Linear Algebra and its Applications 435 (2011) 1920–1944

Contents lists available at ScienceDirect

Linear Algebra and its Applications

journal homepage: www.elsevier.com/locate/laa

Parameter estimation and inference in the linear mixed model

F.N. Gumedze [*], T.T. Dunne

Department of Statistical Sciences, University of Cape Town, Private Bag Rondebosch 7701, South Africa

ARTICLE INFO

Article history:
Received 5 August 2010
Accepted 7 April 2011
Available online 14 May 2011

Submitted by H. Schneider

AMS classification:
Primary: 62J99
Secondary: 62J99

Keywords:
Average information algorithm
Information matrix
Linear mixed model
Likelihood ratio test
REML
Score test
Wald test

ABSTRACT

The paper reviews the linear mixed model with a focus on parameter estimation and inference. Parameter estimation for the different components of the model are reviewed, with an emphasis on variance parameter estimation. Inferential procedures for the fixed effects, random effects or a combination of both fixed and random effects are also discussed.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

The paper is structured as follows. First, we introduce the standard linear mixed model and its assumptions. This introduction is followed by a discussion of parameter estimation and inferential procedures for the various components of the model; the fixed effects parameters, random effects, variance parameters (or ratios). We also discuss inferential procedures for the estimated fixed effects and the variance parameter estimates. An illustration is given using a real data set.

[*] Corresponding author. Tel.: +27 (0)21 6504783; fax: +27 (0)21 6504773.
E-mail address: freedom.gumedze@uct.ac.za (F.N. Gumedze).

0024-3795/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.laa.2011.04.015

2. The linear mixed model

Linear mixed models provide a powerful and flexible tool for the analysis of a broad variety of data including clustered data such as longitudinal data, repeated measures, blocked or multilevel data [11,15,36,52,69], spatial and geostatistics [17,66], and bioinformatics data [57,63].

The linear mixed model is given by

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{y}$ is a $n \times 1$ vector of responses, $\boldsymbol{X}$ is an $n \times p$ known design matrix for the fixed effects, $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector of fixed effects, $\boldsymbol{Z} = [\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_b]$, where $\boldsymbol{Z}_i$ is an $n \times q_i$ design matrix for the ith random effects factor, $\boldsymbol{u} = [\boldsymbol{u}_1', \ldots, \boldsymbol{u}_b']'$ is a $q \times 1$ vector of random effects where $\boldsymbol{u}_i$ is a $q_i \times 1$ vector such that $q = \sum_{i=1}^{b} q_i$, and $\boldsymbol{e}$ is an $n \times 1$ vector of random errors, with $\mathrm{E}(\boldsymbol{u}) = \boldsymbol{0}$ and $\mathrm{E}(\boldsymbol{e}) = \boldsymbol{0}$. In addition it is assumed that $\boldsymbol{u}$ and $\boldsymbol{e}$ follow independent and multivariate Gaussian distributions such that

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{e} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \ \sigma^2 \begin{bmatrix} \boldsymbol{G}(\boldsymbol{\gamma}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}(\boldsymbol{\rho}) \end{bmatrix} \right), \tag{2}$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\rho}$ are $r \times 1$ and $s \times 1$ (with $s \leqslant n(n+1)/2$) vectors of unknown variance parameters corresponding to $\boldsymbol{u}$ and $\boldsymbol{e}$, respectively. If the random terms are correlated then the dimension of $\boldsymbol{\gamma}$ may exceed q, i.e. $\boldsymbol{\gamma}$ may be of dimension $r \leqslant q(q+1)/2$. Following Patterson and Thompson [50] we write the variance–covariance matrix of the data, $\boldsymbol{y}$, as

$$\mathrm{var}(\boldsymbol{y}) = \sigma^2(\boldsymbol{ZGZ}' + \boldsymbol{R}) = \sigma^2\boldsymbol{H}, \tag{3}$$

where

$$\boldsymbol{H} = \boldsymbol{ZGZ}' + \boldsymbol{R}. \tag{4}$$

The appeal of the parameterization (3), i.e. the factoring of the residual variance σ^2 out of the variance matrix for the data, is that it reduces the t-dimensional REML log-likelihood maximization problem by unity [6], where $t = (r + s + 1)$ is the number of variance parameters in model (1). This variance matrix parameterization can also often be useful for establishing overall scaling. However, it may not be useful in multivariate analysis of variance (MANOVA) problems where σ^2 has no meaningful interpretation. An alternative parameterization is when the model (1) is parameterized in terms of the variance components. For instance assuming $\boldsymbol{R} = \boldsymbol{I}$, the variance matrix is written as

$$\mathrm{var}(\boldsymbol{y}) = \boldsymbol{V} = \boldsymbol{ZG}^{v}\boldsymbol{Z}' + \sigma^2\boldsymbol{I}, \tag{5}$$

where $\boldsymbol{G}^{v}$ contains the variance components for each random effect factor and σ^2 is the residual error variance.

The matrix $\boldsymbol{H}$ consists of two components that are used to model heteroscedasticity and correlation: a random effects component $\boldsymbol{ZGZ}'$ and a within-group component $\boldsymbol{R}$. In some applications, the within-group component $\boldsymbol{R}$ is used to directly model the variance–covariance matrix of the data without the need to incorporate random effects in the model to account for dependence among observations.

3. Joint estimation of fixed and random effects

Once the model has been formulated, methods are needed to estimate the model parameters. In this section we first deal with the joint estimation of the fixed effects ($\boldsymbol{\beta}$) and random effects ($\boldsymbol{u}$) and then with estimation of the variance parameters ($\boldsymbol{\gamma}$, $\boldsymbol{\rho}$ and σ^2). There are many methods for obtaining the estimates of the fixed and random effects simultaneously [62, Section 7.4c, 56]. These methods include Henderson's mixed model equations [25], Goldberger's [20] approach of predicting a future observation, techniques based on two-stage regression, linearity in $\boldsymbol{y}$, partitioning of $\boldsymbol{y}$ and Bayes estimation. In this section we describe estimation using Henderson's mixed model equations because it produces sampling variances for the estimators and because it has a connection with maximum likelihood estimation of the variance parameters.

Henderson [25] (also see [27]) assumed $\boldsymbol{u}$ and $\boldsymbol{y}$ to be jointly Gaussian distributed as

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{y} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{X\beta} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{G} & \boldsymbol{GZ}' \\ \boldsymbol{ZG} & \boldsymbol{H} \end{bmatrix}\right) \tag{6}$$

Thus $\boldsymbol{y}$ has the marginal probability density function $N[\boldsymbol{X\beta}, \sigma^2\boldsymbol{H}]$, where $\boldsymbol{H}$ is as defined in (4) with $\boldsymbol{G}$ and $\boldsymbol{R}$ assumed known. Henderson [25] maximized the log-joint distribution of $(\boldsymbol{y}, \boldsymbol{u})$ to obtain estimators of $\boldsymbol{\beta}$ and $\boldsymbol{u}$. However, this logarithmic function is not a log-likelihood function as $\boldsymbol{u}$ is not observed. The marginal distribution is $\boldsymbol{u}$ from (6) is

$$\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{G})$$

and the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{u}$ is

$$\boldsymbol{y}|\boldsymbol{u} \sim N(\boldsymbol{X\beta} + \boldsymbol{Zu}, \sigma^2\boldsymbol{R}).$$

Hence the log-joint distribution of $(\boldsymbol{y}, \boldsymbol{u})$ is given by

$$\begin{aligned} \log f(\boldsymbol{y}, \boldsymbol{u}) &= \log f(\boldsymbol{y}|\boldsymbol{u}) + \log f(\boldsymbol{u}) \\ &= -\frac{1}{2}\left\{n \log \sigma^2 + \log \boldsymbol{R} + (\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})/\sigma^2\right\} \\ &\quad - \frac{1}{2}\left\{q \log \sigma^2 + \log \boldsymbol{G} + \boldsymbol{u}'\boldsymbol{G}^{-1}\boldsymbol{u}/\sigma^2\right\} \\ &= -\frac{1}{2}\left\{(n+q) \log \sigma^2 + \log \boldsymbol{R} + \log \boldsymbol{G} + (\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})/\sigma^2\right\} \\ &\quad - \frac{1}{2\sigma^2}\left\{\boldsymbol{u}'(\boldsymbol{ZR}^{-1}\boldsymbol{Z}' + \boldsymbol{G}^{-1})\boldsymbol{u} - 2(\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{R}^{-1}\boldsymbol{Zu}\right\}. \end{aligned}$$

This function coincides with the h-likelihood function of Lee and Nelder [38] for correlated Gaussian data, with the random effects also having a Gaussian distribution (i.e. linear mixed model). However, Lee and Nelder's [38] approach can also handle correlated non-Gaussian data with conjugate distributions assumed for the random effects.

Estimates for $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are obtained by solving the score equations

$$\begin{aligned} &\boldsymbol{X}'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) - \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\tilde{\boldsymbol{u}} = \boldsymbol{0}, \\ &\boldsymbol{Z}'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) - (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})\tilde{\boldsymbol{u}} = \boldsymbol{0}. \end{aligned}$$

These equations are called the mixed model equations (MMEs) as proposed by Henderson [25] and Henderson et al. [27]. They wrote the equations compactly in matrix form as

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix}. \tag{7}$$

Gilmour et al. [19] rewrote the mixed model equations (7) as

$$\boldsymbol{C\psi} = \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{y}, \tag{8}$$

where $\boldsymbol{W} = [\boldsymbol{X}\,\boldsymbol{Z}]$, $\boldsymbol{\psi} = (\boldsymbol{\beta}', \boldsymbol{u}')'$ and

$$\boldsymbol{C} = \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{W} + \boldsymbol{G}^{*+}$$

with

$$\boldsymbol{G}^{*} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{G} \end{bmatrix} \quad \text{and} \quad \boldsymbol{G}^{*+} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{G}^{-1} \end{bmatrix},$$

where the superscript '+' denotes the Moore–Penrose inverse.

For the model (1) we have $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$ and $\mathrm{var}(\boldsymbol{y}) = \sigma^2\boldsymbol{H}$. Assuming $\boldsymbol{H}$ is known, the fixed effects parameters $\boldsymbol{\beta}$ can be estimated by GLS to obtain

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{y}, \tag{9}$$

which is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$. If $\boldsymbol{X}$ is not full rank, then any generalized inverse $(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-}$ is used instead of $(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}$ to obtain a solution for $\boldsymbol{\beta}$. The resulting solution for $\boldsymbol{\beta}$ is not unique and is no longer unbiased. However, $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is unique and unbiased for $\boldsymbol{X}\boldsymbol{\beta}$.

The computational challenge of using GLS to estimate $\boldsymbol{\beta}$ is that it requires the inverse of $\boldsymbol{H}$ which is an $n \times n$ matrix. In contrast the joint estimators for $\boldsymbol{\beta}$ and $\boldsymbol{u}$ can be obtained by solving either (7) or (8), i.e.

$$\tilde{\boldsymbol{\psi}} = \boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{y}, \tag{10}$$

where $\tilde{\boldsymbol{\psi}} = (\hat{\boldsymbol{\beta}}', \tilde{\boldsymbol{u}}')'$ and $\boldsymbol{C}^{-1}$ is given in Lemma 4 of Appendix A. It must be noted that (10) requires simply the inversion of $\boldsymbol{C}$, a $(p+q) \times (p+q)$ matrix, which is easier than finding the inverse of $\boldsymbol{H}$. We also note that although $\boldsymbol{R}^{-1}$ in (10) is also an $n \times n$ matrix, it usually has a structure that can be exploited (for example independence between subjects) which makes its computation easier.

Lemma 1. *The solutions for* $\boldsymbol{\beta}$ *and* $\boldsymbol{u}$ *from solving the MMEs, for* $\boldsymbol{G}$ *and* $\boldsymbol{R}$ *known, are given by*

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{y} \tag{11}$$

$$\tilde{\boldsymbol{u}} = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}), \tag{12}$$

with corresponding variance matrices

$$\begin{aligned} var(\hat{\boldsymbol{\beta}}) &= \sigma^2[(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}] \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1} \end{aligned} \tag{13}$$

and

$$\begin{aligned} var(\tilde{\boldsymbol{u}}) &= \sigma^2\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{H}\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G} \\ &= \sigma^2\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G}, \end{aligned} \tag{14}$$

respectively, where $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}$.

We also have that

$$var(\tilde{\boldsymbol{u}} - \boldsymbol{u}) = \sigma^2\boldsymbol{G} - var(\tilde{\boldsymbol{u}}), \tag{15}$$

which unlike (14) takes into account the variability of $\boldsymbol{u}$ *and can therefore be useful for constructing confidence intervals for* $\boldsymbol{u}$*.*

Proof. The proof of the lemma follows from the MMEs (7) and the matrix results given in Appendix A (Lemma B.4). □

The predictor $\tilde{\boldsymbol{u}}$ is known as the best linear unbiased predictor (BLUP). It can also be viewed as the estimator of the conditional mean of $\boldsymbol{u}$ given $\boldsymbol{y}$. Applying Result 6 directly to (6) gives

$$\boldsymbol{u}|\boldsymbol{y} \sim N\left[\boldsymbol{0} + \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \sigma^2(\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G})\right].$$

Thus

$$\mathrm{E}(\boldsymbol{u}|\boldsymbol{y}) = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

and

$$\mathrm{var}(\boldsymbol{u}|\boldsymbol{y}) = \sigma^2[\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G}],$$

which can be rewritten as

$$\begin{aligned}\mathrm{var}(\boldsymbol{u}|\boldsymbol{y}) &= \sigma^2[\boldsymbol{G} - (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{G}]\\ &= \sigma^2[\boldsymbol{G} - (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} - \boldsymbol{G}^{-1})\boldsymbol{G}]\\ &= \sigma^2(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}.\end{aligned}$$

Though $\boldsymbol{u}$ is unobserved, (15) implies the reduced variation associated with the recovery of some information about $\boldsymbol{u}$ in $\tilde{\boldsymbol{u}}$.

The estimator $\tilde{\boldsymbol{u}}$ is also referred to as the Empirical Bayes estimator for $\boldsymbol{u}$. This label is justified by recognizing the random effects $\boldsymbol{u}$ as random variables and therefore the likelihood function $l(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\kappa}, \sigma^2; \boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{u})p(\boldsymbol{u})$ corresponds to a complete density function, so that $p(\boldsymbol{u})$ is interpretable as the prior distribution of $\boldsymbol{u}$ and hence under the Gaussian assumptions of Result A.6 the posterior distribution of $\boldsymbol{u}|\boldsymbol{y}$ is Gaussian with mean $\tilde{\boldsymbol{u}}$ and variance $\sigma^2(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}$ [45].

The expressions in Lemma 1 assume that the variance parameters are known, but if estimates of the variance parameters $\boldsymbol{\gamma}$, $\boldsymbol{\rho}$ and σ^2 are not known, $\boldsymbol{G}$, $\boldsymbol{R}$ and σ^2 can be replaced by the estimates $\hat{\boldsymbol{G}}$, $\hat{\boldsymbol{R}}$ and $\hat{\sigma}^2$ to obtain the estimates of the fixed effects and random effects and their standard errors using the expressions in Lemma 1. However, such standard errors of the fixed effects and of the random effects do not take into account the variability introduced by estimating $\boldsymbol{\gamma}$, $\boldsymbol{\rho}$ and σ^2, and so underestimate the variability of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}$. In the following section we discuss methods for estimating the variance parameters $\boldsymbol{\gamma}$, $\boldsymbol{\rho}$ and σ^2.

4. Variance parameter estimation

Several methods for variance parameter estimation in linear mixed models are discussed in Searle et al. [62, Chapters 5 and 11]. These methods include the ANOVA method for balanced data which uses the expected mean squares approach. However, this method is difficult to apply when the data are unbalanced or when we wish to model the variation in the data using a more complex variance structure. Searle [59,60] give a general discussion of the problems associated with estimating variance parameters using ANOVA methods in unbalanced data.

For unbalanced data, Rao [54] proposed the minimum norm quadratic estimation (MINQUE) method for estimating the variance parameters, so-named because it produces quadratic unbiased estimators which have the minimum norm (MINQUE) property, i.e. the resulting estimates are translation invariant under unbiased quadratic forms of the observations. Earlier Henderson [26] had proposed three methods for estimating variance parameters known as Henderson's methods I, II and III. Method I uses quadratic forms which are analogous to the sums of squares of generally balanced designs; Method II is an adaptation of Method I and takes account of the fixed effects in the model; Method III (also called fitting constants (FITCON) method, see [60]) uses sums of squares from fitting the full mixed models as though all terms were fixed effects. A detailed account of Henderson's methods is given by Searle et al. [62, Section 5.3]. Lee and Nelder [39] give another way of estimating variance using extended quasi-likelihood, i.e. using gamma-log generalized linear models.

Maximum likelihood (ML) and Residual Maximum Likelihood (REML), also known as restricted maximum likelihood, are now standard methods for estimating variance parameters for both balanced and unbalanced data. The main attraction of these methods is that they can handle a much wider class of variance models than simple variance components. ML estimators of the variance parameters (ratios) are biased downwards, especially in small samples, because they do not take into account the degrees of freedom lost in the estimation of the fixed effects [41,67]. Hence REML estimation of the variance parameters (or ratios) is preferable to ML estimation. ML estimation of the variance parameters (ratios) have been discussed by several researchers (e.g. [22,29,42,62,69]). Below, we describe both ML and REML estimation for variance parameters (or ratios) in linear mixed models.

4.1. Maximum likelihood

The marginal distribution of $\boldsymbol{y}$ in the linear mixed model is given by $N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{H})$ and hence the marginal log-likelihood function of $\boldsymbol{y}$ is [22]

$$l_{\mathrm{ML}}(\boldsymbol{\beta}, \boldsymbol{\phi}; \boldsymbol{y}) = -\frac{1}{2}\left\{n\log(2\pi) + n\log\sigma^2 + \log|\boldsymbol{H}| + \frac{(\boldsymbol{y}-\boldsymbol{X\beta})'\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})}{\sigma^2}\right\}, \tag{16}$$

where $\boldsymbol{\phi} = (\boldsymbol{\kappa}', \sigma^2)'$, $\boldsymbol{\kappa} = (\boldsymbol{\gamma}\prime, \boldsymbol{\rho}')'$ and the subscript ML denotes marginal log-likelihood. In the following the abbreviation ML is used interchangeably to refer to the marginal log-likelihood or maximum likelihood estimation.

We illustrate, albeit briefly, maximum likelihood estimation of the variance components. Differentiating the marginal log-likelihood function with respect to $\boldsymbol{\beta}$, σ^2 and $\kappa_j, j = 1, \ldots, r+s$ yields the partial derivatives

$$\frac{\partial l_{\mathrm{ML}}(\boldsymbol{\beta}, \boldsymbol{\phi}; \boldsymbol{y})}{\partial\boldsymbol{\beta}} = -\frac{1}{\sigma^2}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X\beta} - \boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{y}), \tag{17a}$$

$$\frac{\partial l_{\mathrm{ML}}(\boldsymbol{\beta}, \boldsymbol{\phi}; \boldsymbol{y})}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{(\boldsymbol{y}-\boldsymbol{X\beta})'\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})}{2\sigma^4}, \tag{17b}$$

$$\frac{\partial l_{\mathrm{ML}}(\boldsymbol{\beta}, \boldsymbol{\phi}; \boldsymbol{y})}{\partial\kappa_j} = -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_j\right) + \frac{(\boldsymbol{y}-\boldsymbol{X\beta})'\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_j\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})}{2\sigma^2}, \tag{17c}$$

where $\dot{\boldsymbol{H}}_j = \partial\boldsymbol{H}/\partial\kappa_j$.

Setting the Eqs. (17a)–(17c) equal to zero gives

$$\boldsymbol{X}'\hat{\boldsymbol{H}}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\hat{\boldsymbol{H}}^{-1}\boldsymbol{y}, \tag{18a}$$

$$n\hat{\sigma}^2 = (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'\hat{\boldsymbol{H}}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}), \tag{18b}$$

$$\mathrm{tr}\left(\hat{\boldsymbol{H}}^{-1}\hat{\dot{\boldsymbol{H}}}_j\right) = \frac{1}{\hat{\sigma}^2}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'\hat{\boldsymbol{H}}^{-1}\hat{\dot{\boldsymbol{H}}}_j\hat{\boldsymbol{H}}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}), \tag{18c}$$

where $\hat{\boldsymbol{H}}$ and $\hat{\dot{\boldsymbol{H}}}_j$ involve the MLE's of $\kappa_j, j = 1, \ldots, r+s$, rather than known κ_j, The number of variance parameters in the model including σ^2 is $t = r + s + 1$.

Solving the Eqs. (18a) and (18b) yields the maximum likelihood estimators

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\hat{\boldsymbol{H}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{H}}^{-1}\boldsymbol{y}, \tag{19a}$$

$$\hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'\hat{\boldsymbol{H}}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{19b}$$

The generalized least squares estimator (19a) is equivalent to the estimator for $\boldsymbol{\beta}$ given in Lemma 1. Solutions for κ_j's must be found by solving (18c) which depends on $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$, and therefore the equations must be solved iteratively as follows

Step 0. Assign initial values to $\boldsymbol{\kappa}_{(0)} = (\kappa_1, \ldots, \kappa_{r+s})'$.
Step 1. At each mth iteration substitute $\hat{\boldsymbol{\kappa}}_{(m-1)}$ in (19a) and (19b) and solve for $\hat{\boldsymbol{\beta}}_{(m)}$ and $\hat{\sigma}^2_{(m)}$.
Step 2. Use the results from steps 0 and 1, i.e. $\hat{\boldsymbol{\kappa}}_{(m-1)}$, $\hat{\boldsymbol{\beta}}_{(m)}$ and $\hat{\sigma}^2_{(m)}$ by substitution in (17c) to calculate new $\hat{\boldsymbol{\kappa}}_{(m)}$ that make $\partial l_{\mathrm{ML}}/\partial\kappa_j$ from (17c) approach zero.
Step 3. Repeat steps 0, 1 and 2 until convergence.

4.2. Residual maximum likelihood

The downward biasedness of ML estimators of the variance parameters (or ratios), hidden in $\boldsymbol{H}$, can be overcome by using residual maximum likelihood (REML) estimation [2,50]. REML maximizes the likelihood of those linearly independent error contrasts, i.e. independent contrasts of linear combinations of the data $\boldsymbol{y}$, orthogonal to the design matrix $\boldsymbol{X}$. The linear combinations are chosen as $\boldsymbol{K}'\boldsymbol{y}$ so that $\boldsymbol{K}'\boldsymbol{y}$ is of maximal rank but is free of the fixed effects $\boldsymbol{\beta}$. These linear combinations are the residuals obtained after fitting the fixed effects hence the name residual maximum likelihood. Therefore $\mathrm{E}(\boldsymbol{K}'\boldsymbol{y}) = 0$ which is true if and only if $\boldsymbol{K}'\boldsymbol{X} = 0$. This device results in performing maximum likelihood on $\boldsymbol{K}'\boldsymbol{y}$ instead of $\boldsymbol{y}$. Verbeke and Molenberghs [69, Section 5.3.1, pp. 43] illustrate the use of REML to obtain the estimate of σ^2 for a single Gaussian distributed random sample of size n and show that this estimate is restricted to $n-1$ error contrasts instead of the n contrasts used to obtain the MLE of σ^2 hence the name restricted maximum likelihood. In the context of the linear mixed model the MLE estimate of σ^2 is RSS/n, where RSS denotes the residual sums of squares, while the REML estimate is RSS/$(n-p)$ (also see Eq. (25) below). From a Bayesian view point, Harville [23] showed that using only error contrasts to make inferences on the variance parameters is equivalent to ignoring any prior information on the fixed effects parameters. Verbyla [71] shows that REML log-likelihood may also be regarded as a marginal likelihood, while Barndoff-Nielsen [4] takes it as a modified profile log-likelihood. Lee et al. [37] view the REML log-likelihood function as a conditional likelihood by assuming asymptotic (multivariate) Gaussian distribution for the fixed effects estimates given fixed variance parameter values. The REML log-likelihood also coincides with the conditional profile likelihood of Cox and Reid [10].

For $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{H})$ and $\boldsymbol{K}'\boldsymbol{X} = 0$ we have

$$\boldsymbol{K}'\boldsymbol{y} \sim N(0, \sigma^2\boldsymbol{K}'\boldsymbol{HK}) \tag{20}$$

and the residual (REML) log-likelihood function is

$$\begin{aligned} l_R(\boldsymbol{\phi}; \boldsymbol{K}'\boldsymbol{y}) = -\frac{1}{2}\Big\{&(n-p)\log(2\pi) + (n-p)\log\sigma^2 + \log\left|\boldsymbol{K}'\boldsymbol{H}^{-1}\boldsymbol{K}\right| \\ &+\frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{K}(\boldsymbol{K}'\boldsymbol{H}^{-1}\boldsymbol{K})^{-1}\boldsymbol{K}'\boldsymbol{y},\Big\} \end{aligned} \tag{21}$$

where $\boldsymbol{\phi} = (\boldsymbol{\kappa}', \sigma^2)', \boldsymbol{\kappa} = (\boldsymbol{\gamma}\prime, \boldsymbol{\rho}')'$. Patterson and Thompson [50] derived the probability distribution of $\boldsymbol{K}'\boldsymbol{y}$ by carefully choosing $\boldsymbol{K}'$ as an $(n-p) \times n$ matrix whose rows are $n-p$ linearly independent rows of $\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Since $\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is symmetric, idempotent and has rank $n-p$, it can be expressed as $\boldsymbol{KK}'$ such that $\boldsymbol{K}'\boldsymbol{K} = \boldsymbol{I}$. Patterson and Thompson [50] argued that since $\mathrm{E}(\boldsymbol{K}'\boldsymbol{y}) = 0$, $\boldsymbol{K}'\boldsymbol{y}$ lies in the error space, and hence contains no information about the fixed effects ($\boldsymbol{\beta}$), but it does contain information about the variance parameters. Then the REML log-likelihood function (ignoring constants) for the model is

$$l_{\mathrm{R}}(\boldsymbol{\phi}; \boldsymbol{y}) = -\frac{1}{2}\left\{(n-p)\log\sigma^2 + \log|\boldsymbol{H}| + \log|\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}| + \frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{\sigma^2}\right\} \tag{22}$$

$$= -\frac{1}{2}\left\{(n-p)\log\sigma^2 + \log|\boldsymbol{H}| + \log|\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}| + \frac{\boldsymbol{y}'\boldsymbol{Py}}{\sigma^2}\right\} \tag{23}$$

where $\hat{\boldsymbol{\beta}}$, the GLS estimate of $\boldsymbol{\beta}$, and $\boldsymbol{P}$ are given in Lemma 1. Khatri [34] and Searle et al. [62, pp. 15–18] showed that if $\boldsymbol{K}'\boldsymbol{X} = \boldsymbol{0}$, where $\boldsymbol{K}'$ has maximum row rank, and $\boldsymbol{H}$ is positive definite then

$$\boldsymbol{K}(\boldsymbol{K}'\boldsymbol{H}^{-1}\boldsymbol{K})^{-1}\boldsymbol{K}' = \boldsymbol{P}$$

so that (21) and (23) are equivalent.

The equivalence between (22) and (23) is based on the relation

$$\begin{aligned}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) &= \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{y} \\ &= \left(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\right)\boldsymbol{y} \\ &= \boldsymbol{HPy},\end{aligned}$$

and hence by Lemma B.2 of Appendix A

$$\begin{aligned}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) &= \boldsymbol{y}'\boldsymbol{PHH}^{-1}\boldsymbol{HPy} \\ &= \boldsymbol{y}'\boldsymbol{Py}.\end{aligned}$$

Differentiating the REML log-likelihood function (23) with respect to σ^2 and $\kappa_j, j = 1, \ldots, r + s$ gives [19]

$$\frac{\partial l_{\mathrm{R}}(\boldsymbol{\phi}; \boldsymbol{y})}{\partial \sigma^2} = -\frac{n-p}{2\sigma^2} + \frac{\boldsymbol{y}'\boldsymbol{Py}}{2\sigma^4} \tag{24a}$$

$$\frac{\partial l_{\mathrm{R}}(\boldsymbol{\phi}; \boldsymbol{y})}{\partial \kappa_j} = -\frac{1}{2}\left\{\mathrm{tr}(\boldsymbol{P}\dot{\boldsymbol{H}}_j) - \frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{Py}\right\}. \tag{24b}$$

Setting the Eq. (24a) equal to zero and solving gives a REML estimator for the error variance as

$$\hat{\sigma}^2 = \frac{\boldsymbol{y}'\hat{\boldsymbol{P}}\boldsymbol{y}}{n-p}, \tag{25}$$

which must be computed iteratively since it depends on $\hat{\boldsymbol{\kappa}}$ through $\hat{\boldsymbol{P}}$. The REML estimate for $\boldsymbol{\kappa}$ must also be found iteratively (see [19,32]). Searle et al. [62, Section 6.6, pp. 251–254] give an iterative scheme for obtaining the REML estimates based on the variance parameters rather the variance ratios.

Result 1. The REML log-likelihood function (23) can be rewritten as [19]

$$l_{\mathrm{R}}(\boldsymbol{\phi}; \boldsymbol{y}) = -\frac{1}{2}\left\{(n-p)\log\sigma^2 + \log|\boldsymbol{C}| + \log|\boldsymbol{R}| + \log|\boldsymbol{G}| + \frac{\boldsymbol{y}'\boldsymbol{Py}}{\sigma^2}\right\} \tag{26}$$

where $\boldsymbol{C}$ is coefficient matrix in the MMEs (7).

Proof. The proof uses matrix results given in Appendix A, and is not shown here. It can also be shown that the log-likelihood functions (23) and (26) are equivalent. □

4.3. Comparison between ML and REML estimation

Searle et al. [62, Section 6.8] discuss the advantages and disadvantages of ML and REML estimators for variance parameters (or ratios). It must be noted that compared with the marginal (unrestricted) log-likelihood function (l_{ML}), the REML (restricted) log-likelihood function (l_{R}) includes an extra term to take into account the degrees of freedom lost to estimating the fixed effects. Another important difference between the marginal log-likelihood function and the REML log-likelihood function is that the former is invariant to one-to-one reparameterization of the fixed effects. The REML log-likelihood function is not a function of $\boldsymbol{\beta}$ and so cannot be used to compare linear mixed models with different fixed effects structures. In particular, likelihood ratio tests are not valid under these circumstances (discussed in detail in Section 5.1).

A REML estimator of $\boldsymbol{\beta}$, which is BLUE, can be obtained by replacing variance components or variance ratios in $\boldsymbol{G}$ by their REML estimates. Jiang [31] established the asymptotic normality of this estimator and also proved that the empirical distributions of the predictors of the random effects (BLUPs), with the unknown variance components replaced by the REML estimates, converge to the true distributions of the corresponding random effects.

4.4. Iterative schemes

Below we describe four related iterative procedures that are used for the calculation of ML or REML estimates of the variance parameters (or ratios), namely: Newton–Raphson (NR), Fisher Scoring (FS) and the Average Information (AI) algorithms. The FS and AI algorithms are variations of the NR algorithm. Some variants of these algorithms have been explored by several authors for estimation of variance parameters in linear mixed models, for example Hemmerle and Hartley [24], Corbeil and Searle [7], Jennrich and Schluchter [30], Lindstrom and Bates [42] and Callanan and Harville [6].

4.4.1. Newton–Raphson algorithm

The Newton–Raphson (NR) algorithm [68, Section 4.2.2] uses the first-order expansion of the score function around the current estimate $\boldsymbol{\phi}_{(m)}$ to produce the next estimate $\boldsymbol{\phi}_{(m+1)}$. This algorithm assumes concavity of log-likelihood function to get the quadratic approximation to the function. Each NR iteration requires the calculation of the score function and its derivative. Briefly, the NR procedure can be described as follows. Consider the log-likelihood function $l(\boldsymbol{\phi})$ for which we want to find the maximum at $\boldsymbol{\phi}$ with

$$\frac{\partial l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \mathbf{0}.$$

By first-order expansion we have the vector equation

$$\frac{\partial l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = U(\boldsymbol{\phi}) \approx U(\boldsymbol{\phi}_{(0)}) + \frac{\partial^2 l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'}(\boldsymbol{\phi} - \boldsymbol{\phi}_{(0)}). \tag{27}$$

Equating (27) to zero, and solving we have

$$U(\boldsymbol{\phi}_{(0)}) + \frac{\partial^2 l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'}(\boldsymbol{\phi} - \boldsymbol{\phi}_{(0)}) = \mathbf{0},$$

which gives

$$\boldsymbol{\phi} = \boldsymbol{\phi}_{(0)} - \left[\frac{\partial^2 l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'}\right]^{-1} U(\boldsymbol{\phi}_{(0)}).$$

This equation can be used iteratively to refine the estimate of the maximum on the $(m+1)$th iteration:

$$\begin{aligned}\boldsymbol{\phi}_{(m+1)} &= \boldsymbol{\phi}_{(m)} - \left[\frac{\partial^2 l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'}\right]^{-1} U(\boldsymbol{\phi}_{(m)}) \\ &= \boldsymbol{\phi}_{(m)} + \left[\mathcal{I}_{\mathcal{O}(m)}\right]^{-1} U(\boldsymbol{\phi}_{(m)}),\end{aligned}$$

starting from a pre-specified initial value $\boldsymbol{\phi}_{(0)}$. $\mathcal{I}_{\mathcal{O}(m)}$ is the observed information matrix evaluated at $\boldsymbol{\phi}_{(m)}$.

4.4.2. Fisher Scoring algorithm

The Fisher Scoring (FS) algorithm replaces the observed information matrix by the expected information matrix, $\mathrm{E}\left[-\frac{\partial^2 l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'}\right]$, in the NR algorithm.

4.4.3. Expectation Maximization algorithm

Dempster et al. [12] introduced the EM algorithm for parameter estimation in models with incomplete data. Dempster et al. [13] showed how the EM algorithm can be used to obtain maximum likelihood estimates of variance components in the linear mixed model. Implementation of this method in the linear mixed model is based on viewing the random effects as unobserved or missing data. The EM algorithm consists in essence of two steps: an expectation step (E-step) and a maximization step (M-step). The steps at the $(m+1)$th iteration can be described as follows:

E-step. Use the mth estimate, $\boldsymbol{\phi}_{(m)}$, to evaluate the log-likelihood of the conditional distribution $\boldsymbol{u}|\boldsymbol{y}$ and compute the expectation of the log-likelihood for a new value of $\boldsymbol{\phi}$ given this conditional distribution.

M-step. Maximize the expectation from the E-step with respect to $\boldsymbol{\phi}$ to produce $\boldsymbol{\phi}_{(m+1)}$.

The EM procedure is completed by iterating between the E and M steps until convergence.

Searle et al. [62, Section 8.3] describe three EM procedures for computing ML and REML estimates in a linear mixed model. Foulley et al. [16] give an illustration of the EM algorithm for computing REML estimates of variance components for various structures of $\boldsymbol{G}$ and $\boldsymbol{R}$.

4.4.4. Average Information algorithm

More recently, Gilmour et al. [19] and Johnson and Thompson [32] introduced the Average Information (AI) algorithm for the estimation of variance parameters in a linear mixed model. The AI algorithm can be regarded as a modified Fisher Scoring algorithm since it replaces the expected information matrix in the FS algorithm with an average of the observed and expected information matrices called the average information matrix. This information matrix avoids the evaluation of trace terms in the observed and expected information matrix by approximating the trace terms by sums of squares with correct expected values, i.e. the use of the average information matrix is motivated by computational efficiency because the sums of squares terms are easier to calculate than the trace terms. Similar to the NR and FS algorithms, the AI algorithm is based on finding an efficient solution of the mixed model equations. At each iteration the current values for $\boldsymbol{\phi}$ are used to solve mixed model equations (8). Gilmour et al. [19] describes how this solution is achieved using sparse matrix techniques and an absorption and backsubstitution procedure which maximizes computational efficiency by avoiding calculation of unnecessary terms in $\boldsymbol{C}$ (and $\boldsymbol{C}^{-1}$) which come from the absorption process.

In the following we present the score functions for the elements of $\boldsymbol{\phi}$ as well as the observed, expected, and (approximate) average information matrices for $\boldsymbol{\phi}$. The proofs for these results are given in Gilmour et al. [19] (also see [32]). These score statistics and information matrices are required for the implementation of the iterative schemes described above and also to estimate the variance–covariance matrix of the variance parameters. The score functions and information matrices for the variance parameters also play an important role in the construction of inferential procedures for the variance parameters. Traditionally, the observed information and expected information matrices are used to obtain the variance–covariance matrix of parameters of a model. Efron and Hinkley [14] give a comparison of the two methods when the observations are independent and identically distributed. They showed that in these situations the observed information is better than the expected information (also see [51, pp. 245–250]). In this paper we also define the exact average information matrix which is an evenly-weighted average of the observed and expected information matrices, i.e. the exact information matrix is constructed as a simple average of the observed and expected information matrix elements which involves evaluation of trace terms in the observed and expected information matrices. Hence our exact average information matrix differs from the average information matrix of Gilmour et al. [19]. We expect the exact average information matrix to give similar variance estimates to the approximate average information matrix as an indication of whether the approximate average information matrix adequately approximates the average of the observed and expected information matrices, i.e. whether the approximate average matrix approximates the trace terms in the observed and expected matrices adequately.

Result 2. The score function for κ_j is given by

$$U(\kappa_j) = \frac{\partial l_R(\boldsymbol{\phi}; \boldsymbol{y})}{\partial \kappa_j} = -\frac{1}{2}\left\{\text{tr}(\boldsymbol{P}\dot{\boldsymbol{H}}_j) - \frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}\right\},$$

where $l_R(\boldsymbol{\phi}; \boldsymbol{y})$ is the REML log-likelihood function (23), $\boldsymbol{\phi} = (\boldsymbol{\kappa}', \sigma^2)'$ and $\dot{\boldsymbol{H}}_j = \partial \boldsymbol{H}/\partial \kappa_j$; for $j = 1, \ldots, r+s$, where $r+s$ is the number of variance parameters in $\boldsymbol{\kappa}$. The number of variance parameters in the model including σ^2, i.e. the number of parameters in $\boldsymbol{\phi}$, is $t = r + s + 1$.

Result 3. The score function for σ^2 is given by

$$U(\sigma^2) = \frac{\partial l_R(\boldsymbol{\phi}; \boldsymbol{y})}{\partial \sigma^2} = -\frac{1}{2}\left\{\frac{(n-p)}{\sigma^2} - \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{\sigma^4}\right\}.$$

Result 4. The elements of the observed information matrix for the variance parameters, κ_j and σ^2 are

$$\begin{aligned}
\mathcal{I}_{\mathcal{O}}(\kappa_j, \kappa_k) &= \frac{1}{2}\text{tr}\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\right) - \frac{1}{2}\text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\right) + \frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\boldsymbol{P}\boldsymbol{y} \\
&\quad - \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\boldsymbol{P}\boldsymbol{y} \\
\mathcal{I}_{\mathcal{O}}(\sigma^2, \kappa_j) &= \frac{\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}}{2\sigma^4} \\
\mathcal{I}_{\mathcal{O}}(\sigma^2, \sigma^2) &= -\frac{(n-p)}{2\sigma^4} + \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{\sigma^6}.
\end{aligned}$$

where $\ddot{\boldsymbol{H}}_{jk} = \partial^2\boldsymbol{H}/\partial\kappa_j\kappa_k$.

Result 5. The elements of the expected information matrix for the variance parameters, κ_j and σ^2 are

$$\begin{aligned}
\mathcal{I}_{\mathcal{E}}(\kappa_j, \kappa_k) &= \frac{1}{2}\text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\right) \\
\mathcal{I}_{\mathcal{E}}(\sigma^2, \kappa_j) &= \frac{1}{2\sigma^2}\text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\right) \\
\mathcal{I}_{\mathcal{E}}(\sigma^2, \sigma^2) &= \frac{(n-p)}{2\sigma^4}.
\end{aligned}$$

Result 6. The elements of the approximate average information matrix for the variance parameters, κ_j and σ^2 are

$$\begin{aligned}
\mathcal{I}_{\mathcal{A}}(\kappa_j, \kappa_k) &= \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\boldsymbol{P}\boldsymbol{y} \\
\mathcal{I}_{\mathcal{A}}(\sigma^2, \kappa_j) &= \frac{1}{2\sigma^4}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y} \\
\mathcal{I}_{\mathcal{A}}(\sigma^2, \sigma^2) &= \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{2\sigma^6}.
\end{aligned}$$

Table 1
Comparison of iterative methods for obtaining estimates of variance components.

Method	Disadvantages	Advantages
EM	Slow rate of convergence Does not give standard errors for estimates	Numerically stable
NR	Computationally intensive compared to EM (unstable when far from the maximum)	Faster convergence than EM Gives asymptotic standard errors for estimates
	Tendency to converge to values outside parameter space	
FS	Computationally intensive compared to EM (unstable when far from the maximum)	Faster convergence than EM Gives asymptotic standard errors errors for estimates
	Tendency to converge to values outside parameter space	Robust to poor starting values than NR
AI		Requires fewer iterations (faster) compared to EM

Result 7. The elements of the exact average information matrix for variance parameters are obtained by taking equally-weighted averages within the three pairs of terms in Results 4 and 5 and are given by

$$\begin{aligned}\mathcal{I}_{\mathcal{A}e}(\kappa_j, \kappa_k) &= \frac{1}{4}\mathrm{tr}\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\right) + \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\boldsymbol{P}\boldsymbol{y} \\ &\quad - \frac{1}{4\sigma^2}\boldsymbol{y}'\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\boldsymbol{P}\boldsymbol{y} \\ \mathcal{I}_{\mathcal{A}e}(\sigma^2, \kappa_j) &= \frac{\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}}{4\sigma^4} + \frac{1}{4\sigma^2}\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\right) \\ \mathcal{I}_{\mathcal{A}e}(\sigma^2, \sigma^2) &= \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{2\sigma^6}.\end{aligned}$$

We will refer to the average information matrix of Gilmour et al. [19] as the approximate average information matrix to reflect the nature of the weighting of the observed and expected information matrix terms. Throughout this paper we will use the subscripts $\mathcal{O}$, $\mathcal{E}$, $\mathcal{A}$, and $\mathcal{A}_e$ to denote quantities relating to the observed, expected, approximate average and exact average information matrices, respectively. Results for the linear variance parameterization of model (1), i.e. when variance of the data, $\boldsymbol{V}$, is in terms of the variance components as defined in (5), are given in Appendix C.

4.5. Comparison of iterative schemes

Table 1 gives a comparison of iterative schemes for obtaining estimates of variance components either using ML and REML estimation. Although the EM algorithm does not provide standard errors for the estimated parameters. Jennrich and Schluchter [30] suggest finding standard errors by taking a single NR or FS step after the EM algorithm has converged. Meng and Rubin [46] and Meng and van Dyk [47] also propose remedies for some of the limitations of the EM algorithm. In particular, Liu et al. [43] introduced the parameter-expanded EM as a variant of the EM which they formulated in order to reduce the number of iterations in the original EM algorithm. To deal with the stability problem of the NR and FS algorithms [52] implement the idea of Baker [3] of a hybrid approach which starts with an EM algorithm and then switches to NR. Here, the EM iterations can be regarded as refinements of the starting values of the estimates before commencement of the optimization routine.

Jennrich and Sampson [29] report that Fisher Scoring is also more robust to poor starting values than the NR algorithm. They recommend an iterative algorithm which starts by using Fisher Scoring for the first few steps and then switches to NR algorithm.

4.6. Starting values

Much of the difficulty in estimating variance parameters (or ratios), using the algorithms just described, is centered on obtaining good starting values. Derivative-based algorithms, such as the AI, EM, Fisher Scoring and Newton–Raphson algorithms can be unreliable when estimating variance parameters, especially for models with complex variance structures, unless good starting values are available. Poor starting values may result in divergence of the algorithm or slow convergence. Thisted [68, Section 4.2.5] provides a general discussion of guidelines for starting values and convergence criteria of algorithms based on iterative schemes.

Searle et al. [62] suggest the use of ordinary least squares estimates for starting values of the fixed effects and ANOVA estimators the variance parameters as starting values. Another method of obtaining starting values of variance parameters is a variant of the MINQUE of Rao [55], namely MIVQUE0 [21, 64]. Corbeil and Searle [8] used MIVQUE0 to obtain starting values for REML estimation of variance parameters. Jennrich and Schluchter [30] used MIVQUE0 estimates as starting values for the NR and FS algorithms for computing maximum likelihood estimates of the variance parameters. Jennrich and Schluchter [30] and Laird et al. [35] give further suggestions for starting values for variance parameters.

4.7. Convergence criteria

The most commonly used criteria of convergence are based on the relative change in either the variance parameter values between successive iterations or score functions or information matrices, or differences between successive log-likelihood functions. For instance, the AI algorithm (in GenStat and ASReml), uses the relative change in the deviance as a check for convergence, whereas the FS method checks for changes in the variance parameter values (in GenStat). For assessing changes in variance parameter values, a measure that involves a multiplier of 0.005 is used. So, for convergence, the change in every variance parameter must be less than 0.005. When assessing change in deviance, convergence is declared when the absolute change in the deviance is less than 10^{-3}. Bates and Watts [5] argue that these criteria may indicate lack of progress rather than convergence. They suggest a convergence criterion based on the relative Hessian (second derivative of the REML log-likelihood) matrix. Their criterion is defined as

$$U(\boldsymbol{\phi}^{(m)})' \left[\frac{\partial^2 l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'} \right]^{-1} U(\boldsymbol{\phi}^{(m)}) [l(\boldsymbol{\phi}^{(m)})]^{-1},$$

where $l(\boldsymbol{\phi}^{(m)})$ is the REML log-likelihood function at the mth iteration and $U(\boldsymbol{\phi}^{(m)})$ is the score function at the mth iteration. This criterion can be used for all three of the NR, FS and AI algorithms and has the advantage that it can be calculated from the information available at each iteration.

5. Statistical inference

This section discusses the theory of inferential procedures used for the estimated parameters in the linear mixed model.

5.1. Inference for fixed effects

When the variance parameters are estimated using maximum likelihood, two nested models with different fixed effects structures but with the same variance structure can be compared using a the likelihood ratio test.

$$MLRT = -2(l_{\mathrm{ML}_0} - l_{\mathrm{ML}_1}), \tag{32}$$

where l_{ML_i} is the marginal log-likelihood function for model i, for $i = 0, 1$ and where l_{ML_1} includes an extra k fixed effects parameters. In general $MLRT$ asymptotically follows a chi-squared distribution with degrees of freedom k [9, Chapter 9]. However, to compare two nested models with different fixed effects structures, a likelihood ratio test based on REML cannot be used. This difficulty arises because when the variance parameters are estimated using REML, the two models being compared use different error contrasts $\boldsymbol{K}'\boldsymbol{y}$. Hence the corresponding REML log-likelihood functions are no longer comparable since they are based on different observations. Welham and Thompson [75] proposed an adjusted likelihood ratio test statistic for the comparison of two models with nested fixed effects, fitted using REML. An alternative would be to use the Wald test statistic [73]. Consider testing the hypothesis

$$H_0 : \boldsymbol{L}'\boldsymbol{\beta} = \boldsymbol{l} \quad \text{vs} \quad H_A : \boldsymbol{L}'\boldsymbol{\beta} \neq \boldsymbol{l},$$

where $\boldsymbol{L}'$ is an $c \times p$ matrix and $\boldsymbol{l}$ is an $c \times 1$ vector. Then the Wald test statistic is given by

$$\begin{aligned} W &= (\boldsymbol{L}'\hat{\boldsymbol{\beta}} - \boldsymbol{l})'[\text{var}(\boldsymbol{L}'\hat{\boldsymbol{\beta}} - \boldsymbol{l})]^{-1}(\boldsymbol{L}'\hat{\boldsymbol{\beta}} - \boldsymbol{l}) \\ &\doteq \frac{(\boldsymbol{L}'\hat{\boldsymbol{\beta}} - \boldsymbol{l})'[\boldsymbol{L}'(\boldsymbol{X}'\hat{\boldsymbol{H}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{L}]^{-1}(\boldsymbol{L}'\hat{\boldsymbol{\beta}} - \boldsymbol{l})}{\hat{\sigma}^2} \end{aligned} \tag{33}$$

where $\hat{\sigma}^2\boldsymbol{L}'(\boldsymbol{X}'\hat{\boldsymbol{H}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{L}$ is the approximate covariance matrix of $\boldsymbol{L}'\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2\hat{\boldsymbol{H}}$ is the REML estimate for $\sigma^2\boldsymbol{H}$. Under H_0, W has an approximate chi-squared distribution with ν degrees of freedom, where $\nu = r_{\boldsymbol{L}}$.

The asymptotic property of the Wald test statistic is based on the assumption that the variance $\sigma^2\boldsymbol{H}$ is known without error, but $\sigma^2\boldsymbol{H}$ is not known and is estimated from the data using REML. This estimation introduces additional variability in the fixed effect estimates. In this way the Wald test statistic underestimates the variability in $\boldsymbol{L}'\hat{\boldsymbol{\beta}}$, so that the test statistic tends to be anti-conservative in small samples, i.e. the test indicates that an effect may be important more often than expected under the null hypothesis of no effect. Lill et al. [40] reported little effect on the nominal size of the Wald test after replacing the unknown variance parameters by their REML estimates. Kenward and Roger [33] suggested a scaled Wald statistic which is based on an adjusted covariance estimate, to account for the extra variability introduced by estimating the variance parameters, $\boldsymbol{\phi}$, using REML. This scaled Wald statistic improves the small sample behaviour of the test. They showed that the finite sampling distribution of the scaled Wald statistic was approximately an F distribution with denominator degrees of freedom estimated by a Satterthwaite-approximation method [58]. Zucker and Manor [77] investigated the small sample performance of several procedures for testing a given fixed effect in a mixed linear model.

5.2. Inference for variance parameters

Fixed effect parameters are usually the focus of scientific interest in the linear mixed model. However, it is important to correctly specify the covariance structure to obtain valid statistical inferences for the fixed effects. Altham [1] noted that overparameterization of the covariance structure may lead to inefficient estimates and poor standard errors for the fixed effects whereas a too restrictive parameterization of the covariance structure renders the inferences about the fixed effect invalid. Verbeke and Molenberghs [69, Chapter 9] and Wolfinger [76] give strategies for model building and covariance structure selection in linear mixed models.

Since the REML estimators of the variance parameters are asymptotically Gaussian, we may use approximate Wald tests for testing for their statistical significance. An alternative measure for comparing nested models with different variance parameters but with the same fixed effects is the likelihood ratio test which we describe below.

Lemma 2. *The Residual Maximum Likelihood Ratio Test (REMLRT) statistic for comparing two nested models R_0 and R_1 where R_1 includes an extra k variance parameters is given by*

$$\mathit{REMLRT} = -2(l_{R_0} - l_{R_1}), \tag{34}$$

where l_{R_i} is the REML log-likelihood function for model i, for $i = 0, 1$.

The REMLRT statistic is asymptotically chi-squared distributed with k degrees of freedom. However, when the null hypothesis is on the boundary of the parameter space, for example testing $H_0 : \sigma_a^2 = 0$ against $H_A : \sigma_a^2 > 0$, where σ_a^2 is the random effects variance, the standard asymptotic theory no longer holds, as regularity conditions are not met. For significance testing the $0.5\chi_0^2 + 0.5\chi_1^2$ mixture distribution of Self and Liang [65] has been used. The distribution χ_0^2 represents a distribution with a point mass at 0. Morrell [49] compared the REMLRT (2) with its ML version in terms of type I errors using the $0.5\chi_0^2 + 0.5\chi_1^2$ mixture distribution. He found that the REML test statistic performed better than the ML statistic, i.e. on average, the empirical type I errors were closer to the nominal levels for the REML statistic than for the ML statistic. He did not compare these statistics in terms of type II errors.

The score test statistic [9, Section 9.3] can also be used for testing the significance of variance parameters instead of the likelihood ratio test statistic. The score test only involves the score vector and information matrix under the null hypothesis, i.e. with covariance parameter estimates obtained under the model that is to be tested. Its main advantage over the likelihood ratio test statistic is that it does not require fitting the model specified under the alternate hypothesis; only the null model fit is required to obtain the quantities involved in its calculation.

Lemma 3. *The score test statistic for comparing the model* (1) *with the model in which some of the specific variance parameters are equal to zero, i.e.* $H_0 : \boldsymbol{\kappa}_0 = \mathbf{0}$ *against* $H_A : \boldsymbol{\kappa}_0 \neq \mathbf{0}$, *where* $\boldsymbol{\kappa}_0$ *is a* $k \times 1$ $(k < r + s)$ *vector of variance parameters of interest, is given by*

$$S(\boldsymbol{\kappa}_0) = \boldsymbol{U}(\boldsymbol{\kappa}_0)'\mathcal{I}^{\boldsymbol{\kappa}_0\boldsymbol{\kappa}_0}\boldsymbol{U}(\boldsymbol{\kappa}_0)|_{\boldsymbol{\kappa}_0=\mathbf{0}}, \tag{35}$$

where $\mathcal{I}^{\boldsymbol{\kappa}_0\boldsymbol{\kappa}_0}$ *is the portion of the inverse of the expected information matrix associated with* $\boldsymbol{\kappa}_0$, $\boldsymbol{U}(\boldsymbol{\kappa}_0)$ *is the score vector for* $\boldsymbol{\kappa}_0$ *and* $r + s$ *is the number of variance parameters in* $\boldsymbol{G}$ *and* $\boldsymbol{R}$. *Note that all terms in* $S(\boldsymbol{\kappa}_0)$ *are evaluated at* $\boldsymbol{\kappa}_0 = \mathbf{0}$.

The score test statistic (35) also has an asymptotic chi-squared distribution under the null hypothesis with k degrees of freedom, in line with the likelihood ratio test. It has been used for variance parameter testing in linear mixed models [28,48,70]. The score test suffers from the same boundary problem (when the null hypothesis is on the boundary of the parameter space) as the likelihood ratio and so the $0.5\chi_0^2 + 0.5\chi_1^2$ mixture distribution is used to assess the significance of the test.

The expected information matrix in the score test (35) may be replaced by other information matrices resulting in different score tests. The properties of the resulting score tests have not been studies in detail in the linear mixed model literature.

In some situations, it may be of interest to distinguish between non-nested variance models, with same fixed effects, we may also use the AIC and the BIC measures based on the REML log-likelihood.

$$AIC = -2l_{R_i} + 2t$$
$$BIC = -2l_{R_i} + t\log(n - p)$$

where l_{R_i} is the REML log-likelihood function for model i, for $i = 1, \ldots, m$. Note that the effective sample size used in the BIC is $n^* = n - p$ and not the total sample size n since REML is based on a set of $n - p$ error contrasts.

5.3. Inference for random effects

Earlier we showed that the BLUP of $\boldsymbol{u}$, $\tilde{\boldsymbol{u}}$, was analogous to the conditional mean of the posterior distribution of $\boldsymbol{u}|\boldsymbol{y}$ with variance $\sigma^2[\boldsymbol{GZ}'\boldsymbol{PZG}]$. We also showed that $\mathrm{var}(\tilde{\boldsymbol{u}} - \boldsymbol{u}) = \sigma^2[\boldsymbol{G} - \boldsymbol{GZ}'\boldsymbol{PZG}]$.

It must be noted that $\mathrm{var}(\tilde{\boldsymbol{u}})$ underestimates the true variability in $(\tilde{\boldsymbol{u}} - \boldsymbol{u})$ since it ignores the variation of $\boldsymbol{u}$. Nevertheless, following [36], we can still base inference on $\boldsymbol{u}$ using $\widehat{\mathrm{var}}(\tilde{\boldsymbol{u}} - \boldsymbol{u})$ as an estimator for the variation in $(\tilde{\boldsymbol{u}} - \boldsymbol{u})$. In conducting inference of $\boldsymbol{u}$, Verbyla et al. [72] argue that tests involving equality do not make sense and suggest comparing u_i and u_j, $i \neq j$, using the probability statement,

$$\Pr(u_i > u_j|\boldsymbol{y}) = 1 - \Phi\left(\frac{\tilde{u}_i - \tilde{u}_j}{\sigma\sqrt{\boldsymbol{a}'[\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G}]\boldsymbol{a}}}\right),$$

where $\boldsymbol{a}$ is a $q \times 1$ vector of zeroes, except for $a_i = 1$ and $a_j = -1$.

However, both $\widehat{\text{var}}(\tilde{\boldsymbol{u}})$ and $\widehat{\text{var}}(\tilde{\boldsymbol{u}} - \boldsymbol{u})$ underestimate the true variability in $\tilde{\boldsymbol{u}}$ because the unknown variance parameters $\boldsymbol{\phi}$ are replaced by their ML or REML estimates in calculating the variance estimates for $\tilde{\boldsymbol{u}}$ (i.e. $\widehat{\text{var}}(\tilde{\boldsymbol{u}})$ and $\widehat{\text{var}}(\tilde{\boldsymbol{u}} - \boldsymbol{u})$). Similar to inference for fixed effects, inference on $\boldsymbol{u}$ can then be based on approximate t-tests or F-tests with denominator degrees of freedom estimated via a Satterthwaite-approximation method. This problem, Satterthwaite-approximation for random effects, has not been addressed in the literature on inference for linear mixed models. Lee et al. [37, Section 5.4.1] assert that their h-likelihood approach to parameter estimation in the linear mixed model (including generalized linear mixed models) gives the necessary correction for the extra variability due to estimation of the fixed effects, that is otherwise suppressed, in the variance of $(\tilde{\boldsymbol{u}} - \boldsymbol{u})$. This correction makes it possible to construct confidence intervals for unknown $\boldsymbol{u}$.

5.4. Inference on a combination of fixed and random effects

Inference on combinations of the fixed effects $\boldsymbol{\beta}$ and random effects could follow the approach to inference for random effects above [72]. For combination of fixed and random effects $\boldsymbol{a}_1'\boldsymbol{\beta} + \boldsymbol{a}_2'\boldsymbol{u}$ we have that

$$(\boldsymbol{a}_1'\hat{\boldsymbol{\beta}} + \boldsymbol{a}_2'\tilde{\boldsymbol{u}}) - (\boldsymbol{a}_1'\boldsymbol{\beta} + \boldsymbol{a}_2'\boldsymbol{u}) \sim N(0, \sigma^2\boldsymbol{a}'\boldsymbol{C}^{-1}\boldsymbol{a}) \tag{36}$$

where $\boldsymbol{a} = [\boldsymbol{a}_1', \boldsymbol{a}_2']'$, and here the special cases $\boldsymbol{a}_1$ as a $p \times 1$ vector of zeroes and ones for the combination of fixed effects parameters and $\boldsymbol{a}_2$ as a $q \times 1$ vector of zeroes and ones for the combination of random effects parameters, may be of interest.

The above distribution can then be used to make probability statements about an arbitrary combination $\boldsymbol{a}_1'\boldsymbol{\beta} + \boldsymbol{a}_2'\boldsymbol{u}$. A computational challenge in using (36) to form predictions is that the dimensions of $\boldsymbol{a}$ and $\boldsymbol{C}$ are usually large, which makes the construction of the predictions and their error variances difficult. Gilmour et al. [18] discuss the general principles of prediction in linear mixed models and give an efficient algorithm for obtaining predictions and the prediction error variances from the fitted linear mixed model (also see the companion paper [74]). Their algorithm has been implemented in the package ASReml and is also used in GenStat.

6. Example: the orthodont data

The data are taken from Potthoff and Roy [53]. The data consist of measurements of the distance in millimetres from the center of the pituitary to the pterygomaxillary fissure at ages 8, 10, 12 and 14 years on 16 boys and 11 girls. The purpose of the study is to model the relationship between distance and age, with investigation of gender differences.

Fig. 1 comprises plots of the distances by age for the boys and girls separately. Generally, the change in distance is approximately linear over the range 8-14 years. The data for boys appear more variable than for girls. The response profiles vary considerably between subjects. We use the prefixes “M" and “F" to number the male and female subjects, respectively. Subjects number 9 and 13 among the boys (M9 and M13) seem to have possible outlying observations. Male subject 4 appears to have a reduced slope while male subject 10 seems to have a higher intercept. Subject number 10 among the girls (F10) appears to have a suppressed response profile compared to other females, while subject number 11 among the girls (F11) seems to have an elevated response profile compared to other subjects in the

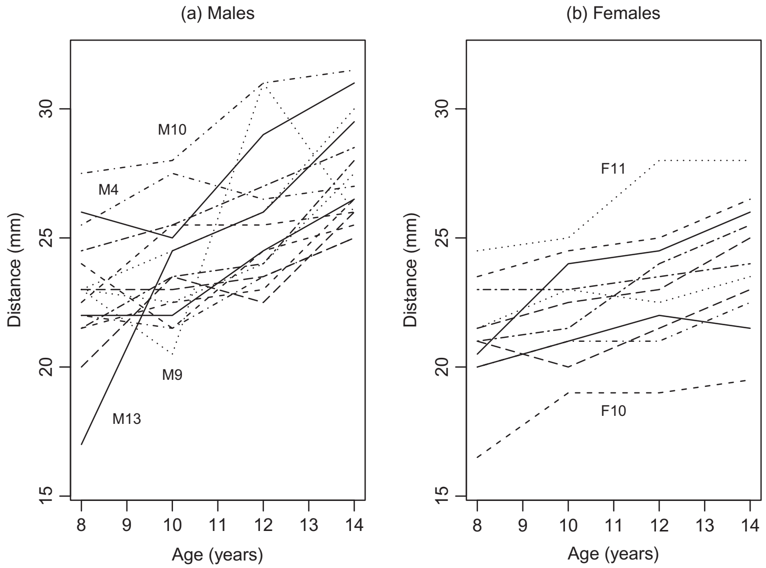

Fig. 1. Plots of distance against age for orthodont data.

group. Therefore any statistical modelling of these data would need to take account of the subject variation and possibly account for the presence of outliers within subjects and outlying subjects.

Following Pinheiro and Bates [52] we fit the following linear mixed model to the data

$$\boldsymbol{y}_{jk} = (\mu + \beta_{0k} + u_{0jk})\mathbf{1}_4 + (\beta_1 + \beta_{1k} + u_{1jk})\boldsymbol{x} + \boldsymbol{e}_{jk}, \tag{37}$$

where, $\boldsymbol{y}_{jk}$ is the vector of distances for the jth subject of gender $k, j = 1, \ldots, 27; k = 1, 0$ with 1 for males and 0 for females, $\boldsymbol{x} = \{x_l - 11 : l = 1, \ldots, 4\}$, x_l is the age at measurement l, μ is the overall mean, β_{0k} is the intercept shift for gender k, β_1 is the overall slope, β_{1k} is the slope for gender k, u_{0jk} is the random additive effect of the jth subject of gender k and u_{1jk} is the random slope effect of the jth subject of gender k, and finally $\boldsymbol{e}_{jk}$ is the random error vector for subject j of gender k. The centering of the explanatory variable for age reduces the correlation between the slope and intercept. The random effects vector for the jth subject $\boldsymbol{u}'_{jk} = (u_{0jk}, u_{1jk})'$ is assumed to be Gaussian distributed with mean zero and variance matrix given by

$$\boldsymbol{G}_{\text{subject}} = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix}$$

where γ_{11} and γ_{22} are the variance ratios for the random intercepts and random slopes, respectively, γ_{21} is the correlation between the variance ratios. and the corresponding error vector $\boldsymbol{e}_{jk}$ is assumed to have a Gaussian distribution with mean zero and variance matrix $\sigma^2\boldsymbol{I}_4$, independently of $\boldsymbol{u}_{jk}$. The matrix $\boldsymbol{G}_{\text{subject}}$ specifies the subject variance structure and the identity matrix specifies random error structure. Then the matrices $\boldsymbol{G}$, $\boldsymbol{R}$, $\boldsymbol{X}$ and $\boldsymbol{Z}$ matrices, defined earlier, are given by

$$\boldsymbol{G} = \boldsymbol{I}_{27} \otimes \boldsymbol{G}_{\text{subject}},$$
$$\boldsymbol{R} = \boldsymbol{I}_{27} \otimes \sigma^2\boldsymbol{I}_4 = \sigma^2\boldsymbol{I}_{108},$$
$$\boldsymbol{X} = \mathbf{1}_{27} \otimes [\mathbf{1}_4 : \boldsymbol{x}]$$

and

$$\boldsymbol{Z} = \boldsymbol{I}_{27} \otimes [\mathbf{1}_4 : \boldsymbol{x}].$$

where $\otimes$ is the Kronecker product of rectangular matrices.

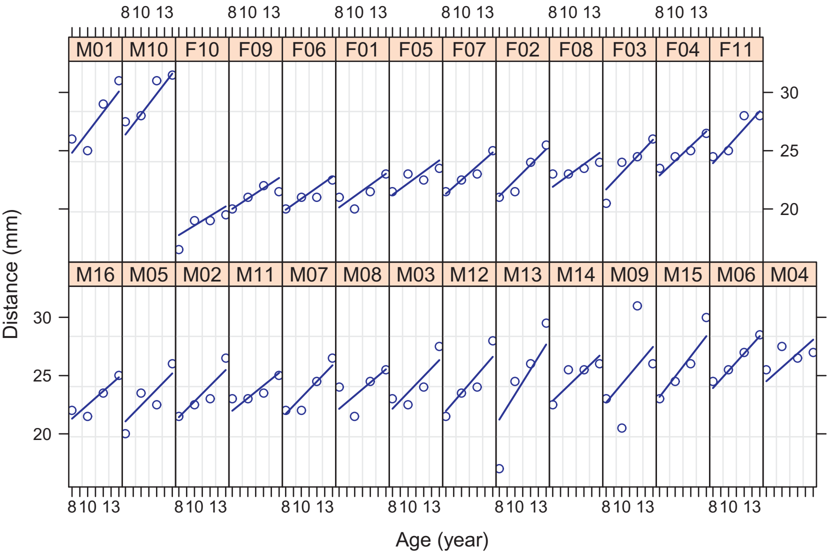

Fig. 2. Scatter plots of distance against age for each subject with fitted lines superimposed for orthodont data.

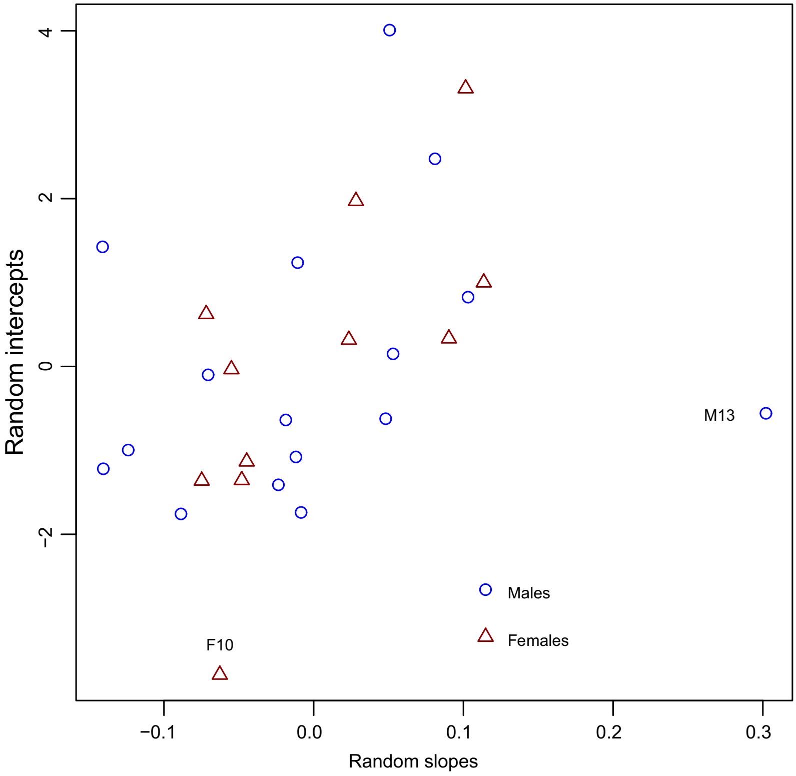

Fig. 3. Scatter plot of random intercepts against random slopes for orthodont data.

The variance–covariance matrix for the data is of the form

$$\mathrm{var}(\boldsymbol{y}) = \sigma^2(\boldsymbol{ZGZ}' + \boldsymbol{I}_{108}).$$

In the following we present results from the fitted model. To index the observations we use the notation $j.l$ to label the lth observation within the jth subject, $j = 1, \dots, 27$: $j = 1, \dots, 16$ for boys and $j = 17, \dots, 27$ for girls. Fig. 2 shows scatter plots of distance against age with fitted values superimposed for each subject. The scatter plots for boys are labelled as $M01, \dots, M16$ (the first 64 observations) and the plots for girls are labelled as $F01, \dots, F10$. Fig. 3 is a scatter plot of the estimated random intercepts against the estimated random slopes and suggests that female 10 has the smallest random intercepts and male 13 has a large slope and may be quite different from other subjects.

7. Summary

In summary, we have reviewed parameter estimation and inference for the linear mixed model, for the variance parameters in particular. We prefer REML for the estimation of the variance parameters using either Fisher scoring or the average information algorithm of Gilmour et al. [19] since it gives unbiased variance parameter estimates. We draw attention to the fact that different types of information matrices (observed, expected, approximate average and exact average) are available for use in the iterative schemes for estimating the variance parameters in the linear mixed model. These different information matrices may also be used in the computation of test statistics for the variance parameters, for example score test statistics or one-step likelihood ratio tests.

A computational challenge for the iterative schemes used for obtaining the variance parameters (variance ratios or variance components) discussed in 4.4 is that the variance components may be near zero or negative when the variance matrices are singular, especially in models with complex variance structures such as random coeffcient regression. This problem requires further research.

Acknowledgements

We thank the University of Cape Town Research Council who provided funding for this work. We also thank the two reviewers for their comments and suggestions, which led to substantial improvement of the manuscript.

Appendix A. Useful (matrix) results and identities

Below is a summary of known results in matrix algebra which we use in this paper. Also included are some fundamental statistical results which are used in the derivation of some of the proofs in this paper. These results can be found in Mardia et al. [44] and Searle [61].

Result A.1. For matrices $\boldsymbol{B}^{p\times n}$ and $\boldsymbol{D}^{n\times p}$ and for non-singular matrices $\boldsymbol{C}^{n\times n}$ and $\boldsymbol{A}^{p\times p}$, from Rao [55, pp. 33] we have

$$(\boldsymbol{A} + \boldsymbol{BCD})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{C}^{-1} + \boldsymbol{DA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{DA}^{-1}.$$

Result A.2. Consider the matrix $\boldsymbol{A}^{m\times m}$ of full rank which is partitioned as

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \tag{A.1}$$

Then the inverse of $\boldsymbol{A}$ is partitioned conformably with $\boldsymbol{A}$ as

$$\boldsymbol{A}^{-1} = \begin{bmatrix} \boldsymbol{A}^{11} & \boldsymbol{A}^{12} \\ \boldsymbol{A}^{21} & \boldsymbol{A}^{22} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1} & -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}^{22} \\ -\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}^{11} & (\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})^{-1} \end{bmatrix},$$

provided $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{22}$ are non-singular.

Result A.3. If $\boldsymbol{A}$ is symmetric and $\boldsymbol{A}_{22}$ and $\boldsymbol{Q} = \boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}$ are non-singular, then $\boldsymbol{A}^{-1}$ can be written as

$$\boldsymbol{A}^{-1} = \begin{bmatrix} \boldsymbol{Q}^{-1} & -\boldsymbol{Q}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{22}^{-1} \\ \boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}\boldsymbol{Q}^{-1} & \boldsymbol{A}_{22}^{-1} + \boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}\boldsymbol{Q}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{22}^{-1} \end{bmatrix}.$$

Result A.4. Using the definition of $\boldsymbol{A}$ in (A.1), the determinant of $\boldsymbol{A}$ can be expressed as

$$|\boldsymbol{A}| = |\boldsymbol{A}_{11}||\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}| = |\boldsymbol{A}_{22}||\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}|,$$

for matrices $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{22}$ non-singular. The notation $|\boldsymbol{A}|$ denotes the determinant of the matrix $\boldsymbol{A}$.

Result A.5. For matrices $\boldsymbol{B}^{p\times n}$ and $\boldsymbol{C}^{n\times p}$, and for non-singular matrix $\boldsymbol{A}^{p\times p}$,

$$|\boldsymbol{A} + \boldsymbol{BC}| = |\boldsymbol{A}||\boldsymbol{I}_p + \boldsymbol{A}^{-1}\boldsymbol{BC}| = |\boldsymbol{A}||\boldsymbol{I}_n + \boldsymbol{CA}^{-1}\boldsymbol{B}|.$$

Result A.6. Let $\boldsymbol{y}$ be multivariate Gaussian, with mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$. Partitioning $\boldsymbol{y}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ conformably as

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

the multivariate Gaussian normal distribution can be written as

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right).$$

Then the conditional distribution of $\boldsymbol{y}_1$ given $\boldsymbol{y}_2$ is also Gaussian and

$$\boldsymbol{y}_1|\boldsymbol{y}_2 \sim N\left[\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right].$$

Result A.7. Quadratic forms

A quadratic form of the vector $\boldsymbol{y}$ is given by $\boldsymbol{y}'\boldsymbol{Ay}$ for some symmetric matrix $\boldsymbol{A}$. If $\boldsymbol{A}$ is not symmetric then the quadratic form is given by

$$\boldsymbol{y}'\boldsymbol{By} = \boldsymbol{y}'\left(\frac{\boldsymbol{A}'}{2} + \frac{\boldsymbol{A}}{2}\right)\boldsymbol{y}$$

If $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then

(i) $\mathrm{E}(\boldsymbol{y}'\boldsymbol{\Sigma}\boldsymbol{y}) = \mathrm{tr}[\boldsymbol{A}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}')] = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}$.

(ii) $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y} \sim \chi^2_b(\frac{1}{2}\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu})$ if and only if $\boldsymbol{A}\boldsymbol{\Sigma}$ is idempotent, i.e. $(\boldsymbol{A}\boldsymbol{\Sigma})^2 = \boldsymbol{A}\boldsymbol{\Sigma}$), where $b = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) = \mathrm{rank}(\boldsymbol{A})$ since $\boldsymbol{\Sigma}$ is non-singular and $\frac{1}{2}\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}$ is the non-centrality parameter.

(iii) $\mathrm{var}(\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y}) = 2\mathrm{tr}[(\boldsymbol{A}\boldsymbol{\Sigma})^2] + 4\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\mu}$.

(iv) $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y}$ and $\boldsymbol{y}'\boldsymbol{B}\boldsymbol{y}$ are independent if and only if $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{0}$.

(v) $\mathrm{cov}(\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y}, \boldsymbol{y}'\boldsymbol{B}\boldsymbol{y}) = 2\mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{\Sigma})$.

Appendix B. Variance ratio parameterization

The following lemmas hold for the linear mixed model (1) namely,

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}\\ &\sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{H}\right).\end{aligned}$$

These lemmas assume that $\boldsymbol{G}, \boldsymbol{H}$ and $\boldsymbol{R}$ are positive definite and that design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are of full rank. A convenient reparameterization of $\boldsymbol{\beta}$ can change $\boldsymbol{X}$ to say, $\boldsymbol{X}^*$ so that $\boldsymbol{X}^*$ is also of full-column rank.

Lemma B.1. *Using Result 1 the inverse of the variance–covariance matrix $\boldsymbol{H} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R}$ is given by*

$$\boldsymbol{H}^{-1} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1}. \tag{B.1}$$

Lemma B.2. *Let $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}$ then $\boldsymbol{P}\boldsymbol{H}\boldsymbol{P} = \boldsymbol{P}$.*

Proof. Since $\boldsymbol{H}\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}$ is idempotent,
$\boldsymbol{H}\boldsymbol{P}\boldsymbol{H}\boldsymbol{P} = \boldsymbol{H}\boldsymbol{P}$, premultiplying by $\boldsymbol{H}^{-1}$ gives

$$\boldsymbol{P}\boldsymbol{H}\boldsymbol{P} = \boldsymbol{P}. \quad \square$$

Lemma B.3. *The partial derivative of $\boldsymbol{P}$ with respect to the variance parameters $\phi_j \in \boldsymbol{\phi}$ is given by*

$$\frac{\partial \boldsymbol{P}}{\partial \phi_j} = -\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P},$$

where $\dot{\boldsymbol{H}}_j = \dfrac{\partial \boldsymbol{H}}{\partial \phi_j}$ and $\boldsymbol{\phi}$ is the vector of variance parameters contained in $\boldsymbol{P}$ through $\boldsymbol{H}$.

Proof. Using matrix differentiation we obtain

$$\begin{aligned}\frac{\partial \boldsymbol{P}}{\partial \phi_j} &= -\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_j\boldsymbol{H}^{-1} + \boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_j\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\\ &\quad - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_j\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\\ &\quad + \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_j\boldsymbol{H}^{-1}\\ &= -\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_j\boldsymbol{P} + \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_j\boldsymbol{P}\\ &= -\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\end{aligned}$$

which proves the Lemma. □

Lemma B.4. *It can be shown that*

$$\boldsymbol{C}^{-1} = \begin{bmatrix} (\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1} & -(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G} \\ -\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1} & \boldsymbol{S} + \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G} \end{bmatrix},$$

where

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{C}_{XX} & \boldsymbol{C}_{XZ} \\ \boldsymbol{C}_{ZX} & \boldsymbol{C}_{ZZ} \end{bmatrix} \tag{B.2}$$

is assumed to be of full rank, i.e. $\boldsymbol{X}$ *is of full column rank, and*

$$\boldsymbol{S} = (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}.$$

Proof. If we let $\boldsymbol{C} = \boldsymbol{A}$ ($\boldsymbol{A}$ as defined in Result 3) then

$$\begin{aligned} \boldsymbol{Q} &= \boldsymbol{C}_{XX} - \boldsymbol{C}_{XZ}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{C}_{ZX} \\ &= \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} - \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} \\ &= \boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}. \end{aligned}$$

Noting that

$$\begin{aligned} (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})\boldsymbol{G}\boldsymbol{Z}' &= \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{Z}' \\ &= \boldsymbol{Z}'\boldsymbol{R}^{-1}(\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R}) \\ &= \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{H}. \end{aligned}$$

Hence

$$(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1} = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}.$$

Thus

$$\begin{aligned} \boldsymbol{C}_{ZZ}^{-1}\boldsymbol{C}_{ZX} &= (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} \\ &= \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}\boldsymbol{X}. \end{aligned}$$

Therefore from Result 3

$$\boldsymbol{C}^{-1} = \begin{bmatrix} (\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1} & -(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G} \\ -\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1} & \boldsymbol{S} + \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G}. \end{bmatrix} \tag{B.3}$$

Using

$$\boldsymbol{H}^{-1} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1},$$

the lower right-hand matrix of $\boldsymbol{C}^{-1}$ can be simplified as follows

$$\begin{aligned} \boldsymbol{K} &= \boldsymbol{C}_{ZZ}^{-1} + \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G} \\ &= (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1} + \boldsymbol{G}\boldsymbol{Z}'(\boldsymbol{H}^{-1} - \boldsymbol{P})\boldsymbol{Z}\boldsymbol{G} \\ &= (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1} \\ &\quad + \boldsymbol{G}\boldsymbol{Z}'[\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}\boldsymbol{R}^{-1}]\boldsymbol{Z}\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G}. \end{aligned}$$

Writing $\boldsymbol{L} = \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}$ gives

$$\begin{aligned} \boldsymbol{K} &= (\boldsymbol{L} + \boldsymbol{G}^{-1})^{-1} + \boldsymbol{G}\boldsymbol{L}\boldsymbol{G} - \boldsymbol{G}\boldsymbol{L}(\boldsymbol{L} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{L}\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G} \\ &= (\boldsymbol{L} + \boldsymbol{G}^{-1})^{-1} - \boldsymbol{G}\boldsymbol{L}(\boldsymbol{L} + \boldsymbol{G}^{-1})^{-1}(\boldsymbol{L} + \boldsymbol{G}^{-1} - \boldsymbol{L})\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G} \\ &= (\boldsymbol{I} + \boldsymbol{G}\boldsymbol{L})(\boldsymbol{L} + \boldsymbol{G}^{-1})^{-1} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G} \\ &= \boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G}. \end{aligned}$$

Thus an alternative expression for the inverse of $\boldsymbol{C}$ is

$$\boldsymbol{C}^{-1}=\begin{bmatrix}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1} & -(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{ZG}\\ -\boldsymbol{GZ}'\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1} & \boldsymbol{G}-\boldsymbol{GZ}'\boldsymbol{PZG}\end{bmatrix}. \quad \square \tag{B.4}$$

Lemma B.5. *An alternative expression for* $\boldsymbol{P}$ *is given by*

$$\boldsymbol{P}=\boldsymbol{R}^{-1}-\boldsymbol{R}^{-1}\boldsymbol{WC}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1}, \tag{B.5}$$

where $\boldsymbol{W}=[\boldsymbol{X}\ \boldsymbol{Z}]$.

Proof. We show that Eq. (B.5) is equivalent to $\boldsymbol{P}$ as given in Lemma 2.

$$\begin{aligned}\boldsymbol{P}&=\boldsymbol{R}^{-1}-\boldsymbol{R}^{-1}\boldsymbol{WC}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1}\\ &=\boldsymbol{R}^{-1}-\boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1}+\boldsymbol{R}^{-1}\boldsymbol{ZSZ}'\boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1}\\ &\quad+\boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{ZSZ}'\boldsymbol{R}^{-1}-\boldsymbol{R}^{-1}\boldsymbol{ZSZ}'\boldsymbol{R}^{-1}\\ &\quad-\boldsymbol{R}^{-1}\boldsymbol{ZSZ}'\boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{ZSZ}'\boldsymbol{R}^{-1}\\ &=\boldsymbol{R}^{-1}-\boldsymbol{R}^{-1}\boldsymbol{ZSZ}'\boldsymbol{R}^{-1}-(\boldsymbol{R}^{-1}-\boldsymbol{R}^{-1}\boldsymbol{ZSZ}'\boldsymbol{R}^{-1})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1}\\ &\quad+(\boldsymbol{R}^{-1}-\boldsymbol{R}^{-1}\boldsymbol{ZSZ}'\boldsymbol{R}^{-1})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{ZSZ}'\boldsymbol{R}^{-1}\\ &=\boldsymbol{H}^{-1}-(\boldsymbol{R}^{-1}-\boldsymbol{R}^{-1}\boldsymbol{ZSZ}'\boldsymbol{R}^{-1})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{R}^{-1}-\boldsymbol{R}^{-1}\boldsymbol{ZSZ}'\boldsymbol{R}^{-1})\\ &=\boldsymbol{H}^{-1}-\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\\ &=\boldsymbol{P},\end{aligned}$$

where $\boldsymbol{S}=(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1})^{-1}$. □

Appendix C. Linear variance parameterization

The following results hold when the linear mixed model (1) is reparameterized as

$$\begin{aligned}\boldsymbol{y}&=\boldsymbol{X\beta}+\boldsymbol{Zu}+\boldsymbol{e}\\ &\sim N\left(\boldsymbol{X\beta},\boldsymbol{V}\right),\end{aligned}$$

where $\boldsymbol{V}$ is as defined in (5).

The REML log-likelihood function (ignoring constants) for the reparameterized model is

$$l_R(\boldsymbol{\sigma}^2;\boldsymbol{y})=C-\frac{1}{2}\Big\{\log|\boldsymbol{V}|+\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}|+\boldsymbol{y}'\boldsymbol{Py}\Big\},$$

where $\boldsymbol{P}=\boldsymbol{V}^{-1}-\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}$ and $\boldsymbol{\sigma}^2$ contains the variance components.

Result C.1. The score functions for the variance components (including the residual variance) in the model are:

$$U(\sigma_j^2)=\frac{\partial l_R(\boldsymbol{\sigma}^2;\boldsymbol{y})}{\partial\sigma_j^2}=-\frac{1}{2}\mathrm{tr}\left(\frac{\partial\boldsymbol{V}}{\partial\sigma_j^2}\boldsymbol{P}\right)+\boldsymbol{y}'\boldsymbol{P}\frac{\partial\boldsymbol{V}}{\partial\sigma_j^2}\boldsymbol{Py}.$$

Result C.2. The elements of the observed information matrix for the variance parameters are:

$$\begin{aligned}\mathcal{I}_{\mathcal{O}}(\sigma_j^2, \sigma_k^2) &= \frac{\partial^2 l_R(\boldsymbol{\sigma}^2; \boldsymbol{y})}{\partial\sigma_j^2\partial\sigma_k^2} \\ &= -\frac{1}{2}\text{tr}\left(\frac{\partial \boldsymbol{V}}{\partial\sigma_j^2}\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\sigma_k^2}\boldsymbol{P}\right) + \boldsymbol{y}'\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\sigma_j^2}\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\sigma_k^2}\boldsymbol{P}\boldsymbol{y}.\end{aligned}$$

Result C.3. The elements of the expected information matrix for the variance parameters are:

$$\mathcal{I}_{\mathcal{E}}(\sigma_j^2, \sigma_k^2) = \mathrm{E}\left(\frac{\partial^2 l_R(\boldsymbol{\sigma}^2; \boldsymbol{y})}{\partial\sigma_j^2\partial\sigma_k^2}\right) = \frac{1}{2}\text{tr}\left(\frac{\partial \boldsymbol{V}}{\partial\sigma_j^2}\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\sigma_k^2}\boldsymbol{P}\right).$$

Result C.4. The elements of the average information matrix for the variance parameters are:

$$\begin{aligned}\mathcal{I}_{\mathcal{A}}(\sigma_j^2, \sigma_k^2) &= \frac{1}{2}\left[\frac{\partial^2 l_R(\boldsymbol{\sigma}^2; \boldsymbol{y})}{\partial\sigma_j^2\partial\sigma_k^2} + \mathrm{E}\left(\frac{\partial^2 l_R(\boldsymbol{\sigma}^2; \boldsymbol{y})}{\partial\sigma_j^2\partial\sigma_k^2}\right)\right] \\ &= \frac{1}{2}\boldsymbol{y}'\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\sigma_j^2}\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\sigma_k^2}\boldsymbol{P}\boldsymbol{y}.\end{aligned}$$

References

[1] P.M.E. Altham, Improving the precision of estimation by fitting a model, J. R. Stat. Soc. Ser. B 46 (1984) 118–119.

[2] R.L. Anderson, T.A. Bancroft, Statistical Theory in Research, McGraw-Hill, New York, 1952.

[3] S.G. Baker, Regression analysis of grouped survival data with incomplete covariates: non-ignorable missing-data and censoring mechanisms, Biometrics 50 (1994) 821–826.

[4] O. Barndoff-Nielsen, On a formula for the distribution of the maximum likelihood estimator, Biometrika 70 (1983) 343–365.

[5] D.M. Bates, D.G. Watts, Nonlinear Regression Analysis and its Applications, Wiley, New York, 1988.

[6] T.P. Callanan, D.A. Harville, Some new algorithms for computing restricted maximum likelihood estimates of variance components, J. Statist. Comput. Simulation 38 (1991) 239–259.

[7] R.R. Corbeil, R. Searle, A comparison of variance components estimators, Biometrics 32 (1976) 779–791.

[8] R.R. Corbeil, R. Searle, Restricted maximum likelihood (REML) estimation of variance components in the mixed model, Technometrics 18 (1976) 31–38.

[9] D.R. Cox, D.V. Hinkley, Theoretical Statistics, Chapman and Hall, London, 1990.

[10] D.R. Cox, N. Reid, Parameter orthogonality and approximate conditional inference (with discussion), J. R. Stat. Soc. Ser. B 49 (1987) 1–39.

[11] E. Demidenko, Mixed Models Theory and Applications, Wiley, New York, 2004.

[12] A.P. Dempster, L.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B 39 (1977) 1–22.

[13] A.P. Dempster, D.B. Rubin, R.K. Tsukatawa, Estimation in covariance components models, J. Amer. Statist. Assoc. 76 (1981) 341–353.

[14] B. Efron, D.V. Hinkley, Assessing the accuracy of the maximum likelihood estimator: observed versus expected fisher information, Biometrika 65 (1978) 457–487.

[15] G. Fitzmaurice, N. Laird, J. Ware, Applied Longitudinal Analysis, John Wiley and Sons, New York, 2004.

[16] J.L. Foulley, F. Jaffrézic, C. Robert-Granie, EM-REML estimates of covariance parameters in Gaussian mixed models for longitudinal data analysis, Genet. Sel. Evol. 32 (2000) 129–141.

[17] A.R. Gilmour, B.R. Cullis, A.P. Verbyla, Accounting for natural and extraneous variation in the analysis of field experiments, J. Agric. Biol. Environ. Stat. 51 (1997) 269–273.

[18] A.R. Gilmour, B.R. Cullis, S.J. Welham, B.J. Gogel, R. Thompson, An efficient computing strategy for prediction in mixed linear models, Comput. Stat. Data Anal. 44 (2004) 571–586.

[19] A.R. Gilmour, R. Thompson, B.R. Cullis, Average Information REML: an efficient algorithm for variance parameter estimation in linear mixed models, Biometrics 51 (1995) 1440–1450.

[20] A.S. Goldberger, Best linear unbiased prediction in general linear regression, J. Amer. Statist. Assoc. 57 (1962) 369–375.

[21] J.H. Goodnight, Computing MIVQUEO Estimates of Variance Components, Technical Report 105, SAS Institute, Cary, NC, 1978.

[22] R. Hartley, J.N.K. Rao, Maximum likelihood estimation for the mixed analysis of variance model, Biometrika 54 (1967) 93–108.

[23] D.A. Harville, Bayesian inference for variance components using only error contrasts, Biometrika 61 (1974) 383–385.

[24] W. Hemmerle, H. Hartley, Computing maximum likelihood estimates for the linear A.O.V. model using W-transformation, Technometrics 15 (1973) 819–831.

[25] C.R. Henderson, Estimation of genetic parameters (abstract), Ann. Math. Stat. 21 (1950) 309–310.

[26] C.R. Henderson, Estimation of variance and covariance components, Biometrics 9 (1953) 226–252.

[27] C.R. Henderson, O. Kempthorne, S.R. Searle, C.N. Von Krosig, Estimation of environmental and genetic trends from records subject to culling, Biometrics 15 (1959) 192–218.

[28] F. Jaffrézic, I.M.S. White, R. Thompson, Use of the score test as a goodness of fit measure of the covariance structure in genetic analysis of longitudinal data, Genet. Sel. Evol. 35 (2003) 185–198.

[29] R.I. Jennrich, P.F. Sampson, Newton–Raphson and related algorithms for maximum likelihood estimation of variance components, Technometrics 18 (1976) 11–18.

[30] R.I. Jennrich, M. Schluchter, Unbalanced repeated-measures models with structured covariance matrices, Biometrics 42 (1986) 805–820.

[31] J. Jiang, Asymptotic properties of the empirical BLUP and BLUE in mixed linear models, Statist. Sinica 8 (1998) 861–885.

[32] D. Johnson, R. Thompson, Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and a new quasi-Newton procedure, J. Dairy Sci. 78 (1995) 449–456.

[33] M. Kenward, J. Roger, Small sample inferences for fixed effects from restricted maximum likelihood, Biometrics 53 (1997) 983–997.

[34] G.C. Khatri, A note on a MANOVA model applied to problems in growth curves, Ann. Inst. Statist. Math. 18 (1966) 75–78.

[35] N.M. Laird, N. Lange, D. Stram, Maximum likelihood computations with repeated measures: application of the EM algorithm, J. Amer. Statist. Assoc. 82 (1987) 97–105.

[36] N.M. Laird, J.H. Ware, Random-effects models for longitudinal data, Biometrics 38 (1982) 963–974.

[37] J. Lee, J.A. Nelder, Y. Pawitan, Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood, Chapman and Hall, London, 2006.

[38] Y. Lee, J.A. Nelder, Hierarchical generalized linear models (with discussion), J. R. Stat. Soc. Ser. B 58 (1996) 619–678.

[39] Y. Lee, J.A. Nelder, Generalized linear models for the analysis of quality-improvement experiments, Canad. J. Stat. 26 (1998) 95–105.

[40] W.J. Lill, A.C. Gleeson, B.R. Cullis, Relative accuracy of a neighbour model for field trials, J. Agric. Sci. (Camb.) 11 (1988) 339–346.

[41] C.Y. Lin, A.J. McAllister, Monte Carlo comparisons of four methods for estimation of genetic parameters in the univariate case, J. Dairy Sci. 67 (1984) 2389–2398.

[42] M.J. Lindstrom, D.M. Bates, Newton–Raphson and EM algorithms for linear mixed-effects models for repeated measures data, J. Amer. Statist. Assoc. 83 (1988) 1014–1022.

[43] C. Liu, D.B. Rubin, Y. Wu, Parameter expansion to accelerate EM: the PX-EM algorithm, Biometrika 85 (1998) 755–770.

[44] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, London, 2003.

[45] C.E. McCulloch, S.R. Searle, Generalized, Linear and Mixed Models, John Wiley and Sons, New York, 2001.

[46] X. Meng, D.B. Rubin, Using EM to obtain asymptotic variance covariance matrices: SEM algorithm, J. Amer. Statist. Assoc. 86 (1991) 899–909.

[47] X. Meng, D. van Dyk, Fast EM-type implementations for mixed effects models, J. R. Stat. Soc. Ser. B 60 (1998) 559–578.

[48] G. Molenberghs, G. Verbeke, Likelihood ratio, score and Wald tests in a constrained parameter space, Amer. Statist. 61 (2007) 22–27.

[49] C.H. Morrell, Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood, Biometrics 54 (1998) 1560–1568.

[50] H.D. Patterson, R. Thompson, Recovery of inter-block information when block sizes are unequal, Biometrika 58 (1971) 545–554.

[51] Y. Pawitan, Clarendon Press, Oxford, 2001.

[52] J.C. Pinheiro, D.M. Bates, Mixed-effects Models in S and S-plus, Springer, New York, 2000.

[53] R.F. Potthoff, S.N. Roy, A generalized multivariate analysis of variance model useful for growth curve problems, Biometrika 51 (1964) 313–326.

[54] C.R. Rao, Estimation of variance and covariance components – MINQUE theory, J. Multivariate Anal. 1 (1971) 257–275.

[55] C.R. Rao, Linear Statistical Inference and its Applications, second ed., John Wiley and Sons, New York, 1973.

[56] G.K. Robinson, That BLUP is a good thing: the estimation of random effects (with discussion), Statist. Sci. 6 (1991) 15–51.

[57] S.I. Rodriguez-Zas, B.R. Southey, Linear mixed effects models for microarray gene expression data, in: Proceedings of the Seventh World Congress on Genetics Applied to Livestock Production, vol. 16, 2002, pp. 04.

[58] F. Satterthwaite, An approximate distribution of estimates of variance components, Biom. Bull. 2 (1946) 110–114.

[59] R. Searle, An overview of variance component estimation, Metrika 42 (1995) 215–230.

[60] S.R. Searle, Linear Models, John Wiley and Sons, New York, 1971.

[61] S.R. Searle, Matrix Algebra Useful for Statistics, John Wiley and Sons, New York, 1982.

[62] S.R. Searle, G. Casella, C.E. McCulloch, Variance Components, John Wiley and Sons, New York, 1992.

[63] G.R. Seaton, C.S. Haley, S.A. Knott, P.M. Vischer, Qtl expression: rapid and user-friendly mapping of quantitative trait loci in livestock, in: Proceedings of the Seventh World Congress on Genetics Applied to Livestock Production, vol. 28, 2002, pp. 11.

[64] J. Seely, Quadratic subspaces and completeness, Ann. Math. Stat. 42 (1971) 710–721.

[65] S.G. Self, K.-Y. Liang, Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, J. Amer. Statist. Assoc. 82 (1987) 605–610.

[66] M.L. Stein, Interpolation of Spatial Data, Springer, New York, 1999.

[67] W.H. Swallow, J.F. Monahan, Monte Carlo comparison of Anova, MINVQUE, REML, and ML estimators of variance components, Technometrics 26 (1984) 47–57.

[68] R. Thisted, Elements of Statistical Computing, Chapman and Hall, New York, 1988.

[69] G. Verbeke, G. Molenberghs, Linear Mixed Models for Longitudinal Data, Springer, New York, 2000.

[70] G. Verbeke, G. Molenberghs, The use of score tests for inference on variance components, Biometrics 59 (2003) 254–262.

[71] A. Verbyla, A conditional derivation of residual maximum likelihood, Aust. J. Stat. 32 (1990) 227–230.

[72] A.P. Verbyla, B.R. Cullis, A.B. Smith, R. Thompson, S.J. Welham, Mixed Models for Data Analysts, Unpublished manuscript, 2009.

[73] A. Wald, Tests of statistical hypotheses concerning several parameters when the number of observations is large, Trans. Amer. Math. Soc. 54 (1943) 426–482.

[74] S.J. Welham, B.R. Cullis, B.J. Gogel, A.R. Gilmour, R. Thompson, Prediction in linear mixed models, Aust. N. Z. J. Stat. 46 (2004) 325–347.

[75] S. Welham, R. Thompson, Likelihood ratio tests for fixed model terms using residual maximum likelihood, J. R. Stat. Soc. Ser. B 59 (1997) 701–714.

[76] R. Wolfinger, Covariance structure selection in general mixed models, Comm. Statist. Simulation Comput. 22 (1993) 1079–1106.

[77] D.M. Zucker, O. Manor, Small inference for the fixed effects in the mixed linear model, Comput. Statist. Data Anal. 46 (2004) 801–817.