ij}}$ depends on the specific form of $\boldsymbol{\Sigma}_*$. If we define the update for $\phi_{*_{ij}}$ as $\phi_{*_{ij}}^{(w+1)}$ then applying the reduction function we have $\phi_{ij}^{(w+1)} = \phi_{*_{ij}}^{(w+1)}$.

Update for $\boldsymbol{\lambda}$

To form an update for $\boldsymbol{\lambda}$ we re-write $Q_X^{(P)}(\sigma_*^2, \boldsymbol{\phi}_*, \boldsymbol{\lambda}; \boldsymbol{\mathcal{K}}^{(w)})$ using the identity

$$\begin{aligned} \boldsymbol{\epsilon}^T \boldsymbol{R}_*^{-1} \boldsymbol{\epsilon} &= (\boldsymbol{y}_o - \boldsymbol{X}\boldsymbol{\beta}_* - \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{f})^T \boldsymbol{R}_*^{-1} (\boldsymbol{y}_o - \boldsymbol{X}\boldsymbol{\beta}_* - \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{f}) \\ &= (\boldsymbol{y}_o - \boldsymbol{X}\boldsymbol{\beta}_*)^T \boldsymbol{R}_*^{-1} (\boldsymbol{y}_o - \boldsymbol{X}\boldsymbol{\beta}_*) - 2(\boldsymbol{y}_o - \boldsymbol{X}\boldsymbol{\beta}_*)^T \boldsymbol{R}_*^{-1} \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{f} \\ &\quad + \boldsymbol{f}^T \boldsymbol{\Lambda}^T \boldsymbol{Z}^T \boldsymbol{R}_*^{-1} \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{f}. \end{aligned} \tag{5.5}$$

Performing the REML PX EM E-step on the right hand side of (5.5) requires the joint conditional distribution

$$\begin{pmatrix} \boldsymbol{y}_o - \boldsymbol{X}\boldsymbol{\beta} \\ \boldsymbol{u} \end{pmatrix} \middle| \boldsymbol{y}_2 \sim N\left\{ \begin{pmatrix} \boldsymbol{H}\boldsymbol{P}\boldsymbol{y}_o \\ \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{y}_o \end{pmatrix}, \begin{pmatrix} \boldsymbol{X}\boldsymbol{C}^{XX}\boldsymbol{X}^T & -\boldsymbol{X}\boldsymbol{C}^{XZ} \\ -\boldsymbol{C}^{ZX}\boldsymbol{X}^T & \boldsymbol{C}^{ZZ} \end{pmatrix} \right\}.$$

The update for $\boldsymbol{\lambda}$ can be expressed as

$$\boldsymbol{\lambda}^{(w+1)} = \boldsymbol{A}^{-1^{(w)}} \boldsymbol{b}^{(w)},$$

where the elements of the $b^2 \times b^2$ matrix $\boldsymbol{A}^{(w)}$ are

$$a_{gh,qr}^{(w)} = \tilde{\boldsymbol{u}}^{(w)^T} \frac{\partial \boldsymbol{\Lambda}^{(w)^T}}{\partial \lambda_{gh}^{(w)}} \boldsymbol{Z}^T \boldsymbol{R}^{-1^{(w)}} \boldsymbol{Z} \frac{\partial \boldsymbol{\Lambda}^{(w)}}{\partial \lambda_{qr}^{(w)}} \tilde{\boldsymbol{u}}^{(w)} + \mathrm{tr}\left(\frac{\partial \boldsymbol{\Lambda}^{(w)^T}}{\partial \lambda_{gh}^{(w)}} \boldsymbol{Z}^T \boldsymbol{R}^{-1^{(w)}} \boldsymbol{Z} \frac{\partial \boldsymbol{\Lambda}^{(w)}}{\partial \lambda_{qr}^{(w)}} \boldsymbol{C}^{ZZ^{(w)}} \right), \tag{5.6}$$

and the elements of the $b^2 \times 1$ vector $\boldsymbol{b}^{(w)}$ are

$$b_{gh}^{(w)} = \tilde{\boldsymbol{u}}^{(w)} \frac{\partial \boldsymbol{\Lambda}^{(w)^T}}{\partial \lambda_{gh}^{(w)}} \boldsymbol{Z}^T \boldsymbol{R}^{-1^{(w)}} (\boldsymbol{y}_o - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(w)}) - \mathrm{tr}\left(\frac{\partial \boldsymbol{\Lambda}^{(w)^T}}{\partial \lambda_{gh}^{(w)}} \boldsymbol{Z}^T \boldsymbol{R}^{-1^{(w)}} \boldsymbol{X} \boldsymbol{C}^{XZ^{(w)}} \right), \quad (5.7)$$

where $\hat{\boldsymbol{\beta}}^{(w)}$ is the generalised least squares estimate of $\boldsymbol{\beta}$ at the w-th iterate. This completes the M-step.

5.2.4 Rate matrix

Liu et al. (1998) note that the convergence of $\boldsymbol{\kappa}^{(w)}$ determines the rate of convergence of the the PX EM algorithm. The rate matrix for the REML PX EM algorithm using the complete data specification $\boldsymbol{y}_c^{(P)}$ is

$$\mathcal{J}(\boldsymbol{\kappa}^\star, \boldsymbol{\lambda}_0) = \boldsymbol{I}_{d+v} - \{\boldsymbol{\mathcal{I}}_c^{(P)}(\boldsymbol{\kappa}^\star, \boldsymbol{\lambda}_0)\}^{-1} \mathrm{I}_o(\boldsymbol{\kappa}^\star, \boldsymbol{\lambda}_0), \quad (5.8)$$

where $\boldsymbol{\kappa}^{(w)}$ has converged to $\boldsymbol{\kappa}^\star$, and where

$$\mathrm{I}_o(\boldsymbol{\kappa}^\star, \boldsymbol{\lambda}_0) = \begin{pmatrix} \mathrm{I}_o(\boldsymbol{\kappa}^\star) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix},$$

and

$$\boldsymbol{\mathcal{I}}_c^{(P)}(\boldsymbol{\kappa}^\star, \boldsymbol{\lambda}_0) = \begin{pmatrix} \boldsymbol{\mathcal{I}}_{c_{\boldsymbol{\kappa},\boldsymbol{\kappa}}}^{(P)} & \boldsymbol{\mathcal{I}}_{c_{\boldsymbol{\kappa},\boldsymbol{\lambda}}}^{(P)} \\ & \boldsymbol{\mathcal{I}}_{c_{\boldsymbol{\lambda},\boldsymbol{\lambda}}}^{(P)} \end{pmatrix}_{\boldsymbol{\kappa} = \boldsymbol{\kappa}^\star, \boldsymbol{\lambda} = \boldsymbol{\lambda}_0}.$$

The terms $\boldsymbol{\mathcal{I}}_{c_{\boldsymbol{\kappa},\boldsymbol{\kappa}}}^{(P)}$ and $\mathrm{I}_o(\boldsymbol{\kappa}^\star)$ can be calculated using the complete and missing data information matrices presented in Section 3.7 and Section 3.8. The elements of $\boldsymbol{\mathcal{I}}_{c_{\boldsymbol{\kappa},\boldsymbol{\lambda}}}^{(P)}$, which comprise $\boldsymbol{\mathcal{I}}_{c_{\sigma^2,\boldsymbol{\lambda}}}^{(P)}$, $\boldsymbol{\mathcal{I}}_{c_{\gamma,\boldsymbol{\lambda}}}^{(P)}$, and $\boldsymbol{\mathcal{I}}_{c_{\boldsymbol{\phi},\boldsymbol{\lambda}}}^{(P)}$, and the elements of $\boldsymbol{\mathcal{I}}_{c_{\boldsymbol{\lambda},\boldsymbol{\lambda}}}^{(P)}$ are presented below.

Elements of $\boldsymbol{\mathcal{I}}_{c_{\sigma^2,\boldsymbol{\lambda}}}^{(P)}$

For $g = 1, \ldots, b$, $h = 1, \ldots, b$ it can be shown that

$$\mathcal{I}^{(P)}_{c_{\sigma^2},\lambda_{gh}} = \frac{1}{\left(\sigma^{2\star}\right)^2}\Bigg[\mathrm{tr}\Big\{\boldsymbol{\Sigma}^{-1^\star}\boldsymbol{Z}\frac{\partial\boldsymbol{\Lambda}^\star}{\partial\lambda^\star_{gh}}\tilde{\boldsymbol{u}}^{\star^T}(\boldsymbol{y}_o-\boldsymbol{X}\hat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}^{-1^\star}\boldsymbol{Z}\frac{\partial\boldsymbol{\Lambda}^\star}{\partial\lambda^\star_{gh}}\boldsymbol{C}^{ZX^\star}\boldsymbol{X}^T\Big\}$$
$$-\tilde{\boldsymbol{u}}^{\star^T}\boldsymbol{Z}^T\boldsymbol{\Sigma}^{-1^\star}\boldsymbol{Z}\frac{\partial\boldsymbol{\Lambda}^\star}{\partial\lambda^\star_{gh}}\tilde{\boldsymbol{u}}^\star - \mathrm{tr}\Big(\boldsymbol{Z}^T\boldsymbol{\Sigma}^{-1^\star}\boldsymbol{Z}\frac{\partial\boldsymbol{\Lambda}^\star}{\partial\lambda^\star_{gh}}\boldsymbol{C}^{ZZ^\star}\Big)\Bigg].$$

Elements of $\mathcal{I}^{(P)}_{c_{\gamma},\lambda}$

For $g = 1,\ldots,b$, $h = 1,\ldots,b$ it can be shown that

$$\mathcal{I}^{(P)}_{c_{\gamma_{gh}},\lambda_{gh}} = \tilde{\boldsymbol{u}}^{\star^T}\boldsymbol{G}^{-1^\star}\frac{\partial\boldsymbol{G}^\star}{\partial\gamma^\star_{gh}}\boldsymbol{G}^{-1^\star}\frac{\partial\boldsymbol{\Lambda}^\star}{\partial\lambda^\star_{gh}}\tilde{\boldsymbol{u}}^\star + \mathrm{tr}\Big(\boldsymbol{G}^{-1^\star}\frac{\partial\boldsymbol{G}^\star}{\partial\gamma^\star_{gh}}\boldsymbol{G}^{-1^\star}\frac{\partial\boldsymbol{\Lambda}^\star}{\partial\lambda^\star_{gh}}\boldsymbol{C}^{ZZ^\star}\Big).$$

Elements of $\mathcal{I}^{(P)}_{c_{\phi},\lambda}$

For $i = 1,\ldots,n$, $j = 1,\ldots,n$, $g = 1,\ldots,b$, $h = 1,\ldots,b$ it can be shown that

$$\mathcal{I}^{(P)}_{c_{\phi_{ij}},\lambda_{gh}} = \frac{1}{\sigma^{2\star}}\Bigg[\mathrm{tr}\Big\{\boldsymbol{\Sigma}^{-1^\star}\frac{\partial\boldsymbol{\Sigma}^\star}{\partial\phi^\star_{ij}}\boldsymbol{\Sigma}^{-1^\star}\boldsymbol{Z}\frac{\partial\boldsymbol{\Lambda}^\star}{\partial\lambda^\star_{gh}}\tilde{\boldsymbol{u}}^{\star^T}(\boldsymbol{y}_o-\boldsymbol{X}\hat{\boldsymbol{\beta}})$$
$$-\boldsymbol{\Sigma}^{-1^\star}\frac{\partial\boldsymbol{\Sigma}^\star}{\partial\phi^\star_{ij}}\boldsymbol{\Sigma}^{-1^\star}\boldsymbol{Z}\frac{\partial\boldsymbol{\Lambda}^\star}{\partial\lambda^\star_{gh}}\boldsymbol{C}^{ZX^\star}\boldsymbol{X}^T\Big\}$$
$$-\tilde{\boldsymbol{u}}^{\star^T}\boldsymbol{Z}^T\boldsymbol{\Sigma}^{-1^\star}\frac{\partial\boldsymbol{\Sigma}^\star}{\partial\phi^\star_{ij}}\boldsymbol{\Sigma}^{-1^\star}\boldsymbol{Z}\frac{\partial\boldsymbol{\Lambda}^\star}{\partial\lambda^\star_{gh}}\tilde{\boldsymbol{u}}^\star$$
$$-\mathrm{tr}\Big(\boldsymbol{Z}^T\boldsymbol{\Sigma}^{-1^\star}\frac{\partial\boldsymbol{\Sigma}^\star}{\partial\phi^\star_{ij}}\boldsymbol{\Sigma}^{-1^\star}\boldsymbol{Z}\frac{\partial\boldsymbol{\Lambda}^\star}{\partial\lambda^\star_{gh}}\boldsymbol{C}^{ZZ^\star}\Big)\Bigg].$$

Elements of $\mathcal{I}^{(P)}_{c_{\lambda},\lambda}$

We assume that the elements of the real invertible matrix $\boldsymbol{\Lambda}$ are λ_{gh}, $g = 1,\ldots,b$, $h = 1,\ldots,b$ and that the second order partial derivative of $\boldsymbol{\Lambda}$ can be expressed as

$$\frac{\partial^2\boldsymbol{\Lambda}}{\partial\lambda_{gh}\partial\lambda_{qr}},$$

where $q = 1,\ldots,b$, $r = 1,\ldots,b$. It can be shown that

$$\begin{aligned}\mathcal{I}^{(P)}_{c_{\lambda_{gh},\lambda_{qr}}} &= \tilde{\boldsymbol{u}}^{\star T}\frac{\partial \boldsymbol{\Lambda}^{\star T}}{\partial \lambda^{\star}_{qr}}\boldsymbol{Z}^T\boldsymbol{R}^{-1^{\star}}\boldsymbol{Z}\frac{\partial \boldsymbol{\Lambda}^{\star}}{\partial \lambda^{\star}_{gh}}\tilde{\boldsymbol{u}}^{\star} + \mathrm{tr}\Big(\frac{\partial \boldsymbol{\Lambda}^{\star T}}{\partial \lambda^{\star}_{qr}}\boldsymbol{Z}^T\boldsymbol{R}^{-1^{\star}}\boldsymbol{Z}\frac{\partial \boldsymbol{\Lambda}^{\star}}{\partial \lambda^{\star}_{gh}}\boldsymbol{C}^{ZZ^{\star}}\Big)\\ &+\mathrm{tr}\Big(\frac{\partial \boldsymbol{\Lambda}^{\star}}{\partial \lambda^{\star}_{gh}}\boldsymbol{\Lambda}^{-1^{\star}}\frac{\partial \boldsymbol{\Lambda}^{\star}}{\partial \lambda^{\star}_{qr}}\boldsymbol{\Lambda}^{-1^{\star}}\Big) + \tilde{\boldsymbol{u}}^{\star T}\frac{\partial \boldsymbol{\Lambda}^{\star T}}{\partial \lambda^{\star}_{gh}}\boldsymbol{G}^{-1^{\star}}\frac{\partial \boldsymbol{\Lambda}^{\star}}{\partial \lambda^{\star}_{qr}}\tilde{\boldsymbol{u}}^{\star}\\ &+\mathrm{tr}\Big(\frac{\partial \boldsymbol{\Lambda}^{\star T}}{\partial \lambda^{\star}_{gh}}\boldsymbol{G}^{-1^{\star}}\frac{\partial \boldsymbol{\Lambda}^{\star}}{\partial \lambda^{\star}_{qr}}\boldsymbol{C}^{ZZ^{\star}}\Big).\end{aligned}$$

5.3 REML PX EM algorithm for $\boldsymbol{y}_c^{(U)}$

A REML PX EM algorithm for the linear mixed model using the complete data specification $\boldsymbol{y}_c^{(U)} = (\boldsymbol{y}_2^T, \boldsymbol{f}^T)^T$ has not been considered previously. We present the joint distribution of the incomplete and missing data for the parameter expanded model, the complete data log-density function, the E and M-steps, and the rate matrix for the REML PX EM algorithm using $\boldsymbol{y}_c^{(U)}$.

For the parameter expanded model based on $\boldsymbol{y}_c^{(U)}$ the joint distribution of the incomplete and missing data is

$$\begin{pmatrix}\boldsymbol{y}_2\\ \boldsymbol{f}\end{pmatrix} \sim N\left\{\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{pmatrix}, \begin{pmatrix}\boldsymbol{L}_2^T\boldsymbol{H}_*\boldsymbol{L}_2 & \boldsymbol{L}_2^T\boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{D}\\ \boldsymbol{D}\boldsymbol{\Lambda}^T\boldsymbol{Z}^T\boldsymbol{L}_2 & \boldsymbol{D}\end{pmatrix}\right\} \tag{5.9}$$

where $\boldsymbol{H}_* = \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{D}\boldsymbol{\Lambda}^T\boldsymbol{Z}^T + \boldsymbol{R}_*$. This can be compared to the joint distribution of the incomplete and missing data for the original model presented in (4.1).

5.3.1 Complete data log density function

The complete data log density function for $\boldsymbol{y}_c^{(U)}$ is

$$\begin{aligned}\log\{f_X(\boldsymbol{y}_c^{(U)};\boldsymbol{\mathcal{K}})\} &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{L}_2^T\boldsymbol{R}_*\boldsymbol{L}_2)\} + \log\{\det(\boldsymbol{D})\}\\ &+\boldsymbol{f}^T\boldsymbol{D}^{-1}\boldsymbol{f} + (\boldsymbol{y}_o - \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{f})^T\boldsymbol{S}_*(\boldsymbol{y}_o - \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{f})\Big],\end{aligned}$$

where $\boldsymbol{S}_* = \boldsymbol{R}_*^{-1} - \boldsymbol{R}_*^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{R}_*^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{R}_*^{-1}$. Like the REML EM, a major difference between (5.2) and (5.10) is that the REML PX EM algorithm E-step using

the former requires at least the joint conditional distribution of $\boldsymbol{u}$ and $\boldsymbol{y}_o - \boldsymbol{X}\boldsymbol{\beta}$ given $\boldsymbol{y}_2$ whereas the later only requires $\boldsymbol{u}$ given $\boldsymbol{y}_2$.

5.3.2 REML PX EM algorithm E-step

The complete data log-density function in (5.10) can be partitioned

$$\log\{f_X(\boldsymbol{y}_c^{(U)}; \boldsymbol{\mathcal{K}})\} = \log\{f_X(\boldsymbol{L}_2^T\boldsymbol{\epsilon}; \boldsymbol{\mathcal{K}})\} + \log\{f_X(\boldsymbol{f}; \boldsymbol{d})\},$$

where

$$\begin{aligned}
\log\{f_X(\boldsymbol{L}_2^T\boldsymbol{\epsilon}; \boldsymbol{\mathcal{K}})\} &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{L}_2^T\boldsymbol{R}_*\boldsymbol{L}_2)\} + (\boldsymbol{y}_o - \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{f})^T\boldsymbol{S}_*(\boldsymbol{y}_o - \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{f})\Big], \\
\log\{f_X(\boldsymbol{f}; \boldsymbol{d})\} &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{D})\} + \boldsymbol{f}^T\boldsymbol{D}^{-1}\boldsymbol{f}\Big].
\end{aligned}$$

The E-step of the REML PX EM algorithm involves taking the expectation of (5.10) over $k(\boldsymbol{f}|\boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)})$ where $\boldsymbol{\mathcal{K}}^{(w)} = \{\boldsymbol{\kappa}_*^{(w)^T}, (\boldsymbol{\lambda} = \boldsymbol{\lambda}_0)^T\}^T$. We can write

$$\begin{aligned}
Q_X^{(U)}(\boldsymbol{\mathcal{K}}; \boldsymbol{\mathcal{K}}^{(w)}) &= E\big[\log\{f_X(\boldsymbol{y}_c^{(U)}; \boldsymbol{\mathcal{K}})\}|\boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\big] \\
&= E\big[\log\{f_X(\boldsymbol{L}_2^T\boldsymbol{\epsilon}; \boldsymbol{\mathcal{K}})\}|\boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\big] + E\big[\log\{f_X(\boldsymbol{f}; \boldsymbol{d})\}|\boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\big] \\
&= Q_X^{(U)}(\sigma_*^2, \boldsymbol{\phi}_*, \boldsymbol{\lambda}; \boldsymbol{\mathcal{K}}^{(w)}) + Q_X^{(U)}(\boldsymbol{d}; \boldsymbol{\mathcal{K}}^{(w)}),
\end{aligned}$$

where

$$\begin{aligned}
Q_X^{(U)}(\sigma_*^2, \boldsymbol{\phi}_*, \boldsymbol{\lambda}; \boldsymbol{\mathcal{K}}^{(w)}) &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{L}_2^T\boldsymbol{R}_*\boldsymbol{L}_2)\} \\
&\quad + E\Big\{(\boldsymbol{y}_o - \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{f})^T\boldsymbol{S}_*(\boldsymbol{y}_o - \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{f})|\boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\Big\}\Big], \\
Q_X^{(U)}(\boldsymbol{d}; \boldsymbol{\mathcal{K}}^{(w)}) &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{D})\} + E\big(\boldsymbol{f}^T\boldsymbol{D}^{-1}\boldsymbol{f}|\boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\big)\Big].
\end{aligned}$$

Note that $Q_X^{(U)}(\boldsymbol{d}; \boldsymbol{\mathcal{K}}^{(w)})$ is equivalent to $Q_X^{(P)}(\boldsymbol{d}; \boldsymbol{\mathcal{K}}^{(w)})$. When $\boldsymbol{\Lambda}(\boldsymbol{\lambda}_0) = \boldsymbol{I}_b$ the REML PX EM algorithm E-step is the same as the REML EM algorithm E-step.

5.3.3 REML PX EM algorithm M-step

The M-step of the REML PX EM algorithm involves maximising $Q_X^{(U)}(\boldsymbol{\mathcal{K}}; \boldsymbol{\mathcal{K}}^{(w)})$ with respect to $\boldsymbol{\mathcal{K}} = (\sigma_*^2, \boldsymbol{d}^T, \boldsymbol{\phi}_*^T, \boldsymbol{\lambda}^T)^T$ and then applying the reduction function $\mathcal{R}(\boldsymbol{\mathcal{K}}) = (\sigma_*^2, \text{vech}(\boldsymbol{\Lambda D \Lambda}^T), \boldsymbol{\phi}_*^T)^T$ to obtain estimates of $\boldsymbol{\kappa} = (\sigma^2, \boldsymbol{\gamma}^T, \boldsymbol{\phi}^T)^T$.

Update for σ_*^2

The update for σ_*^2 is the same as that presented in (4.5), i.e.,

$$\sigma_*^{2(w+1)} = \frac{1}{n-t}\left\{(\boldsymbol{y}_o - \boldsymbol{Z}\tilde{\boldsymbol{u}}^{(w)})^T \boldsymbol{U}(\boldsymbol{y}_o - \boldsymbol{Z}\tilde{\boldsymbol{u}}^{(w)}) + \text{tr}\Big(\boldsymbol{Z}^T\boldsymbol{U}\boldsymbol{Z}\boldsymbol{C}^{ZZ^{(w)}}\Big)\right\}, \quad (5.10)$$

where $\boldsymbol{U} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}$. Applying the reduction function we have $\sigma^{2(w+1)} = \sigma_*^{2(w+1)}$.

Update for d_{gh}

Since $Q_X^{(U)}(\boldsymbol{d}; \boldsymbol{\mathcal{K}}^{(w)})$ is equivalent to $Q_X^{(P)}(\boldsymbol{d}; \boldsymbol{\mathcal{K}}^{(w)})$ the update for d_{gh} is the same as that presented in (5.4).

Update for $\phi_{*_{ij}}$

The update for $\phi_{*_{ij}}$ is the same as that for the REML EM algorithm. Differentiating $Q_X^{(U)}(\sigma_*^2, \boldsymbol{\phi}_*, \boldsymbol{\lambda}; \boldsymbol{\mathcal{K}}^{(w)})$ with respect to $\phi_{*_{ij}}$ we have

$$\begin{aligned}\frac{\partial Q_X^{(U)}(\sigma_*^2, \boldsymbol{\phi}_*, \boldsymbol{\lambda}; \boldsymbol{\mathcal{K}}^{(w)})}{\partial \phi_{*_{ij}}} &= -\frac{1}{2}\Bigg[\text{tr}\Big(\boldsymbol{U}_* \frac{\partial \boldsymbol{\Sigma}_*}{\partial \phi_{*_{ij}}}\Big) \\ &\quad -\frac{1}{\sigma_*^2} E\Big\{\Big(\boldsymbol{y}_o - \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{f}\Big)^T \boldsymbol{U}_* \frac{\partial \boldsymbol{\Sigma}_*}{\partial \phi_{*_{ij}}} \boldsymbol{U}_* \Big(\boldsymbol{y}_o - \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{f}\Big) \Big| \boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\Big\}\Bigg].\end{aligned}$$

The updating equation for $\phi_{*_{ij}}$ depends on the specific form of $\boldsymbol{\Sigma}_*$. If we define the update for $\phi_{*_{ij}}$ as $\phi_{*_{ij}}^{(w+1)}$ then applying the reduction function we have $\phi_{ij}^{(w+1)} = \phi_{*_{ij}}^{(w+1)}$.

Update for $\boldsymbol{\lambda}$

To derive an update for the auxillary parameter vector $\boldsymbol{\lambda}$ we consider the $b \times b$ real invertible matrix $\boldsymbol{\Lambda}$ to have elements λ_{gh} and $\boldsymbol{\lambda} = \text{vec}(\boldsymbol{\Lambda})$. To derive an update for $\boldsymbol{\lambda}$ we consider the partial derivative of $Q_X^{(U)}(\sigma_*^2, \boldsymbol{\phi}_*, \boldsymbol{\lambda}; \boldsymbol{\mathcal{K}}^{(w)})$ with respect to λ_{gh}, i.e.,

$$\begin{aligned}\frac{\partial Q_X^{(U)}(\sigma_*^2, \boldsymbol{\phi}_*, \boldsymbol{\lambda}; \boldsymbol{\mathcal{K}}^{(w)})}{\partial \lambda_{gh}} &= -\frac{1}{2}\Big\{-2E\Big(\boldsymbol{y}_o^T \boldsymbol{S}_* \boldsymbol{Z} \frac{\partial \boldsymbol{\Lambda}}{\partial \lambda_{gh}} \boldsymbol{f} | \boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\Big) \\ &\quad +2E\Big(\boldsymbol{f}^T \boldsymbol{\Lambda}^T \boldsymbol{Z}^T \boldsymbol{S}_* \boldsymbol{Z} \frac{\partial \boldsymbol{\Lambda}}{\partial \lambda_{gh}} \boldsymbol{f} | \boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\Big)\Big\}.\end{aligned}$$

Equating to zero we have

$$E\Big(\boldsymbol{f}^T \boldsymbol{\Lambda}^T \boldsymbol{Z}^T \boldsymbol{S}_* \boldsymbol{Z} \frac{\partial \boldsymbol{\Lambda}}{\partial \lambda_{gh}} \boldsymbol{f} | \boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\Big) = E\Big(\boldsymbol{y}_o^T \boldsymbol{S}_* \boldsymbol{Z} \frac{\partial \boldsymbol{\Lambda}}{\partial \lambda_{gh}} \boldsymbol{f} | \boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\Big). \tag{5.11}$$

We can express $\boldsymbol{\Lambda}$ as

$$\boldsymbol{\Lambda} = \sum_{q=1}^{b} \sum_{r=1}^{b} \lambda_{qr} \boldsymbol{J}_{qr},$$

where $\boldsymbol{J}_{qr}$ is a $b \times b$ indicator matrix with a unit entry in row q, column r and zeros elsewhere. We can re-write the left hand side of (5.11) as

$$\begin{aligned}E\Big(\boldsymbol{f}^T \boldsymbol{\Lambda}^T \boldsymbol{Z}^T \boldsymbol{S}_* \boldsymbol{Z} \frac{\partial \boldsymbol{\Lambda}}{\partial \lambda_{gh}} \boldsymbol{f} | \boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\Big) &= E\Big(\boldsymbol{f}^T \sum_{k=1}^{b} \sum_{l=1}^{b} \lambda_{qr} \boldsymbol{J}_{qr}^T \boldsymbol{Z}^T \boldsymbol{S}_* \boldsymbol{Z} \frac{\partial \boldsymbol{\Lambda}}{\partial \lambda_{gh}} \boldsymbol{f} | \boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\Big) \\ &= \sum_{k=1}^{b} \sum_{l=1}^{b} \lambda_{kl} E\Big(\boldsymbol{f}^T \boldsymbol{J}_{qr}^T \boldsymbol{Z}^T \boldsymbol{S}_* \boldsymbol{Z} \frac{\partial \boldsymbol{\Lambda}}{\partial \lambda_{gh}} \boldsymbol{f} | \boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\Big).\end{aligned} \tag{5.12}$$

Substituting (5.12) into (5.11) we have

$$\sum_{q=1}^{b} \sum_{r=1}^{b} \lambda_{qr} E\Big(\boldsymbol{f}^T \boldsymbol{J}_{qr}^T \boldsymbol{Z}^T \boldsymbol{S}_* \boldsymbol{Z} \frac{\partial \boldsymbol{\Lambda}}{\partial \lambda_{gh}} \boldsymbol{f} | \boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\Big) = E\Big(\boldsymbol{y}_o^T \boldsymbol{S}_* \boldsymbol{Z} \frac{\partial \boldsymbol{\Lambda}}{\partial \lambda_{gh}} \boldsymbol{f} | \boldsymbol{y}_2; \boldsymbol{\mathcal{K}}^{(w)}\Big).$$

The update for $\boldsymbol{\lambda}$ can be expressed as

$$\boldsymbol{\lambda}^{(w+1)} = \boldsymbol{A}^{-1^{(w)}}\boldsymbol{b}^{(w)},$$

where the elements of the $b^2 \times b^2$ matrix $\boldsymbol{A}^{(w)}$ are

$$a^{(w)}_{gh,qr} = \tilde{\boldsymbol{u}}^{(w)^T}\boldsymbol{J}^T_{qr}\boldsymbol{Z}^T\boldsymbol{S}^{(w)}\boldsymbol{Z}\frac{\partial\boldsymbol{\Lambda}^{(w)}}{\partial\lambda^{(w)}_{gh}}\tilde{\boldsymbol{u}}^{(w)} + \mathrm{tr}\Big(\boldsymbol{J}^T_{qr}\boldsymbol{Z}^T\boldsymbol{S}^{(w)}\boldsymbol{Z}\frac{\partial\boldsymbol{\Lambda}^{(w)}}{\partial\lambda^{(w)}_{gh}}\boldsymbol{C}^{ZZ^{(w)}}\Big) \tag{5.13}$$

and the elements of the $b^2 \times 1$ vector $\boldsymbol{b}^{(w)}$ are

$$b^{(w)}_{gh} = \boldsymbol{y}^T_o\boldsymbol{S}^{(w)}\boldsymbol{Z}\frac{\partial\boldsymbol{\Lambda}^{(w)}}{\partial\lambda^{(w)}_{gh}}\tilde{\boldsymbol{u}}^{(w)}. \tag{5.14}$$

Note the difference between (5.13), (5.14) and (5.6), (5.7). The latter, based on $\boldsymbol{y}^{(P)}_c$ requires $\boldsymbol{C}^{XZ}$ and $\boldsymbol{C}^{ZZ}$ whereas the former, based on $\boldsymbol{y}^{(U)}_c$ only requires $\boldsymbol{C}^{ZZ}$.

We have derived an update for $\boldsymbol{\lambda}$ where the variance-covariance matrix $\boldsymbol{D}$ is considered to be completely unstructured. When $\boldsymbol{D}$ is structured the update for $\boldsymbol{\lambda}$ will have a simpler form. For the one-way random effects model for example, $\boldsymbol{D} = d\boldsymbol{I}_b$ and therefore $\boldsymbol{\Lambda} = \lambda\boldsymbol{I}_b$. In this case the update for λ has a relatively simple form.

This completes the M-step.

5.3.4 Rate matrix

The rate matrix for the REML PX EM algorithm using the complete data specification $\boldsymbol{y}^{(U)}_c$ is the same as that presented in (5.8) except we substitute $\boldsymbol{\mathcal{I}}^{(U)}_c(\boldsymbol{\kappa}^\star, \boldsymbol{\lambda}_0)$ for $\boldsymbol{\mathcal{I}}^{(P)}_c(\boldsymbol{\kappa}^\star, \boldsymbol{\lambda}_0)$. We define

$$\boldsymbol{\mathcal{I}}^{(U)}_c(\boldsymbol{\kappa}^\star, \boldsymbol{\lambda}_0) = \begin{pmatrix} \boldsymbol{\mathcal{I}}^{(U)}_{c_{\boldsymbol{\kappa},\boldsymbol{\kappa}}} & \boldsymbol{\mathcal{I}}^{(U)}_{c_{\boldsymbol{\kappa},\boldsymbol{\lambda}}} \\ & \boldsymbol{\mathcal{I}}^{(U)}_{c_{\boldsymbol{\lambda},\boldsymbol{\lambda}}} \end{pmatrix}_{\boldsymbol{\kappa}=\boldsymbol{\kappa}^\star, \boldsymbol{\lambda}=\boldsymbol{\lambda}_0.}$$

The elements of $\boldsymbol{\mathcal{I}}^{(U)}_{c_{\boldsymbol{\kappa},\boldsymbol{\lambda}}}$, which comprise $\boldsymbol{\mathcal{I}}^{(U)}_{c_{\sigma^2,\boldsymbol{\lambda}}}$, $\boldsymbol{\mathcal{I}}^{(U)}_{c_{\boldsymbol{\gamma},\boldsymbol{\lambda}}}$, and $\boldsymbol{\mathcal{I}}^{(U)}_{c_{\boldsymbol{\phi},\boldsymbol{\lambda}}}$, and the elements of $\boldsymbol{\mathcal{I}}^{(U)}_{c_{\boldsymbol{\lambda},\boldsymbol{\lambda}}}$ are presented below.

Elements of $\mathcal{I}^{(U)}_{c_{\sigma^2,\boldsymbol{\lambda}}}$

For $g = 1, \ldots, b$, $h = 1, \ldots, b$ it can be shown that

$$\begin{aligned}\mathcal{I}^{(U)}_{c_{\sigma^2,\lambda_{gh}}} &= \frac{1}{\left(\sigma^{2\star}\right)^2}\Bigg\{\boldsymbol{y}_o^T\boldsymbol{U}_\star^\star\boldsymbol{Z}\frac{\partial\boldsymbol{\Lambda}^\star}{\partial\lambda^\star_{gh}}\tilde{\boldsymbol{u}}^\star - \tilde{\boldsymbol{u}}^{\star^T}\boldsymbol{Z}^T\boldsymbol{U}^\star\boldsymbol{Z}\frac{\partial\boldsymbol{\Lambda}^\star}{\partial\lambda^\star_{gh}}\tilde{\boldsymbol{u}}^\star \\ &\quad -\mathrm{tr}\Big(\boldsymbol{Z}^T\boldsymbol{U}^\star\boldsymbol{Z}\frac{\partial\boldsymbol{\Lambda}^\star}{\partial\lambda^\star_{gh}}\boldsymbol{C}^{ZZ^\star}\Big)\Bigg\}.\end{aligned}$$

Elements of $\mathcal{I}^{(U)}_{c_{\gamma,\boldsymbol{\lambda}}}$

The elements for $\mathcal{I}^{(U)}_{c_{\gamma,\boldsymbol{\lambda}}}$ are equivalent to $\mathcal{I}^{(P)}_{c_{\gamma,\boldsymbol{\lambda}}}$, i.e., for $g = 1, \ldots, b$, $h = 1, \ldots, b$ we have

$$\mathcal{I}^{(U)}_{c_{\gamma_{gh},\lambda_{gh}}} = \tilde{\boldsymbol{u}}^{\star^T}\boldsymbol{G}^{-1^\star}\frac{\partial\boldsymbol{G}^\star}{\partial\gamma^\star_{gh}}\boldsymbol{G}^{-1^\star}\frac{\partial\boldsymbol{\Lambda}^\star}{\partial\lambda^\star_{gh}}\tilde{\boldsymbol{u}}^\star + \mathrm{tr}\Big(\boldsymbol{G}^{-1^\star}\frac{\partial\boldsymbol{G}^\star}{\partial\gamma^\star_{gh}}\boldsymbol{G}^{-1^\star}\frac{\partial\boldsymbol{\Lambda}^\star}{\partial\lambda^\star_{gh}}\boldsymbol{C}^{ZZ^\star}\Big).$$

Elements of $\mathcal{I}^{(U)}_{c_{\boldsymbol{\phi},\boldsymbol{\lambda}}}$

For $i = 1, \ldots, n$, $j = 1, \ldots, n$, $g = 1, \ldots, b$, $h = 1, \ldots, b$ it can be shown that

$$\begin{aligned}\mathcal{I}^{(U)}_{c_{\phi_{ij},\lambda_{gh}}} &= \frac{1}{\sigma^{2\star}}\Bigg\{\boldsymbol{y}_o^T\boldsymbol{U}^\star\frac{\partial\boldsymbol{\Sigma}^\star}{\partial\phi_{ij}}\boldsymbol{U}^\star\boldsymbol{Z}\frac{\partial\boldsymbol{\Lambda}^\star}{\partial\lambda^\star_{gh}}\tilde{\boldsymbol{u}}^\star - \tilde{\boldsymbol{u}}^{\star^T}\boldsymbol{Z}^T\boldsymbol{U}^\star\frac{\partial\boldsymbol{\Sigma}^\star}{\partial\phi_{ij}}\boldsymbol{U}^\star\boldsymbol{Z}\frac{\partial\boldsymbol{\Lambda}^\star}{\partial\lambda^\star_{gh}}\tilde{\boldsymbol{u}}^\star \\ &\quad -\mathrm{tr}\Big(\boldsymbol{Z}^T\boldsymbol{U}^\star\frac{\partial\boldsymbol{\Sigma}^\star}{\partial\phi_{ij}}\boldsymbol{U}^\star\boldsymbol{Z}\frac{\partial\boldsymbol{\Lambda}^\star}{\partial\lambda^\star_{gh}}\boldsymbol{C}^{ZZ^\star}\Big)\Bigg\}.\end{aligned}$$

Elements of $\mathcal{I}^{(U)}_{c_{\boldsymbol{\lambda},\boldsymbol{\lambda}}}$

We make the same assumptions as those made for the elements of $\mathcal{I}^{(P)}_{c_{\boldsymbol{\lambda},\boldsymbol{\lambda}}}$. For $g = 1, \ldots, b$, $h = 1, \ldots, b$, $q = 1, \ldots, b$, $r = 1, \ldots, b$ it can be shown that

$$\begin{aligned}\mathcal{I}^{(U)}_{c_{\lambda_{gh},\lambda_{qr}}} &= \tilde{\boldsymbol{u}}^{\star T}\frac{\partial \boldsymbol{\Lambda}^{\star T}}{\partial \lambda^{\star}_{qr}}\boldsymbol{Z}^T\boldsymbol{S}^{\star}\boldsymbol{Z}\frac{\partial \boldsymbol{\Lambda}^{\star}}{\partial \lambda^{\star}_{gh}}\tilde{\boldsymbol{u}}^{\star} + \mathrm{tr}\Big(\frac{\partial \boldsymbol{\Lambda}^{\star T}}{\partial \lambda^{\star}_{qr}}\boldsymbol{Z}^T\boldsymbol{S}^{\star}\boldsymbol{Z}\frac{\partial \boldsymbol{\Lambda}^{\star}}{\partial \lambda^{\star}_{gh}}\boldsymbol{C}^{ZZ^{\star}}\Big)\\ &+\mathrm{tr}\Big(\frac{\partial \boldsymbol{\Lambda}^{\star}}{\partial \lambda^{\star}_{gh}}\boldsymbol{\Lambda}^{-1^{\star}}\frac{\partial \boldsymbol{\Lambda}^{\star}}{\partial \lambda^{\star}_{qr}}\boldsymbol{\Lambda}^{-1^{\star}}\Big) + \tilde{\boldsymbol{u}}^{\star T}\frac{\partial \boldsymbol{\Lambda}^{\star T}}{\partial \lambda^{\star}_{gh}}\boldsymbol{G}^{-1^{\star}}\frac{\partial \boldsymbol{\Lambda}^{\star}}{\partial \lambda^{\star}_{qr}}\tilde{\boldsymbol{u}}^{\star}\\ &+\mathrm{tr}\Big(\frac{\partial \boldsymbol{\Lambda}^{\star T}}{\partial \lambda^{\star}_{gh}}\boldsymbol{G}^{-1^{\star}}\frac{\partial \boldsymbol{\Lambda}^{\star}}{\partial \lambda^{\star}_{qr}}\boldsymbol{C}^{ZZ^{\star}}\Big).\end{aligned}$$

5.4 Example: lamb weight data

We apply a REML PX EM algorithm using the complete data specifications $\boldsymbol{y}_c^{(U)}$ and $\boldsymbol{y}_c^{(P)}$ to the lamb weight data presented in Table 4.1. We fit the same model as that fitted in the previous chapter for the REML EM algorithm, i.e., the overall mean, age, and line are fitted as fixed effects, and sire is fitted as a random effect. For the REML PX EM algorithm this model can be written in matrix notation as

$$\boldsymbol{y}_o = \boldsymbol{X}\boldsymbol{\beta}_* + \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y}_o$, $\boldsymbol{X}$, and $\boldsymbol{Z}$ are defined in the same way as they were in the previous chapter for the REML EM algorithm. The vector $\boldsymbol{\beta}_*$ is defined in the same way as $\boldsymbol{\beta}$. We assume that the 23×1 vector of rescaled random effects $\boldsymbol{f}$, and the 62×1 vector of residuals $\boldsymbol{\epsilon}$, are independent of each other and have the following distributions

$$\begin{aligned}\boldsymbol{f} &\sim N(\boldsymbol{0}, d\boldsymbol{I}_{23}),\\ \boldsymbol{\epsilon} &\sim N(\boldsymbol{0}, \sigma^2_*\boldsymbol{I}_{62}).\end{aligned}$$

The updating equation for σ^2_* is provided in (5.3) and (5.10) for $\boldsymbol{y}_c^{(P)}$ and $\boldsymbol{y}_c^{(U)}$ respectively. The updating equation for d is provided in (5.4) and applies to both complete data specifications. We define $\boldsymbol{\kappa} = (\sigma^2, \sigma^2_s)^T$ as the parameter vector of interest and $\boldsymbol{\mathcal{K}} = (\sigma^2_*, d, \lambda)^T$ as the expanded parameter vector. The reduction function is $\boldsymbol{\kappa} = \mathcal{R}(\boldsymbol{\mathcal{K}}) = (\sigma^2_*, \lambda^2 d)^T$. We use an equivalent set of starting values to those in the previous chapter, i.e., the first set of starting values are $\boldsymbol{\mathcal{K}}^{(0)} = (\sigma^{2(0)}_*, d^{(0)}, \lambda^{(0)})^T = (1, 0.01, 1)^T$ and the second set of starting values are $\boldsymbol{\mathcal{K}}^{(0)} = (\sigma^{2(0)}_*, d^{(0)}, \lambda^{(0)})^T = (1, 5, 1)^T$. The convergence criteria presented in (4.18)

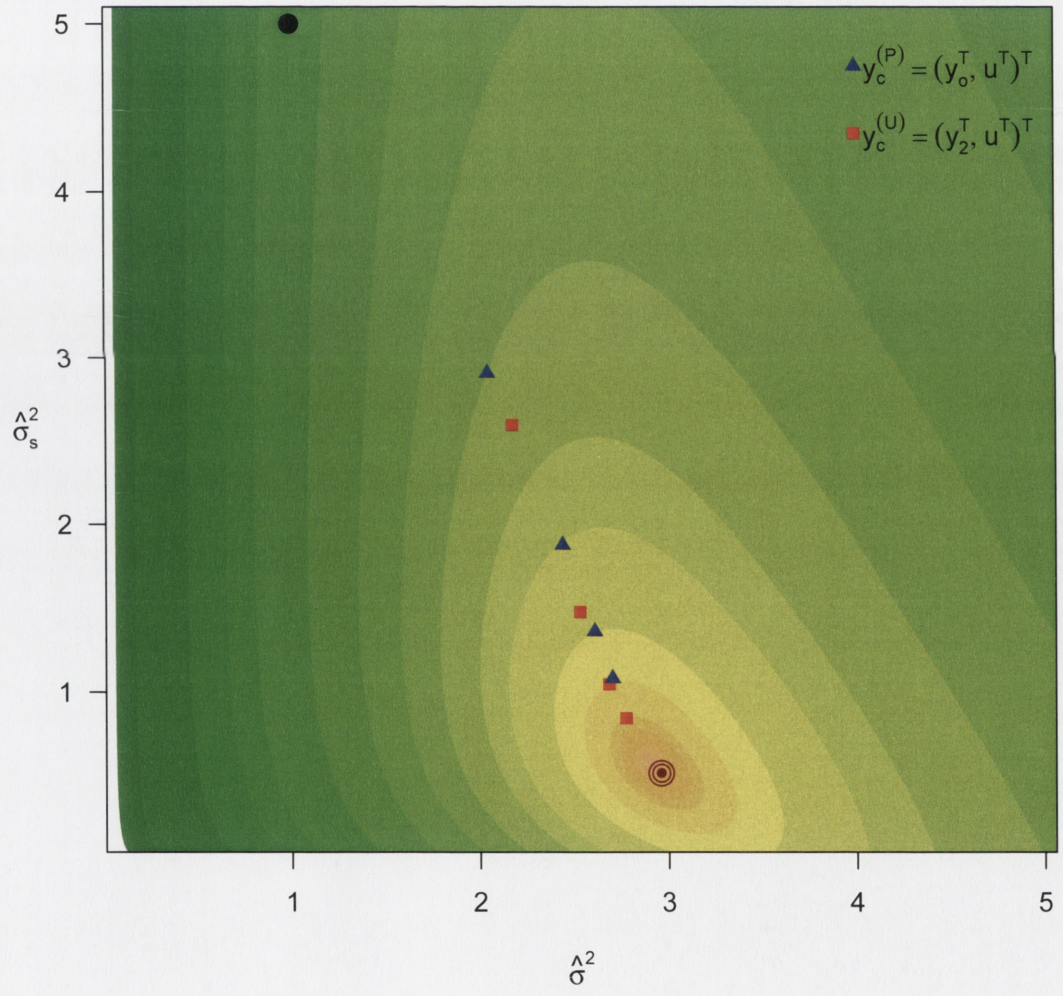
was used to determine if the sequence of iterates $\boldsymbol{\kappa}^{(0)}, \boldsymbol{\kappa}^{(1)}, \ldots$ converged to a (local) maximum of the REML log-likelihood function. The statistical software package R (R Development Core Team, 2011) was used to implement both REML PX EM algorithms. REML estimates of σ^2 and σ_s^2, denoted $\sigma^{2^\star}$ and $\sigma_s^{2^\star}$, the number of iterations until convergence and the rate of convergence for both complete data specifications and both sets of starting values are provided in Table 5.1.

Table 5.1: Results for two REML PX EM algorithms applied to the lamb weight data.

Complete data	$\sigma^{2^\star}$	$\sigma_s^{2^\star}$	Iterations	Rate of convergence
	$\sigma_*^{2^{(0)}} = 1, d^{(0)} = 0.01, \lambda^{(0)} = 1$			
$\boldsymbol{y}_c^{(P)}$	2.961597	0.5170765	83	0.8167880
$\boldsymbol{y}_c^{(U)}$	2.961597	0.5170765	57	0.7434961
	$\sigma_*^{2^{(0)}} = 1, d^{(0)} = 5, \lambda^{(0)} = 1$			
$\boldsymbol{y}_c^{(P)}$	2.961597	0.5170767	78	0.8167879
$\boldsymbol{y}_c^{(U)}$	2.961597	0.5170767	55	0.7434961

The observed information matrix is exactly the same as that obtained using a REML EM algorithm. A comparison of the results in Table 4.2 and Table 5.1 show that the REML PX EM converges in far fewer iterations than the REML EM algorithm. Table 5.1 highlights the benefit in using the complete data specification $\boldsymbol{y}_c^{(U)}$ compared to $\boldsymbol{y}_c^{(P)}$. The rate of convergence is noticeably less and convergence is achieved in approximately 30% fewer iterations for both sets of starting values. A contour plot of the REML log-likelihood surface and the first four iterations of the REML PX EM algorithms using the complete data specification $\boldsymbol{y}_c^{(U)}$ and $\boldsymbol{y}_c^{(P)}$ is provided in Figure 5.1. This figure illustrates the faster rate of convergence associated with using the complete data specification $\boldsymbol{y}_c^{(U)}$. The complete data specification $\boldsymbol{y}_c^{(U)}$ makes a larger initial step towards the REML solution and its' third iterate is approximately equivalent to the fourth iterate of the REML PX EM algorithm using the complete data specification $\boldsymbol{y}_c^{(P)}$. Figure 5.1 also highlights another feature of EM algorithms in general, which is that they often get close to a solution quite quickly but then take a long time to converge.

Figure 5.1: Contour plot of the residual log-likelihood surface for the lamb weight data and the first 4 iterations of the REML PX EM algorithms using the complete data specification $\boldsymbol{y}_c^{(U)}$ and $\boldsymbol{y}_c^{(P)}$. The starting values of $\sigma^{2(0)} = 1$ and $\sigma_s^{2(0)} = 5$ are represented by the black circle. The REML estimates of $\sigma^{2\star} = 2.962$ and $\sigma_s^{2\star} = 0.517$ are represented by the "target".



Chapter 6

Two model specifications for a factor analytic mixed model

In this and the next two chapters we consider REML EM and REML PX EM algorithms for factor analytic mixed models. The difference between linear mixed models and factor analytic mixed models is that in the former the vector of random effects is assumed to have a Gaussian distribution with a mean of zero and a variance defined by a symmetric positive definite variance-covariance matrix. In the later the vector of random effects is modelled using factor analysis. Our aim in the next three chapters is to consider ways in which to reduce the number of iterations until convergence for REML EM (and REML PX EM) algorithms applied to factor analytic mixed models. In all three chapters we consider the efficacy of using a less strict convergence criteria. In this chapter we consider two different model specifications, and in the next two chapters we consider a new parameter expansion and an alternative missing data specification respectively.

The motivation for the work presented in this chapter and the next two chapters is the analysis of multi-environment trial data produced by plant breeding programmes. Factor analytic mixed models were recommended by Smith et al. (2001) for the analysis of this type of data. In Australia, a major source of multi-environment trial data is the National Variety Trial programme which aims to deliver robust, independent results on the performance of recently released cereal grain varieties. The measure most often used to rank performance is grain yield. This programme comprises large numbers of varieties grown over large numbers of trials. As an example of the numbers involved, the 2005 to 2010 non-Durum wheat data

comprises 538 varieties and 928 trials where an "average" trial contains 162 plots. This data set therefore contains approximately 150000 measurements of grain yield. Fitting a factor analytic mixed model of the type advocated by Smith et al. (2001) to a data set of this size and structure is currently infeasible.

Although not applied to multi-environment trial data comprising tens of thousands of records, factor analytic mixed models are routinely applied to multi-environment trial data which contain thousands of records. For example, Thompson et al. (2003) apply a factor analytic mixed model to multi-environment trial data comprising 7613 records, 62 trials, and 3300 variety by trial combinations. One of the challenges in fitting a factor analytic mixed model to data of this size and structure is that modern statistical software packages, such as ASReml-R (Butler et al., 2007), typically implement a Newton-Raphson type algorithm, such as the average information algorithm (Gilmour et al., 1995), and successive iterations of these algorithms don't always increase the residual log-likelihood function, parameter updates don't always remain in their parameter space, and the algorithm can fail to converge. A potential solution to these problems is to use a REML EM or REML PX EM algorithm.

For a REML EM or REML PX EM algorithm to be a practical option for estimating the variance parameters associated with a factor analytic mixed model applied to multi-environment trial data the number of iterations until convergence has to be at a manageable level as each iteration can be computationally expensive in terms of computing time. Depending on the size and structure of the data this might mean having an algorithm that converges in no more than 50 iterations and preferably less than 25 iterations.

In this chapter we consider two different model specifications for a factor analytic mixed model. The first is based on the most commonly employed model specification, where the k factors are assumed to have unit variance. The second model specification is closely related to the PX EM algorithm model for a factor analytic mixed model presented by Liu et al. (1998). In this specification the variance-covariance matrix associated with the k factors is assumed to be unstructured. The main aim of this chapter is to show, using an example, that large gains in the number of iterations until convergence using a REML EM algorithm for factor analytic mixed models can be made by choosing a different model specification to that most commonly employed.

6.1 Model specification

We consider a simplified plant breeding case where we have m varieties grown in p trials. Every variety is grown in every trial and there is r replicates of every variety within a trial, i.e., the total number of observations $n = rmp$. Consider the model

$$\text{Model } \mathcal{O} \qquad \boldsymbol{y}_o = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_g\boldsymbol{u}_g + \boldsymbol{e},$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects representing trial means and $\boldsymbol{X}$ is its associated $n \times p$ design matrix of full column rank. It is assumed that the $n \times 1$ vector of residuals $\boldsymbol{e}$ is distributed $N(\boldsymbol{0}, \boldsymbol{R})$ where $\boldsymbol{R} = \sigma^2\boldsymbol{I}_n$. The $n \times mp$ design matrix $\boldsymbol{Z}_g$ is associated with the $mp \times 1$ vector $\boldsymbol{u}_g$ of variety by trial effects. These effects are assumed to be correlated within a trial but are themselves independent, i.e,

$$V(\boldsymbol{u}_g) = \boldsymbol{G}_e \otimes \boldsymbol{I}_m,$$

where $\boldsymbol{G}_e$ is a $p \times p$ symmetric positive definite matrix and is commonly referred to as the genetic variance-covariance matrix. The diagonal elements of $\boldsymbol{G}_e$ are the genetic variances for each trial and off-diagonal elements are genetic covariances between pairs of trials. A completely general or unstructured form of $\boldsymbol{G}_e$ involves $p(p+1)/2$ parameters. For even moderately large values of p it is computationally difficult to estimate unstructured $\boldsymbol{G}_e$. A more parsimonious representation of $\boldsymbol{G}_e$ can be achieved using factor analysis to model the variety effects within each trial, i.e.,

$$\begin{aligned} \boldsymbol{u}_g &= (\boldsymbol{\lambda}_1 \otimes \boldsymbol{I}_m)\boldsymbol{f}_1 + \ldots + (\boldsymbol{\lambda}_k \otimes \boldsymbol{I}_m)\boldsymbol{f}_k + \boldsymbol{\delta} \\ &= (\boldsymbol{\Lambda} \otimes \boldsymbol{I}_m)\boldsymbol{f} + \boldsymbol{\delta}, \end{aligned}$$

where $\boldsymbol{\Lambda}$ is a $p \times k$ matrix of loadings and $\boldsymbol{f}$ is a $mk \times 1$ vector of varietal scores that can be partitioned as $\boldsymbol{f} = (\boldsymbol{f}_1^T, \ldots, \boldsymbol{f}_k^T)^T$. The matrix of factor loadings $\boldsymbol{\Lambda}$ has elements λ_{jr}, $j = 1, \ldots, p$, $r = 1, \ldots, k$. To ensure the uniqueness of the estimates of factor loadings when $k > 1$, $k(k-1)/2$ independent constraints on the elements of $\boldsymbol{\Lambda}$ are necessary. The constraints to be used are arbitrary. However, Smith et al. (2001) recommend the constraint $\lambda_{jr} = 0$ for $j < r = 2, \ldots, k$.

The vector $\boldsymbol{\delta}$ is a $mp \times 1$ vector of residuals that can be partitioned with respect to trials, i.e., $\boldsymbol{\delta} = (\boldsymbol{\delta}_1^T, \ldots, \boldsymbol{\delta}_p^T)^T$. It is assumed that $\boldsymbol{f}$ and $\boldsymbol{\delta}$ are independent of each other and that

$$\boldsymbol{\delta} \sim N(\mathbf{0}, \boldsymbol{\Psi} \otimes \boldsymbol{I}_m),$$

where $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\boldsymbol{\psi})$ is a $p \times p$ diagonal matrix. The genetic variance-covariance matrix is therefore

$$\boldsymbol{G}_e = \boldsymbol{\Lambda} \boldsymbol{D} \boldsymbol{\Lambda}^T + \boldsymbol{\Psi}.$$

The vector of variance parameters $\boldsymbol{\psi}$ is commonly referred to as the specific variances. For simplicity we will assume that $\boldsymbol{\Psi} = \psi \boldsymbol{I}_p$, i.e., there is a common specific variance.

The two model specifications we consider differ in their assumption regarding the variance of $\boldsymbol{f}$. Using the same notation as Liu et al. (1998) we refer to the first model as Model $\mathcal{O}$ where it is assumed that

$$\boldsymbol{f} \sim N(\mathbf{0}, \boldsymbol{I}_{mk}).$$

We refer to the second model as Model $\mathcal{M}$ where it is assumed that

$$\boldsymbol{f} \sim N(\mathbf{0}, \boldsymbol{D} \otimes \boldsymbol{I}_m),$$

where $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{d})$ is a $k \times k$ symmetric positive definite matrix and $\boldsymbol{d}$ is a vector of variance parameters.

Model $\mathcal{O}$ is a special case of Model $\mathcal{M}$, namely, when $\boldsymbol{D} = \boldsymbol{I}_k$. Therefore, from here on we only present the development of a REML EM algorithm for Model $\mathcal{M}$ as the complete data density function and the E and M-steps for Model $\mathcal{O}$ can be obtained by substituting an identity matrix of order k for $\boldsymbol{D}$.

For Model $\mathcal{M}$ the distribution of the observed data is

$$\boldsymbol{y}_o \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{H}),$$

where $\boldsymbol{H} = \boldsymbol{Z}_g(\boldsymbol{G}_e \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T + \boldsymbol{R}$ where $\boldsymbol{R} = \sigma^2\boldsymbol{I}_n$ and the vector of variance parameters to be estimated is $\boldsymbol{\Gamma} = (\sigma^2, \boldsymbol{\lambda}^T, \psi, \boldsymbol{d}^T)^T$. The estimate of the genetic variance-covariance matrix, denoted $\widehat{\boldsymbol{G}}_e$, and the best linear unbiased predictor of variety by trial effects, denoted $\tilde{\boldsymbol{u}}_g$, are the quantities most often reported to plant breeders. The former for providing the genetic correlation between trials and the later for determining which varieties are broadly adaptable across environments.

It should be noted that $\boldsymbol{\Lambda}$ and $\boldsymbol{D}$ are not unique for Model $\mathcal{M}$. We don't consider this an impediment to the implementation of Model $\mathcal{M}$ for the analysis of plant breeding data as $\widehat{\boldsymbol{G}}_e$ and $\tilde{\boldsymbol{u}}_g$ are unique and are the quantities of primary interest.

6.2 Complete data density function

We define the complete data to be $\boldsymbol{y}_c = (\boldsymbol{y}_2^T, \boldsymbol{f}^T, \boldsymbol{\delta}^T)^T$ where the missing data vector is $\boldsymbol{y}_m = (\boldsymbol{f}^T, \boldsymbol{\delta}^T)^T$. Model $\mathcal{O}$ and Model $\mathcal{M}$ could be considered new in the sense that we consider the incomplete data to be the transformed observed data vector associated with the marginal log-density function in the Verbyla (1990) conditional derivation of REML. This has not been considered previously for REML EM or REML PX EM algorithms applied to factor analytic mixed models.

To form the complete data log-density function requires the joint distribution

$$\begin{pmatrix} \boldsymbol{y}_2 \\ \boldsymbol{f} \\ \boldsymbol{\delta} \end{pmatrix} \sim N\left\{ \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2 & \boldsymbol{L}_2^T\boldsymbol{Z}_g(\boldsymbol{\Lambda}\boldsymbol{D} \otimes \boldsymbol{I}_m) & \boldsymbol{L}_2^T\boldsymbol{Z}_g(\boldsymbol{\Psi} \otimes \boldsymbol{I}_m) \\ (\boldsymbol{D}\boldsymbol{\Lambda}^T \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{L}_2 & \boldsymbol{D} \otimes \boldsymbol{I}_m & \boldsymbol{0} \\ (\boldsymbol{\Psi} \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{L}_2 & \boldsymbol{0} & \boldsymbol{\Psi} \otimes \boldsymbol{I}_m \end{pmatrix} \right\}. \tag{6.1}$$

The complete data log-density function can be expressed as

$$\log\{f(\boldsymbol{y}_c; \boldsymbol{\Gamma})\} = \log\{f(\boldsymbol{y}_2|\boldsymbol{f}, \boldsymbol{\delta}; \boldsymbol{\Gamma})\} + \log\{f(\boldsymbol{f}, \boldsymbol{\delta}; \boldsymbol{\Gamma})\}.$$

Consider $\log\{f(\boldsymbol{y}_2|\boldsymbol{f}, \boldsymbol{\delta}; \boldsymbol{\Gamma})\}$. Using (6.1) and Result A.1 it can be shown that

$$\boldsymbol{y}_2|\boldsymbol{f}, \boldsymbol{\delta} \sim N\{\boldsymbol{L}_2^T\boldsymbol{Z}_g(\boldsymbol{\Lambda} \otimes \boldsymbol{I}_m)\boldsymbol{f} + \boldsymbol{L}_2^T\boldsymbol{Z}_g\boldsymbol{\delta}, \boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2\},$$

and the log-density function is

$$\log\{f(\boldsymbol{y}_2|\boldsymbol{f},\boldsymbol{\delta};\boldsymbol{\Gamma})\} = -\frac{1}{2}\Big[\log\{\det(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2)\} + \{\boldsymbol{y}_o - \boldsymbol{Z}_g(\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)\boldsymbol{f} - \boldsymbol{Z}_g\boldsymbol{\delta}\}^T\boldsymbol{S}\{\boldsymbol{y}_o - \boldsymbol{Z}_g(\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)\boldsymbol{f} - \boldsymbol{Z}_g\boldsymbol{\delta}\}\Big],$$

where

$$\begin{aligned}\boldsymbol{S} &= \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T \\ &= \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{R}^{-1}.\end{aligned}$$

The marginal log-density function for the missing data is

$$\begin{aligned}\log\{f(\boldsymbol{f},\boldsymbol{\delta};\boldsymbol{\Gamma})\} &= \log\{f(\boldsymbol{f};\boldsymbol{d})\} + \log\{f(\boldsymbol{\delta};\boldsymbol{\psi})\} \\ &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{D}\otimes\boldsymbol{I}_m)\} + \log\{\det(\boldsymbol{\Psi}\otimes\boldsymbol{I}_m)\} \\ &\quad + \boldsymbol{f}^T(\boldsymbol{D}^{-1}\otimes\boldsymbol{I}_m)\boldsymbol{f} + \boldsymbol{\delta}^T(\boldsymbol{\Psi}^{-1}\otimes\boldsymbol{I}_m)\boldsymbol{\delta}\Big].\end{aligned}$$

The complete data log-density function is therefore

$$\begin{aligned}\log\{f(\boldsymbol{y}_c;\boldsymbol{\Gamma})\} = -\frac{1}{2}\Big[&\log\{\det(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2)\} + \log\{\det(\boldsymbol{D}\otimes\boldsymbol{I}_m)\} + \log\{\det(\boldsymbol{\Psi}\otimes\boldsymbol{I}_m)\} \\ &+\{\boldsymbol{y}_o - \boldsymbol{Z}_g(\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)\boldsymbol{f} - \boldsymbol{Z}_g\boldsymbol{\delta}\}^T\boldsymbol{S}\{\boldsymbol{y}_o - \boldsymbol{Z}_g(\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)\boldsymbol{f} - \boldsymbol{Z}_g\boldsymbol{\delta}\} \\ &+\boldsymbol{f}^T(\boldsymbol{D}^{-1}\otimes\boldsymbol{I}_m)\boldsymbol{f} + \boldsymbol{\delta}^T(\boldsymbol{\Psi}^{-1}\otimes\boldsymbol{I}_m)\boldsymbol{\delta}\Big].\end{aligned} \tag{6.2}$$

6.3 REML EM algorithm E-step for Model $\mathcal{M}$

The E-step involves taking the conditional expectation of (6.2) over $k(\boldsymbol{f},\boldsymbol{\delta}|\boldsymbol{y}_2;\boldsymbol{\Gamma}^{(w)})$ where $\boldsymbol{\Gamma}^{(w)}$ is the w-th iterate of $\boldsymbol{\Gamma}$. We can write

$$\begin{aligned}Q(\boldsymbol{\Gamma};\boldsymbol{\Gamma}^{(w)}) &= E[\log\{f(\boldsymbol{y}_c;\boldsymbol{\Gamma})\}|\boldsymbol{y}_2;\boldsymbol{\Gamma}^{(w)}] \\ &= E[\log\{f(\boldsymbol{y}_2|\boldsymbol{f},\boldsymbol{\delta};\boldsymbol{\Gamma})\}|\boldsymbol{y}_2;\boldsymbol{\Gamma}^{(w)}] + E[\log\{f(\boldsymbol{f};\boldsymbol{d})\}|\boldsymbol{y}_2;\boldsymbol{\Gamma}^{(w)}] \\ &\quad + E[\log\{f(\boldsymbol{\delta};\boldsymbol{\psi})\}|\boldsymbol{y}_2;\boldsymbol{\Gamma}^{(w)}] \\ &= Q(\sigma^2,\boldsymbol{\lambda};\boldsymbol{\Gamma}^{(w)}) + Q(\boldsymbol{d};\boldsymbol{\Gamma}^{(w)}) + Q(\boldsymbol{\psi};\boldsymbol{\Gamma}^{(w)}).\end{aligned} \tag{6.3}$$

To compute the $Q(.)$ functions in (6.3) requires the joint conditional distribution of $\boldsymbol{f}$ and $\boldsymbol{\delta}$ given $\boldsymbol{y}_2$. Using (6.1) and Result A.1 we have

$$\left.\begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{\delta} \end{pmatrix}\right| \boldsymbol{y}_2 \sim N\left\{\begin{pmatrix} (\boldsymbol{D}\boldsymbol{\Lambda}^T \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{P}\boldsymbol{y}_o \\ (\boldsymbol{\Psi} \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{P}\boldsymbol{y}_o \end{pmatrix}, \begin{pmatrix} \boldsymbol{T}_{ff} & \boldsymbol{T}_{f\delta} \\ \boldsymbol{T}_{\delta f} & \boldsymbol{T}_{\delta\delta} \end{pmatrix}\right\},$$

where

$$\begin{aligned}
\boldsymbol{T}_{ff} &= (\boldsymbol{D} \otimes \boldsymbol{I}_m) - (\boldsymbol{D}\boldsymbol{\Lambda}^T \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{P}\boldsymbol{Z}_g(\boldsymbol{\Lambda}\boldsymbol{D} \otimes \boldsymbol{I}_m), \\
\boldsymbol{T}_{f\delta} &= \boldsymbol{T}_{\delta f}^T = -(\boldsymbol{\Psi} \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{P}\boldsymbol{Z}_g(\boldsymbol{\Lambda}\boldsymbol{D} \otimes \boldsymbol{I}_m), \\
\boldsymbol{T}_{\delta\delta} &= (\boldsymbol{\Psi} \otimes \boldsymbol{I}_m) - (\boldsymbol{\Psi} \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{P}\boldsymbol{Z}_g(\boldsymbol{\Psi} \otimes \boldsymbol{I}_m)
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{P} &= \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T \\
&= \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}.
\end{aligned}$$

We can write

$$\begin{aligned}
&Q(\sigma^2, \boldsymbol{\lambda}; \boldsymbol{\Gamma}^{(w)}) = -\frac{1}{2}\Big[\log\{\det(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2)\} \\
&+ \{\boldsymbol{y}_o - \boldsymbol{Z}_g(\boldsymbol{\Lambda} \otimes \boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} - \boldsymbol{Z}_g\tilde{\boldsymbol{\delta}}^{(w)}\}^T\boldsymbol{S}\{\boldsymbol{y}_o - \boldsymbol{Z}_g(\boldsymbol{\Lambda} \otimes \boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} - \boldsymbol{Z}_g\tilde{\boldsymbol{\delta}}^{(w)}\} \\
&+ \mathrm{tr}\Big\{(\boldsymbol{\Lambda}^T \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{S}\boldsymbol{Z}_g(\boldsymbol{\Lambda} \otimes \boldsymbol{I}_m)\boldsymbol{T}_{ff}^{(w)}\Big\} + \mathrm{tr}\Big\{\boldsymbol{Z}_g^T\boldsymbol{S}\boldsymbol{Z}_g\boldsymbol{T}_{\delta\delta}^{(w)}\Big\} \\
&- 2\mathrm{tr}\Big\{\boldsymbol{Z}_g^T\boldsymbol{S}\boldsymbol{Z}_g(\boldsymbol{\Lambda} \otimes \boldsymbol{I}_m)E\big(\boldsymbol{f}\boldsymbol{\delta}^T|\boldsymbol{y}_2; \boldsymbol{\Gamma}^{(w)}\big)\Big\}\Big], \\
&Q(\boldsymbol{d}; \boldsymbol{\Gamma}^{(w)}) = -\frac{1}{2}\Big[\log\{\det(\boldsymbol{D} \otimes \boldsymbol{I}_m)\} + \tilde{\boldsymbol{f}}^{(w)^T}(\boldsymbol{D}^{-1} \otimes \boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} + \mathrm{tr}\big(\boldsymbol{T}_{ff}^{(w)}\big)\Big], \\
&Q(\boldsymbol{\psi}; \boldsymbol{\Gamma}^{(w)}) = -\frac{1}{2}\Big[\log\{\det(\boldsymbol{\Psi} \otimes \boldsymbol{I}_m)\} + \tilde{\boldsymbol{\delta}}^{(w)^T}(\boldsymbol{\Psi}^{-1} \otimes \boldsymbol{I}_m)\tilde{\boldsymbol{\delta}}^{(w)} + \mathrm{tr}\big(\boldsymbol{T}_{\delta\delta}^{(w)}\big)\Big],
\end{aligned}$$

where

$$\begin{aligned}
&\tilde{\boldsymbol{f}}^{(w)} = (\boldsymbol{D}^{(w)}\boldsymbol{\Lambda}^{(w)^T} \otimes \boldsymbol{I}_m)\boldsymbol{P}^{(w)}\boldsymbol{y}_o, \\
&E(\boldsymbol{f}\boldsymbol{\delta}^T|\boldsymbol{y}_2; \boldsymbol{\Gamma}^{(w)}) = \tilde{\boldsymbol{f}}^{(w)}\tilde{\boldsymbol{\delta}}^{(w)^T} + \boldsymbol{T}_{f\delta}^{(w)} \qquad \text{using Result A.2.}
\end{aligned}$$

This completes the E-step.

6.4 REML EM algorithm M-step for Model $\mathcal{M}$

The M-step involves maximising $Q(\boldsymbol{\Gamma};\boldsymbol{\Gamma}^{(w)})$ with respect to $\boldsymbol{\Gamma}$ to find an update for $\boldsymbol{\Gamma} = (\sigma^2, \boldsymbol{\lambda}^T, \psi, \boldsymbol{d}^T)^T$. The partitioning of $Q(\boldsymbol{\Gamma};\boldsymbol{\Gamma}^{(w)})$ into $Q(\sigma^2, \boldsymbol{\lambda};\boldsymbol{\Gamma}^{(w)})$, $Q(\boldsymbol{d};\boldsymbol{\Gamma}^{(w)})$ and $Q(\psi;\boldsymbol{\Gamma}^{(w)})$ simplifies the process.

6.4.1 Update for σ^2

Differentiating $Q(\sigma^2, \boldsymbol{\lambda};\boldsymbol{\Gamma}^{(w)})$ with respect to σ^2 we have

$$\begin{aligned}\frac{\partial Q(\sigma^2,\boldsymbol{\lambda};\boldsymbol{\Gamma}^{(w)})}{\partial\sigma^2} &= -\frac{1}{2}\Bigg[\frac{n-p}{\sigma^2}\\ &- \frac{1}{(\sigma^2)^2}\{\boldsymbol{y}_o - \boldsymbol{Z}_g(\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} - \boldsymbol{Z}_g\tilde{\boldsymbol{\delta}}^{(w)}\}^T\boldsymbol{K}\{\boldsymbol{y}_o - \boldsymbol{Z}_g(\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} - \boldsymbol{Z}_g\tilde{\boldsymbol{\delta}}^{(w)}\}\\ &- \frac{1}{(\sigma^2)^2}\mathrm{tr}\Big\{(\boldsymbol{\Lambda}^T\otimes\boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{K}\boldsymbol{Z}_g(\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)\boldsymbol{T}_{ff}^{(w)}\Big\}\\ &- \frac{1}{(\sigma^2)^2}\mathrm{tr}\Big(\boldsymbol{Z}_g^T\boldsymbol{K}\boldsymbol{Z}_g\boldsymbol{T}_{\delta\delta}^{(w)}\Big) - \frac{2}{(\sigma^2)^2}\mathrm{tr}\Big\{\boldsymbol{Z}_g^T\boldsymbol{K}\boldsymbol{Z}_g(\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)E\big(\boldsymbol{f}\boldsymbol{\delta}^T|\boldsymbol{y}_2;\boldsymbol{\Gamma}^{(w)}\big)\Big\}\Bigg],\end{aligned}$$

where $\boldsymbol{K} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Equating to zero and solving for σ^2 we have

$$\begin{aligned}\sigma^{2^{(w+1)}} = \frac{1}{n-p}\Bigg[&\{\boldsymbol{y}_o - \boldsymbol{Z}_g(\boldsymbol{\Lambda}^{(w)}\otimes\boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} - \boldsymbol{Z}_g\tilde{\boldsymbol{\delta}}^{(w)}\}^T\boldsymbol{K}\{\boldsymbol{y}_o - \boldsymbol{Z}_g(\boldsymbol{\Lambda}^{(w)}\otimes\boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} - \boldsymbol{Z}_g\tilde{\boldsymbol{\delta}}^{(w)}\}\\ &+\mathrm{tr}\Big\{(\boldsymbol{\Lambda}^{(w)^T}\otimes\boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{K}\boldsymbol{Z}_g(\boldsymbol{\Lambda}^{(w)}\otimes\boldsymbol{I}_m)\boldsymbol{T}_{ff}^{(w)}\Big\}\\ &+\mathrm{tr}\Big(\boldsymbol{Z}_g^T\boldsymbol{K}\boldsymbol{Z}_g\boldsymbol{T}_{\delta\delta}^{(w)}\Big) + 2\mathrm{tr}\Big\{\boldsymbol{Z}_g^T\boldsymbol{K}\boldsymbol{Z}_g(\boldsymbol{\Lambda}^{(w)}\otimes\boldsymbol{I}_m)E\big(\boldsymbol{f}\boldsymbol{\delta}^T|\boldsymbol{y}_2;\boldsymbol{\Gamma}^{(w)}\big)\Big\}\Bigg].\end{aligned}$$

6.4.2 Update for λ_{jr}

We denote the elements of the $p\times k$ matrix of factor loadings $\boldsymbol{\Lambda}$ as λ_{jr}, $j = 1,\ldots,p$, $r = 1,\ldots,k$ and $\boldsymbol{\lambda} = \mathrm{vec}(\boldsymbol{\Lambda})$. Finding an update for $\boldsymbol{\lambda}$ is similar to finding an update for the auxillary parameter in the model presented in (5.1).

We can write $Q(\sigma^2, \boldsymbol{\lambda}; \boldsymbol{\Gamma}^{(w)})$ as

$$\begin{aligned} Q(\sigma^2, \boldsymbol{\lambda}; \boldsymbol{\Gamma}^{(w)}) &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{L}_2^T \boldsymbol{R} \boldsymbol{L}_2\}] + \boldsymbol{y}_o^T \boldsymbol{S} \boldsymbol{y}_o \\ &\quad -2E\big\{\boldsymbol{y}_o^T \boldsymbol{S} \boldsymbol{Z}_g (\boldsymbol{\Lambda} \otimes \boldsymbol{I}_m) \boldsymbol{f} | \boldsymbol{y}_2; \boldsymbol{\Gamma}^{(w)}\big\} \\ &\quad -2E\big\{\boldsymbol{\delta}^T \boldsymbol{Z}_g^T \boldsymbol{S} \boldsymbol{Z}_g (\boldsymbol{\Lambda} \otimes \boldsymbol{I}_m) \boldsymbol{f} | \boldsymbol{y}_2; \boldsymbol{\Gamma}^{(w)}\big\} \\ &\quad +E\big\{\boldsymbol{f}^T (\boldsymbol{\Lambda}^T \otimes \boldsymbol{I}_m) \boldsymbol{Z}_g^T \boldsymbol{S} \boldsymbol{Z}_g (\boldsymbol{\Lambda} \otimes \boldsymbol{I}_m) \boldsymbol{f} | \boldsymbol{y}_2; \boldsymbol{\Gamma}^{(w)}\big\}\Big]. \end{aligned}$$

Differentiating $Q(\sigma^2, \boldsymbol{\lambda}; \boldsymbol{\Gamma}^{(w)})$ with respect to λ_{jr} we have

$$\begin{aligned} \frac{\partial Q(\sigma^2, \boldsymbol{\lambda}; \boldsymbol{\Gamma}^{(w)})}{\partial \lambda_{jr}} &= -\frac{1}{2}\Big[-2E\big\{\boldsymbol{y}_o^T \boldsymbol{S} \boldsymbol{Z}_g (\frac{\partial \boldsymbol{\Lambda}}{\partial \lambda_{jr}} \otimes \boldsymbol{I}_m) \boldsymbol{f} | \boldsymbol{y}_2; \boldsymbol{\Gamma}^{(w)}\big\} \\ &\quad +2\mathrm{tr}\big\{\boldsymbol{Z}_g^T \boldsymbol{S} \boldsymbol{Z}_g (\frac{\partial \boldsymbol{\Lambda}}{\partial \lambda_{jr}} \otimes \boldsymbol{I}_m) E\big(\boldsymbol{f} \boldsymbol{\delta}^T | \boldsymbol{y}_2; \boldsymbol{\Gamma}^{(w)}\big)\big\} \\ &\quad +2E\big\{\boldsymbol{f}^T (\boldsymbol{\Lambda}^T \otimes \boldsymbol{I}_m) \boldsymbol{Z}_g^T \boldsymbol{S} \boldsymbol{Z}_g (\frac{\partial \boldsymbol{\Lambda}}{\partial \lambda_{jr}} \otimes \boldsymbol{I}_m) \boldsymbol{f} | \boldsymbol{y}_2; \boldsymbol{\Gamma}^{(w)}\big\}\Big]. \end{aligned}$$

Equating to zero and using the identity

$$\boldsymbol{\Lambda}^T = \sum_{l=1}^{p} \sum_{q=1}^{k} \lambda_{lq} \boldsymbol{J}_{lq}^T$$

where $\boldsymbol{J}_{lq}$ is a $p \times k$ indicator matrix having a 1 in row l, column q and zeroes elsewehere we obtain

$$\begin{aligned} &\sum_{l=1}^{p} \sum_{q=1}^{k} \lambda_{lq} E\big\{\boldsymbol{f}^T (\boldsymbol{J}_{lq}^T \otimes \boldsymbol{I}_m) \boldsymbol{Z}_g^T \boldsymbol{S} \boldsymbol{Z}_g (\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m) \boldsymbol{f} | \boldsymbol{y}_2; \boldsymbol{\Gamma}^{(w)}\big\} \\ &= \sum_{l=1}^{p} \sum_{q=1}^{k} \lambda_{lq} \Big[E\big\{\boldsymbol{y}_o^T \boldsymbol{S} \boldsymbol{Z}_g (\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m) \boldsymbol{f} | \boldsymbol{y}_2; \boldsymbol{\Gamma}^{(w)}\big\} \\ &-\mathrm{tr}\big\{\boldsymbol{Z}_g^T \boldsymbol{S} \boldsymbol{Z}_g (\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m) E\big(\boldsymbol{f} \boldsymbol{\delta}^T | \boldsymbol{y}_2; \boldsymbol{\Gamma}^{(w)}\big)\big\}\Big]\Big\}, \end{aligned}$$

where $\boldsymbol{J}_{jr}$ is a $p \times k$ indicator matrix having a 1 in row j, column r, and zeroes elsewhere. This equation can be written as

$$\sum_{l=1}^{p}\sum_{q=1}^{k}\lambda_{lq}a_{jr,lq}^{(w)} = b_{jr}^{(w)}$$

and in matrix notation as

$$\boldsymbol{A}^{(w)}\boldsymbol{\lambda} = \boldsymbol{b}^{(w)},$$

where $\boldsymbol{A}^{(w)}$ is a $pk \times pk$ matrix with elements

$$\begin{aligned} a_{jr,lq}^{(w)} &= E\{\boldsymbol{f}^T(\boldsymbol{J}_{lq}^T \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{S}\boldsymbol{Z}_g(\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)\boldsymbol{f}|\boldsymbol{y}_2;\boldsymbol{\Gamma}^{(w)}\} \\ &= \tilde{\boldsymbol{f}}^{(w)^T}(\boldsymbol{J}_{lq}^T \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{S}^{(w)}\boldsymbol{Z}_g(\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} \\ &\quad +\mathrm{tr}\{(\boldsymbol{J}_{lq}^T \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{S}^{(w)}\boldsymbol{Z}_g(\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)\boldsymbol{T}_{ff}^{(w)}\} \end{aligned}$$

and $\boldsymbol{b}^{(w)}$ is a $pk \times 1$ vector with elements

$$\begin{aligned} b_{jr}^{(w)} &= E\{\boldsymbol{y}_o^T\boldsymbol{S}\boldsymbol{Z}_g(\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)\boldsymbol{f}|\boldsymbol{y}_2;\boldsymbol{\Gamma}^{(w)}\} - \mathrm{tr}\{\boldsymbol{Z}_g^T\boldsymbol{S}^{(w)}\boldsymbol{Z}_g(\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)E(\boldsymbol{f}\boldsymbol{\delta}^T|\boldsymbol{y}_2;\boldsymbol{\Gamma}^{(w)})\} \\ &= \boldsymbol{y}_o^T\boldsymbol{S}^{(w)}\boldsymbol{Z}_g(\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} - \mathrm{tr}\{\boldsymbol{Z}_g^T\boldsymbol{S}^{(w)}\boldsymbol{Z}_g(\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)(\tilde{\boldsymbol{f}}^{(w)}\tilde{\boldsymbol{\delta}}^{(w)^T} + \boldsymbol{T}_{f\delta}^{(w)})\}. \end{aligned}$$

The update for $\boldsymbol{\lambda}$ is therefore

$$\boldsymbol{\lambda}^{(w+1)} = \boldsymbol{A}^{(w)^{-1}}\boldsymbol{b}^{(w)}.$$

6.4.3 Update for ψ

Assuming $\boldsymbol{\Psi} = \psi\boldsymbol{I}_p$ we can write $Q(\boldsymbol{\psi};\boldsymbol{\Gamma}^{(w)})$ as

$$Q(\boldsymbol{\psi};\boldsymbol{\Gamma}^{(w)}) = -\frac{1}{2}\Big\{mp\log(\psi) + \frac{1}{\psi}\tilde{\boldsymbol{\delta}}^{(w)^T}\tilde{\boldsymbol{\delta}}^{(w)} + \frac{1}{\psi}\mathrm{tr}\big(\boldsymbol{T}_{\delta\delta}^{(w)}\big)\Big\}.$$

Differentiating with respect to ψ, equating to zero, and solving for ψ we have

$$\psi^{(w+1)} = \frac{1}{mp}\Big\{\tilde{\boldsymbol{\delta}}^{(w)^T}\tilde{\boldsymbol{\delta}}^{(w)} + \mathrm{tr}\big(\boldsymbol{T}_{\delta\delta}^{(w)}\big)\Big\}.$$

6.4.4 Update for d_{qr}

The $k \times k$ matrix $\boldsymbol{D}$ is symmetric with elements d_{qr}, $q = 1, \ldots, k$, $r = 1, \ldots, k$, $q \geq r$, i.e., due to the symmetry of $\boldsymbol{D}$ we only consider the lower left triangle of $\boldsymbol{D}$. We define the elements of $\boldsymbol{D}^{-1}$ to be d^{qr}, $q = 1, \ldots, k$, $r = 1, \ldots, k$, $q \geq r$. Differentiating $Q(\boldsymbol{d}; \boldsymbol{\Gamma}^{(w)})$ with respect to d_{qr} we have

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{d}; \boldsymbol{\Gamma}^{(w)})}{\partial d_{qr}} &= -\frac{1}{2}\Big[\mathrm{tr}\Big\{\Big(\boldsymbol{D}^{-1} \otimes \boldsymbol{I}_m\Big)\Big(\frac{\partial \boldsymbol{D}}{\partial d_{qr}} \otimes \boldsymbol{I}_m\Big)\Big\} \\
&\quad -\tilde{\boldsymbol{f}}^{(w)T}\Big(\boldsymbol{D}^{-1}\frac{\partial \boldsymbol{D}}{\partial d_{qr}}\boldsymbol{D}^{-1} \otimes \boldsymbol{I}_m\Big)\tilde{\boldsymbol{f}}^{(w)} \\
&\quad -\mathrm{tr}\Big\{\Big(\boldsymbol{D}^{-1}\frac{\partial \boldsymbol{D}}{\partial d_{qr}}\boldsymbol{D}^{-1} \otimes \boldsymbol{I}_m\Big)\boldsymbol{T}_{ff}^{(w)}\Big\}\Big] \\
&= -\frac{1}{2}\Big[\mathrm{tr}\Big(\boldsymbol{D}^{-1}\frac{\partial \boldsymbol{D}}{\partial d_{qr}} \otimes \boldsymbol{I}_m\Big) - \mathrm{tr}\Big\{\boldsymbol{V}^{(w)}\Big(\boldsymbol{D}^{-1}\frac{\partial \boldsymbol{D}}{\partial d_{qr}}\boldsymbol{D}^{-1} \otimes \boldsymbol{I}_m\Big)\Big\} \\
&\quad -\mathrm{tr}\Big\{\boldsymbol{T}_{ff}^{(w)}\Big(\boldsymbol{D}^{-1}\frac{\partial \boldsymbol{D}}{\partial d_{qr}}\boldsymbol{D}^{-1} \otimes \boldsymbol{I}_m\Big)\Big\}\Big], \qquad (6.4)
\end{aligned}$$

where $\boldsymbol{V}^{(w)} = \tilde{\boldsymbol{f}}^{(w)}\tilde{\boldsymbol{f}}^{(w)T}$ is symmetric and can be partitioned

$$\boldsymbol{V}^{(w)} = \begin{pmatrix} \boldsymbol{V}_{11}^{(w)} & & \\ \vdots & \ddots & \\ \boldsymbol{V}_{k1}^{(w)} & \cdots & \boldsymbol{V}_{kk}^{(w)} \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{f}}_1^{(w)}\tilde{\boldsymbol{f}}_1^{(w)T} & & \\ \vdots & \ddots & \\ \tilde{\boldsymbol{f}}_k^{(w)}\tilde{\boldsymbol{f}}_1^{(w)T} & \cdots & \tilde{\boldsymbol{f}}_k^{(w)}\tilde{\boldsymbol{f}}_k^{(w)T} \end{pmatrix},$$

where each $\boldsymbol{V}_{gh}^{(w)}$ is a $m \times m$ matrix. Working with the first term in (6.4) we have

$$\begin{aligned}
\mathrm{tr}\Big(\boldsymbol{D}^{-1}\frac{\partial \boldsymbol{D}}{\partial d_{qr}} \otimes \boldsymbol{I}_m\Big) &= m\mathrm{tr}\Big(\boldsymbol{D}^{-1}\frac{\partial \boldsymbol{D}}{\partial d_{qr}}\Big) \\
&= md^{qr}.
\end{aligned}$$

Working with the second term in (6.4) it can be shown that

$$\begin{aligned}
\mathrm{tr}\Big\{\boldsymbol{V}^{(w)}\Big(\boldsymbol{D}^{-1}\frac{\partial \boldsymbol{D}}{\partial d_{qr}}\boldsymbol{D}^{-1}\otimes \boldsymbol{I}_m\Big)\Big\} &= \sum_{g=1}^{k}\sum_{h=1}^{k}\mathrm{tr}\big(\boldsymbol{V}_{gh}^{(w)}d^{gq}d^{rh}\big)\\
&= \sum_{g=1}^{k}\sum_{h=1}^{k}d^{gq}\mathrm{tr}\big(\boldsymbol{V}_{gh}^{(w)}\big)d^{rh}\\
&= \sum_{g=1}^{k}\sum_{h=1}^{k}d^{gq}\delta_{gh}^{(w)}d^{rh},
\end{aligned}$$

where $\delta_{gh}^{(w)} = \mathrm{tr}(\boldsymbol{V}_{gh}^{(w)})$ are the elements of the $k \times k$ matrix $\boldsymbol{\Delta}^{(w)}$. Working with the third term and noting that $\boldsymbol{T}_{ff}^{(w)}$ can be partitioned

$$\boldsymbol{T}_{ff}^{(w)} = \begin{pmatrix} \boldsymbol{T}_{ff_{11}}^{(w)} & & \\ \vdots & \ddots & \\ \boldsymbol{T}_{ff_{k1}}^{(w)} & \cdots & \boldsymbol{T}_{ff_{kk}}^{(w)} \end{pmatrix},$$

where each $\boldsymbol{T}_{ff_{gh}}^{(w)}$ is a $m \times m$ matrix, we have

$$\begin{aligned}
\mathrm{tr}\Big\{\boldsymbol{T}_{ff}^{(w)}\Big(\boldsymbol{D}^{-1}\frac{\partial \boldsymbol{D}}{\partial d_{qr}}\boldsymbol{D}^{-1}\otimes \boldsymbol{I}_m\Big)\Big\} &= \sum_{g=1}^{k}\sum_{h=1}^{k}\mathrm{tr}\big(\boldsymbol{T}_{ff_{gh}}^{(w)}d^{gq}d^{rh}\big)\\
&= \sum_{g=1}^{k}\sum_{h=1}^{k}d^{gq}\mathrm{tr}\big(\boldsymbol{T}_{ff_{gh}}^{(w)}\big)d^{rh}\\
&= \sum_{g=1}^{k}\sum_{h=1}^{k}d^{gq}\xi_{gh}^{(w)}d^{rh},
\end{aligned}$$

where $\xi_{gh}^{(w)} = \mathrm{tr}(\boldsymbol{T}_{ff_{gh}}^{(w)})$ are the elements of the $k \times k$ matrix $\boldsymbol{\Xi}^{(w)}$. Equating (6.4) to zero we have

$$md^{qr} = \sum_{g=1}^{k}\sum_{h=1}^{k}d^{gq}\delta_{gh}^{(w)}d^{rh} + \sum_{g=1}^{k}\sum_{h=1}^{k}d^{gq}\xi_{gh}^{(w)}d^{rh},$$

which can be written in matrix notation as

$$m\boldsymbol{D}^{-1} = \boldsymbol{D}^{-1}\boldsymbol{\Delta}^{(w)}\boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{\Xi}^{(w)}\boldsymbol{D}^{-1}.$$

Pre and post multiplying by $\boldsymbol{D}$ we have

$$m\boldsymbol{D} = \boldsymbol{\Delta}^{(w)} + \boldsymbol{\Xi}^{(w)}.$$

The update for d_{qr} is therefore

$$md_{qr}^{(w+1)} = \delta_{qr}^{(w)} + \xi_{qr}^{(w)}.$$

Using the identity

$$\delta_{qr}^{(w)} = \mathrm{tr}\big(\boldsymbol{V}_{qr}^{(w)}\big) = \mathrm{tr}\big(\tilde{\boldsymbol{f}}_r^{(w)}\tilde{\boldsymbol{f}}_q^{(w)^T}\big) = \tilde{\boldsymbol{f}}_q^{(w)^T}\tilde{\boldsymbol{f}}_r^{(w)},$$

the update for d_{qr} can also be expressed as

$$d_{qr}^{(w+1)} = \frac{1}{m}\Big\{\tilde{\boldsymbol{f}}_q^{(w)^T}\tilde{\boldsymbol{f}}_r^{(w)} + \mathrm{tr}\big(\boldsymbol{T}_{ff_{qr}}^{(w)}\big)\Big\}.$$

This completes the M-step.

6.5 Example: National Variety Trial data

The purpose of this example is to investigate algorithm performance. We do not provide an interpretation of the results from the models fitted. The interpretation of factor analytic mixed models applied to plant breeding data can be found in Smith et al. (2001). In the following when we refer to Model $\mathcal{O}$ or Model $\mathcal{M}$ it should be read as "the REML EM algorithm used to estimate the variance parameters associated with Model A " where Model A is Model $\mathcal{O}$ or Model $\mathcal{M}$ as the case may be.

Model $\mathcal{O}$ and Model $\mathcal{M}$ with $k = 1$ factors were applied to the small subset of National Variety Trial grain yield data presented in Table 6.2. The subset of data we consider comprises 588 observations made up of 42 varieties grown in 7 trials with 2 replicates of every variety in each trial. In this and later chapters we do not consider models with $k > 1$ factors for two reasons. Firstly, we will not have to consider the effect of placing arbitrary constraints on factor loadings on algorithm performance. This would have been a concern when we consider a REML PX EM algorithm in the next chapter where the parameter expansion is associated with the

factor loadings. Secondly, fitting factor analytic mixed models where $k > 1$ involves a model building process. The most important stage of this process is fitting a factor analytic mixed model with $k = 1$ factors as the parameter estimates obtained from this model can be used to inform the starting values for subsequent models. We do not believe that restricting ourselves to factor analytic mixed models with $k = 1$ factors diminishes the results and conclusions presented in this and the next two chapters.

For both models we define the starting values $\boldsymbol{\lambda}^{(0)} = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1)^T$, $\psi^{(0)} = 0.1$, and $\sigma^{2(0)} = 0.1$. For Model $\mathcal{M}$ we define $d^{(0)} = 1$. The criterion presented in (4.18), where $\boldsymbol{\kappa}$ is replaced by $\boldsymbol{\Gamma}$, was used to determine if the sequence $\boldsymbol{\Gamma}^{(0)}, \boldsymbol{\Gamma}^{(1)}, \ldots$ converged to a (local) maximum of the residual log-likelihood function. We assess whether or not the stopping rule with $\varepsilon = 10^{-8}$ is too stringent by considering different levels of ε. In Table 6.1 we present the different levels of ε, iterations until convergence, the final residual log-likelihood (ℓ_R), and the Frobenius matrix norm of the estimated genetic variance-covariance matrix when using Model $\mathcal{O}$ and Model $\mathcal{M}$ with $k = 1$ factors.

Model $\mathcal{M}$ converges in fewer iterations than Model $\mathcal{O}$ for all levels of ε. If we consider an accurate estimate of the genetic variance-covariance matrix to be one that has the same Frobenius matrix norm to 5 decimal places as that calculated when $\varepsilon = 10^{-8}$ then, based on the information presented in Table 6.1, it is possible to consider a less stringent convergence criterion for both models. For Model $\mathcal{M}$ there is no cost in terms of the final residual log-likelihood and very little cost in terms of an accurate estimate of genetic variance-covariance matrix by increasing ε to 10^{-5}. Likewise, for Model $\mathcal{O}$ there is little cost in increasing ε to 10^{-6}. At these levels of ε Model $\mathcal{O}$ and Model $\mathcal{M}$ converged in 76 and 38 iterations respectively. Further reductions in the number of iterations until convergence are possible by considering parameter expanded versions of Model $\mathcal{O}$ and Model $\mathcal{M}$. These are considered in the next chapter.

Table 6.1: Iterations until convergence, final residual log-likelihood (ℓ_R), and the Frobenius matrix norm of the estimated genetic variance-covariance matrix when using a REML EM algorithm to fit Model $\mathcal{O}$ and Model $\mathcal{M}$ with $k = 1$ factors and using a stopping rule with different levels of ε. For Model $\mathcal{O}$ and Model $\mathcal{M}$ with $k = 1$ factors there are 9 and 10 variance parameters to be estimated respectively.

	Model $\mathcal{O}$			Model $\mathcal{M}$		
ε	Iter.	ℓ_R	$\|\widehat{\boldsymbol{G}}_e\|_F$	Iter.	ℓ_R	$\|\widehat{\boldsymbol{G}}_e\|_F$
10^{-8}	104	300.9849	0.3180123	66	300.9849	0.3180123
10^{-7}	90	300.9849	0.3180121	56	300.9849	0.3180123
10^{-6}	76	300.9849	0.3180101	47	300.9849	0.3180125
10^{-5}	63	300.9849	0.3179936	38	300.9849	0.3180146
10^{-4}	49	300.9849	0.3178231	29	300.9848	0.3180426
10^{-3}	35	300.9842	0.3161148	21	300.9778	0.3182776
10^{-2}	21	300.8976	0.3002162	15	300.8208	0.3185271

Table 6.2: National Variety Trial grain yield data (tonnes per hectare) for 7 trials comprising 2 replicates, labelled "I" and "II", of 42 varieties.

	Trial 1		Trial 2		Trial 3		Trial 4		Trial 5		Trial 6		Trial 7	
	I	II	I	II	I	II	I	II	I	II	I	II	I	II
V1	5.14	5.27	2.63	0.90	3.18	3.24	2.08	2.18	4.83	5.31	2.85	3.41	3.51	3.26
V2	5.60	5.28	1.91	1.89	3.03	2.66	2.23	2.00	4.13	4.32	2.75	2.52	2.80	2.47
V3	5.22	4.69	2.46	1.90	2.86	2.76	2.09	1.86	4.51	4.40	2.58	2.70	2.80	2.85
V4	4.73	5.23	2.62	1.33	3.08	2.60	1.93	1.73	4.62	4.65	2.69	3.06	2.64	2.13
V5	4.89	5.51	3.24	2.18	2.97	2.87	2.49	2.20	4.67	4.65	2.78	3.18	3.35	3.09
V6	5.21	4.82	2.29	1.99	3.19	3.13	1.81	1.90	4.26	4.26	2.51	2.84	3.38	3.53
V7	5.36	4.93	2.06	1.98	3.18	3.04	2.26	2.34	4.54	5.09	2.62	3.09	3.10	2.83
V8	5.52	5.46	2.32	2.49	3.15	3.11	2.09	2.49	5.01	5.08	2.79	3.28	3.16	3.24
V9	4.93	4.97	2.62	2.82	2.86	3.01	2.46	2.11	4.51	4.67	2.72	2.68	2.81	2.97
V10	5.51	5.22	1.89	2.18	3.03	2.97	1.99	2.04	5.04	4.54	2.35	2.49	2.67	2.60
V11	6.03	5.51	2.89	2.63	2.92	3.27	2.14	2.34	4.52	5.06	3.26	3.34	3.45	3.23
V12	3.83	4.46	0.79	0.70	1.79	1.74	1.86	2.08	2.97	3.48	1.31	1.62	2.09	2.21
V13	5.63	5.07	1.54	1.45	2.51	2.31	1.80	1.81	4.35	4.62	2.84	3.03	2.57	2.76
V14	6.21	5.79	3.37	1.20	2.86	2.86	2.03	2.46	5.11	5.20	2.99	3.19	3.43	2.97
V15	6.09	6.11	2.66	1.65	2.64	2.80	2.16	2.33	4.86	4.95	3.14	3.23	3.29	3.39
V16	6.36	6.08	2.26	2.24	2.73	2.50	2.27	2.52	5.21	5.01	2.99	2.83	2.86	2.88
V17	5.84	5.21	2.13	2.35	2.91	3.20	2.38	2.29	4.60	4.87	3.18	2.90	3.39	3.14
V18	5.56	5.52	2.69	2.05	2.95	2.97	2.06	2.12	4.95	4.91	2.97	2.77	3.31	3.06
V19	5.32	5.50	2.68	2.16	3.33	3.17	2.24	2.08	5.21	5.11	3.01	3.41	3.43	3.25
V20	5.45	5.72	2.51	2.73	3.20	2.89	2.20	2.32	5.05	5.43	2.91	3.13	3.48	3.55
V21	6.03	5.87	2.27	1.89	2.75	2.52	3.05	2.28	4.71	5.01	2.99	3.30	3.39	2.79
V22	5.65	5.55	2.01	1.77	2.85	2.94	2.05	2.26	4.20	4.55	2.73	3.19	3.48	2.82
V23	5.29	5.28	0.69	1.71	2.39	2.55	2.01	2.25	3.38	3.23	2.26	2.26	2.89	2.55
V24	5.92	5.82	1.89	2.07	3.12	2.92	2.14	2.27	4.36	4.64	2.74	3.06	2.96	3.03
V25	5.36	5.74	2.59	2.11	3.08	2.68	2.16	2.61	4.01	4.56	2.76	2.57	3.21	2.98
V26	5.24	5.35	2.32	2.12	2.49	2.73	2.07	2.09	4.13	4.00	2.58	2.92	2.63	2.39
V27	5.27	5.24	1.56	1.99	2.60	2.57	2.16	2.47	5.20	4.88	2.58	3.31	3.40	2.79
V28	5.19	5.34	3.14	1.88	3.23	3.11	2.32	2.17	4.97	4.88	2.99	2.81	2.37	2.37
V29	5.19	4.99	1.25	1.56	2.93	2.94	2.13	1.87	4.36	4.11	2.78	2.39	3.41	3.13
V30	5.33	5.08	2.03	2.04	3.30	2.93	2.36	2.36	5.34	5.08	3.14	3.46	3.17	3.20
V31	5.76	5.58	2.01	1.83	2.98	2.92	2.08	2.22	4.85	4.71	3.03	3.40	2.68	2.86
V32	5.67	5.51	1.90	2.38	2.76	3.45	2.30	2.28	4.89	5.22	2.99	3.07	3.47	3.09
V33	5.17	5.00	2.25	2.15	2.52	2.91	2.54	2.18	4.16	4.39	2.58	2.76	2.81	2.68
V34	5.20	5.06	2.14	2.09	2.61	3.10	1.99	2.39	4.69	5.01	2.42	2.67	2.90	2.87
V35	5.44	5.01	2.39	2.38	3.14	2.99	2.05	2.09	4.45	4.54	2.51	3.40	3.27	3.15
V36	5.68	5.74	2.17	2.49	3.21	3.29	2.17	2.05	4.37	5.06	2.85	3.06	3.28	3.46
V37	6.18	5.86	1.24	1.36	2.30	2.14	2.26	2.17	4.45	4.48	2.86	2.36	2.50	2.51
V38	5.23	5.37	2.04	2.36	3.09	2.69	2.12	2.45	4.30	4.86	3.29	3.16	2.85	2.45
V39	5.76	5.46	2.16	1.26	2.98	2.87	2.08	2.51	5.05	5.56	3.16	2.96	3.44	3.40
V40	5.46	5.17	1.71	1.41	2.68	2.49	2.03	2.43	4.41	4.85	3.04	2.82	3.15	2.97
V41	5.25	4.84	3.02	2.12	2.95	3.17	2.28	2.32	4.46	4.68	2.59	2.42	3.62	3.07
V42	5.24	5.16	2.85	2.17	3.09	2.95	1.95	2.00	4.44	3.89	3.01	3.04	2.21	2.71

Chapter 7

REML PX EM algorithms for two factor analytic mixed model specifications

In this chapter we consider REML PX EM algorithms based on the two factor analytic mixed model specifications considered in the previous chapter, which were referred to as Model $\mathcal{O}$ and Model $\mathcal{M}$. The difference between the two being that for Model $\mathcal{O}$ the variance of the factor scores is the identity matrix whereas for Model $\mathcal{M}$ it comprises an unstructured matrix.

The parameter expansion of Model $\mathcal{O}$ will be referred to as Model $\mathcal{X}$, this is the same notation used by Liu et al. (1998). Model $\mathcal{M}$ and Model $\mathcal{X}$ are very similar and it is of interest to compare the two. However, the main purpose of this chapter is to present a parameter expansion of Model $\mathcal{M}$, which will be referred to as Model $\mathcal{Z}$. The specification of Model $\mathcal{Z}$ is new. In this model the parameter expansion is associated with the specific variances rather than the factor loadings as in Model $\mathcal{X}$. Using the example data considered in the previous chapter it is shown that this model converges in less iterations than either Model $\mathcal{X}$ or Model $\mathcal{M}$. We consider Model $\mathcal{X}$ and Model $\mathcal{Z}$ in turn.

7.1 REML PX EM algorithm - Model $\mathcal{X}$

Using the parameter expansion considered by Liu et al. (1998) we define the parameter expanded model

$$\text{Model } \mathcal{X} \qquad \boldsymbol{y}_o = \boldsymbol{X}\boldsymbol{\beta}_* + \boldsymbol{Z}_g\boldsymbol{u}_g + \boldsymbol{e},$$

where $\boldsymbol{u}_g = (\boldsymbol{\Lambda}_* \otimes \boldsymbol{I}_m)\boldsymbol{f} + \boldsymbol{\delta}$. We assume $\boldsymbol{\delta} \sim N(\boldsymbol{0}, \boldsymbol{\Psi}_* \otimes \boldsymbol{I}_m)$ where $\boldsymbol{\Psi}_* = \psi_*\boldsymbol{I}_p$ and $\boldsymbol{f} \sim N(\boldsymbol{0}, \boldsymbol{D} \otimes \boldsymbol{I}_m)$. We assume that the $n \times 1$ vector of residuals is distributed $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{R}_*)$ where $\boldsymbol{R}_* = \sigma^2_*\boldsymbol{I}_n$. The difference between Model $\mathcal{O}$ and Model $\mathcal{X}$ is the assumed distribution of the $mk \times 1$ vector $\boldsymbol{f}$. In Model $\mathcal{X}$ it is assumed that $\boldsymbol{f} \sim N(\boldsymbol{0}, \boldsymbol{D} \otimes \boldsymbol{I}_m)$ whereas in Model $\mathcal{O}$ it is assumed that $\boldsymbol{f} \sim N(\boldsymbol{0}, \boldsymbol{I}_{mk})$. We define $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{d})$ as a $k \times k$ symmetric positive definite matrix where $\boldsymbol{d}$ is a vector of variance parameters.

We follow the convention established by Liu et al. (1998) and define the expanded parameter set associated with Model $\mathcal{X}$ as $\boldsymbol{\Gamma}_* = (\sigma^2_*, \boldsymbol{\lambda}^T_*, \psi_*, \boldsymbol{d}^T)^T$ where the subscript $*$ is used to distinguish parameters associated with Model $\mathcal{O}$ and Model $\mathcal{X}$. For the parameter expanded model the distribution of the observed data is

$$\boldsymbol{y}_o \sim N(\boldsymbol{X}\boldsymbol{\beta}_*, \boldsymbol{H}_*),$$

where $\boldsymbol{H}_* = (\boldsymbol{\Lambda}_*\boldsymbol{D}\boldsymbol{\Lambda}^T_* + \boldsymbol{\Psi}_*) \otimes \boldsymbol{I}_m$. Using the parameter expanded model we identify the parameters associated with Model $\mathcal{O}$ as $\boldsymbol{\Gamma} = (\sigma^2, \boldsymbol{\lambda}^T, \psi)^T = \mathcal{R}(\boldsymbol{\Gamma}_*) = \left[\sigma^2_*, \text{vec}\{\boldsymbol{\Lambda}_*\text{Chol}(\boldsymbol{D})\}, \psi_*\right]^T$ where Chol denotes the Cholesky decomposition and $\mathcal{R}$ is the reduction function from the expanded parameter set to the original parameter set.

7.1.1 Complete data density function

For Model $\mathcal{X}$ the complete data specification remains the same as that for Model $\mathcal{O}$, i.e., $\boldsymbol{y}_c = (\boldsymbol{y}^T_2, \boldsymbol{f}^T, \boldsymbol{\delta}^T)^T$ where the missing data vector is define as $\boldsymbol{y}_m = (\boldsymbol{f}^T, \boldsymbol{\delta}^T)^T$. However, the complete data log-density function is different in that it contains the so-called auxillary parameter $\boldsymbol{d}$. We use the subscript $\mathcal{X}$ to distinguish between the complete data log-density function and $Q(.)$ function for Model $\mathcal{O}$ and Model $\mathcal{X}$. The complete data log-density function for the latter is

$$\log\{f_{\mathcal{X}}(\boldsymbol{y}_c;\boldsymbol{\Gamma}_*)\} = \log\{f_{\mathcal{X}}(\boldsymbol{y}_2|\boldsymbol{f},\boldsymbol{\delta};\sigma^2_*,\boldsymbol{\lambda}_*)\} + \log\{f_{\mathcal{X}}(\boldsymbol{f};\boldsymbol{d})\} + \log\{f_{\mathcal{X}}(\boldsymbol{\delta};\psi_*)\}, \quad (7.1)$$

where

$$\log\{f_{\mathcal{X}}(\boldsymbol{f};\boldsymbol{d})\} = -\frac{1}{2}\Big[\log\{\det(\boldsymbol{D}\otimes\boldsymbol{I}_m)\} + \boldsymbol{f}^T(\boldsymbol{D}^{-1}\otimes\boldsymbol{I}_m)\boldsymbol{f}\Big],$$
$$\log\{f_{\mathcal{X}}(\boldsymbol{\delta};\psi_*)\} = -\frac{1}{2}\Big[\log\{\det(\boldsymbol{\Psi}_*\otimes\boldsymbol{I}_m)\} + \boldsymbol{\delta}^T(\boldsymbol{\Psi}_*^{-1}\otimes\boldsymbol{I}_m)\boldsymbol{\delta}\Big]$$

and

$$\begin{aligned}&\log\{f_{\mathcal{X}}(\boldsymbol{y}_2|\boldsymbol{f},\boldsymbol{\delta};\sigma^2_*,\boldsymbol{\lambda}_*)\}\\ &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{L}_2^T\boldsymbol{R}_*\boldsymbol{L}_2)\}\\ &+\{\boldsymbol{y}_o - \boldsymbol{Z}_g(\boldsymbol{\Lambda}_*\otimes\boldsymbol{I}_m)\boldsymbol{f} - \boldsymbol{Z}_g\boldsymbol{\delta}\}^T\boldsymbol{S}_*\{\boldsymbol{y}_o - \boldsymbol{Z}_g(\boldsymbol{\Lambda}_*\otimes\boldsymbol{I}_m)\boldsymbol{f} - \boldsymbol{Z}_g\boldsymbol{\delta}\}\Big],\end{aligned}$$

where $\boldsymbol{S}_* = \boldsymbol{R}_*^{-1} - \boldsymbol{R}_*^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{R}_*^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{R}_*^{-1}$. This is the same as the complete data density function for Model $\mathcal{M}$ where the variance parameters σ^2, $\boldsymbol{\lambda}$, and ψ have been substituted with σ^2_*, $\boldsymbol{\lambda}_*$, and ψ_*.

7.2 REML PX EM algorithm E-step for Model $\mathcal{X}$

For Model $\mathcal{X}$ we define $\boldsymbol{\Gamma}_*^{(w)} = (\sigma^2_*, \boldsymbol{\lambda}_*^{(w)^T}, \psi_*^{(w)}, \boldsymbol{d}_0^T)^T$ where $\boldsymbol{D}(\boldsymbol{d}_0) = \boldsymbol{I}_k$. The E-step of the REML PX EM algorithm involves calculating $Q_{\mathcal{X}}(\boldsymbol{\Gamma}_*;\boldsymbol{\Gamma}_*^{(w)})$ which is equivalent to the E-step of the REML PX EM algorithm for Model $\mathcal{O}$. Note that for Model $\mathcal{O}$ and Model $\mathcal{X}$

$$\begin{aligned}\tilde{\boldsymbol{f}}^{(w)} &= (\boldsymbol{\Lambda}^{(w)^T}\otimes\boldsymbol{I}_m)\boldsymbol{P}^{(w)}\boldsymbol{y}_o,\\ \boldsymbol{T}_{ff}^{(w)} &= \boldsymbol{I}_{mk} - (\boldsymbol{\Lambda}^{(w)^T}\otimes\boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{P}^{(w)}\boldsymbol{Z}_g(\boldsymbol{\Lambda}^{(w)}\otimes\boldsymbol{I}_m),\\ \boldsymbol{T}_{f\delta}^{(w)} &= -(\boldsymbol{\Psi}^{(w)}\otimes\boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{P}^{(w)}\boldsymbol{Z}_g(\boldsymbol{\Lambda}^{(w)}\otimes\boldsymbol{I}_m).\end{aligned}$$

7.3 REML PX EM algorithm M-step for Model $\mathcal{X}$

The M-step of the REML PX-EM algorithm involves maximising $Q_{\mathcal{X}}(\boldsymbol{\Gamma}_*; \boldsymbol{\Gamma}_*^{(w)})$ with respect to $\boldsymbol{\Gamma}_* = (\sigma_*^2, \boldsymbol{\lambda}_*^T, \psi_*, \boldsymbol{d}^T)^T$ and then applying the reduction function $\mathcal{R}(\boldsymbol{\Gamma}_*) = [\sigma_*^2, \text{vec}\{\boldsymbol{\Lambda}_* \text{Chol}(\boldsymbol{D})\}, \psi_*]^T$ to obtain estimates of $\boldsymbol{\Gamma}^{(w+1)} = (\sigma^{2^{(w+1)}}, \boldsymbol{\lambda}^{(w+1)^T}, \psi^{(w+1)})^T$.

7.3.1 Update for σ_*^2

The update for σ_*^2 is the same as the update for σ^2 under Model $\mathcal{O}$, i.e.,

$$\begin{aligned}\sigma_*^{2^{(w+1)}} = \frac{1}{n-p}\Big[&\{\boldsymbol{y}_o - \boldsymbol{Z}_g(\boldsymbol{\Lambda}^{(w)} \otimes \boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} - \boldsymbol{Z}_g\tilde{\boldsymbol{\delta}}^{(w)}\}^T \boldsymbol{K}\{\boldsymbol{y}_o - \boldsymbol{Z}_g(\boldsymbol{\Lambda}^{(w)} \otimes \boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} - \boldsymbol{Z}_g\tilde{\boldsymbol{\delta}}^{(w)}\} \\ &+\text{tr}\Big\{(\boldsymbol{\Lambda}^{(w)^T} \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{K}\boldsymbol{Z}_g(\boldsymbol{\Lambda}^{(w)} \otimes \boldsymbol{I}_m)\boldsymbol{T}_{ff}^{(w)}\Big\} + \text{tr}\Big(\boldsymbol{Z}_g^T\boldsymbol{K}\boldsymbol{Z}_g\boldsymbol{T}_{\delta\delta}^{(w)}\Big) \\ &+2\text{tr}\Big\{\boldsymbol{Z}_g^T\boldsymbol{K}\boldsymbol{Z}_g(\boldsymbol{\Lambda}^{(w)} \otimes \boldsymbol{I}_m)(\tilde{\boldsymbol{f}}^{(w)}\tilde{\boldsymbol{\delta}}^{(w)^T} + \boldsymbol{T}_{f\delta}^{(w)})\Big\}\Big].\end{aligned}$$

Applying the reduction function we have $\sigma^{2^{(w+1)}} = \sigma_*^{2^{(w+1)}}$.

7.3.2 Update for $\lambda_{*_{jr}}$

The update for $\lambda_{*_{jr}}$ is the same as that presented for λ_{jr} under Model $\mathcal{O}$ except we substitute $\boldsymbol{\Lambda}_*$ and $\boldsymbol{S}_*$ for $\boldsymbol{\Lambda}$ and $\boldsymbol{S}$ respectively, i.e.,

$$\boldsymbol{\lambda}_*^{(w+1)} = \boldsymbol{A}_*^{(w)^{-1}}\boldsymbol{b}_*^{(w)},$$

where the elements of $\boldsymbol{A}_*^{(w)}$ are

$$\begin{aligned}a_{*_{jr,lq}}^{(w)} &= E\{\boldsymbol{f}^T(\boldsymbol{J}_{lq}^T \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{S}_*\boldsymbol{Z}_g(\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)\boldsymbol{f}|\boldsymbol{y}_2; \boldsymbol{\Gamma}_*^{(w)}\} \\ &= \tilde{\boldsymbol{f}}^{(w)^T}(\boldsymbol{J}_{lq}^T \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{S}^{(w)}\boldsymbol{Z}_g(\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} \\ &\quad +\text{tr}\{(\boldsymbol{J}_{lq}^T \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{S}^{(w)}\boldsymbol{Z}_g(\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)\boldsymbol{T}_{ff}^{(w)}\}\end{aligned}$$

and the elements of $\boldsymbol{b}_*^{(w)}$ are

$$\begin{aligned} b_{*jr}^{(w)} &= E\{\boldsymbol{y}_o^T \boldsymbol{S}_* \boldsymbol{Z}_g (\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m) \boldsymbol{f} | \boldsymbol{y}_2; \boldsymbol{\Gamma}_*^{(w)}\} \\ &\quad -\mathrm{tr}\{E(\boldsymbol{Z}_g^T \boldsymbol{S}_*^{(w)} \boldsymbol{Z}_g (\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m) \boldsymbol{f}\boldsymbol{\delta}^T | \boldsymbol{y}_2; \boldsymbol{\Gamma}_*^{(w)})\} \\ &= \boldsymbol{y}_o^T \boldsymbol{S}^{(w)} \boldsymbol{Z}_g (\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m) \tilde{\boldsymbol{f}}^{(w)} - \mathrm{tr}\{\boldsymbol{Z}_g^T \boldsymbol{S}^{(w)} \boldsymbol{Z}_g (\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)(\tilde{\boldsymbol{f}}^{(w)} \tilde{\boldsymbol{\delta}}^{(w)^T} + \boldsymbol{T}_{f\delta}^{(w)})\}. \end{aligned}$$

Applying the reduction function we have $\boldsymbol{\lambda}^{(w+1)} = \mathrm{vec}\{\boldsymbol{\Lambda}_*^{(w+1)} \mathrm{Chol}(\boldsymbol{D}^{(w+1)})\}$.

7.3.3 Update for ψ_*

The update for ψ_* is the same as the update for ψ under Model $\mathcal{O}$ where ψ_* replaces ψ, i.e.,

$$\psi_*^{(w+1)} = \frac{1}{mp}\left\{\tilde{\boldsymbol{\delta}}^{(w)^T} \tilde{\boldsymbol{\delta}}^{(w)} + \mathrm{tr}\left(\boldsymbol{T}_{\delta\delta}^{(w)}\right)\right\}$$

Applying the reduction function we have $\psi^{(w+1)} = \psi_*^{(w+1)}$.

7.3.4 Update for d_{qr}

The update for d_{qr} is

$$d_{qr}^{(w+1)} = \frac{1}{m}\left\{\tilde{\boldsymbol{f}}_q^{(w)^T} \tilde{\boldsymbol{f}}_r^{(w)} + \mathrm{tr}\left(\boldsymbol{T}_{ff_{qr}}^{(w)}\right)\right\},$$

which is similar to the update for d_{qr} under Model $\mathcal{M}$ presented in the previous chapter. The only difference being the form of $\tilde{\boldsymbol{f}}^{(w)}$ and $\boldsymbol{T}_{ff}^{(w)}$. Under Model $\mathcal{M}$ these contain the unstructured matrix $\boldsymbol{D}$ whereas for the model we are considering, Model $\mathcal{X}$, it is assumed that $\boldsymbol{D} = \boldsymbol{I}_k$.

7.4 REML PX EM algorithm - Model $\mathcal{Z}$

We now consider a new specification of a REML PX EM algorithm for a factor analytic mixed model. Consider the parameter expanded model

$$\text{Model } \mathcal{Z} \qquad \boldsymbol{y}_o = \boldsymbol{X}\boldsymbol{\beta}_* + \boldsymbol{Z}_g \boldsymbol{u}_g + \boldsymbol{e}$$

where $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{R}_*)$, $\boldsymbol{R}_* = \sigma_*^2 \boldsymbol{I}_n$ and $\boldsymbol{u}_g = (\boldsymbol{\Lambda}_* \otimes \boldsymbol{I}_m)\boldsymbol{f} + (\boldsymbol{\Omega} \otimes \boldsymbol{I}_m)\boldsymbol{\delta}$, i.e., we introduce the $p \times p$ positive definite diagonal matrix $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\omega})$. When we assume a common specific variance for $\boldsymbol{\delta}$, i.e., $\boldsymbol{\delta} \sim N(\boldsymbol{0}, \boldsymbol{\Psi}_* \otimes \boldsymbol{I}_m)$ where $\boldsymbol{\Psi}_* = \psi_* \boldsymbol{I}_{mp}$ then $\boldsymbol{\Omega}$ has the form $\boldsymbol{\Omega} = \omega \boldsymbol{I}_p$. We assume that $\boldsymbol{\delta}$ and the $mk \times 1$ vector $\boldsymbol{f} \sim N(\boldsymbol{0}, \boldsymbol{D}_* \otimes \boldsymbol{I}_m)$ are independent of each other. We define $\boldsymbol{D}_* = \boldsymbol{D}_*(\boldsymbol{d}_*)$ as a $k \times k$ symmetric positive definite matrix where $\boldsymbol{d}_*$ is a vector of variance parameters.

We define the expanded parameter set associated with Model $\mathcal{Z}$ as $\boldsymbol{\mathcal{T}} = (\sigma_*^2, \boldsymbol{\lambda}_*^T, \psi_*, \boldsymbol{d}_*^T, \boldsymbol{\omega}^T)^T$ where the subscript $*$ is used to distinguish parameters associated with Model $\mathcal{M}$ and Model $\mathcal{Z}$. For the parameter expanded model the distribution of the observed data is

$$\boldsymbol{y}_o \sim N(\boldsymbol{X}\boldsymbol{\beta}_*, \boldsymbol{H}_*),$$

where $\boldsymbol{H}_* = \boldsymbol{Z}_g\{(\boldsymbol{\Lambda}_*\boldsymbol{D}_*\boldsymbol{\Lambda}_*^T + \boldsymbol{\Omega}\boldsymbol{\Psi}_*\boldsymbol{\Omega}) \otimes \boldsymbol{I}_m\}\boldsymbol{Z}_g^T + \boldsymbol{R}_*$. Using the parameter expanded model we identify the parameters associated with Model $\mathcal{M}$ as $\boldsymbol{\Gamma} = (\sigma^2, \boldsymbol{\lambda}^T, \psi, \boldsymbol{d}^T)^T = \mathcal{R}(\boldsymbol{\mathcal{T}}) = (\sigma_*^2, \boldsymbol{\lambda}_*, \omega^2\psi_*, \boldsymbol{d}_*^T)^T$ where $\mathcal{R}$ is the reduction function from the expanded parameter set to the original parameter set.

The difference between the parameter expanded Model $\mathcal{X}$ and Model $\mathcal{Z}$ is that in the former the auxillary parameter $\boldsymbol{d}$ is associated with the factor loadings in the reduction function, whereas in the later the auxillary parameter $\boldsymbol{\omega}$ is associated with the specific variances.

An important implication of the non-uniqueness of $\boldsymbol{\Lambda}_*$ and $\boldsymbol{D}_*$ for Model $\mathcal{Z}$ is that the REML PX EM algoritm applied to this model is not guaranteed to have a rate of convergence at least as fast as the REML EM algorithm applied to Model $\mathcal{M}$. We don't believe this is an impediment to the implementation of Model $\mathcal{Z}$ for the analysis of plant breeding data for two reasons. Firstly, $\widehat{\boldsymbol{G}}_e$ and $\tilde{\boldsymbol{u}}_g$ are unique; and secondly, the cases in which a REML EM algorithm applied to Model $\mathcal{M}$ would converge faster than a REML PX EM algorithm applied to Model $\mathcal{Z}$ are atypical.

7.4.1 Complete data density function

For Model $\mathcal{Z}$ the complete data specification is the same as that for Model $\mathcal{M}$, i.e., $\boldsymbol{y}_c = (\boldsymbol{y}_2^T, \boldsymbol{f}^T, \boldsymbol{\delta}^T)^T$. We use the subscript $\mathcal{Z}$ to distinguish between the complete data log-density function and $Q(.)$ function for Model $\mathcal{M}$ and Model $\mathcal{Z}$. To form the complete data density function for Model $\mathcal{Z}$ requires the joint distribution

$$\begin{pmatrix} \boldsymbol{y}_2 \\ \boldsymbol{f} \\ \boldsymbol{\delta} \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{L}_2^T \boldsymbol{H}_* \boldsymbol{L}_2 & \boldsymbol{L}_2^T \boldsymbol{Z}_g (\boldsymbol{\Lambda}_* \boldsymbol{D}_* \otimes \boldsymbol{I}_m) & \boldsymbol{L}_2^T \boldsymbol{Z}_g (\boldsymbol{\Omega}\boldsymbol{\Psi}_* \otimes \boldsymbol{I}_m) \\ (\boldsymbol{D}_* \boldsymbol{\Lambda}_*^T \otimes \boldsymbol{I}_m) \boldsymbol{Z}_g^T \boldsymbol{L}_2 & \boldsymbol{D}_* \otimes \boldsymbol{I}_m & \mathbf{0} \\ (\boldsymbol{\Psi}_* \boldsymbol{\Omega} \otimes \boldsymbol{I}_m) \boldsymbol{Z}_g^T \boldsymbol{L}_2 & \mathbf{0} & \boldsymbol{\Psi}_* \otimes \boldsymbol{I}_m \end{pmatrix} \right\}.$$

The complete data log-density function is

$$\log\{f_{\mathcal{Z}}(\boldsymbol{y}_c; \boldsymbol{\mathcal{T}})\} = \log\{f_{\mathcal{Z}}(\boldsymbol{y}_2 | \boldsymbol{f}, \boldsymbol{\delta}; \sigma_*^2, \boldsymbol{\lambda}_*, \omega)\} + \log\{f_{\mathcal{Z}}(\boldsymbol{f}; \boldsymbol{d}_*)\} + \log\{f_{\mathcal{Z}}(\boldsymbol{\delta}; \psi_*)\},$$

where

$$\begin{aligned} \log\{f_{\mathcal{Z}}(\boldsymbol{f}; \boldsymbol{d}_*)\} &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{D}_* \otimes \boldsymbol{I}_m)\} + \boldsymbol{f}^T (\boldsymbol{D}_*^{-1} \otimes \boldsymbol{I}_m) \boldsymbol{f}\Big], \\ \log\{f_{\mathcal{Z}}(\boldsymbol{\delta}; \boldsymbol{\psi}_*, \omega)\} &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{\Psi}_* \otimes \boldsymbol{I}_m)\} + \boldsymbol{\delta}^T (\boldsymbol{\Psi}_*^{-1} \otimes \boldsymbol{I}_m) \boldsymbol{\delta}\Big] \end{aligned}$$

and

$$\begin{aligned} \log\{f_{\mathcal{Z}}(\boldsymbol{y}_2 | \boldsymbol{f}, \boldsymbol{\delta}; \boldsymbol{\lambda}_*, \psi_*, \omega)\} = -\frac{1}{2}\Big[&\log\{\det(\boldsymbol{L}_2^T \boldsymbol{R}_* \boldsymbol{L}_2)\} \\ &+ \{\boldsymbol{y}_o - \boldsymbol{Z}_g(\boldsymbol{\Lambda}_* \otimes \boldsymbol{I}_m)\boldsymbol{f} - \boldsymbol{Z}_g(\boldsymbol{\Omega} \otimes \boldsymbol{I}_m)\boldsymbol{\delta}\}^T \boldsymbol{S}_* \{\boldsymbol{y}_o - \boldsymbol{Z}_g(\boldsymbol{\Lambda}_* \otimes \boldsymbol{I}_m)\boldsymbol{f} - \boldsymbol{Z}_g(\boldsymbol{\Omega} \otimes \boldsymbol{I}_m)\boldsymbol{\delta}\}\Big]. \end{aligned}$$

7.5 REML PX EM algorithm E-step for Model $\mathcal{Z}$

For Model $\mathcal{Z}$ we define $\boldsymbol{\mathcal{T}}^{(w)} = (\sigma_*^2, \boldsymbol{\lambda}_*^{(w)^T}, \psi_*^{(w)}, \boldsymbol{d}_*^{(w)^T}, \omega_0)^T$ where $\boldsymbol{\Omega}(\omega_0) = \boldsymbol{I}_p$. The E-step of the REML PX EM algorithm involves calculating $Q_{\mathcal{Z}}(\boldsymbol{\mathcal{T}}; \boldsymbol{\mathcal{T}}^{(w)})$ which is equivalent to $Q(\boldsymbol{\Gamma}; \boldsymbol{\Gamma}^{(w)})$, i.e., the E-step of the REML PX EM algorithm for Model $\mathcal{Z}$ and the E-step of the REML EM algorithm for Model $\mathcal{M}$ are the same.

7.6 REML PX EM algorithm M-step for Model $\mathcal{Z}$

The M-step of the REML PX EM algorithm involves maximising $Q_{\mathcal{Z}}(\boldsymbol{\mathcal{T}}; \boldsymbol{\mathcal{T}}^{(w)})$ wih respect to $\boldsymbol{\mathcal{T}} = (\sigma_*^2, \boldsymbol{\lambda}_*^T, \psi_*, \boldsymbol{d}_*^T, \omega)^T$ and then applying the reduction function $\mathcal{R}(\boldsymbol{\mathcal{T}})$ to obtain $\boldsymbol{\Gamma} = (\sigma^2, \boldsymbol{\lambda}^T, \psi, \boldsymbol{d}^T)^T$.

7.6.1 Update for σ_*^2

The update for σ_*^2 is the same as the update for σ^2 under Model $\mathcal{M}$, i.e.,

$$\begin{aligned}\sigma_*^{2(w+1)} &= \frac{1}{n-p}\Big[\{\boldsymbol{y}_o - \boldsymbol{Z}_g(\boldsymbol{\Lambda}^{(w)} \otimes \boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} - \boldsymbol{Z}_g\tilde{\boldsymbol{\delta}}^{(w)}\}^T\boldsymbol{K}\{\boldsymbol{y}_o - \boldsymbol{Z}_g(\boldsymbol{\Lambda}^{(w)} \otimes \boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} - \boldsymbol{Z}_g\tilde{\boldsymbol{\delta}}^{(w)}\} \\ &\quad +\mathrm{tr}\Big\{(\boldsymbol{\Lambda}^{(w)^T} \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{K}\boldsymbol{Z}_g(\boldsymbol{\Lambda}^{(w)} \otimes \boldsymbol{I}_m)\boldsymbol{T}_{ff}^{(w)}\Big\} \\ &\quad +\mathrm{tr}\Big(\boldsymbol{Z}_g^T\boldsymbol{K}\boldsymbol{Z}_g\boldsymbol{T}_{\delta\delta}^{(w)}\Big) + 2\mathrm{tr}\Big\{\boldsymbol{Z}_g^T\boldsymbol{K}\boldsymbol{Z}_g(\boldsymbol{\Lambda}^{(w)} \otimes \boldsymbol{I}_m)(\tilde{\boldsymbol{f}}^{(w)}\tilde{\boldsymbol{\delta}}^{(w)^T} + \boldsymbol{T}_{f\delta}^{(w)})\Big\}\Big].\end{aligned}$$

Applying the reduction function we have $\sigma^{2(w+1)} = \sigma_*^{2(w+1)}$.

7.6.2 Update for $\lambda_{*_{jr}}$

The derivation of the update for $\lambda_{*_{jr}}$ is the same as that for λ_{jr} under Model $\mathcal{M}$, i.e.,

$$\boldsymbol{\lambda}_*^{(w+1)} = \boldsymbol{A}_*^{(w)^{-1}}\boldsymbol{b}_*^{(w)},$$

where the elements of the $mk \times mk$ matrix $\boldsymbol{A}_*^{(w)}$ are

$$\begin{aligned}a_{*_{jr,lq}}^{(w)} &= E\{\boldsymbol{f}^T(\boldsymbol{J}_{lq}^T \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{S}_*\boldsymbol{Z}_g(\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)\boldsymbol{f}|\boldsymbol{y}_2;\boldsymbol{\mathcal{T}}^{(w)}\} \\ &= \tilde{\boldsymbol{f}}^{(w)^T}(\boldsymbol{J}_{lq}^T \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{S}^{(w)}\boldsymbol{Z}_g(\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} \\ &\quad +\mathrm{tr}\{(\boldsymbol{J}_{lq}^T \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{S}^{(w)}\boldsymbol{Z}_g(\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)\boldsymbol{T}_{ff}^{(w)}\}\end{aligned}$$

and $\boldsymbol{b}_*^{(w)}$ is a $mk \times 1$ vector with elements

$$\begin{aligned}b_{*_{jr}}^{(w)} &= E\{\boldsymbol{y}_o^T\boldsymbol{S}_*\boldsymbol{Z}_g(\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)\boldsymbol{f}|\boldsymbol{y}_2;\boldsymbol{\mathcal{T}}^{(w)}\} - \mathrm{tr}\{E\big(\boldsymbol{Z}_g^T\boldsymbol{S}_*^{(w)}\boldsymbol{Z}_g(\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)\boldsymbol{f}\boldsymbol{\delta}^T|\boldsymbol{y}_2;\boldsymbol{\mathcal{T}}^{(w)}\big)\} \\ &= \boldsymbol{y}_o^T\boldsymbol{S}^{(w)}\boldsymbol{Z}_g(\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} - \mathrm{tr}\{\boldsymbol{Z}_g^T\boldsymbol{S}^{(w)}\boldsymbol{Z}_g(\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)(\tilde{\boldsymbol{f}}^{(w)}\tilde{\boldsymbol{\delta}}^{(w)^T} + \boldsymbol{T}_{f\delta}^{(w)})\}.\end{aligned}$$

Applying the reduction function we have $\boldsymbol{\lambda}^{(w+1)} = \boldsymbol{\lambda}_*^{(w+1)}$.

7.6.3 Update for ψ_*

The update for ψ_* is the same as that presented for ψ under Model $\mathcal{M}$, i.e.,

$$\psi_*^{(w+1)} = \frac{1}{mp}\Big\{\tilde{\boldsymbol{\delta}}^{(w)^T}\tilde{\boldsymbol{\delta}}^{(w)} + \mathrm{tr}\big(\boldsymbol{T}_{\delta\delta}^{(w)}\big)\Big\}.$$

Applying the reduction function we have $\psi^{(w+1)} = \omega^{2(w+1)}\psi_*^{(w+1)}$.

7.6.4 Update for $d_{*_{qr}}$

The update for $d_{*_{qr}}$ is the same as that for d_{qr} under Model $\mathcal{M}$, i.e.,

$$d_{*_{qr}}^{(w+1)} = \frac{1}{m}\Big\{\tilde{\boldsymbol{f}}_q^{(w)^T}\tilde{\boldsymbol{f}}_r^{(w)} + \mathrm{tr}\big(\boldsymbol{T}_{ff_{qr}}^{(w)}\big)\Big\}.$$

Applying the reduction function we have $d_{qr}^{(w+1)} = d_{*_{qr}}^{(w+1)}$.

7.6.5 Update for ω

Differentiating $Q_{\mathcal{Z}}(\boldsymbol{\mathcal{T}};\boldsymbol{\mathcal{T}}^{(w)})$ with respect to ω we have

$$\begin{aligned}\frac{\partial Q_{\mathcal{Z}}(\boldsymbol{\mathcal{T}};\boldsymbol{\mathcal{T}}^{(w)})}{\partial\omega} &= -\frac{1}{2}\Big\{-2\boldsymbol{y}_o^T\boldsymbol{S}\boldsymbol{Z}_g\tilde{\boldsymbol{\delta}}^{(w)} \\ &\quad +2\mathrm{tr}\big\{(\boldsymbol{\Lambda}^T\otimes\boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{S}\boldsymbol{Z}_g(\tilde{\boldsymbol{f}}^{(w)}\tilde{\boldsymbol{\delta}}^{(w)^T} + \boldsymbol{T}_{f\delta}^{(w)})\big\} \\ &\quad +2\omega\tilde{\boldsymbol{\delta}}^{(w)^T}\boldsymbol{Z}_g^T\boldsymbol{S}\boldsymbol{Z}_g\tilde{\boldsymbol{\delta}}^{(w)} + 2\omega\mathrm{tr}\big(\boldsymbol{Z}_g^T\boldsymbol{S}\boldsymbol{Z}_g\boldsymbol{T}_{\delta\delta}^{(w)}\big)\Big\}\end{aligned}$$

Equating to zero and solving for ω we have

$$\omega^{(w+1)} = \frac{\boldsymbol{y}_o^T\boldsymbol{S}^{(w)}\boldsymbol{Z}_g\tilde{\boldsymbol{\delta}}^{(w)} - \mathrm{tr}\big\{(\boldsymbol{\Lambda}^{(w)^T}\otimes\boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{S}^{(w)}\boldsymbol{Z}_g(\tilde{\boldsymbol{f}}^{(w)}\tilde{\boldsymbol{\delta}}^{(w)^T} + \boldsymbol{T}_{f\delta}^{(w)})\big\}}{\tilde{\boldsymbol{\delta}}^{(w)^T}\boldsymbol{Z}_g^T\boldsymbol{S}^{(w)}\boldsymbol{Z}_g\tilde{\boldsymbol{\delta}}^{(w)} + \mathrm{tr}\big(\boldsymbol{Z}_g^T\boldsymbol{S}^{(w)}\boldsymbol{Z}_g\boldsymbol{T}_{\delta\delta}^{(w)}\big)}$$

7.7 Example: National Variety Trial data

In the following when we refer to Model $\mathcal{O}$ or Model $\mathcal{M}$ it should be read as "the REML EM algorithm used to estimate the variance parameters associated with

Model A " where Model A is Model $\mathcal{O}$ or Model $\mathcal{M}$ as the case may be. When we refer to Model $\mathcal{X}$ or Model $\mathcal{Z}$ it should be read as "the REML PX EM algorithm used to estimate the variance parameters associated with Model B " where Model B is Model $\mathcal{X}$ or Model $\mathcal{Z}$ as the case may be. We consider the small subset of National Variety Trial data used in the previous two chapters and presented in Table 6.2.

For Model $\mathcal{X}$ and Model $\mathcal{Z}$ with $k = 1$ factors we define the starting values $\boldsymbol{\lambda}^{(0)} = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1)^T$, $\psi^{(0)} = 0.1$, $\sigma^{2(0)} = 0.1$, and $d^{(0)} = 1$. For Model $\mathcal{Z}$ we define $\omega^{(0)} = 1$. In Table 7.1 iterations until convergence and the final residual log-likelihood are considered for different levels of ε for both models.

If we compare the Table 7.1 and Table 6.1 we see that Model $\mathcal{X}$ converges in significantly fewer iterations than Model $\mathcal{O}$ and Model $\mathcal{Z}$ converges in 1 or 2 iterations less than Model $\mathcal{M}$. Of particular interest is the comparison between Model $\mathcal{X}$ and Model $\mathcal{M}$ as they differ in only one major respect, i.e., the later incorporates the variance parameter d in the E-step. For this particular set of data Model $\mathcal{M}$ converges in significantly fewer iterations than Model $\mathcal{X}$ which suggests that having the unstructured positive definite matrix $\boldsymbol{D}$ as part of the E-step can improve algorithm performance in terms of the number of iterations until convergence.

In regard to a less stringent convergence criteria, there is little or no cost in increasing ε to 10^{-5} for both models. Our preferred model and stopping rule is Model $\mathcal{Z}$ using $\varepsilon = 10^{-5}$. With this combination convergence is obtained in 36 iterations compared to the 45 iterations it takes Model $\mathcal{X}$ using $\varepsilon = 10^{-5}$ to converge. For smaller data sets such as the one considered here having an algorithm converge in 36 iterations might be acceptable. However, for larger data sets this number of iterations until convergence would be unsatisfactory in terms of elapsed time. In the next chapter we consider a new missing data specification for factor analytic mixed models with the aim of making further gains in REML PX EM algorithm performance.

Table 7.1: Iterations until convergence, final residual log-likelihood (ℓ_R), and the Frobenius matrix norm of the estimated genetic variance-covariance matrix when using a REML PX EM algorithm to fit Model $\mathcal{X}$ and Model $\mathcal{Z}$ with $k = 1$ factors and using a stopping rule with different levels of ε. For Model $\mathcal{X}$ and Model $\mathcal{Z}$ where $k = 1$ there are 9 and 10 variance parameters to be estimated respectively.

	Model $\mathcal{X}$			Model $\mathcal{Z}$		
ε	Iter.	ℓ_R	$\|\widehat{\boldsymbol{G}}_e\|_F$	Iter.	ℓ_R	$\|\widehat{\boldsymbol{G}}_e\|_F$
10^{-8}	74	300.9849	0.3180123	65	300.9849	0.3180123
10^{-7}	64	300.9849	0.3180123	55	300.9849	0.3180123
10^{-6}	54	300.9849	0.3180123	46	300.9849	0.3180123
10^{-5}	45	300.9849	0.3180126	36	300.9849	0.3180126
10^{-4}	35	300.9849	0.3180178	27	300.9847	0.3180158
10^{-3}	26	300.9844	0.3180826	19	300.9741	0.3179943
10^{-2}	17	300.9276	0.3186011	14	300.8342	0.3169530

Chapter 8

A new missing data specification for factor analytic mixed models

In Chapter 6 we considered two model specifications for a factor analytic mixed model and showed that one of these specifications, Model $\mathcal{M}$, could greatly reduce the number of iterations until convergence when using a REML EM algorithm. In Chapter 7 we considered a new REML PX EM algorithm for factor analytic mixed models where the parameter expansion is associated with the specific variances. This model was referred to as Model $\mathcal{Z}$. In both these chapters the distribution of the missing data vector was considered to be the joint distribution of the vector of factor scores and the vector of residuals associated with the factor analytic model.

In this chapter we consider an alternative missing data specification and hence complete data specification for factor analytic mixed models. The new missing data specification we consider is based on the joint distribution of the vector of factor scores and the vector of variety by trial effects. This new missing data specification is motivated by the "dependent formulation" of factor analytic variance models presented in Thompson et al. (2003).

This chapter proceeds by showing that the complete data log-density function associated with the new complete data specfication for Model $\mathcal{M}$ is the same as the complete data density function for this model presented in Chapter 6. We consider Model $\mathcal{M}$ and Model $\mathcal{Z}$ using the new missing data specification and demonstrate, using the example considered in the previous two chapters and a simulation study, that this new missing data specification can further reduce the number of iterations until convergence.

8.1 REML EM algorithm - Model $\mathcal{M}$

We considered Model $\mathcal{M}$ in Chapter 6. We provide a succinct overview here for convenience. We define

$$\text{Model } \mathcal{M} \qquad \boldsymbol{y}_o = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_g\boldsymbol{u}_g + \boldsymbol{e},$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects representing trial means and $\boldsymbol{X}$ is its associated $n \times p$ design matrix of full column rank. It is assumed that the $n \times 1$ vector of residuals $\boldsymbol{e}$ is distributed $N(\boldsymbol{0}, \boldsymbol{R})$ where $\boldsymbol{R} = \sigma^2\boldsymbol{I}_n$. The $n \times mp$ design matrix $\boldsymbol{Z}_g$ is associated with the $mp \times 1$ vector $\boldsymbol{u}_g$ of variety by trial effects. Factor analysis is used to model the variety effects within each trial, i.e.,

$$\boldsymbol{u}_g = (\boldsymbol{\Lambda} \otimes \boldsymbol{I}_m)\boldsymbol{f} + \boldsymbol{\delta},$$

where $\boldsymbol{\Lambda}$ is a $p \times k$ matrix of loadings and $\boldsymbol{f}$ is a $mk \times 1$ vector of varietal scores that can be partitioned as $\boldsymbol{f} = (\boldsymbol{f}_1^T, \ldots, \boldsymbol{f}_k^T)^T$. It is assumed that $\boldsymbol{f} \sim N(\boldsymbol{0}, \boldsymbol{D} \otimes \boldsymbol{I}_m)$ where $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{d})$ is a $k \times k$ symmetric positive definite matrix and $\boldsymbol{d}$ is a vector of variance parameters. The vector $\boldsymbol{\delta}$ is a $mp \times 1$ vector of residuals that can be partitioned with respect to trials, i.e., $\boldsymbol{\delta} = (\boldsymbol{\delta}_1^T, \ldots, \boldsymbol{\delta}_p^T)^T$ and is independent of $\boldsymbol{f}$. It is assumed that $\boldsymbol{\delta} \sim N(\boldsymbol{0}, \boldsymbol{\Psi} \otimes \boldsymbol{I}_m)$ where $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\boldsymbol{\psi})$ is a $p \times p$ diagonal matrix. The vector of variance parameters $\boldsymbol{\psi}$ is commonly referred to as the specific variances. For simplicity we will assume that $\boldsymbol{\Psi} = \psi\boldsymbol{I}_p$, i.e., there is a common specific variance. It is assumed that the distribution of the vector of variety effects within trials is

$$\boldsymbol{u}_g \sim N(\boldsymbol{0}, \boldsymbol{G}_e \otimes \boldsymbol{I}_m),$$

where $\boldsymbol{G}_e = \boldsymbol{\Lambda}\boldsymbol{D}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$ and is referred to as the genetic variance-covariance matrix. For Model $\mathcal{M}$ the distribution of the observed data is

$$\boldsymbol{y}_o \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{H}),$$

where $\boldsymbol{H} = \boldsymbol{Z}_g(\boldsymbol{G}_e \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T + \boldsymbol{R}$ and the vector of variance parameters to be estimated is $\boldsymbol{\Gamma} = (\sigma^2, \boldsymbol{\lambda}^T, \psi, \boldsymbol{d}^T)^T$.

8.1.1 Complete data density function for $\boldsymbol{y}_c^{(D)} = (\boldsymbol{y}_2^T, \boldsymbol{f}^T, \boldsymbol{u}_g^T)^T$

We define the new missing data specification as

$$\boldsymbol{y}_m^{(D)} = (\boldsymbol{f}^T, \boldsymbol{u}_g^T)^T,$$

where the superscript D is used to distinguish between this missing data specification and $\boldsymbol{y}_m = (\boldsymbol{f}^T, \boldsymbol{\delta}^T)^T$. Likewise, the new complete data specification is defined as

$$\boldsymbol{y}_c^{(D)} = (\boldsymbol{y}_2^T, \boldsymbol{f}^T, \boldsymbol{u}_g^T)^T.$$

To form the complete data log-likelihood function for $\boldsymbol{y}_c^{(D)}$ requires the joint distribution

$$\begin{pmatrix} \boldsymbol{y}_2 \\ \boldsymbol{f} \\ \boldsymbol{u}_g \end{pmatrix} \sim N\left\{ \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} & \boldsymbol{V}_{13} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} & \boldsymbol{V}_{23} \\ \boldsymbol{V}_{31} & \boldsymbol{V}_{32} & \boldsymbol{V}_{33} \end{pmatrix} \right\}, \tag{8.1}$$

where

$$\begin{aligned}
\boldsymbol{V}_{11} &= \boldsymbol{L}_2^T \boldsymbol{H} \boldsymbol{L}_2, \\
\boldsymbol{V}_{22} &= \boldsymbol{D} \otimes \boldsymbol{I}_m, \\
\boldsymbol{V}_{33} &= (\boldsymbol{\Lambda} \boldsymbol{D} \boldsymbol{\Lambda}^T + \boldsymbol{\Psi}) \otimes \boldsymbol{I}_m, \\
\boldsymbol{V}_{12} &= \boldsymbol{V}_{21}^T = \boldsymbol{L}_2^T \boldsymbol{Z}_g (\boldsymbol{\Lambda} \boldsymbol{D} \otimes \boldsymbol{I}_m), \\
\boldsymbol{V}_{13} &= \boldsymbol{V}_{31}^T = \boldsymbol{L}_2^T \boldsymbol{Z}_g \{ (\boldsymbol{\Lambda} \boldsymbol{D} \boldsymbol{\Lambda}^T + \boldsymbol{\Psi}) \otimes \boldsymbol{I}_m \}, \\
\boldsymbol{V}_{23} &= \boldsymbol{V}_{32}^T = \boldsymbol{D} \boldsymbol{\Lambda}^T \otimes \boldsymbol{I}_m.
\end{aligned}$$

The complete data log-density function can be expressed as

$$\log\{f(\boldsymbol{y}_c^{(D)}; \boldsymbol{\Gamma})\} = \log\{f(\boldsymbol{y}_2 | \boldsymbol{y}_m^{(D)}; \boldsymbol{\Gamma})\} + \log\{f(\boldsymbol{y}_m^{(D)}; \boldsymbol{\Gamma})\}.$$

Consider $\log\{f(\boldsymbol{y}_2 | \boldsymbol{y}_m^{(D)}; \boldsymbol{\Gamma})\}$. Using (8.6) and Result A.1 it can be shown that

$$\boldsymbol{y}_2 | \boldsymbol{y}_m^{(D)} \sim N(\boldsymbol{L}_2^T \boldsymbol{Z}_g \boldsymbol{u}_g, \boldsymbol{L}_2^T \boldsymbol{R} \boldsymbol{L}_2)$$

and the log-density function for $\boldsymbol{y}_2$ given $\boldsymbol{y}_m^{(D)}$ is (excluding constants)

$$\begin{aligned}\log\{f(\boldsymbol{y}_2|\boldsymbol{y}_m^{(D)};\boldsymbol{\Gamma})\} &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2)\} \\ &\quad +(\boldsymbol{y}_o-\boldsymbol{Z}_g\boldsymbol{u}_g)^T\boldsymbol{S}(\boldsymbol{y}_o-\boldsymbol{Z}_g\boldsymbol{u}_g)\Big], \qquad (8.2)\end{aligned}$$

where $\boldsymbol{S}=\boldsymbol{R}^{-1}-\boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{R}^{-1}$. Now consider the log-density function $\log\{f(\boldsymbol{y}_m^{(D)};\boldsymbol{\Gamma})\}$. We define

$$\mathcal{V}=\begin{Bmatrix}\boldsymbol{D}\otimes\boldsymbol{I}_m & \boldsymbol{D}\boldsymbol{\Lambda}^T\otimes\boldsymbol{I}_m \\ \boldsymbol{\Lambda}\boldsymbol{D}\otimes\boldsymbol{I}_m & (\boldsymbol{\Lambda}\boldsymbol{D}\boldsymbol{\Lambda}^T+\boldsymbol{\Psi})\otimes\boldsymbol{I}_m\end{Bmatrix}.$$

The log-density function for $\boldsymbol{y}_m^{(D)}$ is

$$\log\{f(\boldsymbol{y}_m^{(D)};\boldsymbol{\Gamma})\}=-\frac{1}{2}\Big[\log\{\det(\mathcal{V})\}+\boldsymbol{y}_m^{(D)^T}\mathcal{V}^{-1}\boldsymbol{y}_m^{(D)}\Big]. \qquad (8.3)$$

To express (8.3) in terms of $\boldsymbol{f}$ and $\boldsymbol{u}_g$ and their associated variance parameters we require $\det(\mathcal{V})$ and $\mathcal{V}^{-1}$. Using the fourth identity in Result A.3.3 we have

$$\begin{aligned}\det(\mathcal{V}) &= \det(\boldsymbol{D}\otimes\boldsymbol{I}_m)\det\big[\{(\boldsymbol{\Lambda}\boldsymbol{D}\boldsymbol{\Lambda}^T+\boldsymbol{\Psi})\otimes\boldsymbol{I}_m\} \\ &\quad -\{(\boldsymbol{\Lambda}\boldsymbol{D}\otimes\boldsymbol{I}_m)(\boldsymbol{D}^{-1}\otimes\boldsymbol{I}_m)(\boldsymbol{D}\boldsymbol{\Lambda}^T\otimes\boldsymbol{I}_m)\}\big] \\ &= \det(\boldsymbol{D}\otimes\boldsymbol{I}_m)\det\big[\{(\boldsymbol{\Lambda}\boldsymbol{D}\boldsymbol{\Lambda}^T+\boldsymbol{\Psi})\otimes\boldsymbol{I}_m\}-(\boldsymbol{\Lambda}\boldsymbol{D}\boldsymbol{\Lambda}^T\otimes\boldsymbol{I}_m)\big] \\ &= \det(\boldsymbol{D}\otimes\boldsymbol{I}_m)\det(\boldsymbol{\Psi}\otimes\boldsymbol{I}_m).\end{aligned}$$

Using Result A.3.6 we can express $\mathcal{V}^{-1}$ as

$$\mathcal{V}^{-1}=\begin{pmatrix}\mathcal{V}^{11} & \mathcal{V}^{12} \\ \mathcal{V}^{21} & \mathcal{V}^{22}\end{pmatrix},$$

where

$$\begin{aligned}\mathcal{V}^{11} &= (\boldsymbol{D}^{-1}+\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})\otimes\boldsymbol{I}_m, \\ \mathcal{V}^{12} &= \mathcal{V}^{21^T} \\ &= -(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\otimes\boldsymbol{I}_m), \\ \mathcal{V}^{22} &= \boldsymbol{\Psi}^{-1}\otimes\boldsymbol{I}_m.\end{aligned}$$

We can express (8.3) as

$$\begin{aligned}
\log\{f(\boldsymbol{y}_m^{(D)};\boldsymbol{\Gamma})\} &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{D}\otimes\boldsymbol{I}_m)\}+\log\{\det(\boldsymbol{\Psi}\otimes\boldsymbol{I}_m)\} \\
&\quad +\boldsymbol{f}^T\{(\boldsymbol{D}^{-1}+\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})\otimes\boldsymbol{I}_m\}\boldsymbol{f}-2\boldsymbol{u}_g^T(\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)\boldsymbol{f} \\
&\quad +\boldsymbol{u}_g^T(\boldsymbol{\Psi}^{-1}\otimes\boldsymbol{I}_m)\boldsymbol{u}_g\Big] \\
&= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{D}\otimes\boldsymbol{I}_m)\}+\log\{\det(\boldsymbol{\Psi}\otimes\boldsymbol{I}_m)\} \\
&\quad +\boldsymbol{f}^T(\boldsymbol{D}^{-1}\otimes\boldsymbol{I}_m)\boldsymbol{f}+\boldsymbol{f}^T(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)\boldsymbol{f} \\
&\quad -2\boldsymbol{u}_g^T(\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)\boldsymbol{f}+\boldsymbol{u}_g^T(\boldsymbol{\Psi}^{-1}\otimes\boldsymbol{I}_m)\boldsymbol{u}_g\Big]. \qquad (8.4)
\end{aligned}$$

8.1.2 Equivalence of $\boldsymbol{y}_c^{(D)}=(\boldsymbol{y}_2^T,\boldsymbol{f}^T,\boldsymbol{u}_g^T)^T$ and $\boldsymbol{y}_c=(\boldsymbol{y}_2^T,\boldsymbol{f}^T,\boldsymbol{\delta}^T)^T$

If we substitute $\boldsymbol{u}_g=(\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)\boldsymbol{f}+\boldsymbol{\delta}$ into (8.2) then it is easily seen that $\log\{f(\boldsymbol{y}_2|\boldsymbol{y}_m^{(D)};\boldsymbol{\Gamma})\}=\log\{f(\boldsymbol{y}_2|\boldsymbol{y}_m;\boldsymbol{\Gamma})\}$. Making the same substitution into (8.4) we have

$$\begin{aligned}
\log\{f(\boldsymbol{y}_m^{(D)};\boldsymbol{\Gamma})\} &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{D}\otimes\boldsymbol{I}_m)\}+\log\{\det(\boldsymbol{\Psi}\otimes\boldsymbol{I}_m)\} \\
&\quad +\boldsymbol{f}^T(\boldsymbol{D}^{-1}\otimes\boldsymbol{I}_m)\boldsymbol{f}+\boldsymbol{f}^T(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)\boldsymbol{f} \\
&\quad -2\{\boldsymbol{f}^T(\boldsymbol{\Lambda}^T\otimes\boldsymbol{I}_m)+\boldsymbol{\delta}^T\}(\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)\boldsymbol{f} \\
&\quad +\{\boldsymbol{f}^T(\boldsymbol{\Lambda}^T\otimes\boldsymbol{I}_m)+\boldsymbol{\delta}^T\}(\boldsymbol{\Psi}^{-1}\otimes\boldsymbol{I}_m)\{(\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)\boldsymbol{f}+\boldsymbol{\delta}\}\boldsymbol{f}\Big] \\
&= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{D}\otimes\boldsymbol{I}_m)\}+\log\{\det(\boldsymbol{\Psi}\otimes\boldsymbol{I}_m)\} \\
&\quad +\boldsymbol{f}^T(\boldsymbol{D}^{-1}\otimes\boldsymbol{I}_m)\boldsymbol{f}+\boldsymbol{f}^T(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)\boldsymbol{f} \\
&\quad -2\boldsymbol{f}^T(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)\boldsymbol{f}-2\boldsymbol{\delta}^T(\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)\boldsymbol{f} \\
&\quad +\boldsymbol{f}^T(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)\boldsymbol{f}+2\boldsymbol{\delta}^T(\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\otimes\boldsymbol{I}_m)\boldsymbol{f} \\
&\quad +\boldsymbol{\delta}^T(\boldsymbol{\Psi}^{-1}\otimes\boldsymbol{I}_m)\boldsymbol{\delta}]\Big] \\
&= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{D}\otimes\boldsymbol{I}_m)\}+\log\{\det(\boldsymbol{\Psi}\otimes\boldsymbol{I}_m)\} \\
&\quad +\boldsymbol{f}^T(\boldsymbol{D}^{-1}\otimes\boldsymbol{I}_m)\boldsymbol{f}+\boldsymbol{\delta}^T(\boldsymbol{\Psi}^{-1}\otimes\boldsymbol{I}_m)\boldsymbol{\delta}\Big] \\
&= \log\{f(\boldsymbol{y}_m;\boldsymbol{\Gamma})\}.
\end{aligned}$$

Since $\log\{f(\boldsymbol{y}_m^{(D)};\boldsymbol{\Gamma})\}=\log\{f(\boldsymbol{y}_m;\boldsymbol{\Gamma})\}$ and $\log\{f(\boldsymbol{y}_2|\boldsymbol{y}_m^{(D)};\boldsymbol{\Gamma})\}=\log\{f(\boldsymbol{y}_2|\boldsymbol{y}_m;\boldsymbol{\Gamma})\}$ then it follows that $\log\{f(\boldsymbol{y}_c^{(D)};\boldsymbol{\Gamma})\}=\log\{f(\boldsymbol{y}_c;\boldsymbol{\Gamma})\}$.

8.2 REML EM algorithm E-step for Model $\mathcal{M}$ using $\boldsymbol{y}_c^{(D)}$

The E-step involves taking the conditional expectation of $\log\{f(\boldsymbol{y}_c^{(D)};\boldsymbol{\Gamma})\}$ over $k(\boldsymbol{f},\boldsymbol{u}_g|\boldsymbol{y}_2;\boldsymbol{\Gamma}^{(w)})$ where $\boldsymbol{\Gamma}^{(w)}$ is the w-th iterate of $\boldsymbol{\Gamma}$. We can write

$$\begin{aligned} Q^{(D)}(\boldsymbol{\Gamma};\boldsymbol{\Gamma}^{(w)}) &= E[\log\{f(\boldsymbol{y}_c^{(D)};\boldsymbol{\Gamma})\}|\boldsymbol{y}_2;\boldsymbol{\Gamma}^{(w)}] \\ &= E[\log\{f(\boldsymbol{y}_2|\boldsymbol{y}_m^{(D)};\boldsymbol{\Gamma})\}|\boldsymbol{y}_2;\boldsymbol{\Gamma}^{(w)}] + E[\log\{f(\boldsymbol{y}_m^{(D)};\boldsymbol{\Gamma})\}|\boldsymbol{y}_2;\boldsymbol{\Gamma}^{(w)}] \\ &= Q^{(D)}(\sigma^2;\boldsymbol{\Gamma}^{(w)}) + Q^{(D)}(\boldsymbol{\lambda},\psi;\boldsymbol{\Gamma}^{(w)}) + Q^{(D)}(\boldsymbol{d};\boldsymbol{\Gamma}^{(w)}). \end{aligned} \tag{8.5}$$

We use the superscript D on the $Q(.)$ function to distinguish between the $Q(.)$ function using $\boldsymbol{y}_c^{(D)} = (\boldsymbol{y}_2^T, \boldsymbol{f}^T, \boldsymbol{u}_g^T)^T$ and the $Q(.)$ function using $\boldsymbol{y}_c = (\boldsymbol{y}_2^T, \boldsymbol{f}^T, \boldsymbol{\delta}^T)^T$. To compute the $Q(.)$ functions in (8.5) requires the joint conditional distribution of $\boldsymbol{f}$ and $\boldsymbol{u}_g$ given $\boldsymbol{y}_2$. Using (8.6) and Result A.1 we have

$$\left.\begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{u}_g \end{pmatrix}\right| \boldsymbol{y}_2 \sim N\left\{\begin{pmatrix} (\boldsymbol{\Lambda}^T \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{P}\boldsymbol{y}_o \\ (\boldsymbol{G}_e \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{P}\boldsymbol{y}_o \end{pmatrix}, \begin{pmatrix} \boldsymbol{T}_{ff} & \boldsymbol{T}_{fu_g} \\ \boldsymbol{T}_{u_gf} & \boldsymbol{T}_{u_gu_g} \end{pmatrix}\right\},$$

where

$$\begin{aligned} \boldsymbol{T}_{ff} &= (\boldsymbol{D} \otimes \boldsymbol{I}_m) - (\boldsymbol{D}\boldsymbol{\Lambda}^T \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{P}\boldsymbol{Z}_g(\boldsymbol{\Lambda}\boldsymbol{D} \otimes \boldsymbol{I}_m), \\ \boldsymbol{T}_{fu_g} &= \boldsymbol{T}_{u_gf}^T = (\boldsymbol{D}\boldsymbol{\Lambda}^T \otimes \boldsymbol{I}_m) - (\boldsymbol{D}\boldsymbol{\Lambda}^T \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{P}\boldsymbol{Z}_g(\boldsymbol{G}_e \otimes \boldsymbol{I}_m), \\ \boldsymbol{T}_{u_gu_g} &= (\boldsymbol{G}_e \otimes \boldsymbol{I}_m) - (\boldsymbol{G}_e \otimes \boldsymbol{I}_m)\boldsymbol{Z}_g^T\boldsymbol{P}\boldsymbol{Z}_g(\boldsymbol{G}_e \otimes \boldsymbol{I}_m) \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{P} &= \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T \\ &= \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}. \end{aligned}$$

We can write

$$Q^{(D)}(\sigma^2;\boldsymbol{\Gamma}^{(w)}) = -\frac{1}{2}\Big[\log\{\det(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2)\} + E\big\{(\boldsymbol{y}_o - \boldsymbol{Z}_g\boldsymbol{u}_g)^T\boldsymbol{S}(\boldsymbol{y}_o - \boldsymbol{Z}_g\boldsymbol{u}_g)|\boldsymbol{y}_2;\boldsymbol{\Gamma}^{(w)}\big\}\Big]$$

and

$$\begin{aligned} Q^{(D)}(\boldsymbol{\lambda}, \psi; \boldsymbol{\Gamma}^{(w)}) &= -\frac{1}{2}\Big[mp\log(\psi) + E\big[\boldsymbol{f}^T\{(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \otimes \boldsymbol{I}_m\}\boldsymbol{f}|\boldsymbol{y}_2; \boldsymbol{\Gamma}^{(w)}\big] \\ &\quad -2E\big\{\boldsymbol{f}^T(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1} \otimes \boldsymbol{I}_m)\boldsymbol{u}_g|\boldsymbol{y}_2; \boldsymbol{\Gamma}^{(w)}\big\} \\ &\quad +E\big\{\boldsymbol{u}_g^T(\boldsymbol{\Psi}^{-1} \otimes \boldsymbol{I}_m)\boldsymbol{u}_g|\boldsymbol{y}_2; \boldsymbol{\Gamma}^{(w)}\big\}\Big] \end{aligned}$$

and

$$Q^{(D)}(\boldsymbol{d}; \boldsymbol{\Gamma}^{(w)}) = -\frac{1}{2}\Big[\log\{\det(\boldsymbol{D} \otimes \boldsymbol{I}_m)\} + E\big[\boldsymbol{f}^T(\boldsymbol{D}^{-1} \otimes \boldsymbol{I}_m)\boldsymbol{f}|\boldsymbol{y}_2; \boldsymbol{\Gamma}^{(w)}\big]\Big].$$

8.3 REML EM algorithm M-step for Model $\mathcal{M}$ using $\boldsymbol{y}_c^{(D)}$

8.3.1 Update for σ^2

Differentiating $Q^{(D)}(\sigma^2; \boldsymbol{\Gamma}^{(w)})$ with respect to σ^2 we have

$$\begin{aligned} \frac{\partial Q^{(D)}(\sigma^2; \boldsymbol{\Gamma}^{(w)})}{\partial \sigma^2} &= -\frac{1}{2}\Big\{\frac{n-p}{\sigma^2} - \frac{1}{(\sigma^2)^2}(\boldsymbol{y}_o - \boldsymbol{Z}_g\tilde{\boldsymbol{u}}_g{}^{(w)})^T\boldsymbol{K}(\boldsymbol{y}_o - \boldsymbol{Z}_g\tilde{\boldsymbol{u}}_g{}^{(w)}) \\ &\quad -\frac{1}{(\sigma^2)^2}\mathrm{tr}\big(\boldsymbol{Z}_g^T\boldsymbol{K}\boldsymbol{Z}_g\boldsymbol{T}_{u_g u_g}^{(w)}\big)\Big\}, \end{aligned}$$

where $\boldsymbol{K} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Equating to zero and solving for σ^2 we have

$$\sigma^{2(w+1)} = \frac{1}{n-p}\Big\{(\boldsymbol{y}_o - \boldsymbol{Z}_g\tilde{\boldsymbol{u}}_g{}^{(w)})^T\boldsymbol{K}(\boldsymbol{y}_o - \boldsymbol{Z}_g\tilde{\boldsymbol{u}}_g{}^{(w)}) + \mathrm{tr}\big(\boldsymbol{Z}_g^T\boldsymbol{K}\boldsymbol{Z}_g\boldsymbol{T}_{u_g u_g}^{(w)}\big)\Big\}.$$

8.3.2 Update for λ_{jr}

Differentiating $Q^{(D)}(\boldsymbol{\lambda}, \psi; \boldsymbol{\Gamma}^{(w)})$ with respect to λ_{jr} we have

$$\frac{\partial Q^{(D)}(\boldsymbol{\lambda}, \psi; \boldsymbol{\Gamma}^{(w)})}{\partial \lambda_{jr}} = -\frac{1}{2}\Big[2E\Big\{\boldsymbol{f}^T\Big(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\frac{\partial \boldsymbol{\Lambda}}{\partial \lambda_{jr}} \otimes \boldsymbol{I}_m\Big)\boldsymbol{f}|\boldsymbol{y}_2; \boldsymbol{\Gamma}^{(w)}\Big\} -2\mathrm{tr}\Big\{\Big(\boldsymbol{\Psi}^{-1}\frac{\partial \boldsymbol{\Lambda}}{\partial \lambda_{jr}} \otimes \boldsymbol{I}_m\Big)E\Big(\boldsymbol{f}\boldsymbol{u}_g^T|\boldsymbol{y}_2; \boldsymbol{\Gamma}^{(w)}\Big)\Big\}\Big].$$

Equating to zero and using the identity

$$\boldsymbol{\Lambda}^T = \sum_{l=1}^{p}\sum_{q=1}^{k}\lambda_{lq}\boldsymbol{J}_{lq}^T,$$

where $\boldsymbol{J}_{lq}$ is a $p \times k$ indicator matrix having a 1 in row l, column q and zeroes elsewehere we obtain

$$\sum_{l=1}^{p}\sum_{q=1}^{k}\lambda_{lq}E\{\boldsymbol{f}^T(\boldsymbol{J}_{lq}^T\boldsymbol{\Psi}^{-1}\boldsymbol{J}_{jr}\otimes\boldsymbol{I}_m)\boldsymbol{f}|\boldsymbol{y}_2; \boldsymbol{\Gamma}^{(w)}\} = \mathrm{tr}\{(\boldsymbol{\Psi}^{-1}\boldsymbol{J}_{jr}\otimes\boldsymbol{I}_m)E(\boldsymbol{f}\boldsymbol{u}_g^T|\boldsymbol{y}_2; \boldsymbol{\Gamma}^{(w)})\}.$$

In matrix notation this can be expressed as

$$\boldsymbol{A}^{(w)}\boldsymbol{\lambda} = \boldsymbol{b}^{(w)},$$

where $\boldsymbol{A}^{(w)}$ is a $mk \times mk$ matrix with elements

$$a_{jr,lq}^{(w)} = \tilde{\boldsymbol{f}}^{(w)^T}(\boldsymbol{J}_{lq}^T\boldsymbol{\Psi}^{(w)^{-1}}\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} + \mathrm{tr}\{(\boldsymbol{J}_{lq}^T\boldsymbol{\Psi}^{(w)^{-1}}\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)\boldsymbol{T}_{ff}^{(w)}\}$$

and $\boldsymbol{b}^{(w)}$ is a $mk \times 1$ vector with elements

$$b_{jr}^{(w)} = \mathrm{tr}\{(\boldsymbol{\Psi}^{(w)^{-1}}\boldsymbol{J}_{jr} \otimes \boldsymbol{I}_m)(\tilde{\boldsymbol{f}}^{(w)}\tilde{\boldsymbol{u}}_g^{(w)^T} + \boldsymbol{T}_{fu_g}^{(w)})\}.$$

The update for $\boldsymbol{\lambda}$ is therefore

$$\boldsymbol{\lambda}^{(w+1)} = \boldsymbol{A}^{(w)^{-1}}\boldsymbol{b}^{(w)}.$$

8.3.3 Update for ψ

Differentiating $Q^{(D)}(\boldsymbol{\lambda}, \psi; \boldsymbol{\Gamma}^{(w)})$ with respect to ψ we have

$$\begin{aligned}\frac{\partial Q^{(D)}(\boldsymbol{\lambda}, \psi; \boldsymbol{\Gamma}^{(w)})}{\partial \psi} &= -\frac{1}{2}\Big[\frac{mp}{\psi} - \frac{1}{\psi^2}\tilde{\boldsymbol{f}}^{(w)^T}(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda} \otimes \boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} \\ &\quad -\frac{1}{\psi^2}\mathrm{tr}\big\{(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda} \otimes \boldsymbol{I}_m)\boldsymbol{T}_{ff}^{(w)}\big\} \\ &\quad -\frac{1}{\psi^2}\tilde{\boldsymbol{u}}_g^{(w)^T}\tilde{\boldsymbol{u}}_g^{(w)} - \frac{1}{\psi^2}\mathrm{tr}\big(\boldsymbol{T}_{u_g u_g}^{(w)}\big) \\ &\quad +\frac{2}{\psi^2}\mathrm{tr}\big\{(\boldsymbol{\Lambda} \otimes \boldsymbol{I}_m)(\tilde{\boldsymbol{f}}^{(w)}\tilde{\boldsymbol{u}}_g^{(w)^T} + \boldsymbol{T}_{fu_g}^{(w)})\big\}\Big].\end{aligned}$$

Equating to zero and solving for ψ we have

$$\begin{aligned}\psi^{(w+1)} &= \frac{1}{mp}\Big[\tilde{\boldsymbol{f}}^{(w)^T}(\boldsymbol{\Lambda}^{(w)^T}\boldsymbol{\Lambda}^{(w)} \otimes \boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} + \mathrm{tr}\big\{(\boldsymbol{\Lambda}^{(w)^T}\boldsymbol{\Lambda}^{(w)} \otimes \boldsymbol{I}_m)\boldsymbol{T}_{ff}^{(w)}\big\} \\ &\quad +\tilde{\boldsymbol{u}}_g^{(w)^T}\tilde{\boldsymbol{u}}_g^{(w)} + \mathrm{tr}\big(\boldsymbol{T}_{u_g u_g}^{(w)}\big) - 2\mathrm{tr}\big\{(\boldsymbol{\Lambda}^{(w)} \otimes \boldsymbol{I}_m)(\tilde{\boldsymbol{f}}^{(w)}\tilde{\boldsymbol{u}}_g^{(w)^T} + \boldsymbol{T}_{fu_g}^{(w)})\big\}\Big].\end{aligned}$$

8.3.4 Update for d_{qr}

Since $Q^{(D)}(\boldsymbol{d}; \boldsymbol{\Gamma}^{(w)})$ is equivalent to $Q(\boldsymbol{d}; \boldsymbol{\Gamma}^{(w)})$ the update for d_{qr}, $q = 1, \ldots, k$, $r = 1, \ldots, j$, $q \geq r$ is the same as that presented in Chapter 6 for Model $\mathcal{M}$, i.e.,

$$d_{qr}^{(w+1)} = \frac{1}{m}\Big\{\tilde{\boldsymbol{f}}_q^{(w)^T}\tilde{\boldsymbol{f}}_r^{(w)} + \mathrm{tr}\big(\boldsymbol{T}_{ff_{qr}}^{(w)}\big)\Big\}.$$

8.4 REML PX EM algorithm - Model $\mathcal{Z}$

We considered Model $\mathcal{Z}$ in Chapter 7. We provide a succinct overview here for convenience. We define

$$\text{Model } \mathcal{Z} \qquad \boldsymbol{y}_o = \boldsymbol{X}\boldsymbol{\beta}_* + \boldsymbol{Z}_g\boldsymbol{u}_g + \boldsymbol{e},$$

where $\boldsymbol{e}$ is distributed $N(\boldsymbol{0}, \boldsymbol{R}_*)$ and $\boldsymbol{R}_* = \sigma_*^2\boldsymbol{I}_n$. For Model $\mathcal{Z}$ we consider

$$\boldsymbol{u}_g = (\boldsymbol{\Lambda}_* \otimes \boldsymbol{I}_m)\boldsymbol{f} + (\boldsymbol{\Omega} \otimes \boldsymbol{I}_m)\boldsymbol{\delta},$$

where $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\omega})$ is a $p \times p$ diagonal positive definite diagonal matrix. When we assume common specific variances $\boldsymbol{\Omega}$ has the form $\boldsymbol{\Omega} = \omega \boldsymbol{I}_p$. We assume $\boldsymbol{f} \sim N(\boldsymbol{0}, \boldsymbol{D}_* \otimes \boldsymbol{I}_m)$ where $\boldsymbol{D}_* = \boldsymbol{D}_*(\boldsymbol{d}_*)$ and $\boldsymbol{\delta} \sim N(\boldsymbol{0}, \boldsymbol{\Psi}_* \otimes \boldsymbol{I}_m)$ where $\boldsymbol{\Psi}_* = \psi_* \boldsymbol{I}_p$. The $mk \times 1$ vector $\boldsymbol{f}$ and the $mp \times 1$ vector $\boldsymbol{\delta}$ are assumed to be independent of each other. We define the expanded parameter set associated with Model $\mathcal{Z}$ as $\boldsymbol{\mathcal{T}} = (\sigma_*^2, \boldsymbol{\lambda}_*^T, \psi_*, \boldsymbol{d}_*^T, \boldsymbol{\omega}^T)^T$ where the subscript $*$ is used to distinguish parameters associated with Model $\mathcal{M}$ and Model $\mathcal{Z}$. For Model $\mathcal{Z}$ the distribution of the observed data is

$$\boldsymbol{y}_o \sim N(\boldsymbol{X}\boldsymbol{\beta}_*, \boldsymbol{H}_*),$$

where

$$\boldsymbol{H}_* = \boldsymbol{Z}_g \boldsymbol{G}_{e_*} \boldsymbol{Z}_g^T + \boldsymbol{R}_*$$

and

$$\boldsymbol{G}_{e_*} = \{(\boldsymbol{\Lambda}_* \boldsymbol{D}_* \boldsymbol{\Lambda}_*^T + \omega^2 \psi_* \boldsymbol{I}_p) \otimes \boldsymbol{I}_m\}.$$

Using the parameter expanded model we identify the parameters associated with Model $\mathcal{M}$ as $\boldsymbol{\Gamma} = (\sigma^2, \boldsymbol{\lambda}^T, \psi, \boldsymbol{d}^T)^T = \mathcal{R}(\boldsymbol{\mathcal{T}}) = (\sigma_*^2, \boldsymbol{\lambda}_*, \omega^2\psi_*, \boldsymbol{d}_*^T)^T$ where $\mathcal{R}$ is the reduction function from the expanded parameter set to the original parameter set.

8.4.1 Complete data density function

For Model $\mathcal{Z}$ the complete data specification is the same as that for Model $\mathcal{M}$, i.e., $\boldsymbol{y}_c = (\boldsymbol{y}_2^T, \boldsymbol{f}^T, \boldsymbol{u}_g^T)^T$. We use the subscript $\mathcal{Z}$ to distinguish between the complete data log-density function and $Q(.)$ function for Model $\mathcal{M}$ and Model $\mathcal{Z}$. To form the complete data density function for Model $\mathcal{Z}$ requires the joint distribution

$$\begin{pmatrix} \boldsymbol{y}_2 \\ \boldsymbol{f} \\ \boldsymbol{u}_g \end{pmatrix} \sim N\left\{ \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{V}_{11_*} & \boldsymbol{V}_{12_*} & \boldsymbol{V}_{13_*} \\ \boldsymbol{V}_{21_*} & \boldsymbol{V}_{22_*} & \boldsymbol{V}_{23_*} \\ \boldsymbol{V}_{31_*} & \boldsymbol{V}_{32_*} & \boldsymbol{V}_{33_*} \end{pmatrix} \right\}, \tag{8.6}$$

where

$$\begin{aligned}
\boldsymbol{V}_{11_*} &= \boldsymbol{L}_2^T \boldsymbol{H}_* \boldsymbol{L}_2, \\
\boldsymbol{V}_{22_*} &= \boldsymbol{D}_* \otimes \boldsymbol{I}_m, \\
\boldsymbol{V}_{33_*} &= \boldsymbol{G}_{e_*} \otimes \boldsymbol{I}_m, \\
\boldsymbol{V}_{12_*} &= \boldsymbol{V}_{21_*}^T = \boldsymbol{L}_2^T \boldsymbol{Z}_g (\boldsymbol{\Lambda}_* \boldsymbol{D}_* \otimes \boldsymbol{I}_m), \\
\boldsymbol{V}_{13_*} &= \boldsymbol{V}_{31_*}^T = \boldsymbol{L}_2^T \boldsymbol{Z}_g (\boldsymbol{G}_{e_*} \otimes \boldsymbol{I}_m) \\
\boldsymbol{V}_{23_*} &= \boldsymbol{V}_{32_*}^T = \boldsymbol{D}_* \boldsymbol{\Lambda}_*^T \otimes \boldsymbol{I}_m.
\end{aligned}$$

The log-density function for $\boldsymbol{y}_2$ given $\boldsymbol{y}_m^{(D)}$ is (excluding constants)

$$\begin{aligned}
\log\{f_{\mathcal{Z}}(\boldsymbol{y}_2 | \boldsymbol{y}_m^{(D)}; \boldsymbol{\Gamma})\} &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{L}_2^T \boldsymbol{R}_* \boldsymbol{L}_2)\} \\
&\quad + (\boldsymbol{y}_o - \boldsymbol{Z}_g \boldsymbol{u}_g)^T \boldsymbol{S}_* (\boldsymbol{y}_o - \boldsymbol{Z}_g \boldsymbol{u}_g)\Big],
\end{aligned}$$

where $\boldsymbol{S}_* = \boldsymbol{R}_*^{-1} - \boldsymbol{R}_*^{-1}\boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{R}_*^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{R}_*^{-1}$. Now consider the log-density function $\log\{f_{\mathcal{Z}}(\boldsymbol{y}_m^{(D)}; \boldsymbol{\Gamma})\}$. We define

$$\mathcal{V}_* = \begin{Bmatrix} \boldsymbol{D}_* \otimes \boldsymbol{I}_m & \boldsymbol{D}_* \boldsymbol{\Lambda}_*^T \otimes \boldsymbol{I}_m \\ \boldsymbol{\Lambda}_* \boldsymbol{D}_* \otimes \boldsymbol{I}_m & \boldsymbol{G}_{e_*} \otimes \boldsymbol{I}_m \end{Bmatrix}.$$

The log-density function for $\boldsymbol{y}_m^{(D)}$ is

$$\log\{f_{\mathcal{Z}}(\boldsymbol{y}_m^{(D)}; \boldsymbol{\Gamma})\} = -\frac{1}{2}\Big[\log\{\det(\mathcal{V}_*)\} + \boldsymbol{y}_m^{(D)^T} \mathcal{V}_*^{-1} \boldsymbol{y}_m^{(D)}\Big].$$

Using the fourth identity in Result A.3.3 we can write

$$\det(\mathcal{V}_*) = \det(\boldsymbol{D}_* \otimes \boldsymbol{I}_m) \det(\boldsymbol{\Omega}\boldsymbol{\Psi}_*\boldsymbol{\Omega} \otimes \boldsymbol{I}_m).$$

Using Result A.3.6 we can write

$$\mathcal{V}_*^{-1} = \begin{pmatrix} \mathcal{V}_*^{11} & \mathcal{V}_*^{12} \\ \mathcal{V}_*^{21} & \mathcal{V}_*^{22} \end{pmatrix},$$

where

$$\begin{aligned}\mathcal{V}_*^{11} &= \{\boldsymbol{D}_*^{-1} + \boldsymbol{\Lambda}_*^T(\boldsymbol{\Omega}\boldsymbol{\Psi}_*\boldsymbol{\Omega})^{-1}\boldsymbol{\Lambda}_*\} \otimes \boldsymbol{I}_m,\\ \mathcal{V}_*^{12} &= \mathcal{V}_*^{21^T} = -\{\boldsymbol{\Lambda}_*^T(\boldsymbol{\Omega}\boldsymbol{\Psi}_*\boldsymbol{\Omega})^{-1} \otimes \boldsymbol{I}_m\},\\ \mathcal{V}_*^{22} &= (\boldsymbol{\Omega}\boldsymbol{\Psi}_*\boldsymbol{\Omega})^{-1} \otimes \boldsymbol{I}_m.\end{aligned}$$

Therefore,

$$\begin{aligned}\log\{f_{\mathcal{Z}}(\boldsymbol{y}_m^{(D)};\boldsymbol{\Gamma})\} &= -\frac{1}{2}\Big[\log\{\det(\boldsymbol{D}_* \otimes \boldsymbol{I}_m)\} + \log\{\det(\boldsymbol{\Omega}\boldsymbol{\Psi}_*\boldsymbol{\Omega} \otimes \boldsymbol{I}_m)\}\\ &\quad + \boldsymbol{f}^T\big[\{\boldsymbol{D}_*^{-1} + \boldsymbol{\Lambda}_*^T(\boldsymbol{\Omega}\boldsymbol{\Psi}_*\boldsymbol{\Omega})^{-1}\boldsymbol{\Lambda}\} \otimes \boldsymbol{I}_m\big]\boldsymbol{f} - 2\boldsymbol{u}_g^T\{(\boldsymbol{\Omega}\boldsymbol{\Psi}_*\boldsymbol{\Omega})^{-1}\boldsymbol{\Lambda} \otimes \boldsymbol{I}_m\}\boldsymbol{f}\\ &\quad + \boldsymbol{u}_g^T\{(\boldsymbol{\Omega}\boldsymbol{\Psi}_*\boldsymbol{\Omega})^{-1} \otimes \boldsymbol{I}_m\}\boldsymbol{u}_g\Big].\end{aligned}$$

8.5 REML PX EM algorithm E-step for Model $\mathcal{Z}$ using $\boldsymbol{y}_c^{(D)}$

For Model $\mathcal{Z}$ we define $\boldsymbol{\mathcal{T}}^{(w)} = (\sigma_*^2, \boldsymbol{\lambda}_*^{(w)^T}, \psi_*^{(w)}, \boldsymbol{d}_*^{(w)^T}, \omega_0)^T$ where $\boldsymbol{\Omega}(\omega_0) = \boldsymbol{I}_p$. The E-step REML PX EM algorithm for Model $\mathcal{Z}$ and the E-step of the REML EM algorithm for Model $\mathcal{M}$ are the same.

8.6 REML PX EM algorithm M-step for Model $\mathcal{Z}$ using $\boldsymbol{y}_c^{(D)}$

The M-step of the REML PX EM algorithm involves maximising $Q_{\mathcal{Z}}(\boldsymbol{\mathcal{T}};\boldsymbol{\mathcal{T}}^{(w)})$ wih respect to $\boldsymbol{\mathcal{T}} = (\sigma_*^2, \boldsymbol{\lambda}_*^T, \psi_*, \boldsymbol{d}_*^T, \omega)^T$ and then applying the reduction function $\mathcal{R}(\boldsymbol{\mathcal{T}}) = (\sigma_*^2, \boldsymbol{\lambda}_*, \omega^2\psi_*, \boldsymbol{d}_*^T)^T$ to obtain $\boldsymbol{\Gamma} = (\sigma^2, \boldsymbol{\lambda}^T, \psi, \boldsymbol{d}^T)^T$.

8.6.1 Update for σ_*^2

The update for σ_*^2 is the same as the update for σ^2 under Model $\mathcal{M}$, i.e.,

$$\sigma_*^{2(w+1)} = -\frac{1}{n-p}\Big\{(\boldsymbol{y}_o - \boldsymbol{Z}_g\tilde{\boldsymbol{u}}_g{}^{(w)})^T\boldsymbol{K}(\boldsymbol{y}_o - \boldsymbol{Z}_g\tilde{\boldsymbol{u}}_g{}^{(w)}) + \mathrm{tr}\big(\boldsymbol{Z}_g^T\boldsymbol{K}\boldsymbol{Z}_g\boldsymbol{T}_{u_gu_g}^{(w)}\big)\Big\}.$$

Applying the reduction function we have $\sigma^{2(w+1)} = \sigma_*^{2(w+1)}$.

8.6.2 Update for $\lambda_{*_{jr}}$

The update for $\lambda_{*_{jr}}$ is the same as that for λ_{jr} under Model $\mathcal{M}$, i.e.,

$$\boldsymbol{\lambda}_*^{(w+1)} = \boldsymbol{A}_*^{(w)^{-1}}\boldsymbol{b}_*^{(w)},$$

where the elements of the $mk \times mk$ matrix $\boldsymbol{A}_*^{(w)}$ are

$$a_{jr,lq}^{(w)} = \tilde{\boldsymbol{f}}^{(w)^T}(\boldsymbol{J}_{lq}^T\boldsymbol{\Psi}^{(w)^{-1}}\boldsymbol{J}_{jr}\otimes\boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} + \mathrm{tr}\big\{(\boldsymbol{J}_{lq}^T\boldsymbol{\Psi}^{(w)^{-1}}\boldsymbol{J}_{jr}\otimes\boldsymbol{I}_m)\boldsymbol{T}_{ff}^{(w)}\big\}$$

and $\boldsymbol{b}^{(w)}$ is a $mk \times 1$ vector with elements

$$b_{jr}^{(w)} = \mathrm{tr}\big\{\boldsymbol{\Psi}^{-1}\boldsymbol{J}_{jr}\otimes\boldsymbol{I}_m)(\tilde{\boldsymbol{f}}^{(w)}\tilde{\boldsymbol{u}}_g^{(w)^T} + \boldsymbol{T}_{fu_g}^{(w)})\big\}.$$

Applying the reduction function we have $\boldsymbol{\lambda}^{(w+1)} = \boldsymbol{\lambda}_*^{(w+1)}$.

8.6.3 Update for ψ_*

The update for ψ_* is the same as that for ψ under Model $\mathcal{M}$, i.e.,

$$\begin{aligned}\psi_*^{(w+1)} = \frac{1}{mp}\Big[&\tilde{\boldsymbol{f}}^{(w)^T}(\boldsymbol{\Lambda}^{(w)^T}\boldsymbol{\Lambda}^{(w)}\otimes\boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} + \mathrm{tr}\big\{(\boldsymbol{\Lambda}^{(w)^T}\boldsymbol{\Lambda}^{(w)}\otimes\boldsymbol{I}_m)\boldsymbol{T}_{ff}^{(w)}\big\}\\ &+\tilde{\boldsymbol{u}}_g^{(w)^T}\tilde{\boldsymbol{u}}_g^{(w)} + \mathrm{tr}\big(\boldsymbol{T}_{u_gu_g}^{(w)}\big) - 2\mathrm{tr}\big\{(\boldsymbol{\Lambda}^{(w)}\otimes\boldsymbol{I}_m)(\tilde{\boldsymbol{f}}^{(w)}\tilde{\boldsymbol{u}}_g^{(w)^T} + \boldsymbol{T}_{fu_g}^{(w)})\big\}\Big].\end{aligned}$$

Applying the reduction function we have $\psi^{(w+1)} = \omega^{2^{(w+1)}}\psi_*^{(w+1)}$.

8.6.4 Update for $d_{*_{qr}}$

The update for $d_{*_{qr}}$ is the same as that for d_{qr} under Model $\mathcal{M}$, i.e.,

$$d_{*_{qr}}^{(w+1)} = \frac{1}{m}\Big\{\tilde{\boldsymbol{f}}_q^{(w)^T}\tilde{\boldsymbol{f}}_r^{(w)} + \mathrm{tr}\big(\boldsymbol{T}_{ff_{qr}}^{(w)}\big)\Big\}.$$

Applying the reduction function we have $d_{qr}^{(w+1)} = d_{*_{qr}}^{(w+1)}$.

8.6.5 Update for ω

Differentiating $Q_{\mathcal{Z}}(\boldsymbol{\mathcal{T}};\boldsymbol{\mathcal{T}}^{(w)})$ with respect to ω we have

$$\begin{aligned}\frac{\partial Q_{\mathcal{Z}}(\boldsymbol{\mathcal{T}};\boldsymbol{\mathcal{T}}^{(w)})}{\partial\omega} &= -\frac{1}{2}\Big[\frac{2mp}{\omega} - \frac{2}{\psi\omega^3}E\big\{\boldsymbol{f}^T(\boldsymbol{\Lambda}_*^T\boldsymbol{\Lambda}_*\otimes\boldsymbol{I}_m)\boldsymbol{f}|\boldsymbol{y}_2;\boldsymbol{\mathcal{T}}^{(w)}\big\}\\ &\quad -\frac{2}{\psi\omega^3}E\big(\boldsymbol{u}_g^T\boldsymbol{u}_g|\boldsymbol{y}_2;\boldsymbol{\mathcal{T}}^{(w)}\big) + \frac{4}{\psi\omega^3}E\big\{\boldsymbol{u}_g^T(\boldsymbol{\Lambda}_*\otimes\boldsymbol{I}_m)\boldsymbol{f}|\boldsymbol{y}_2;\boldsymbol{\mathcal{T}}^{(w)}\big\}\Big].\end{aligned}$$

Equating to zero and solving for ω^2 we have

$$\begin{aligned}\omega^{2^{(w+1)}} &= \frac{1}{mp\psi^{(w)}}\Big[\tilde{\boldsymbol{f}}^{(w)^T}(\boldsymbol{\Lambda}^{(w)^T}\boldsymbol{\Lambda}^{(w)}\otimes\boldsymbol{I}_m)\tilde{\boldsymbol{f}}^{(w)} + \mathrm{tr}\big\{(\boldsymbol{\Lambda}^{(w)^T}\boldsymbol{\Lambda}^{(w)}\otimes\boldsymbol{I}_m)\boldsymbol{T}_{ff}^{(w)}\big\}\\ &\quad +\tilde{\boldsymbol{u}}_g^{(w)^T}\tilde{\boldsymbol{u}}_g^{(w)} + \mathrm{tr}\big(\boldsymbol{T}_{u_gu_g}^{(w)}\big) - 2\mathrm{tr}\big\{(\boldsymbol{\Lambda}^{(w)}\otimes\boldsymbol{I}_m)(\tilde{\boldsymbol{f}}^{(w)}\tilde{\boldsymbol{u}}_g^{(w)^T} + \boldsymbol{T}_{fu_g}^{(w)})\big\}\Big]\end{aligned}$$

8.7 Example: National Variety Trial data

In the following example and simulation study when we refer to Model $\mathcal{O}$ or Model $\mathcal{M}$ it should be read as "the REML EM algorithm used to estimate the variance parameters associated with Model A " where Model A is Model $\mathcal{O}$ or Model $\mathcal{M}$ as the case may be. When we refer to Model $\mathcal{X}$ or Model $\mathcal{Z}$ it should be read as "the REML PX EM algorithm used to estimate the variance parameters associated with Model B " where Model B is Model $\mathcal{X}$ or Model $\mathcal{Z}$ as the case may be. We consider the small subset of National Variety Trial data used in the previous two chapters and presented in Table 6.2. For every model fitted to the simulated data we define the starting values $\boldsymbol{\lambda}^{(0)} = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1)^T$, $\psi^{(0)} = 0.1$, $\sigma^{2^{(0)}} = 0.1$, and $d^{(0)} = 1$, $\omega^{(0)} = 1$ when required.

Results for Model $\mathcal{O}$, $\mathcal{M}$, $\mathcal{X}$, and Model $\mathcal{Z}$ with $k = 1$ factors and using the new missing data specification $\boldsymbol{y}_m^{(D)} = (\boldsymbol{f}^T, \boldsymbol{u}_g^T)^T$ are provided in Table 8.1. In the interests of brevity, parameter updates for Model $\mathcal{O}$ and Model $\mathcal{X}$ using the missing data specification $\boldsymbol{y}_m^{(D)} = (\boldsymbol{f}^T, \boldsymbol{u}_g^T)^T$ were omitted from this chapter, however they can be derived using information in this and previous chapters.

Comparing the results in Table 8.1 with the results in Table 6.1 and Table 7.1 we can see that for all models and levels of ε that an algorithm using the new missing data specification $\boldsymbol{y}_m^{(D)} = (\boldsymbol{f}^T, \boldsymbol{u}_g^T)^T$ outperforms an algorithm using the missing data specification $\boldsymbol{y}_m = (\boldsymbol{f}^T, \boldsymbol{\delta}^T)^T$ in terms of iterations until convergence. If we had to choose a model and a value of ε then we would choose Model $\mathcal{Z}$ based on the new missing data specification and $\varepsilon = 10^{-5}$. This model results in significantly fewer iterations to convergence than the next best option. When the stopping rule is based on $\varepsilon = 10^{-5}$ there is no loss in terms of the final residual log-likelihood and the Frobenious matrix norm of the variance-covariance matrix is accurate to 5 decimal places. This model and stopping rule combination converges in 23 iterations which is acceptable from a practical point of view. In terms of iterations until convergence and accuracy of the estimated genetic variance-covariance matrix, Model $\mathcal{Z}$ using the new missing data specification compares favourably to the next best model options, which are Model $\mathcal{M}$ and Model $\mathcal{X}$ using the new missing data specification.

8.8 Simulation study

Model $\mathcal{Z}$ using the new missing data specification $\boldsymbol{y}_m^{(D)} = (\boldsymbol{f}^T, \boldsymbol{u}_g^T)^T$ was identified as the best model when applied to the small subset of National Variety Trial data presented in Table 6.2 and a stopping rule with $\varepsilon = 10^{-5}$ was identified as being sufficient in determining convergence. We consider a simulation study to confirm these findings.

The simulation study involved Model $\mathcal{O}$, $\mathcal{M}$, $\mathcal{X}$, and Model $\mathcal{Z}$ using both $\boldsymbol{y}_m^{(D)} = (\boldsymbol{f}^T, \boldsymbol{u}_g^T)^T$ and $\boldsymbol{y}_m = (\boldsymbol{f}^T, \boldsymbol{\delta}^T)^T$. Each model was fitted to 3 lots of 100 randomly generated data sets. Each lot of 100 were generated using $\sigma^2 = 0.05$ and $d = 1$ but differed in the specification of the genetic variance-covariance matrix. The three different genetic variance-covariance matrices used to generate the data can be broadly classified as consisting of high, medium, and low correlations. They were generated by considering the same vector of factor loadings, i.e.,
$\boldsymbol{\lambda} = (0.1, 0.5, -0.05, -0.1, 0.2, 0.05, 0.005)^T$ but three different values for the com-

Table 8.1: Iterations until convergence when using a stopping rule with varying levels of ε, final residual log-likelihood (ℓ_R), and the Frobenius matrix norm of the estimated genetic variance-covariance matrix when using a REML EM algorithm to fit Model $\mathcal{O}$ and Model $\mathcal{M}$ and a REML PX EM algorithm to fit Model $\mathcal{X}$ and $\mathcal{Z}$ with $k = 1$ factors to the small subset of National Variety Trial data presented in Table 6.2. All algorithms use the missing data specification $\boldsymbol{y}_m^{(D)} = (\boldsymbol{f}^T, \boldsymbol{u}_g^T)^T$.

	Model $\mathcal{O}$			Model $\mathcal{X}$			Model $\mathcal{M}$			Model $\mathcal{Z}$		
ε	Iter.	ℓ_R	$\|\widehat{\boldsymbol{G}}_e\|_F$	Iter.	ℓ_R	$\|\widehat{\boldsymbol{G}}_e\|_F$	Iter.	ℓ_R	$\|\widehat{\boldsymbol{G}}_e\|_F$	Iter.	ℓ_R	$\|\widehat{\boldsymbol{G}}_e\|_F$
10^{-8}	66	300.9849	0.3180123	54	300.9849	0.3180123	54	300.9849	0.3180123	40	300.9849	0.3180123
10^{-7}	57	300.9849	0.3180122	46	300.9849	0.3180124	46	300.9849	0.3180124	35	300.9849	0.3180123
10^{-6}	48	300.9849	0.3180113	38	300.9849	0.3180129	38	300.9849	0.3180129	29	300.9849	0.3180124
10^{-5}	39	300.9849	0.3180019	30	300.9849	0.3180189	30	300.9849	0.3180189	23	300.9849	0.3180132
10^{-4}	30	300.9849	0.3179079	22	300.9849	0.3180807	23	300.9849	0.3180634	18	300.9849	0.3180216
10^{-3}	21	300.9842	0.3169733	15	300.9802	0.3185419	16	300.9824	0.3184078	13	300.9830	0.3180960
10^{-2}	13	300.9369	0.3101104	10	300.8763	0.3200647	10	300.8763	0.3200647	9	300.9220	0.3178401

mon specific variance, i.e., $\psi = 0.01$, $\psi = 0.05$, and $\psi = 0.25$. The three genetic variance-covariance (and correlation) matrices that were used to generate the data are presented in Table 8.2. When using the common specific variance $\psi = 0.01$, $\psi = 0.05$, and $\psi = 0.25$ the correlations in the genetic correlation matrix have ranges of (-0.69, 0.88), (-0.37, 0.61), and (-0.14, 0.26) respectively.

For every model fitted to the simulated data we define the starting values $\boldsymbol{\lambda}^{(0)} = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1)^T$, $\psi^{(0)} = 0.1$, $\sigma^{2(0)} = 0.1$, and $d^{(0)} = 1$, $\omega^{(0)} = 1$ when required. The minimum, median, and maximum number of iterations until convergence for all models using the two different missing data specifications and varying levels of ε are provided in Table 8.3. Model $\mathcal{Z}$ is the preferred model for both missing data specifications and at all levels of ε, regardless of the level of ψ. Model $\mathcal{Z}$ using the new missing data specification $\boldsymbol{y}_m^{(D)} = (\boldsymbol{f}^T, \boldsymbol{u}_g^T)^T$ is preferred to Model $\mathcal{Z}$ when using the missing data specification $\boldsymbol{y}_m = (\boldsymbol{f}^T, \boldsymbol{\delta}^T)^T$ at all levels of ε and when $\psi = 0.05$ and $\psi = 0.25$, i.e., when the genetic variance-covariance matrix consists of low to medium correlations.

An interesting outcome of the simulation study is the performance of Model $\mathcal{Z}$ using the missing data specification $\boldsymbol{y}_m = (\boldsymbol{f}^T, \boldsymbol{\delta}^T)^T$ when $\psi = 0.01$, i.e., when the genetic variance-covariance matrix consists of high correlations. This model and missing data specification significantly outperforms all other models. For levels of ε smaller than 10^{-5} the performance gain, in terms of iterations until convergence, is between $\frac{1}{2}$ to almost $\frac{1}{3}$ the number of iterations until convergence of the next best model, which is Model $\mathcal{Z}$ using the new missing data specification. A caveat on this result is that performance is judged only on the number of iterations until convergence and does not consider computing time. Although we have not considered a comprehensive computing strategy for each model in this thesis, the time per iteration when using the missing data specification $\boldsymbol{y}_m = (\boldsymbol{f}^T, \boldsymbol{\delta}^T)^T$ is approximately 2.5 times greater than time per iteration when using the new missing data specification $\boldsymbol{y}_m^{(D)} = (\boldsymbol{f}^T, \boldsymbol{u}_g^T)^T$. Therefore, the two have comparable performance in terms of elapsed time to convergence when $\psi = 0.01$. In all other cases Model $\mathcal{Z}$ using the new missing data specification is far superior in terms of elapsed time to convergence than all other models. The reduction in time per iteration for models using the new missing data specification concurs with results presented in Thompson et al. (2003). Thompson et al. (2003) found that time per iteration was substantially reduced when applying the average information algorithm to a "dependent" formulation of the factor analytic mixed model.

In the small subset of National Variety Trial data example it was found that using the less stringent convergence criteria $\varepsilon = 10^{-5}$ had no effect on the final residual log-likelihood and little effect on the estimate of the genetic variance-covariance matrix. In Table 8.4 summary statistics for the absolute difference between the Frobenius matrix norm of the estimated genetic variance-covariance matrix using $\varepsilon = 10^{-8}$ and the Frobenius matrix norm of the estimated genetic variance covariance matrix using another level of ε are presented. If we consider the standard to be the Frobenius matrix norm of the estimated variance-covariance matrix obtained when using $\varepsilon = 10^{-8}$ then the data in Table 8.4 supports an increase in ε to 10^{-5}. For Model $\mathcal{Z}$ and $\varepsilon = 10^{-5}$ the worst case is when using the new missing data specification and when $\psi = 0.01$. In this case the median value of the difference between the Frobenius matrix norm of the estimated variance-covariance matrix using $\varepsilon = 10^{-8}$ and $\varepsilon = 10^{-5}$ is 0.000028. This difference at the other two levels of ψ is less than 0.00001. These differences indicate that from a practical point of view there is little loss in terms of an accurate estimate of the genetic variance-covariance matrix in moving to a less stringent stopping rule based on $\varepsilon = 10^{-5}$.

We recommend Model $\mathcal{Z}$ using the new missing data specification when fitting a factor analytic mixed model. In Figure 8.1 we compare the number of iterations until convergence for Model $\mathcal{Z}$ using $\boldsymbol{y}_m = (\boldsymbol{f}^T, \boldsymbol{\delta}^T)^T$ and $\boldsymbol{y}_m^{(D)} = (\boldsymbol{f}^T, \boldsymbol{u}_g^T)^T$. The convergence criteria for both was based on $\varepsilon = 10^{-5}$. Model $\mathcal{Z}$ using the new missing data specification and applied to data generated using $\psi = 0.05$ and $\psi = 0.25$ converges (on average) in 54% and 16% respectively of the number of iterations it takes Model $\mathcal{Z}$ using $\boldsymbol{y}_m = (\boldsymbol{f}^T, \boldsymbol{\delta}^T)^T$ to converge. When $\psi = 0.05$ and $\psi = 0.25$ the median (50th percentile in Figure 8.1) number of iterations to convergence is 15 and 25 iterations respectively for Model $\mathcal{Z}$ using the new missing data specification. Both are acceptable from a practical point of view. When $\psi = 0.01$ the median number of iterations to convergence is 56. This is unacceptably high and greater than the median number of iterations to convergence for Model $\mathcal{Z}$ using $\boldsymbol{y}_m = (\boldsymbol{f}^T, \boldsymbol{\delta}^T)^T$. However, when $\psi = 0.01$ the two missing data specifications for Model $\mathcal{Z}$ are comparable in terms of elapsed time to convergence. It is also worth noting that increasing ε to 10^{-4} reduces the median number of iterations until convergence of Model $\mathcal{Z}$ using $\boldsymbol{y}_m^{(D)} = (\boldsymbol{f}^T, \boldsymbol{u}_g^T)^T$ to 28 iterations. This is a feasible alternative if we are prepared to accept an estimate of the genetic variance-covariance matrix which is slightly less accurate than that obtained when using $\varepsilon = 10^{-5}$.

Table 8.2: The three genetic variance-covariance matrices based on $\boldsymbol{\lambda} = (0.1, 0.5, -0.05, -0.1, 0.2, 0.05, 0.005)^T$ and $\psi = 0.01$, $\psi = 0.05$, and $\psi = 0.25$ used to generate the three lots of 100 data sets. Variances are on the diagonals, covariances in the lower triangle, and correlations in the upper triangle.

(a) $\psi = 0.01$

0.0200	0.69	-0.32	-0.50	0.63	0.32	0.04
0.0500	0.2600	-0.44	-0.69	0.88	0.44	0.05
-0.0050	-0.0250	0.0125	0.32	-0.40	-0.20	-0.02
-0.0100	-0.0500	0.0050	0.0200	-0.63	-0.32	-0.04
0.0200	0.1000	-0.0100	-0.0200	0.0500	0.40	0.04
0.0050	0.0250	-0.0025	-0.0050	0.0100	0.0125	0.02
0.0005	0.0025	-0.0002	-0.0005	0.0010	0.0002	0.0100

(b) $\psi = 0.05$

0.0600	0.37	-0.09	-0.17	0.27	0.09	0.01
0.0500	0.3000	-0.20	-0.37	0.61	0.20	0.02
-0.0050	-0.0250	0.0525	0.09	-0.15	-0.05	-0.00
-0.0100	-0.0500	0.0050	0.0600	-0.27	-0.09	-0.01
0.0200	0.1000	-0.0100	-0.0200	0.0900	0.15	0.01
0.0050	0.0250	-0.0025	-0.0050	0.0100	0.0525	0.00
0.0005	0.0025	-0.0002	-0.0005	0.0010	0.0002	0.0500

(c) $\psi = 0.25$

0.2600	0.14	-0.02	-0.04	0.07	0.02	0.00
0.0500	0.5000	-0.07	-0.14	0.26	0.07	0.01
-0.0050	-0.0250	0.2525	0.02	-0.04	-0.01	-0.00
-0.0100	-0.0500	0.0050	0.2600	-0.07	-0.02	-0.00
0.0200	0.1000	-0.0100	-0.0200	0.2900	0.04	0.00
0.0050	0.0250	-0.0025	-0.0050	0.0100	0.2525	0.00
0.0005	0.0025	-0.0002	-0.0005	0.0010	0.0002	0.2500

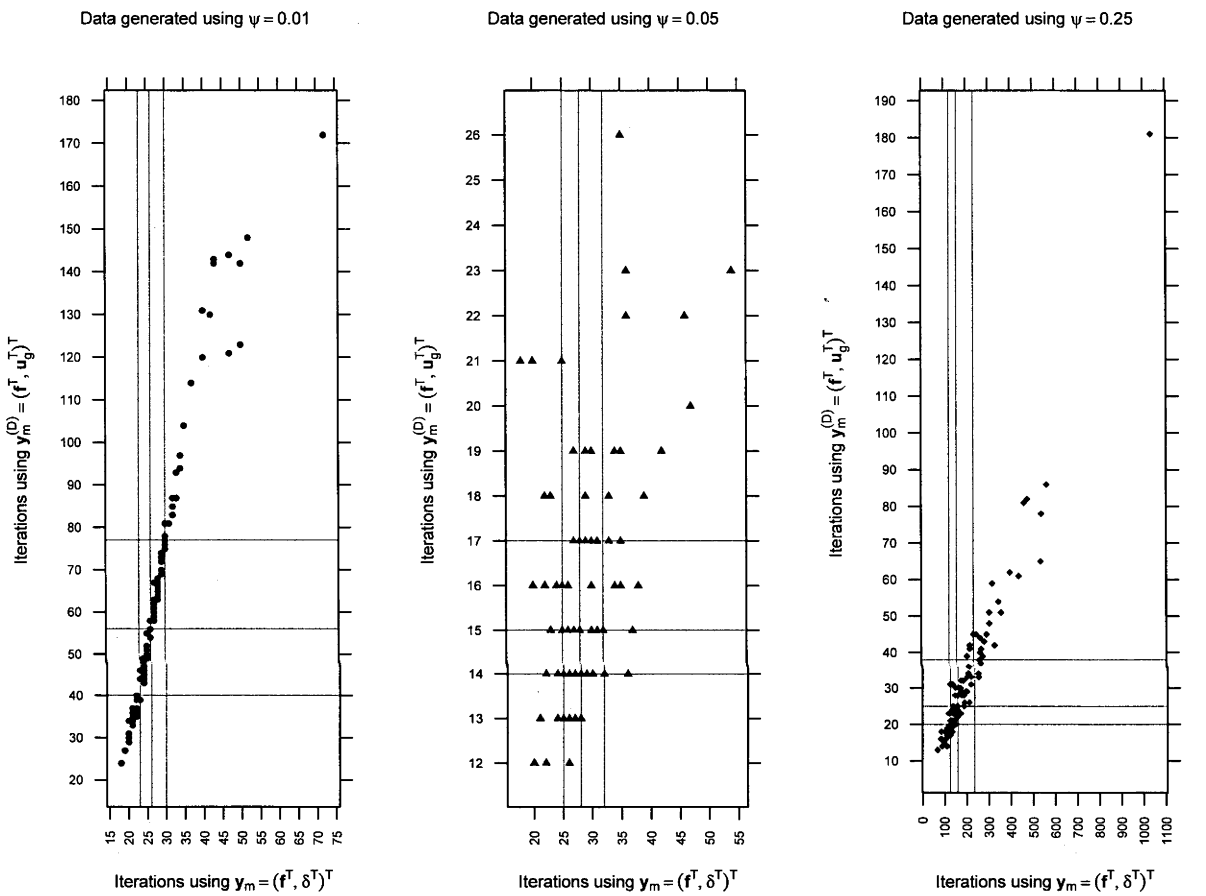
Table 8.3: Minimum, median, and maximum number of iterations until convergence for Model $\mathcal{O}$, $\mathcal{M}$, $\mathcal{X}$, and $\mathcal{Z}$ using two different missing data specifications and different levels of ε. A 100 data sets were generated for each level of the common specific variance, i.e., $\psi = 0.01$, $\psi = 0.05$, and $\psi = 0.25$.

		Missing data specification $\boldsymbol{y}_m = (\boldsymbol{f}^T, \boldsymbol{\delta}^T)^T$												Missing data specification $\boldsymbol{y}_m^{(D)} = (\boldsymbol{f}^T, \boldsymbol{u}_g^T)^T$									
		Model $\mathcal{O}$			Model $\mathcal{M}$			Model $\mathcal{X}$			Model $\mathcal{Z}$			Model $\mathcal{O}$			Model $\mathcal{M}$			Model $\mathcal{X}$			Model $\mathcal{Z}$		
		min	med	max	min	med	max	min	med	max	min	med	max	min	med	max	min	med	max	min	med	max	min	med	max
$\varepsilon = 10^{-8}$																									
	$\psi = 0.01$	119	299	2500	94	289	2500	94	295	2500	36	57	816	137	339	2500	93	290	2500	93	294	2500	51	157	2500
	$\psi = 0.05$	86	144	235	48	58	114	51	64	130	34	54	114	50	76	121	33	44	77	32	44	77	23	28	50
	$\psi = 0.25$	265	447	2500	145	387	2500	177	451	2500	146	387	2500	35	50	364	26	47	339	27	50	364	26	46	340
$\varepsilon = 10^{-7}$																									
	$\psi = 0.01$	100	232	2500	77	222	2500	76	229	2500	29	46	479	115	276	2500	76	224	2500	75	228	2500	41	123	2500
	$\psi = 0.05$	73	122	198	40	48	93	42	54	109	28	45	93	42	64	102	27	36	63	27	36	63	19	23	41
	$\psi = 0.25$	224	370	2492	116	315	2067	145	376	2496	117	315	2058	30	41	311	22	38	285	23	42	311	21	38	286
$\varepsilon = 10^{-6}$																									
	$\psi = 0.01$	82	167	1126	61	157	873	60	163	1125	24	36	208	94	214	1550	59	159	970	59	162	1124	33	89	758
	$\psi = 0.05$	61	101	162	33	39	74	34	45	90	23	36	74	36	54	84	22	30	50	22	30	50	16	19	33
	$\psi = 0.25$	184	300	1982	91	231	1557	118	306	1986	91	231	1548	25	35	258	18	32	232	19	35	258	17	32	233
$\varepsilon = 10^{-5}$																									
	$\psi = 0.01$	64	104	249	44	92	203	44	98	247	18	26	72	72	154	651	43	93	223	42	97	246	24	56	172
	$\psi = 0.05$	49	81	126	25	31	54	28	37	70	18	28	54	29	43	66	17	23	37	17	23	37	12	15	26
	$\psi = 0.25$	144	226	1472	67	157	1046	92	225	1476	67	158	1037	21	28	205	15	25	180	15	28	205	13	25	181
$\varepsilon = 10^{-4}$																									
	$\psi = 0.01$	47	59	74	28	41	60	28	44	63	13	16	24	51	93	229	27	42	66	26	43	62	16	28	48
	$\psi = 0.05$	38	60	90	19	22	35	21	28	51	13	20	35	23	32	47	12	17	25	12	16	25	9	11	18
	$\psi = 0.25$	104	156	961	44	90	536	68	155	965	44	91	528	16	22	153	10	19	127	11	22	153	10	19	128
$\varepsilon = 10^{-3}$																									
	$\psi = 0.01$	29	35	46	13	15	19	13	15	19	8	9	10	30	43	59	12	15	20	12	14	18	8	11	16
	$\psi = 0.05$	26	39	54	13	15	20	14	20	32	9	13	19	16	21	29	8	10	15	8	11	16	6	8	11
	$\psi = 0.25$	59	92	441	28	48	260	44	88	445	29	49	261	11	16	100	7	12	74	8	16	100	7	12	75

Table 8.4: The absolute difference between $\|\widehat{G}_e\|_F$ using $\varepsilon = 10^{-8}$ and $\|\widehat{G}_e\|_F$ using another level of ε was calculated for the 100 generated data sets for each level of the common specific variance, i.e., $\psi = 0.01$, $\psi = 0.05$, and $\psi = 0.25$. The minimum, median, and maximum of these absolute differences are provided below. We have used the notation $|(8) - (a)|$ where (8) refers to $\varepsilon = 10^{-8}$ and (a) is one of the other levels of ε. These summary statistics have been calculated for Model $\mathcal{O}$, $\mathcal{M}$, $\mathcal{X}$, and $\mathcal{Z}$ using the two different missing data specifications.

Missing data specification $\boldsymbol{y}_m = (\boldsymbol{f}^T, \boldsymbol{\delta}^T)^T$												
	Model $\mathcal{O}$			Model $\mathcal{M}$			Model $\mathcal{X}$			Model $\mathcal{Z}$		
	min	med	max	min	med	max	min	med	max	min	med	max
$\psi = 0.01$												
$\|(8) - (7)\|$	0.000000	0.000000	0.000008	0.000000	0.000001	0.000016	0.000000	0.000000	0.000009	0.000000	0.000000	0.000003
$\|(8) - (6)\|$	0.000000	0.000004	0.000143	0.000001	0.000006	0.000211	0.000001	0.000005	0.000145	0.000000	0.000001	0.000028
$\|(8) - (5)\|$	0.000000	0.000041	0.000524	0.000012	0.000062	0.000677	0.000012	0.000051	0.000559	0.000005	0.000011	0.000140
$\|(8) - (4)\|$	0.000006	0.000336	0.001168	0.000113	0.000488	0.001591	0.000113	0.000434	0.001418	0.000043	0.000107	0.000513
$\|(8) - (3)\|$	0.000649	0.004479	0.012572	0.000885	0.001959	0.003221	0.000760	0.002009	0.003221	0.000014	0.000581	0.001471
$\psi = 0.05$												
$\|(8) - (7)\|$	0.000000	0.000001	0.000001	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
$\|(8) - (6)\|$	0.000001	0.000006	0.000015	0.000000	0.000000	0.000002	0.000000	0.000000	0.000002	0.000000	0.000000	0.000002
$\|(8) - (5)\|$	0.000009	0.000061	0.000155	0.000000	0.000003	0.000023	0.000000	0.000001	0.000015	0.000000	0.000002	0.000018
$\|(8) - (4)\|$	0.000084	0.000602	0.001539	0.000000	0.000035	0.000191	0.000000	0.000005	0.000152	0.000000	0.000027	0.000195
$\|(8) - (3)\|$	0.000965	0.005895	0.015257	0.000004	0.000277	0.001178	0.000000	0.000056	0.001086	0.000004	0.000452	0.001667
$\psi = 0.25$												
$\|(8) - (7)\|$	0.000000	0.000000	0.000003	0.000000	0.000000	0.000001	0.000000	0.000000	0.000000	0.000000	0.000000	0.000001
$\|(8) - (6)\|$	0.000000	0.000001	0.000033	0.000000	0.000001	0.000007	0.000000	0.000000	0.000001	0.000000	0.000001	0.000008
$\|(8) - (5)\|$	0.000000	0.000020	0.000331	0.000000	0.000011	0.000072	0.000000	0.000001	0.000006	0.000000	0.000012	0.000077
$\|(8) - (4)\|$	0.000002	0.000616	0.003154	0.000005	0.000155	0.000533	0.000000	0.000013	0.000061	0.000000	0.000158	0.000540
$\|(8) - (3)\|$	0.000177	0.008385	0.029612	0.000458	0.002320	0.009164	0.000008	0.000199	0.000758	0.000483	0.002363	0.009201
Missing data specification $\boldsymbol{y}_m^{(D)} = (\boldsymbol{f}^T, \boldsymbol{u}_g^T)^T$												
	Model $\mathcal{O}$			Model $\mathcal{M}$			Model $\mathcal{X}$			Model $\mathcal{Z}$		
	min	med	max	min	med	max	min	med	max	min	med	max
$\psi = 0.01$												
$\|(8) - (7)\|$	0.000000	0.000001	0.000004	0.000000	0.000001	0.000014	0.000000	0.000000	0.000009	0.000000	0.000000	0.000012
$\|(8) - (6)\|$	0.000000	0.000011	0.000305	0.000001	0.000006	0.000194	0.000001	0.000005	0.000145	0.000000	0.000003	0.000169
$\|(8) - (5)\|$	0.000004	0.000122	0.002836	0.000011	0.000056	0.000644	0.000013	0.000050	0.000559	0.000006	0.000028	0.000529
$\|(8) - (4)\|$	0.000408	0.001283	0.015645	0.000107	0.000440	0.001515	0.000124	0.000430	0.001415	0.000053	0.000245	0.001234
$\|(8) - (3)\|$	0.004177	0.011778	0.062254	0.000808	0.001862	0.003068	0.000931	0.001962	0.003187	0.000315	0.001167	0.002429
$\psi = 0.05$												
$\|(8) - (7)\|$	0.000000	0.000000	0.000001	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
$\|(8) - (6)\|$	0.000000	0.000004	0.000008	0.000000	0.000001	0.000001	0.000000	0.000001	0.000001	0.000000	0.000000	0.000001
$\|(8) - (5)\|$	0.000005	0.000036	0.000093	0.000000	0.000006	0.000015	0.000000	0.000007	0.000015	0.000000	0.000002	0.000014
$\|(8) - (4)\|$	0.000048	0.000363	0.000920	0.000002	0.000061	0.000149	0.000000	0.000060	0.000149	0.000000	0.000029	0.000127
$\|(8) - (3)\|$	0.000567	0.003547	0.009039	0.000012	0.000392	0.001092	0.000012	0.000360	0.001312	0.000000	0.000592	0.004037
$\psi = 0.25$												
$\|(8) - (7)\|$	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
$\|(8) - (6)\|$	0.000000	0.000000	0.000005	0.000000	0.000000	0.000001	0.000000	0.000000	0.000001	0.000000	0.000000	0.000001
$\|(8) - (5)\|$	0.000000	0.000001	0.000040	0.000000	0.000001	0.000009	0.000000	0.000000	0.000008	0.000000	0.000001	0.000019
$\|(8) - (4)\|$	0.000000	0.000018	0.000481	0.000000	0.000013	0.000312	0.000000	0.000003	0.000141	0.000000	0.000012	0.000145
$\|(8) - (3)\|$	0.000000	0.000390	0.004412	0.000002	0.000244	0.004239	0.000000	0.000055	0.001695	0.000009	0.000189	0.004155

Figure 8.1: Model $\mathcal{Z}$ number of iterations until convergence using the new missing data specification $\boldsymbol{y}_m^{(D)} = (\boldsymbol{f}^T, \boldsymbol{u}_g^T)^T$ versus iterations until convergence using the missing data specification $\boldsymbol{y}_m^{(D)} = (\boldsymbol{f}^T, \boldsymbol{\delta}^T)^T$. Iterations until convergence for both is based on $\varepsilon = 10^{-5}$. Each data point represents 1 generated data set out of 100 for each of $\psi = 0.01$, $\psi = 0.05$, and $\psi = 0.25$. Grid lines represent the 25th, 50th, and 75th percentiles.



Chapter 9

Conclusion

In this thesis we have advanced the theory underlying the application of REML EM and REML PX EM algorithms to linear mixed models and factor analytic mixed models. A new incomplete data specification has been introduced based on the transformed observed data vector associated with the marginal distrubution in the conditional derivation of REML (Verbyla, 1990). This transformed observed data vector has been referred to as $\boldsymbol{y}_2$. There are two benefits in defining the incomplete data in a REML EM or REML PX EM algorithm as $\boldsymbol{y}_2$ rather than the observed data vector $\boldsymbol{y}_o$. Firstly, the new incomplete data specification results in a REML EM or REML PX EM algorithm that is computationally more efficient. Secondly, in our experience the new incomplete data specification will result in a REML EM or REML PX EM algorithm having a faster rate of convergence. We have used Loewner partial ordering to describe the conditions necessary for the new incomplete data specification to result in a faster rate of convergence.

There are two related aspects to the improvement in computational efficiency when specifying the incomplete data as $\boldsymbol{y}_2$. Firstly, the E-step of a REML EM or REML PX EM algorithm using the new incomplete data specification only requires knowledge of one conditional distribution, namely $\boldsymbol{u}|\boldsymbol{y}_2$. In contrast, the E-step of a REML EM and REML PX EM algorithm using $\boldsymbol{y}_o$ as the incomplete data requires two conditional distributions, namely $\boldsymbol{u}|\boldsymbol{y}_2$ and typically $\boldsymbol{e}|\boldsymbol{y}_2$. This aspect of the new incomplete data specification may have an application in situations where the E-step is intractable. This is a potential area of future research.

The second aspect to the improvement in computational efficiency relates to the first. REML EM and REML PX EM algorithm parameter updates involve computing the

trace of a matrix. For the new incomplete data specification this matrix is either the $b \times b$ matrix $\boldsymbol{C}^{ZZ}$ or the $b \times b$ matrix $\boldsymbol{Z}^T\boldsymbol{U}\boldsymbol{Z}\boldsymbol{C}^{ZZ}$ where $\boldsymbol{U}$ is a projection matrix. When the incomplete data is specified as $\boldsymbol{y}_o$ this matrix is either $\boldsymbol{C}^{ZZ}$ or the $n \times n$ matrix $\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}^T$ where $\boldsymbol{W} = (\boldsymbol{X} \ \boldsymbol{Z})$. The latter involves the inverse of the entire coefficient matrix associated with Henderson's mixed model equations whereas the former only involves a smaller partition of this matrix. It is computing the trace of $\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}^T$ which makes a REML EM or REML PX EM algorithm where the incomplete data is specified as $\boldsymbol{y}_o$ less computationally efficient than a REML EM or REML PX EM algorithm based on the new incomplete data specification.

Although the REML EM and REML PX EM algorithm for linear mixed models based on the new incomplete data specification is a step forward in terms of computational efficiency further work is required before either of these two algorithms can be easily implemented within existing software packages using Newton-Raphson type algorithms. In the case of the average information algorithm (Gilmour et al., 1995) the only term computed in terms of the inverse of the coefficient matrix associated with Henderson's mixed model equations is $\boldsymbol{C}^{ZZ}$. The REML EM and REML PX EM algorithm require computing $\boldsymbol{Z}^T\boldsymbol{U}\boldsymbol{Z}\boldsymbol{C}^{ZZ}$. A potential area of future research would be to study approximations of the trace of this matrix only using terms that form part of the average information algorithm computing strategy.

Factor analytic mixed models are routinely applied to data arising from plant improvement programmes. In statistical software packages using Newton-Raphson type algorithms these models can be extremely difficult to fit. Successive iterations of these algorithms are not guaranteed to increase the residual log-likelihood and parameter updates may be outside their parameter space. REML EM or REML PX EM algorithms for factor analytic mixed models solve these problems. However, they do suffer from slow convergence. Besides specifying the incomplete data for these types of models as $\boldsymbol{y}_2$ this thesis has addressed the problem of slow convergence by considering model specification, a new parameter expansion, a new missing data specification, and the efficacy of a less stringent stopping rule.

Factor analytic mixed models with $k = 1$ factors and a common specific variance were considered in an example and a simulation study. The recommendation of this thesis is to use a REML PX EM algorithm based on a model specification where the variance of the factor scores is assumed to be an unstructured positive definite matrix rather than the identity matrix, a new missing data specification based on the "dependent" formulation of factor analytic variance models presented in Thomp-

son et al. (2003), a new parameter expansion which is associated with the specific variances, and a less stringent convergence criterion based on $\varepsilon = 10^{-5}$ rather than $\varepsilon = 10^{-8}$. In many instances this results in iterations to convergence being less than 10% of the number of iterations to convergence using an alternative model specification, missing data specification, parameter expansion, and convergence criterion. Future research is required to determine if these results apply to factor analytic mixed models which don't have a common specific variance and with $k > 1$ factors. The research presented in this thesis suggests that a REML PX EM algorithm has the potential to be a practical option when fitting factor analytic mixed models.

Appendix A

Vector and matrix results

A.1 Conditional distribution between two vectors

If $\boldsymbol{x} = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T)^T$ is distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T)^T$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

where $\det(\boldsymbol{\Sigma}_{22}) > 0$ then for a given value of $\boldsymbol{x}_2$

$$\boldsymbol{x}_1 | \boldsymbol{x}_2 \sim N\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\boldsymbol{x}_2 - \boldsymbol{\mu}_2\right), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)$$

A.2 Covariance between two vectors

If $\boldsymbol{x} = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T)^T$ is distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T)^T$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

then we can write

$$\boldsymbol{\Sigma}_{12} = E(\boldsymbol{x}_1\boldsymbol{x}_2^T) - E(\boldsymbol{x}_1)E(\boldsymbol{x}_2)^T.$$

Rearranging we have

$$\begin{aligned} E(\boldsymbol{x}_1\boldsymbol{x}_2^T) &= E(\boldsymbol{x}_1)E(\boldsymbol{x}_2)^T + \boldsymbol{\Sigma}_{12} \\ &= \boldsymbol{\mu}_1\boldsymbol{\mu}_2^T + \boldsymbol{\Sigma}_{12}. \end{aligned}$$

A.3 Matrix identities

A.3.1 Trace of a matrix

The following is taken from Mardia et al. (1979). For a $n \times n$ matrix $\boldsymbol{A}$ with elements a_{ij} the trace of $\boldsymbol{A}$ is

$$\text{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii} = \sum_{i=1}^{n} \lambda_i$$

where λ_i is the i-th eigenvalue of $\boldsymbol{A}$.

A.3.2 Quadratic forms

The following result is taken from Searle et al. (1992). For $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$ and symmetric matrix $\boldsymbol{A}$,

$$E(\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu} + \text{tr}(\boldsymbol{A}\boldsymbol{V})$$

A.3.3 Determinants

The following results are taken from Harville (1997).

Matrix multiplied by a scalar

For any $n \times n$ matrix $\boldsymbol{A}$ and scalar k

$$\det(k\boldsymbol{A}) = k^n \det(\boldsymbol{A})$$

Invertible matrix

If $\boldsymbol{A}$ is non-singular then

$$\det(\boldsymbol{A}^{-1}) = \frac{1}{\det(\boldsymbol{A})}$$

Matrix multiplication

For $n \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$

$$\det(\boldsymbol{AB}) = \det(\boldsymbol{A}) \det(\boldsymbol{B})$$

Partitioned matrix

For a non-singular matrix $\boldsymbol{A}$ partitioned as

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}$$

where $\boldsymbol{A}_{11}$ is a $m \times m$ matrix, $\boldsymbol{A}_{12}$ is a $m \times n$ matrix, $\boldsymbol{A}_{21}$ is a $n \times m$ matrix and $\boldsymbol{A}_{22}$ is a $n \times n$ matrix. The determinant of $\boldsymbol{A}$ is

$$\det(\boldsymbol{A}) = \det(\boldsymbol{A}_{11}) \det(\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})$$

A.3.4 Inverses of matrices of the form $\boldsymbol{A} + \boldsymbol{B}$

The following result is taken from Harville (1997). Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be $n \times n$ matrices and $\boldsymbol{A} + \boldsymbol{B}$ be nonsingular then

$$\begin{aligned} (\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{A} &= \boldsymbol{I}_n - (\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{B} \\ \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1} &= \boldsymbol{I}_n - \boldsymbol{B}(\boldsymbol{A}+\boldsymbol{B})^{-1} \end{aligned}$$

A.3.5 Inverses of matrices of the form $A + BCD$

The following result is taken from Harville (1997). Let $\boldsymbol{A}$ be a nonsingular $n \times n$ matrix, $\boldsymbol{B}$ a $n \times m$ matrix, $\boldsymbol{C}$ a nonsingular $m \times m$ matrix, and $\boldsymbol{D}$ a $m \times n$ matrix then

$$(\boldsymbol{A} + \boldsymbol{BCD})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{C}^{-1} + \boldsymbol{DA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{DA}^{-1}$$

A.3.6 Inverses of partitioned matrices

The following results are taken from Harville (1997). Define a partitioned non-singular matrix $\boldsymbol{A}$ as

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}$$

where $\boldsymbol{A}_{11}$ is a $m \times m$ matrix, $\boldsymbol{A}_{12}$ is a $m \times n$ matrix, $\boldsymbol{A}_{21}$ is a $n \times m$ matrix and $\boldsymbol{A}_{22}$ is a $n \times n$ matrix. Define the the non-singular matrix $\boldsymbol{A}^{-1}$ as

$$\boldsymbol{A}^{-1} = \begin{pmatrix} \boldsymbol{A}^{11} & \boldsymbol{A}^{12} \\ \boldsymbol{A}^{21} & \boldsymbol{A}^{22} \end{pmatrix}$$

where $\boldsymbol{A}^{11}, \boldsymbol{A}^{12}, \boldsymbol{A}^{21}, \boldsymbol{A}^{22}$ have the same dimensions as $\boldsymbol{A}_{11}, \boldsymbol{A}_{12}, \boldsymbol{A}_{21}, \boldsymbol{A}_{22}$ respectively. If $\boldsymbol{A}_{11}$ is non-singular then $\boldsymbol{A}^{11}$ is non-singular and

$$\begin{aligned}
\boldsymbol{A}^{11} &= (\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1} \\
\boldsymbol{A}^{12} &= -(\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1} \\
\boldsymbol{A}^{21} &= -\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}(\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1} \\
\boldsymbol{A}^{22} &= \boldsymbol{A}_{22}^{-1} + \boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}(\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}
\end{aligned}$$

A.3.7 $GZ^T(R+ZGZ^T)^{-1} = (G^{-1}+Z^TR^{-1}Z)^{-1}Z^TR^{-1}$

Using Result A.3.5 and Result A.3.4 we have

$$\begin{aligned}
(R+ZGZ^T)^{-1} &= R^{-1} - R^{-1}Z(G^{-1}+Z^TR^{-1}Z)^{-1}Z^TR^{-1} \\
R(R+ZGZ^T)^{-1} &= I_n - Z(G^{-1}+Z^TR^{-1}Z)^{-1}Z^TR^{-1} \\
I_n - ZGZ^T(R+ZGZ^T)^{-1} &= I_n - Z(G^{-1}+Z^TR^{-1}Z)^{-1}Z^TR^{-1} \\
ZGZ^T(R+ZGZ^T)^{-1} &= Z(G^{-1}+Z^TR^{-1}Z)^{-1}Z^TR^{-1} \\
GZ^T(R+ZGZ^T)^{-1} &= (G^{-1}+Z^TR^{-1}Z)^{-1}Z^TR^{-1}
\end{aligned}$$

as required.

A.3.8 $P = L_2(L_2^THL_2)^{-1}L_2^T = H^{-1} - H^{-1}X(X^TH^{-1}X)^{-1}X^TH^{-1}$

If $L_2^TX = 0$ and H is symmetric positive definite then it is possible to define a projection matrix P where

$$\begin{aligned}
P &= L_2(L_2^THL_2)^{-1}L_2^T \\
&= H^{-1} - H^{-1}X(X^TH^{-1}X)^{-1}X^TH^{-1}
\end{aligned}$$

Part of the proof is presented below and is taken from Searle et al. (1992). We start by noting that $L_2(L_2^TL_2)^{-1}L_2^T$ and $X(X^TX)^{-1}X^T$ are symmetric and idempotent and that it is possible to show that

$$I_n - X(X^TX)^{-1}X^T = L_2(L_2^TL_2)^{-1}L_2^T$$

Since H is symmetric positive definite a matrix $H^{\frac{1}{2}}$ exists such that $H = (H^{\frac{1}{2}})^2$ and we can write

$$I_n - H^{-\frac{1}{2}}X(X^TH^{-1}X)^{-1}X^TH^{-\frac{1}{2}} = H^{\frac{1}{2}}L_2(L_2^THL_2)^{-1}L_2^TH^{\frac{1}{2}}$$

Pre and post-multiplying both sides by $H^{-\frac{1}{2}}$ gives

$$\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1} = \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T$$

Since $\boldsymbol{R}$ and $\boldsymbol{\Sigma}$ are symmetric positive definite it follows that we can also define

$$\boldsymbol{S} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{R}^{-1} = \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T$$
$$\boldsymbol{U} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1} = \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{\Sigma}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T$$

A.3.9 $\boldsymbol{P} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}^T\boldsymbol{R}^{-1}$

From Result A.3.8 we have

$$\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}$$

Now consider

$$\begin{aligned}\boldsymbol{H}^{-1} &= \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1} \\ &= \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}\end{aligned}$$

and

$$\begin{aligned}&\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1} \\ &= (\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1})\boldsymbol{X}\boldsymbol{C}^{XX}\boldsymbol{X}^T(\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}) \\ &= \boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{C}^{XX}\boldsymbol{X}^T\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{C}^{XX}\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1} \\ &\quad -\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{C}^{XX}\boldsymbol{X}^T\boldsymbol{R}^{-1} \\ &\quad +\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{C}^{XX}\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1} \\ &= \boldsymbol{R}^{-1}\big(\boldsymbol{X}\boldsymbol{C}^{XX}\boldsymbol{X}^T - \boldsymbol{X}\boldsymbol{C}^{XX}\boldsymbol{C}_{XZ}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{Z}^T - \boldsymbol{Z}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{C}_{ZX}\boldsymbol{C}^{XX}\boldsymbol{X}^T \\ &\quad +\boldsymbol{Z}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{C}_{ZX}\boldsymbol{C}^{XX}\boldsymbol{C}_{XZ}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{Z}^T\big)\boldsymbol{R}^{-1}\end{aligned}$$

Therefore,

$$
\begin{aligned}
\boldsymbol{P} &= \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1} \\
&= \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\big(\boldsymbol{X}\boldsymbol{C}^{XX}\boldsymbol{X}^T - \boldsymbol{X}\boldsymbol{C}^{XX}\boldsymbol{C}_{XZ}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{Z}^T \\
&\quad -\boldsymbol{Z}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{C}_{ZX}\boldsymbol{C}^{XX}\boldsymbol{X}^T + \boldsymbol{Z}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{C}_{ZX}\boldsymbol{C}^{XX}\boldsymbol{C}_{XZ}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{Z}^T\big)\boldsymbol{R}^{-1} \\
&= \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\big(\boldsymbol{Z}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{Z}^T + \boldsymbol{X}\boldsymbol{C}^{XX}\boldsymbol{X}^T - \boldsymbol{X}\boldsymbol{C}^{XX}\boldsymbol{C}_{XZ}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{Z}^T \\
&\quad -\boldsymbol{Z}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{C}_{ZX}\boldsymbol{C}^{XX}\boldsymbol{X}^T + \boldsymbol{Z}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{C}_{ZX}\boldsymbol{C}^{XX}\boldsymbol{C}_{XZ}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{Z}^T\big)\boldsymbol{R}^{-1} \\
&= \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}^T\boldsymbol{R}^{-1}
\end{aligned}
$$

where $\boldsymbol{W} = (\boldsymbol{X} \;\; \boldsymbol{Z})$.

A.3.10 $\boldsymbol{P} = \boldsymbol{S} - \boldsymbol{S}\boldsymbol{Z}\boldsymbol{C}^{ZZ}\boldsymbol{Z}^T\boldsymbol{S}$

Begin by noting that $\boldsymbol{P}$ can be written as

$$
\begin{aligned}
&\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T \\
&= \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2 + \boldsymbol{L}_2\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T \\
&= \boldsymbol{L}_2\big\{(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2)^{-1} \\
&\quad -(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T\boldsymbol{Z}\{\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T\boldsymbol{Z}\}^{-1}\boldsymbol{Z}^T\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2)^{-1}\big\}\boldsymbol{L}_2^T \\
&\quad \text{using Result A.3.5} \\
&= \boldsymbol{S} - \boldsymbol{S}\boldsymbol{Z}(\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{S} \\
&= \boldsymbol{S} - \boldsymbol{S}\boldsymbol{Z}\boldsymbol{C}^{ZZ}\boldsymbol{Z}^T\boldsymbol{S}
\end{aligned}
$$

If we define $\boldsymbol{L}_2^T\boldsymbol{y}_o = \boldsymbol{y}_2$ then we can use the result presented above to re-express $\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G}$ as

$$
\begin{aligned}
\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G} &= \boldsymbol{G} - \boldsymbol{G}\left\{\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z}(\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z}\right\}\boldsymbol{G} \\
&= \boldsymbol{G} - \boldsymbol{G}\left\{\boldsymbol{I}_b - \boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z}(\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z})^{-1}\right\}\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z}\boldsymbol{G} \\
&= \boldsymbol{G} - \boldsymbol{G}\left[\boldsymbol{I}_b - \left\{\boldsymbol{I}_b - \boldsymbol{G}^{-1}(\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z})^{-1}\right\}\right]\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z}\boldsymbol{G} \\
&\quad \text{using Result A.3.4} \\
&= \boldsymbol{G} - (\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z}\boldsymbol{G} \\
&= \left\{\boldsymbol{I}_b - (\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z}\right\}\boldsymbol{G} \\
&= \left[\boldsymbol{I}_b - \left\{\boldsymbol{I}_b - (\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z})^{-1}\boldsymbol{G}^{-1}\right\}\right]\boldsymbol{G} \\
&\quad \text{using Result A.3.4} \\
&= (\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z})^{-1} \\
&= \boldsymbol{C}^{ZZ}
\end{aligned}
$$

A.3.11 $\boldsymbol{PHP} = \boldsymbol{P}$

Note that $\boldsymbol{P} = \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T$. Therefore,

$$
\begin{aligned}
\boldsymbol{P}\boldsymbol{H}\boldsymbol{P} &= \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T \\
&= \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T \\
&= \boldsymbol{P}
\end{aligned}
$$

as required. Similarly $\boldsymbol{S}\boldsymbol{R}\boldsymbol{S} = \boldsymbol{S}$ and $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U} = \boldsymbol{U}$.

A.3.12 $\boldsymbol{PHS} = \boldsymbol{S}$

Noting that

$$
\begin{aligned}
\boldsymbol{P} &= \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T \\
\boldsymbol{S} &= \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T
\end{aligned}
$$

then

$$\begin{aligned} \boldsymbol{PHS} &= \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T \\ &= \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T \\ &= \boldsymbol{S} \end{aligned}$$

as required.

A.3.13 $\boldsymbol{PRS} = \boldsymbol{P}$

Noting that

$$\begin{aligned} \boldsymbol{P} &= \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T \\ \boldsymbol{S} &= \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T \end{aligned}$$

then

$$\begin{aligned} \boldsymbol{PRS} &= \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{R}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T \\ &= \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T \\ &= \boldsymbol{P} \end{aligned}$$

as required.

A.3.14 $\mathrm{tr}(\boldsymbol{HP}) = n - t$

If we define $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}$ then

$$\begin{aligned} \mathrm{tr}(\boldsymbol{HP}) &= \mathrm{tr}\big[\boldsymbol{H}\{\boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\}\big] \\ &= \mathrm{tr}\big\{\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\big\} \\ &= \mathrm{tr}(\boldsymbol{I}_n) - \mathrm{tr}(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}) \\ &= n - \mathrm{tr}\big\{(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}\big\} \\ &= n - \mathrm{tr}(\boldsymbol{I}_t) \\ &= n - t \end{aligned}$$

as required. Similarly $\text{tr}(\boldsymbol{RS}) = \text{tr}(\boldsymbol{\Sigma U}) = n - t$.

A.4 Matrix differentiation results

A.4.1 Derivatives of determinants

The following result is taken from Harville (1997). If $\boldsymbol{A}$ is a $p \times p$ matrix where $\boldsymbol{A} = \boldsymbol{A}(x) = \{b_{ij}\}$ is a non-singular matrix whose elements are functions of x and is continuously differentiable at every x then

$$\frac{\partial \log\{\det(\boldsymbol{A})\}}{\partial x_j} = \text{tr}\left(\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x_j}\right)$$

A.4.2 Derivatives of inverses

The following result is taken from Harville (1997). If $\boldsymbol{A}$ is a $p \times p$ matrix where $\boldsymbol{A} = \boldsymbol{A}(x) = \{b_{ij}\}$ is a non-singular matrix whose elements are functions of x and is continuously differentiable at every x then

$$\frac{\partial \boldsymbol{A}^{-1}}{\partial x_j} = -\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x_j}\boldsymbol{A}^{-1}$$

A.4.3 Derivatives of $\boldsymbol{P}$

Define a vector of variance parameters $\boldsymbol{\kappa} = (\sigma^2, \boldsymbol{\phi}^T, \boldsymbol{\gamma}^T)^T$ and consider the partial derivative of $\boldsymbol{P} = \boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T$ (Result A.3.8) with respect to κ_i, i.e.,

$$\begin{aligned}
\frac{\partial \boldsymbol{P}}{\partial \kappa_i} &= \frac{\partial}{\partial \kappa_i}\left\{\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T\right\} \\
&= \boldsymbol{L}_2\frac{\partial(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}}{\partial \kappa_i}\boldsymbol{L}_2^T \\
&= -\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T\frac{\partial \boldsymbol{H}}{\partial \kappa_i}\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T \qquad \text{using Result A.4.2} \\
&= -\boldsymbol{P}\frac{\partial \boldsymbol{H}}{\partial \kappa_i}\boldsymbol{P}
\end{aligned}$$

Likewise,

$$\begin{aligned}
\frac{\partial \boldsymbol{U}}{\partial \phi_{ij}} &= \frac{\partial}{\partial \phi_{ij}}\left\{\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{\Sigma}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T\right\} \\
&= \boldsymbol{L}_2\frac{\partial(\boldsymbol{L}_2^T\boldsymbol{\Sigma}\boldsymbol{L}_2)^{-1}}{\partial \phi_{ij}}\boldsymbol{L}_2^T \\
&= -\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{\Sigma}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T\frac{\partial \boldsymbol{\Sigma}}{\partial \phi_{ij}}\boldsymbol{L}_2(\boldsymbol{L}_2^T\boldsymbol{\Sigma}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2^T \qquad \text{using Result A.4.2} \\
&= -\boldsymbol{U}\frac{\partial \boldsymbol{\Sigma}}{\partial \phi_{ij}}\boldsymbol{U}
\end{aligned}$$

Appendix B

Derivation of identity in Oakes (1999)

A key result in Oakes (1999) is the identity

$$E\left[\frac{\partial^2 \log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}}{\partial\boldsymbol{\theta}^{(w)}\partial\boldsymbol{\theta}^{(w)T}}\middle|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)}\right]$$
$$= -E\left(\frac{\partial}{\partial\boldsymbol{\theta}^{(w)}}\left[\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}\right]\frac{\partial}{\partial\boldsymbol{\theta}^{(w)T}}\left[\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}\right]\middle|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)}\right)$$

Noting that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^T$, defining $i = 1, \ldots, d$ and $j = 1, \ldots, d$ we can write

$$\begin{aligned}\frac{\partial^2 \log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}}{\partial\theta_i^{(w)}\partial\theta_j^{(w)}} &= \frac{\partial}{\partial\theta_i^{(w)}}\left[\frac{\partial \log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}}{\partial\theta_j^{(w)}}\right] \\ &= \frac{\partial}{\partial\theta_i^{(w)}}\left[\frac{\frac{\partial}{\partial\theta_j^{(w)}}\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}}{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})}\right]\end{aligned}$$

Using the quotient rule we have

$$\frac{\partial^2 \log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}}{\partial\theta_i^{(w)}\partial\theta_j^{(w)}}$$
$$= \frac{\frac{\partial^2}{\partial\theta_i^{(w)}\partial\theta_j^{(w)}}\left\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\right\}k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)}) - \frac{\partial}{\partial\theta_j^{(w)}}\left\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\right\}\frac{\partial}{\partial\theta_i^{(w)}}\left\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\right\}}{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})^2}$$
$$= \frac{\frac{\partial^2}{\partial\theta_i^{(w)}\partial\theta_j^{(w)}}\left\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\right\}}{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})} - \frac{\partial}{\partial\theta_j^{(w)}}\left[\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}\right]\frac{\partial}{\partial\theta_i^{(w)}}\left[\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}\right]$$

Which can be written using matrix notation as

$$\frac{\partial^2 \log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}}{\partial\boldsymbol{\theta}^{(w)}\partial\boldsymbol{\theta}^{(w)T}}$$
$$= \frac{\frac{\partial^2}{\partial\boldsymbol{\theta}^{(w)}\partial\boldsymbol{\theta}^{(w)T}}\left\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\right\}}{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})} - \frac{\partial}{\partial\boldsymbol{\theta}^{(w)}}\left[\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}\right]\frac{\partial}{\partial\boldsymbol{\theta}^{(w)T}}\left[\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}\right]$$

Taking expectations on both sides over $k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})$ we have

$$E\left[\frac{\partial^2 \log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}}{\partial\boldsymbol{\theta}^{(w)}\partial\boldsymbol{\theta}^{(w)T}}\middle|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)}\right]$$
$$= E\left[\frac{\frac{\partial^2}{\partial\boldsymbol{\theta}^{(w)}\partial\boldsymbol{\theta}^{(w)T}}\left\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\right\}}{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})}\middle|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)}\right]$$
$$-E\left(\frac{\partial}{\partial\boldsymbol{\theta}^{(w)}}\left[\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}\right]\frac{\partial}{\partial\boldsymbol{\theta}^{(w)T}}\left[\log\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\}\right]\middle|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)}\right)$$

The required identity is obtained if the first term on the right hand side is zero. We have

$$
\begin{aligned}
& E\left[\left.\frac{\frac{\partial^2}{\partial\boldsymbol{\theta}^{(w)}\partial\boldsymbol{\theta}^{(w)^T}}\left\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\right\}}{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})}\right|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)}\right] \\
&= \int_{S(\boldsymbol{y}_m)} \frac{\frac{\partial^2}{\partial\boldsymbol{\theta}^{(w)}\partial\boldsymbol{\theta}^{(w)^T}}\left\{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})\right\}}{k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})} k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})d\boldsymbol{y}_m \\
&= \frac{\partial^2}{\partial\boldsymbol{\theta}^{(w)}\partial\boldsymbol{\theta}^{(w)^T}}\left\{\int_{S(\boldsymbol{y}_m)} k(\boldsymbol{y}_m|\boldsymbol{y}_o;\boldsymbol{\theta}^{(w)})d\boldsymbol{y}_m\right\} \\
&= \frac{\partial^2}{\partial\boldsymbol{\theta}^{(w)}\partial\boldsymbol{\theta}^{(w)^T}}\left(\mathbf{1}\right) \\
&= \mathbf{0}
\end{aligned}
$$

as required.

Appendix C

Loewner partial ordering of symmetric positive semidefinite matrices

Loewner (sometimes referred to as Löwner) partial ordering of matrices is analagous to ordering numbers on the real number line using the inequalities $>$, $\geq$, $<$, and $\leq$. Loewner partial ordering has its' mathematical foundations in convex geometry. Much of what follows has been taken from Horn and Johnson (1990)

We begin by defining $\mathbb{S}_+^n$ as the convex cone containing the set of all positive semidefinite symmetric $n \times n$ matrices and the interior of $\mathbb{S}_+^n$, denoted $\mathbb{S}_{++}^n$, as the set of all positive definite symmetric $n \times n$ matrices. For $\boldsymbol{A} \in \mathbb{S}_+^n$ we can write

$$\boldsymbol{A} \succeq 0 \iff \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \geq 0 \qquad \forall \boldsymbol{x} \in \mathbb{R}^n$$

where $\mathbb{R}^n$ refers to Euclidean n-dimensional real vector space and the symbols $\in$, $\forall$ and $\iff$ can be read as "in", "for all" and "material equivalence" respectively. Later we introduce the symbol $\implies$ which can be read as "material implication". For $\boldsymbol{A} \in \mathbb{S}_{++}^n$ we can write

$$\boldsymbol{A} \succ 0 \iff \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} > 0 \qquad \forall \boldsymbol{x} \in \mathbb{R}^n$$

If $\boldsymbol{A} - \boldsymbol{B} \succeq 0$ then $\boldsymbol{A} \succeq \boldsymbol{B}$ where $\boldsymbol{A} \succeq \boldsymbol{B}$ defines a so-called Loewner partial ordering (see for example Jarre (2000) or Dattorro (2005) Appendix A). The corollaries we

are particularly interested in are (Horn and Johnson (1990) Chapter 7)

Corollary C.1. *If* $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{S}^n_{++}$, *then*
(a) $\boldsymbol{A} \succeq \boldsymbol{B} \iff \boldsymbol{A}^{-1} \preceq \boldsymbol{B}^{-1}$;
(b) $\boldsymbol{A} \succeq \boldsymbol{B} \implies \det(\boldsymbol{A}) \geq \det(\boldsymbol{B})$; *and*
(c) $\boldsymbol{A} \succeq \boldsymbol{B} \implies \boldsymbol{\lambda}_k(\boldsymbol{A}) \geq \boldsymbol{\lambda}_k(\boldsymbol{B})$ *for all* $k = 1, 2, \ldots, n$ *if the respective eigenvalues of* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *are arranged in the same (increasing or decreasing) order.*

Note that Corollary C.1 (b) and (c) imply that for $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{S}^n_{++}$ then

$$\det(\boldsymbol{A}) \geq \det(\boldsymbol{B}) \iff \lambda_k(\boldsymbol{A}) \geq \lambda_k(\boldsymbol{B}) \qquad \text{for all } k = 1, 2, \ldots, n$$

According to Dattorro (2005) the requirement that $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive definite is only required for Corollary C.1 (a) whereas $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{S}^N_+$ is sufficient for Corollary C.1 (b) and (c).

Bibliography

Butler, D., Cullis, B., Gilmour, A., and Gogel, B. (2007). *ASReml-R reference manual.* ISSN 0812-0005.

Callanan, T. P. and Harville, D. A. (1991). Some new algorithms for computing restricted maximum likelihood estimates of variance components. *Journal of Statistical Computation and Simulation*, 38:239–259.

Cullis, B., Smith, A., and Thompson, R. (2004). *Perspectives of ANOVA, REML, and a general linear mixed model.* Methods and Models in Statistics - In honour of Professor John Nelder, FRS, Editors N. Adams, M. Crowder, D. J. Hand and D. Stephens. Imperial College Press, London.

Dattorro, J. (2005). *Convex Optimization and Euclidean Distance Geometry.* Meboo Publishing. v2009.09.22.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Foulley, J.-L., Jaffrezic, F., and Robert-Cranie, C. (2000). EM-REML estimation of covariance parameters in Gaussian mixed models for longitudinal data analysis. *Genetics Selection Evolution*, 32:129–141.

Foulley, J.-L. and Van Dyk, D. (2000). The PX-EM algorithm for fast stable fitting of Henderson's mixed model. *Genetics Selection Evolution*, 32:129–141.

Gilmour, A., Thompson, R., and Cullis, B. (1995). AI, an efficient algorithm for REML estimation in linear mixed models. *Biometrics*, 51:1440–1450.

Harville, D. (1997). *Matrix algebra from a statistician's perspective.* Springer-Verlag, New York, Inc.

Henderson, C. (1953). Estimation of variance and covariance components. *Biometrics*, 9(2):226–252.

Henderson, C., Kempthorne, O., Searle, S., and von Krosigk, C. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15(2):192–218.

Horn, R. and Johnson, C. (1990). *Matrix analysis.* Cambridge University Press.

Jarre, F. (2000). *Convex analysis on symmetric matrices*, chapter 2. The geometry of semidefinite programming, Editors H. Wolkowicz, R. Saigal and L. Vandenberghe. Kluwer.

Kalbfleisch, J. and Sprott, A. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society, Series B*, 32(2):175–208.

Knight, E. (2008). *Improved iterative schemes for REML estimation of variance parameters in linear mixed models.* PhD thesis, The University of Adelaide, School of Agriculture, Food and Wine.

Laird, N. and Ware, J. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, 38:963–974.

Liu, C., Rubin, D., and Wu, Y. (1998). Parameter expansion to accelerated EM: The PX-EM algorithm. *Biometrika*, 85:755–770.

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, B*, 44:226–233.

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate analysis.* Academic Press (London) Limited.

McLachlan, G. and Krishnan, K. (1997). *The EM Algorithm and Extensions.* John-Wiley and Sons Inc.

Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society, B*, 61:479–482.

Orchard, T. and Woodbury, M. (1972). A missing information principle: theory an applications. In *Proceedings of the 6th Berkely Symposium on Mathematical Statistics and Probability*, volume 1, pages 697–715, Berkeley, California. University of California Press.

Patterson, H. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Searle, S., Casella, G., and McCulloch, C. (1992). *Variance Components*. John-Wiley and Sons Inc.

Smith, A., Cullis, B., and Thompson, R. (2001). Analyzing variety by environment data using multiplicative mixed models and adjustment for spatial field trend. *Biometrics*, 57(4):1138–1147.

Thompson, R., Cullis, B., Smith, A., and Gilmour, A. (2003). A sparse implementation of the average information algorithm for factor analytic and reduced rank variance models. *Australian and New Zealand Journal of Statistics*, 45:445–459.

Verbyla, A. (1990). A conditional derivation of residual maximum likelihood. *Australian Journal of Statistics*, 32(2):227–230.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, 20:595.