Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT®

Computational Statistics & Data Analysis 44 (2004) 571–586

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

An efficient computing strategy for prediction in mixed linear models

Arthur Gilmour[a,*], Brian Cullis[b], Sue Welham[c,1], Beverley Gogel[d], Robin Thompson[c]

[a] *Orange Agricultural Institute, Orange NSW Australia 2800*
[b] *Wagga Agricultural Institute, Wagga Wagga NSW Australia 2650*
[c] *IACR-Rothamsted, Harpenden, UK AL5 2JQ*
[d] *Queensland Department of Primary Industries, Yeerongpilly, Queensland, Australia 4299*

Received 19 January 2002; received in revised form 19 August 2002

Abstract

After estimation of effects from a linear mixed model, it is often useful to form predicted values for certain factor/variate combinations. This process has been well-defined for linear models, but the introduction of random effects means that a decision has to be made about the inclusion or exclusion of random model terms from the predictions, including the residual error. For spatially correlated data, kriging then becomes prediction from the fitted model. In many cases, the size of the matrices required to calculate predictions and their covariance matrix directly can be prohibitive. An efficient computational strategy for calculating predictions and their standard errors is given, which includes the ability to detect the invariance of predictions to the parameterisation used in the model.

Keywords: REML; BLUP; Linear mixed models; Prediction

1. Introduction

Linear mixed models are a rich and flexible tool for the analysis of data arising in many applications. Recent developments have extended the range of variance

[*] Corresponding author.
E-mail address: arthur.gilmour@agric.nsw.gov.au (A. Gilmour).
[1] Present address: Medical Statistics Unit, London School of Hygiene and Tropical Medicine, Keppel St, London WC1E 7HT.

0167-9473/$ - see front matter © 2002 Elsevier B.V. All rights reserved.
PII: S0167-9473(02)00258-X

models available in many common statistical packages including GenStat (Welham and Thompson, 2000), S-PLUS (Pinheiro and Bates, 2000) and SAS (Littell et al., 1996), as well as specialist packages such as ASReml (Gilmour et al., 1999). It is often desirable to construct predicted values from the effects fitted to explore the relationships established in the analysis. Such predictions may be fitted values from a multiple regression, or summaries such as treatment means in the analysis of designed experiments, fitted curves from the analysis of longitudinal data using cubic smoothing splines (Verbyla et al., 1999) or factor means from a series of trials using factor analysis (Smith et al., 2001). Residual maximum likelihood (REML) estimation of variance parameters in these models is assumed throughout this paper.

Lane and Nelder (1982) describe a general approach for forming predictions in general(ised) linear models. Briefly, their approach involves forming the fitted values for all combinations of the variables in the model, then taking marginal means across the variables not relevant to the current prediction. Their approach has been implemented in GenStat. Some computational limitations with the calculation of the standard errors of predicted values have been recently removed (Lane, 1998). This algorithm however is not generally suitable for use in linear mixed models. An alternative approach suited to the class of balanced linear mixed models with several random terms which can be analysed by ANOVA, is to replace predictions by treatment means. This approach may not be suitable for unbalanced or non-orthogonal data sets. Where random effects are present in the model, a decision must be made about how to treat these terms in prediction, which may differ according to the purpose of a particular prediction. For correlated random effects, information on effects present in the data may be used to predict effects not present in the data set, with prediction standard errors allowing for the extra uncertainty associated with the effect not being observed. The application of this principle to the residual error gives the kriging predictions used in geostatistics.

Welham et al. (2002) is a companion to this paper, in which we outline the principles of prediction in mixed linear models, using four examples which illustrate the need for some flexibility in a prediction algorithm. This paper details the functionality required with an efficient algorithm for the calculation of predictions and their covariance matrix from a linear mixed model.

The paper is arranged as follows. In Section 2 we briefly present some basic results for the linear mixed model and set up the notation for subsequent developments. In particular we consider the mixed model equations and their role in the average information algorithm (Gilmour et al., 1995), as these are used in the prediction calculations. In Section 3 we formally define the prediction process, describe the prediction algorithm and present an outline of the implementation in mixed models software. The issue of estimability and prediction invariance is discussed in Section 4. Different approaches to averaging are addressed in Section 5 and the details of prediction for new observations are given in Section 6. The final section assesses the efficiency of the proposed algorithm.

2. Preliminaries

2.1. Linear mixed model

If $\boldsymbol{y}$ denotes the $n \times 1$ vector of observations, the linear mixed model can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{\tau}$ is the $t \times 1$ vector of fixed effects, $\boldsymbol{X}$ is an $n \times t$ design matrix which associates observations with the appropriate combination of fixed effects, $\boldsymbol{u}$ is the $q \times 1$ vector of random effects, $\boldsymbol{Z}$ is the $n \times q$ design matrix which associates observations with the appropriate combination of random effects, and $\boldsymbol{e}$ is the $n \times 1$ vector of residual errors.

The model (1) is called a linear mixed model or linear mixed-effects model. It is assumed

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

where the covariance matrices $\boldsymbol{G}$ and $\boldsymbol{R}$ for the random effects and residual are functions of parameters γ and $\boldsymbol{\phi}$, respectively. We can then define

$$\boldsymbol{H} = \mathrm{var}(\boldsymbol{y}) = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R}.$$

2.2. REML estimation

Gilmour et al. (1995) describe an algorithm for obtaining REML estimates of the variance parameters in (2), namely $(\gamma', \boldsymbol{\phi}')$, by maximising the residual log-likelihood (Patterson and Thompson, 1971). The residual log-likelihood is the log-likelihood of $\boldsymbol{L}_2'\boldsymbol{y}$, where $\boldsymbol{L}_2$ is an $n \times [n - rank(\boldsymbol{X})]$ matrix of rank $n - rank(\boldsymbol{X})$ with $\boldsymbol{L}_2'\boldsymbol{X} = \boldsymbol{0}$ (Verbyla, 1990). Empirical best linear unbiased predictors of random effects and generalised least squares estimates of fixed effects are obtained as the solution to the mixed model equations (evaluated at the REML estimate of the variance parameters). The mixed model equations are given by

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\tau}} \\ \tilde{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix}. \tag{3}$$

These can be written as

$$\boldsymbol{C}\tilde{\boldsymbol{\beta}} = \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{y}, \tag{4}$$

where $\boldsymbol{C} = \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{W} + \boldsymbol{G}^*$, $\boldsymbol{W} = [\boldsymbol{X}\ \boldsymbol{Z}]$, $\tilde{\boldsymbol{\beta}} = [\hat{\boldsymbol{\tau}}'\ \tilde{\boldsymbol{u}}']'$ is of length $p = t + q$ and

$$\boldsymbol{G}^* = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}^{-1} \end{bmatrix}.$$

For the moment we assume that $\boldsymbol{C}$ is of full rank, i.e. $\boldsymbol{X}$ is of full column rank. Estimation and estimability for the case when $\boldsymbol{X}$ is not full rank will be discussed in Section 4.

The mixed model equations can be solved by absorption and back-substitution of the mixed model matrix

$$\begin{bmatrix} \boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{y} & \boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{W} \\ \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{y} & \boldsymbol{C} \end{bmatrix}.$$

The algorithm presented by Gilmour et al. (1995) is known as an average information (AI) algorithm, so-called as it replaces the expected information in the updating formula for the variance parameters by a matrix which is approximately the average of the observed and expected information matrices. Its convergence properties are similar to the Fisher scoring algorithm but the elements of the AI matrix are much easier, and hence faster, to compute than the corresponding elements of either the observed or expected information matrices.

The AI algorithm (in common with other derivative-based algorithms) is built around the efficient solution of the mixed model equations. For each iteration (3) is solved using the current values for $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$. Gilmour et al. (1995) describes how this is achieved using sparse matrix methods and a modified absorption and backsubstitution routine which avoids calculation of unnecessary terms in $\boldsymbol{C}$ (and $\boldsymbol{C}^{-1}$).

3. The prediction model

We define a prediction to be a linear function of the best linear unbiased predictor of random effects with the best linear unbiased estimate of fixed effects in the model. A prediction is typically formed as the predicted response from an experiment for a subset of explanatory variables at given values, with the remaining explanatory variables in the model being either averaged over, ignored, or taking a specified value. Welham et al. (2002) consider the possible roles of fixed and random model terms in prediction and conclude that while fixed model terms can never be ignored, random terms may be either included (for a conditional prediction) or ignored (to obtain a marginal prediction). In addition, they show that there must also be flexibility in the averaging process which allows for different weighting schemes over factors, or combinations of factors. The algorithm must also be able to recognise aliasing and nesting, and to check for predictions affected by aliasing, i.e. to check whether the predicted value is invariant to the parameterisation used.

3.1. Steps in the prediction process

Before presenting the algorithm it is useful to consider the conceptual steps involved in the prediction process. The four main steps are

(1) Choosing the explanatory variable(s) and their respective values for which predictive margins are required; the variables involved will be referred to as the *classify* set.

(2) Determine which variables should be averaged over to form predictions. The values to be averaged over must also be defined for each variable; the variables involved

will be referred to as the *averaging* set. The combination of the classify set with these averaging variables defines a multiway *hyper-table*. Formally, variables to be evaluated at a single specified value within the prediction, eg. a covariate evaluated at its mean value, can be equivalently included as a member of either the classify or averaging sets.

At this point, there may be some explanatory variables in the model that do not classify the hyper-table. These variables will normally only occur in random terms that are ignored when forming the fitted values.

(3) Determine which terms from the linear mixed model are to be used in forming predictions for each cell in the multiway hyper-table.

(4) Choose the weighting for forming means over each dimension (or combination of dimensions) of the hyper-table.

3.2. Prediction process

Prediction involves forming a linear function of $\tilde{\boldsymbol{\beta}}$. If we denote the vector of predictive margins of interest by $\tilde{\boldsymbol{\pi}}$, then

$$\tilde{\boldsymbol{\pi}} = \boldsymbol{D}\tilde{\boldsymbol{\beta}} \tag{5}$$

say, for some $d \times p$ matrix $\boldsymbol{D}$. It follows that

$$\boldsymbol{D}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \boldsymbol{D}\begin{bmatrix} \hat{\boldsymbol{\tau}} - \boldsymbol{\tau} \\ \tilde{\boldsymbol{u}} - \boldsymbol{u} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \ \boldsymbol{D}\boldsymbol{C}^{-1}\boldsymbol{D}' \right). \tag{6}$$

Consideration of the values required for forming confidence intervals make it clear that it is the prediction error variance, i.e. var$(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$, rather than the variance of the estimator, var$(\tilde{\boldsymbol{\beta}})$, that is usually of interest.

The matrices $\boldsymbol{D}$ and $\boldsymbol{C}$ are often prohibitively large, so that it is not practical to explicitly compute the matrix products involved in the evaluation of $\tilde{\boldsymbol{\pi}}$ and the prediction error variance of $\tilde{\boldsymbol{\pi}}$. It is however instructive to decompose $\boldsymbol{D}$ into its component matrices, where each component matrix relates to a step in the prediction process described in the previous section. We can write $\boldsymbol{D}$ as

$$\boldsymbol{D} = \boldsymbol{A}\boldsymbol{W}_M\boldsymbol{M}\boldsymbol{S}, \tag{7}$$

where

- $\boldsymbol{S}$ (r rows $\times$ p columns) is a binary matrix which $\underline{\text{s}}$elects the elements of $\boldsymbol{\beta}$ which are used to form the predictions for each cell of the hyper-table—this relates to step 3. Note that p is the dimension of $\boldsymbol{\beta}$ and $r \leqslant p$, the number of effects used in forming the fitted values, is in general less than p.
- $\boldsymbol{M}$ ($c \times r$) is a ‘design’ matrix which forms (a portion of) the $\underline{\text{m}}$ultiway hyper-table for the specified combinations of the classify set plus the averaging set—this relates to step 2. Note that c is the number of values in the hyper-table, (usually) equal to the product of the number of combinations in the classify set with the number of combinations in the averaging set.
- $\boldsymbol{W}_M$ ($c \times c$) is a diagonal matrix of weights—this relates to step 4.

- $\boldsymbol{A}$ $(d \times c)$ is a matrix which when combined with $\boldsymbol{W}_M$ averages the multiway table to produce the predictive margins—this relates to steps 1 and 4. Note that d is the number of predicted values, equal to the number of combinations of factor and covariate values in the classify set.

Operationally, the matrices $\boldsymbol{A}$ and $\boldsymbol{W}_M$ may be combined, however it is helpful to keep them separate here to reflect our intention to control the type of averaging of the multiway hyper-table. This will be particularly important for problems in which aliasing has occurred. Lane (1998) discusses this issue and indicates that aliasing may occur either as a result of linear dependencies in the explanatory variables or because of non-representation of some combinations of factor levels in the dataset. The latter may occur by chance or through the intrinsic structure of the data, for example, where locations are nested within regions. Care must be taken in this case to ensure sensible averaging occurs.

At this level, the major difference between our algorithm and the algorithm proposed by Lane and Nelder (1982) is the presence of the matrix $\boldsymbol{S}$ in $\boldsymbol{D}$.

3.3. Computing strategy

One of the major obstacles with the implementation of the Lane and Nelder (1982) algorithm in GenStat has been limits on the size of the model and or dataset for which predictions and associated standard errors can be readily obtained. The following strategy attempts to minimise the computational load by use of sparse matrix methods and judicious formation of $\boldsymbol{D}$ and the matrix of prediction error variances. The strategy has been designed to fit in with the algorithm already used in ASReml but could easily be adapted for other packages.

3.3.1. Initialisation of the component matrices

For full flexibility, the user must be able to specify factor/covariate combinations for which predictions are required (the classify set), combinations of factors/covariates to be averaged over (the averaging set), methods of averaging for each factor (or combination of factors) in the averaging set and model terms to be used when forming predictions. However, given the basic information, sensible default values can be determined to minimise user input. For example, in the terms which define the full multiway hyper-table, sensible default values would be the mean value for covariates, all levels for factors and knot points for spline terms. Note that where a single variable defines several derived terms (e.g. linear and quadratic trend) care must be taken to maintain the link with the underlying variable.

3.3.2. Forming $\boldsymbol{D}$

While we have defined $\boldsymbol{D}$ as the product $\boldsymbol{A}\boldsymbol{W}_M\boldsymbol{M}\boldsymbol{S}$ to highlight the operations involved, we propose forming $\boldsymbol{D}$ directly, one row at a time.

Recall that each row of $\boldsymbol{D}$ relates to a unique combination of the levels of the factors and values of the covariates in the classify set. These rows are successively formed using a modified version of the subroutine which generates the design matrix for the linear mixed model (see 4), $\boldsymbol{W}$. Each row of the prediction design matrix generates one predicted value. Columns corresponding to the predicted combination will be set to the appropriate value (1 for a factor level, specified value for a covariate). Columns corresponding to averaging factors will contain weights dependent on the averaging process (although a slightly different procedure is used for weights depending on data presence, see Section 5). Columns corresponding to model terms ignored in the prediction process will be set to zero and the matrix $\boldsymbol{D}$ is stored in a linklist sparse form.

3.3.3. *Calculation of predictions and prediction error variances*

The major computational issue is the formation of the prediction error variance matrix. In the following we present an approach for simultaneously computing both $\tilde{\boldsymbol{\pi}}$ and the prediction error variance matrix of $\tilde{\boldsymbol{\pi}}$. The approach involves forming an augmented set of mixed model equations, which can be manipulated during the final iteration of the AI algorithm. That is, let $\boldsymbol{Q}$ be the augmented mixed model matrix, given by

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{y} & \boldsymbol{0} & \boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{W} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D} \\ \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{y} & \boldsymbol{D}' & \boldsymbol{C} \end{bmatrix}.$$

Absorption of $\boldsymbol{C}$ gives

$$\boldsymbol{Q}^* = \begin{bmatrix} \boldsymbol{y}'\boldsymbol{P}\boldsymbol{y} & -\tilde{\boldsymbol{\pi}}' \\ -\tilde{\boldsymbol{\pi}} & -\boldsymbol{D}\boldsymbol{C}^{-1}\boldsymbol{D}' \end{bmatrix},$$

where $\boldsymbol{P}=\boldsymbol{R}^{-1}-\boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1}=\boldsymbol{H}^{-1}-\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}$. The absorption is performed using a reordering of the mixed model matrix designed to retain a high degree of sparsity (Gilmour et al., 1995).

It is advantageous to have control over the formation of the elements of $\boldsymbol{D}\boldsymbol{C}^{-1}\boldsymbol{D}'$ since this will be a very large matrix ($d(d+1)/2$ elements) if there are many predicted values (d). For example, where standard errors of differences (SEDs) are required, the full matrix must be calculated. However an ‘average’ SED can be calculated by inserting an extra column (prediction) in $\boldsymbol{D}$, being the sum of the original columns (predictions) of $\boldsymbol{D}$, and then calculating only the diagonal elements of $\boldsymbol{D}\boldsymbol{C}^{-1}\boldsymbol{D}$. The variance of the extra prediction with the variances of the original predictions can be used to calculate an average covariance. This leads to an SED based on the average variance of differences, which for unbalanced situations is not identical to the average of the individual SEDs. Thus, individual SEs and an average SED can be calculated without forming the full covariance matrix. In addition, we allow $\boldsymbol{D}$ to be defined in sections, i.e. as several separate sets of predictions. We do not form the covariance terms in $\boldsymbol{D}\boldsymbol{C}^{-1}\boldsymbol{D}$ between sections and can form average SEDs within sections.

4. Prediction in models not of full rank

There are often situations in which the fixed effects design matrix, $\boldsymbol{X}$, is not of full column rank. These can be classified according to the cause of aliasing:

(1) linear dependencies between explanatory variables due to over parameterisation of factor terms,
(2) no data present for some factor combinations, so that the corresponding effects cannot be estimated,
(3) linear dependencies due to other, usually unexpected, structure in the data.

The first type of aliasing is imposed by the parameterisation chosen and can be determined from the model. The second type of aliasing can be detected when setting up the design matrix for parameter estimation (which may require revision of imposed constraints). The third type can then be detected during absorption of the mixed model matrix. Dependencies (aliasing) can be dealt with in several ways and we wish to check that predictions are invariant to the method used. This can be ensured by checking that the function of parameters being predicted is estimable in the sense defined by Searle (1971, pp. 160,180).

4.1. Parameter estimation

After absorption of the rows of the mixed model matrix associated with the random effects, the fixed effects are estimated as

$$\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}\hat{\boldsymbol{\tau}} = \boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{y}. \tag{8}$$

If $\boldsymbol{X}$ is not full rank, then there is no unique solution to (8). To obtain a solution, say $\hat{\boldsymbol{\tau}}_0$, we compute

$$\hat{\boldsymbol{\tau}}_0 = (\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{y}$$

for some generalised inverse $(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-}$ of $\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}$. We note that $\hat{\boldsymbol{\tau}}_0$ is not an unbiased estimator of $\boldsymbol{\tau}$, since

$$\mathrm{E}(\hat{\boldsymbol{\tau}}_0) = (\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}\boldsymbol{\tau}$$

which in general is not identical to $\boldsymbol{\tau}$, depending on the generalised inverse used.

Since $\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}$ is symmetric, there exists an orthogonal permutation matrix $\boldsymbol{L}$ such that

$$\boldsymbol{L}'\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}\boldsymbol{L} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{12}' & \boldsymbol{A}_{22} \end{bmatrix},$$

where $\boldsymbol{A}_{22}$ is a square matrix of full rank, equal to the rank of $\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}$. Further, we define

$$\boldsymbol{X}^* = [\boldsymbol{X}_1^* \ \boldsymbol{X}_2^*] = \boldsymbol{X}\boldsymbol{L},$$

$$\boldsymbol{\tau}^* = \begin{bmatrix} \boldsymbol{\tau}_1^* \\ \boldsymbol{\tau}_2^* \end{bmatrix} = \boldsymbol{L}'\boldsymbol{\tau}$$

and note

$$\boldsymbol{X}^*\boldsymbol{\tau}^* = \boldsymbol{X}\boldsymbol{\tau}.$$

Hence a convenient choice for $(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^-$ is given by

$$\boldsymbol{L}\begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{22}^{-1} \end{bmatrix}\boldsymbol{L}' \tag{9}$$

giving

$$\hat{\boldsymbol{\tau}}_0 = \boldsymbol{L}\begin{bmatrix} \boldsymbol{0} \\ \hat{\boldsymbol{\tau}}_2^* \end{bmatrix},$$

where

$$\hat{\boldsymbol{\tau}}_2^* = \boldsymbol{A}_{22}^{-1}\boldsymbol{X}_2^{*'}\boldsymbol{H}^{-1}\boldsymbol{y}.$$

This is exactly the procedure followed by ASReml to estimate fixed effects: aliased effects are flagged and reordered to the top of the set of equations, so that the fixed effects estimate produced is $\hat{\boldsymbol{\tau}}_2^*$.

4.2. Estimability

We first consider the case of estimability of functions of fixed effects, as this corresponds to the case considered by Searle (1971). The linear function $\boldsymbol{D}_\tau\boldsymbol{\tau}$ is defined to be estimable (Searle, 1971) if

$$\mathrm{E}(\boldsymbol{D}_\tau\hat{\boldsymbol{\tau}}_0) = \boldsymbol{D}_\tau\boldsymbol{\tau}. \tag{10}$$

Note that estimability in this context implies that the value of $\boldsymbol{D}_\tau\hat{\boldsymbol{\tau}}_0$ is *invariant* to the parameterisation (i.e. the generalised inverse of $\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}$) chosen, and also that expectation is taken over the random effects. We have

$$\mathrm{E}(\boldsymbol{D}_\tau\hat{\boldsymbol{\tau}}_0) = \boldsymbol{D}_\tau\boldsymbol{L}\begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{22}^{-1} \end{bmatrix}\boldsymbol{L}'\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}\boldsymbol{\tau}$$

and we define

$$\boldsymbol{D}_\tau^* = [\boldsymbol{D}_{\tau 1}^* \ \boldsymbol{D}_{\tau 2}^*] = \boldsymbol{D}_\tau\boldsymbol{L}$$

so that $\boldsymbol{D}_\tau\boldsymbol{\tau} = \boldsymbol{D}_\tau^*\boldsymbol{\tau}^*$. Then

$$\begin{aligned}\mathrm{E}(\boldsymbol{D}_\tau\hat{\boldsymbol{\tau}}_0) &= \boldsymbol{D}_\tau\boldsymbol{L}\begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{22}^{-1} \end{bmatrix}\boldsymbol{L}'\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}\boldsymbol{L}\boldsymbol{L}'\boldsymbol{\tau} \\ &= [\boldsymbol{D}_{\tau 1}^* \ \boldsymbol{D}_{\tau 2}^*]\begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{22}^{-1} \end{bmatrix}\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{12}' & \boldsymbol{A}_{22} \end{bmatrix}\boldsymbol{\tau}^*\end{aligned}$$

$$= [\boldsymbol{D}_{\tau 1}^* \;\; \boldsymbol{D}_{\tau 2}^*] \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}' & \boldsymbol{I} \end{bmatrix} \boldsymbol{\tau}^*$$

$$= [\boldsymbol{D}_{\tau 2}^*\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}' \;\; \boldsymbol{D}_{\tau 2}^*] \begin{bmatrix} \boldsymbol{\tau}_1^* \\ \boldsymbol{\tau}_2^* \end{bmatrix}. \tag{11}$$

For $\boldsymbol{D}_\tau \boldsymbol{\tau}$ to be estimable (11) must equal

$$\boldsymbol{D}_\tau \boldsymbol{\tau} = \boldsymbol{D}_\tau^* \boldsymbol{\tau}^* = \boldsymbol{D}_{\tau 1}^* \boldsymbol{\tau}_1^* + \boldsymbol{D}_{\tau 2}^* \boldsymbol{\tau}_2^*$$

and so

$$\boldsymbol{D}_{\tau 1}^* - \boldsymbol{D}_{\tau 2}^* \boldsymbol{A}_{22}^{-1} \boldsymbol{A}_{12}' = \mathbf{0}. \tag{12}$$

The other case to consider is of estimability of a linear function of random effects $\boldsymbol{D}_u \tilde{\boldsymbol{u}}$. It can be shown that $\mathrm{E}(\boldsymbol{D}_u \tilde{\boldsymbol{u}})$ is zero, taking expectation over $\boldsymbol{u}$, essentially because the subset of equations corresponding to $\tilde{\boldsymbol{u}}$ in (3) are full rank and $\boldsymbol{X}\hat{\boldsymbol{\tau}}$ is estimable (Searle, 1971). If $\boldsymbol{D}_\tau \hat{\boldsymbol{\tau}}$ and $\boldsymbol{D}_u \tilde{\boldsymbol{u}}$ are estimable, it follows that the linear function $\boldsymbol{D}_\tau \hat{\boldsymbol{\tau}} + \boldsymbol{D}_u \tilde{\boldsymbol{u}}$ is also estimable.

4.3. Computing strategy for determining estimability

A convenient strategy for computing the estimability criteria (12) is now presented. This strategy has been developed so that it can be easily implemented within the framework devised within Section 3.3. Consider the augmented mixed model matrix $\boldsymbol{Q}$, after reordering and absorption of the random effects, which is given by

$$\boldsymbol{Q}_1 = \begin{bmatrix} \boldsymbol{y}'\boldsymbol{H}^{-1}\boldsymbol{y} & \mathbf{0} & \boldsymbol{y}'\boldsymbol{H}^{-1}\boldsymbol{X}_1^* & \boldsymbol{y}'\boldsymbol{H}^{-1}\boldsymbol{X}_2^* \\ \mathbf{0} & \mathbf{0} & \boldsymbol{D}_{\tau 1}^* & \boldsymbol{D}_{\tau 2}^* \\ \boldsymbol{X}_1^{*'}\boldsymbol{H}^{-1}\boldsymbol{y} & \boldsymbol{D}_{\tau 1}^{*'} & \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{X}_2^{*'}\boldsymbol{H}^{-1}\boldsymbol{y} & \boldsymbol{D}_{\tau 2}^{*'} & \boldsymbol{A}_{12}' & \boldsymbol{A}_{22} \end{bmatrix}.$$

Absorption of the last row pertaining to the $\boldsymbol{\tau}_2^*$ leaves the symmetric matrix

$$\boldsymbol{Q}_1^a = \begin{bmatrix} \boldsymbol{y}'\boldsymbol{P}\boldsymbol{y} & -\hat{\boldsymbol{\tau}}_2^{*'}\boldsymbol{D}_{\tau 2}^{*'} & \boldsymbol{y}'\boldsymbol{H}^{-1}\boldsymbol{X}_1^* - \hat{\boldsymbol{\tau}}_2^{*'}\boldsymbol{A}_{12}' \\ -\boldsymbol{D}_{\tau 2}^*\hat{\boldsymbol{\tau}}_2^* & \boldsymbol{D}_{\tau 2}^*\boldsymbol{A}_{22}^{-1}\boldsymbol{D}_{\tau 2}^{*'} & \boldsymbol{D}_{\tau 1}^* - \boldsymbol{D}_{\tau 2}^*\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}' \\ \boldsymbol{X}_1^{*'}\boldsymbol{H}^{-1}\boldsymbol{y} - \boldsymbol{A}_{12}\hat{\boldsymbol{\tau}}_2^* & \boldsymbol{D}_{\tau 1}^{*'} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{D}_{\tau 2}^{*'} & \boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}' \end{bmatrix}.$$

Since the reordering of the vector $\boldsymbol{\tau}$ into the partition $(\boldsymbol{\tau}_1^{*'} \boldsymbol{\tau}_2^{*'})'$ has been established and implemented during the first iteration, then the criteria for determining estimability (invariance to parameterisation) can be assessed during the same absorption process that determines the vector of predictions and the matrix of prediction variances, i.e. invariant predictions are characterised by elements of $\boldsymbol{D}_{\tau 1}^{*'} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{D}_{\tau 2}^{*'}$ in $\boldsymbol{Q}_1^a$ becoming zero during the absorption process.

5. Issues of averaging

Different averaging schemes may be required when forming the marginal table of predictions from the multiway hyper-table. In many examples, averaging over all cells using equal weights will be desirable. In other cases, specific user-supplied weights will be required for some factors, with equal weighting over others. In either case the weight matrix is defined by multiplying together the fixed weights associated with each margin being averaged. A special case, which deserves attention because it requires a slightly modified algorithm, is the case of averaging only over factor combinations that are present in the data. In this case, the prediction matrix $\boldsymbol{D}$ can be written as

$$\boldsymbol{D} = \boldsymbol{A}_D \boldsymbol{W}_{MD} \boldsymbol{A}_F \boldsymbol{W}_{MF} \boldsymbol{M}\boldsymbol{S},$$

where the averaging step is split into averaging over factors with weighting fixed (without reference to the data, subscript F) and factors with weighting determined by data presence, denoted by subscript D. The first step of averaging over factors with fixed weights ($\boldsymbol{A}_F$) is done to reduce the size of the matrices, before checking data presence on the reduced hyper-table.

A general algorithm for prediction should allow specification of the type of weighting (equal, population, data present, user-supplied) on each of the averaging factors, or on any combination of the averaging factors.

6. Prediction of new observations

In some circumstances, it is desirable to predict new observations. This may include the predicted mean for a new experiment or prediction at new points within the data set and requires an extension to the algorithm. Using the model (1), we can write our predicted value as

$$\boldsymbol{y}_p = \boldsymbol{X}_p\boldsymbol{\tau} + \boldsymbol{Z}_{p1}\boldsymbol{u} + \boldsymbol{Z}_{p2}\boldsymbol{u}_p + \boldsymbol{e}_p, \tag{13}$$

where the subscript p denotes design matrices associated with the predicted values. The vectors $\boldsymbol{u}_p$ and $\boldsymbol{e}_p$ denote random effects not present in the observed data set, but drawn from the same population as $\boldsymbol{u}$ and $\boldsymbol{e}$ with

$$\mathrm{var}\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{u}_p \end{pmatrix} = \boldsymbol{G}_a = \begin{pmatrix} \boldsymbol{G} & \boldsymbol{G}_{op} \\ \boldsymbol{G}_{po} & \boldsymbol{G}_{pp} \end{pmatrix} \tag{14}$$

and

$$\mathrm{var}\begin{pmatrix} \boldsymbol{e} \\ \boldsymbol{e}_p \end{pmatrix} = \boldsymbol{R}_a = \begin{pmatrix} \boldsymbol{R} & \boldsymbol{R}_{op} \\ \boldsymbol{R}_{po} & \boldsymbol{R}_{pp} \end{pmatrix}. \tag{15}$$

Note that $\boldsymbol{G}$ and $\boldsymbol{R}$ are the variance matrices from the observed random effects. We do not include residual errors from the current experiment or a general design matrix for the new residual errors $\boldsymbol{e}_p$ here, but the extension is straightforward. It is shown in Appendix A that the best linear unbiased predictor of $\boldsymbol{y}_p$ is

$$\tilde{\boldsymbol{y}}_p = \boldsymbol{X}_p\hat{\boldsymbol{\tau}} + \boldsymbol{Z}_{p1}\tilde{\boldsymbol{u}} + \boldsymbol{Z}_{p2}\tilde{\boldsymbol{u}}_p + \tilde{\boldsymbol{e}}_p, \tag{16}$$

where $\hat{\boldsymbol{\tau}}$, $\tilde{\boldsymbol{u}}$ are estimated as in Section 4, $\tilde{\boldsymbol{e}}=\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\tau}}-\boldsymbol{Z}\tilde{\boldsymbol{u}}$ are the residuals and $\tilde{\boldsymbol{u}}_p$, $\tilde{\boldsymbol{e}}_p$ are the BLUPs of $\boldsymbol{u}_p$ and $\boldsymbol{e}_p$, namely

$$\tilde{\boldsymbol{u}}_p = E(\boldsymbol{u}_p|\boldsymbol{L}_2'\boldsymbol{y}) = \boldsymbol{G}_{po}\boldsymbol{G}^{-1}\tilde{\boldsymbol{u}}$$

and

$$\tilde{\boldsymbol{e}}_p = E(\boldsymbol{e}_p|\boldsymbol{L}_2'\boldsymbol{y}) = \boldsymbol{R}_{po}\boldsymbol{R}^{-1}\tilde{\boldsymbol{e}},$$

where $\boldsymbol{L}_2$ is the matrix used in defining the residual log-likelihood (see Section 2.2). Thus predictions of future values can be written in terms of the original BLUPs as

$$\tilde{\boldsymbol{y}}_p = \boldsymbol{X}_p\hat{\boldsymbol{\tau}} + (\boldsymbol{Z}_{p1} + \boldsymbol{Z}_{p2}\boldsymbol{G}_{po}\boldsymbol{G}^{-1})\tilde{\boldsymbol{u}} + \boldsymbol{R}_{po}\boldsymbol{R}^{-1}\tilde{\boldsymbol{e}}, \tag{17}$$

where $\boldsymbol{Z}_{p2}$ is non-zero, these estimates of $\tilde{\boldsymbol{u}}_p$ can be obtained by simply augmenting the original set of random effects. The same approach may be used for the residual term, by augmenting the data set with extra observations as in Appendix A. The efficiency of this approach will depend on the structure of the data: it may be particularly inefficient where the observed data set can be treated as a direct product structure but the augmented data set cannot, although in some cases data augmentation may be used to improve the structure (Verbyla and Cullis, 1992). Alternatively, the predict absorption step can be extended to include the residual term, as shown in the next section.

The application of this technique to predict the value and variance of new observations gives the usual kriging results. Note that this is different to interpolation of the surface, which does not take account of uncertainty at the new data points.

6.1. Computation for new observations

After incorporating any new random effects into the original model, we can write the predicted observations as

$$\begin{aligned}\tilde{\boldsymbol{y}}_p &= \boldsymbol{X}_p\hat{\boldsymbol{\tau}} + \boldsymbol{Z}_p\tilde{\boldsymbol{u}} + \boldsymbol{R}_{po}\boldsymbol{R}^{-1}\tilde{\boldsymbol{e}}\\ &= \boldsymbol{W}_p\tilde{\boldsymbol{\beta}} + \boldsymbol{R}_{po}\boldsymbol{R}^{-1}\tilde{\boldsymbol{e}}\\ &= (\boldsymbol{W}_p - \boldsymbol{R}_{po}\boldsymbol{R}^{-1}\boldsymbol{W})\tilde{\boldsymbol{\beta}} + \boldsymbol{R}_{po}\boldsymbol{R}^{-1}\boldsymbol{y},\end{aligned} \tag{18}$$

where $\boldsymbol{Z}_p$=$(\boldsymbol{Z}_{p1}\ \boldsymbol{Z}_{p2})$, $\boldsymbol{u}_p$ now represents the full set of observed and predicted random effects, and $\boldsymbol{W}_p = (\boldsymbol{X}_p\ \boldsymbol{Z}_p)$. Note that since

$$\begin{aligned}\begin{pmatrix}\boldsymbol{R} & \boldsymbol{R}_{op}\\ \boldsymbol{R}_{po} & \boldsymbol{R}_{pp}\end{pmatrix}^{-1} &= \begin{pmatrix}\boldsymbol{R}^{-1} + \boldsymbol{R}^{-1}\boldsymbol{R}_{op}(\boldsymbol{R}^{pp})\boldsymbol{R}_{po}\boldsymbol{R}^{-1} & -\boldsymbol{R}^{-1}\boldsymbol{R}_{op}\boldsymbol{R}^{pp}\\ -\boldsymbol{R}^{pp}\boldsymbol{R}_{po}\boldsymbol{R}^{-1} & \boldsymbol{R}^{pp}\end{pmatrix}\\ &= \begin{pmatrix}\boldsymbol{R}^{oo} & \boldsymbol{R}^{op}\\ \boldsymbol{R}^{po} & \boldsymbol{R}^{pp}\end{pmatrix},\end{aligned}$$

where $\boldsymbol{R}^{pp} = (\boldsymbol{R}_{pp} - \boldsymbol{R}_{po}\boldsymbol{R}^{-1}\boldsymbol{R}_{op})^{-1}$, the predicted observations can be written as

$$\tilde{\boldsymbol{y}}_p = (\boldsymbol{W}_p + (\boldsymbol{R}^{pp})^{-1}\boldsymbol{R}^{po}\boldsymbol{W})\tilde{\boldsymbol{\beta}} - (\boldsymbol{R}^{pp})^{-1}\boldsymbol{R}^{po}\boldsymbol{y}.$$

The predicted observations $\boldsymbol{y}_p$ then have prediction error variance

$$\begin{aligned}\mathrm{var}(\tilde{\boldsymbol{y}}_p - \boldsymbol{y}_p) &= \mathrm{var}(\boldsymbol{W}_p(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \boldsymbol{R}_{po}\boldsymbol{R}^{-1}\tilde{\boldsymbol{e}} - \boldsymbol{e}_p)\\ &= \mathrm{var}((\boldsymbol{W}_p - \boldsymbol{R}_{po}\boldsymbol{R}^{-1}\boldsymbol{W})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \boldsymbol{R}_{po}\boldsymbol{R}^{-1}\boldsymbol{e} - \boldsymbol{e}_p)\\ &= (\boldsymbol{W}_p - \boldsymbol{R}_{po}\boldsymbol{R}^{-1}\boldsymbol{W})\boldsymbol{C}^{-1}(\boldsymbol{W}_p - \boldsymbol{R}_{po}\boldsymbol{R}^{-1}\boldsymbol{W})' + (\boldsymbol{R}^{pp})^{-1}\\ &= (\boldsymbol{W}_p + (\boldsymbol{R}^{pp})^{-1}\boldsymbol{R}^{po}\boldsymbol{W})\boldsymbol{C}^{-1}(\boldsymbol{W}_p + (\boldsymbol{R}^{pp})^{-1}\boldsymbol{R}^{po}\boldsymbol{W})' + (\boldsymbol{R}^{pp})^{-1}\end{aligned}$$

since $\mathrm{cov}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}, \boldsymbol{R}_{po}\boldsymbol{R}^{-1}\boldsymbol{e} - \boldsymbol{e}_p) = \boldsymbol{0}$. The predictions are now written in terms of the parameters $\boldsymbol{\beta}$ and the data $\boldsymbol{y}$, and in the spirit of Section 3.3, we can extend the mixed model matrix and calculate predictions and prediction error variances through an absorption step. The augmented mixed model matrix now becomes

$$\begin{pmatrix} \boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{y} & \boldsymbol{y}'\boldsymbol{R}^{op}(\boldsymbol{R}^{pp})^{-1} & \boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{W}\\ (\boldsymbol{R}^{pp})^{-1}\boldsymbol{R}^{po}\boldsymbol{y} & -(\boldsymbol{R}^{pp})^{-1} & \boldsymbol{D}_p\\ \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{y} & \boldsymbol{D}_p' & \boldsymbol{C}\end{pmatrix},$$

where $\boldsymbol{D}_p = \boldsymbol{W}_p + (\boldsymbol{R}^{pp})^{-1}\boldsymbol{R}^{po}\boldsymbol{W}$. Absorption of $\boldsymbol{C}$ gives

$$\begin{pmatrix} \boldsymbol{y}'\boldsymbol{P}\boldsymbol{y} & -\tilde{\boldsymbol{y}}_p'\\ -\tilde{\boldsymbol{y}}_p & -\mathrm{var}(\tilde{\boldsymbol{y}}_p - \boldsymbol{y}_p)\end{pmatrix}.$$

One advantage of this strategy is that the check for prediction invariance given in Section 4 can still be carried out during the absorption process.

7. Assessment of the proposed algorithm

There are several numerical advantages of this algorithm compared to previous algorithms, such as that implemented in Genstat for prediction in generalised linear models as described by Lane and Nelder (1982) and Lane (1998). The first advantage relates to the memory requirement which is at most 2× the size of $\boldsymbol{D}$, which is held in sparse form as a linked list, plus the size of the variance matrix for the predicted values (or its diagonal if covariances are not required). If memory limitations cause a problem, the prediction can be split into several runs, at the cost of losing some SEDs. Previous algorithms have formed predictions (and variances) for the full hyper-table, which is in general much larger than the number of predicted values, and then derived predicted values from the hyper-table. The second advantage of our algorithm is in the number of computations required to form the predictions, which again is related more to the number of predictions than to the size of the hyper-table. The third advantage is that the algorithm does not explicitly calculate the variance matrix $\boldsymbol{C}^{-1}$, this reduces both the memory requirement and number of computations.

Roughly, for our algorithm, the number of operations required to form the variance matrix of the predictions is

$$s_1 p^3/6 + s_2^2 d(d+1)/2 \times s_1 p^2, \tag{19}$$

where p is the total number of effects in the model, d is the number of predictions formed, s_1 is a scaling factor (<1) depending on the sparsity of the matrix $\boldsymbol{C}$ and s_2 is a scaling factor (<1) depending on the sparsity of the matrix $\boldsymbol{D}$. The sparsity factor s_1 is typically order $1/l$, where l is the largest number of levels for a factor term in the model. The sparsity factor s_2 depends on the structure of the prediction. The first term in (19) is (roughly) the number of operations required to form $\boldsymbol{C}^{-1}$ and the second term is the number of operations required to form $\boldsymbol{D}\boldsymbol{C}^{-1}\boldsymbol{D}'$. Previous algorithms using the full hyper-table would use approximately $p^3/3 + c(c+1)/2 \times p^2$ operations, where c is the number of cells in the hyper-table.

In general, predictions will be required after the model has been determined. Since calculation of the predictions takes place during estimation of the fixed and random effects, this requires a single additional absorption of the augmented mixed model matrix, which includes re-estimation of the model effects. This slight inefficiency is outweighed by the overall advantage of the algorithm.

Acknowledgements

We gratefully acknowledge the ongoing financial assistance of the Grains Research and Development Corporation. We would like to thank Alison Smith for many helpful discussions. IACR receives grant aided support from the Biotechnology and Biological Sciences Research Council.

Appendix A. Estimation of new observations

Consider the augmented model

$$\boldsymbol{y}_a = \boldsymbol{F}\boldsymbol{\psi} + \boldsymbol{W}_a\boldsymbol{\beta}_a + \boldsymbol{e}_a, \tag{A.1}$$

where n_p is the number of new observations, and the augmented structures are defined as

$$\boldsymbol{y}_a = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0}_{n_p} \end{pmatrix}, \quad \boldsymbol{W}_a = \begin{pmatrix} \boldsymbol{X} & \boldsymbol{Z} & \boldsymbol{0} \\ \boldsymbol{X}_p & \boldsymbol{Z}_{p1} & \boldsymbol{Z}_{p2} \end{pmatrix}, \quad \boldsymbol{\beta}_a = \begin{pmatrix} \boldsymbol{\tau} \\ \boldsymbol{u} \\ \boldsymbol{u}_p \end{pmatrix} = \begin{pmatrix} \boldsymbol{\tau} \\ \boldsymbol{u}_a \end{pmatrix},$$

$$\boldsymbol{e}_a = \begin{pmatrix} \boldsymbol{e} \\ \boldsymbol{e}_p \end{pmatrix} \quad \text{and} \quad \boldsymbol{F} = \begin{pmatrix} \boldsymbol{0}_n \\ \boldsymbol{I}_{n_p} \end{pmatrix},$$

with $\text{var}(\boldsymbol{u}_a) = \boldsymbol{G}_a$ and $\text{var}(\boldsymbol{e}_a) = \boldsymbol{R}_a$ defined in Section 6 and $\boldsymbol{\psi}$ is a vector of length n_p of fixed effects, with a separate effect for each new observation.

The augmented model can be re-written as

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{\beta} + \boldsymbol{e},$$

$$\boldsymbol{0} = \boldsymbol{\psi} + \boldsymbol{y}_p.$$

As this is still a linear mixed model, then the predictor $\tilde{\boldsymbol{y}}_p$ will be the BLUP of $\boldsymbol{y}_p$ from this model. The second line also shows that $\tilde{\boldsymbol{y}}_p = -\hat{\boldsymbol{\psi}}$. If it can be shown that the augmented model in Eq. (A.1) is equivalent to the original model, then it follows that $\tilde{\boldsymbol{y}}_p = -\hat{\boldsymbol{\psi}}$ is the BLUP for $\boldsymbol{y}_p$ based on the original data and model.

The expanded mixed model matrix for the augmented data and model becomes

$$\begin{pmatrix} \boldsymbol{y}_a'\boldsymbol{R}_a^{-1}\boldsymbol{y}_a & \boldsymbol{y}_a'\boldsymbol{R}_a^{-1}\boldsymbol{W}_a & \boldsymbol{y}_a'\boldsymbol{R}_a^{-1}\boldsymbol{F} \\ \boldsymbol{W}_a'\boldsymbol{R}_a^{-1}\boldsymbol{y}_a & \boldsymbol{W}_a'\boldsymbol{R}_a^{-1}\boldsymbol{W}_a + \boldsymbol{G}_a^* & \boldsymbol{W}_a'\boldsymbol{R}_a^{-1}\boldsymbol{F} \\ \boldsymbol{F}'\boldsymbol{R}_a^{-1}\boldsymbol{y}_a & \boldsymbol{F}'\boldsymbol{R}_a^{-1}\boldsymbol{W}_a & \boldsymbol{F}'\boldsymbol{R}_a^{-1}\boldsymbol{F} \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{y}_a'\boldsymbol{R}_a^{-1}\boldsymbol{y}_a & \boldsymbol{y}_a'\boldsymbol{R}_a^{-1}\boldsymbol{W}_a & \boldsymbol{y}'\boldsymbol{R}^{op} \\ \boldsymbol{W}_a'\boldsymbol{R}_a^{-1}\boldsymbol{y}_a & \boldsymbol{W}_a'\boldsymbol{R}_a^{-1}\boldsymbol{W}_a + \boldsymbol{G}_a^* & \boldsymbol{W}'\boldsymbol{R}^{op} + \boldsymbol{W}_p'\boldsymbol{R}^{pp} \\ \boldsymbol{R}^{po}\boldsymbol{y} & \boldsymbol{R}^{po}\boldsymbol{W} + \boldsymbol{R}^{pp}\boldsymbol{W}_p & \boldsymbol{R}^{pp} \end{pmatrix},$$

where

$$\boldsymbol{G}_a^* = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}_a^{-1} \end{pmatrix}.$$

Absorbing the last section leaves the matrix

$$\begin{pmatrix} \boldsymbol{y}_a'\boldsymbol{P}_F\boldsymbol{y}_a & \boldsymbol{y}_a'\boldsymbol{P}_F\boldsymbol{W}_a \\ \boldsymbol{W}_a'\boldsymbol{P}_F\boldsymbol{y}_a & \boldsymbol{W}_a'\boldsymbol{P}_F\boldsymbol{W}_a + \boldsymbol{G}_a^* \end{pmatrix},$$

where

$$\boldsymbol{P}_F = \boldsymbol{R}_a^{-1} - \boldsymbol{R}_a^{-1}\boldsymbol{F}(\boldsymbol{F}'\boldsymbol{R}_a^{-1}\boldsymbol{F})^{-1}\boldsymbol{F}'\boldsymbol{R}_a^{-1}$$

$$= \begin{pmatrix} \boldsymbol{R}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}$$

so that the absorbed mixed model matrix becomes

$$\begin{pmatrix} \boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{y} & \boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{Z} & \boldsymbol{0} \\ \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} & \boldsymbol{0} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{oo} & \boldsymbol{G}^{op} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{G}^{po} & \boldsymbol{G}^{pp} \end{pmatrix},$$

where

$$\boldsymbol{G}_a^{-1} = \begin{pmatrix} \boldsymbol{G}^{oo} & \boldsymbol{G}^{op} \\ \boldsymbol{G}^{po} & \boldsymbol{G}^{pp} \end{pmatrix}.$$

The final rows give the required solution $\tilde{\boldsymbol{u}}_p = -(\boldsymbol{G}^{pp})^{-1}\boldsymbol{G}^{po}\tilde{\boldsymbol{u}} = \boldsymbol{G}_{po}\boldsymbol{G}^{-1}\tilde{\boldsymbol{u}}$. Absorbing the last section corresponding to $\boldsymbol{u}_p$ leaves

$$\begin{pmatrix} \boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{y} & \boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{W} \\ \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{y} & \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{W} + \boldsymbol{G}^* \end{pmatrix}$$

i.e. the original mixed model matrix. So, the augmented data approach gives the required estimates of model parameters.

The estimate of $\boldsymbol{\psi}$ is found (by back-substitution) from the last line of the expanded mixed model matrix:

$$\boldsymbol{R}^{po}\boldsymbol{y} = (\boldsymbol{R}^{po}\boldsymbol{W} + \boldsymbol{R}^{pp}\boldsymbol{W}_p)\boldsymbol{\beta}_a + \boldsymbol{R}^{pp}\boldsymbol{\psi}$$

which gives as required

$$\begin{aligned}-\hat{\boldsymbol{\psi}} &= -(\boldsymbol{R}^{pp})^{-1}[\boldsymbol{R}^{po}\boldsymbol{y} - (\boldsymbol{R}^{po}\boldsymbol{W} + \boldsymbol{R}^{pp}\boldsymbol{W}_p)\hat{\boldsymbol{\beta}}_a]\\ &= -(\boldsymbol{R}^{pp})^{-1}[\boldsymbol{R}^{po}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}} - \boldsymbol{Z}\tilde{\boldsymbol{u}}) - \boldsymbol{R}^{pp}(\boldsymbol{X}_p\hat{\boldsymbol{\tau}} + \boldsymbol{Z}_{p1}\tilde{\boldsymbol{u}} + \boldsymbol{Z}_{p2}\tilde{\boldsymbol{u}}_p)]\\ &= -(\boldsymbol{R}^{pp})^{-1}\boldsymbol{R}^{po}\tilde{\boldsymbol{e}} + \boldsymbol{X}_p\hat{\boldsymbol{\tau}} + \boldsymbol{Z}_{p1}\tilde{\boldsymbol{u}} + \boldsymbol{Z}_{p2}\tilde{\boldsymbol{u}}_p\\ &= \boldsymbol{X}_p\hat{\boldsymbol{\tau}} + (\boldsymbol{Z}_{p1} + \boldsymbol{Z}_{p2}\boldsymbol{G}_{po}\boldsymbol{G}^{-1})\tilde{\boldsymbol{u}} + (\boldsymbol{R}_{po})\boldsymbol{R}^{-1}\tilde{\boldsymbol{e}}.\end{aligned}$$

References

Gilmour, A.R., Thompson, R., Cullis, B.R., 1995. Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 51, 1440–1450.

Gilmour, A.R., Cullis, B.R., Welham, S.J., Thompson, R., 1999. ASREML reference manual. Biometric Bulletin 3, NSW Agriculture, Locked Bag 21, Orange, NSW 2800, Australia, 210pp.

Lane, P.W., 1998. Predicting from unbalanced linear or generalised linear models. Proceeding of Compstat 98, Contributed Papers.

Lane, P.W., Nelder, J.A., 1982. Analysis of covariance and standardisation as instances of prediction. Biometrics 38 (3), 613–621.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., 1996. SAS System for Mixed Models. SAS Institute Inc, Cary, NC.

Patterson, H.D., Thompson, R., 1971. Recovery of interblock information when block sizes are unequal. Biometrika 31, 100–109.

Pinheiro, J., Bates, D.M., 2000. Mixed Effects Models in S and S-Plus. Springer, New York.

Searle, S.R., 1971. Linear Models. Wiley, New York.

Smith, A.B., Cullis, B.R., Gilmour, A.R., 2001. The analysis of crop variety evaluation data in Australia. Austral. New Zealand J. Statist. 43, 129–145.

Verbyla, A.P., 1990. A conditional derivation of residual maximum likelihood. Austral. J. Statist. 32, 227–230.

Verbyla, A.P., Cullis, B.R., 1992. The analysis of multistratum and spatially correlated repeated measures data. Biometrics 48, 1015–1032.

Verbyla, A.P., Cullis, B.R., Kenward, M.G., Welham, S.J., 1999. The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). Appl. Statist. 48, 269–311.

Welham, S.J., Thompson, R., 2000. The Guide to Genstat, Part 2: Statistics. Ch. REML analysis of mixed models, VSN International Ltd, Wilkinson House, Jordan Hill Road, Oxford, UK, pp. 413–503.

Welham, S., Cullis, B.R., Gogel, B.J., Gilmour, A.R., Thompson, R., 2002. Prediction in mixed linear models, submitted.