University of Wollongong

Research Online

2016

Linear mixed models for genomic selection

Alison Smith
University of Wollongong

Emi Tanaka
University of Wollongong

Brian Cullis
University of Wollongong

Robin Thompson
Rothamsted Research

Recommended Citation

Linear mixed models for genomic selection

Abstract

We commence by considering the analysis of a single trial. Let y denote the n x 1 vector of (phenotypic) data, where n is the number of plots in the trial.

NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong, Australia

Working Paper

12-16

Linear Mixed Models for Genomic
Selection

Alison Smith, Emi Tanaka, Brian Cullis and Robin Thompson

Statistics for the Australian Grains Industry Technical Report Series

Linear mixed models for genomic selection

Alison Smith [1], Emi Tanaka [1], Brian Cullis[12] and Robin Thompson[3]
National Institute for Applied Statistics and Research Australia
School of Mathematics and Applied Statistics
University of Wollongong
[2] Computational Informatics, Canberra
[3] Rothamsted Research

email: alismith@uow.edu.au

April 22, 2015

1 Statistical model for a single trial

We commence by considering the analysis of a single trial. Let $\boldsymbol{y}$ denote the $n \times 1$ vector of (phenotypic) data, where n is the number of plots in the trial. We assume that m_d genotypes were grown in the trial but that we only have marker data (on r markers) for $m < m_d$ genotypes. Pedigree information is available on $m_p > m_d$ genotypes. Using the results in Appendix I and II we can write the model for the data vector as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z_g}\boldsymbol{u_g} + \boldsymbol{Z_p}\boldsymbol{u_p} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{\tau}$ is a vector of fixed effects with associated design matrix $\boldsymbol{X}$; $\boldsymbol{u_g}$ is the $m \times 1$ vector of random genetic effects corresponding to those genotypes with marker data, and has associated $n \times m$ design matrix $\boldsymbol{Z_g}$; $\boldsymbol{u_p}$ is a vector of non-genetic or peripheral random effects with associated design matrix $\boldsymbol{Z_p}$ and $\boldsymbol{e}$ is the $n \times 1$ vector of residuals. The fixed effects are partitioned as $\boldsymbol{\tau} = (\boldsymbol{\tau_0}^\top, \boldsymbol{\tau_g}^\top)^\top$ where $\boldsymbol{\tau_g}$ is the $(m_d - m) \times 1$ vector of fixed effects corresponding to the genotypes without marker data and we let $\boldsymbol{X_g}$ denote the associated $n \times (m_d - m)$ design matrix. Thus $\boldsymbol{X} = [\boldsymbol{X_0} \ \ \boldsymbol{X_g}]$ where $\boldsymbol{X_0}$ is the design matrix associated with the (non-genetic) fixed effects $\boldsymbol{\tau_0}$.

We assume that the vectors of random effects and residuals are mutually independent, and distributed as multivariate Gaussian, with zero means. The variance matrix for $\boldsymbol{u_p}$ is given by $\boldsymbol{G_p}$ and for the residuals is $\boldsymbol{R}$. blah blah blah

We then consider a simple model for $\boldsymbol{u_g}$ given by

$$\boldsymbol{u_g} = \boldsymbol{u_a} + \boldsymbol{u_e} \tag{2}$$

where the two terms represent the additive and non-additive (or residual) genetic effects. Then we propose that the additive genetic effects be modelled as a linear function of the marker covariates so write

$$\begin{aligned} \boldsymbol{u_a} &= \boldsymbol{M}\boldsymbol{\alpha} + \boldsymbol{u_\epsilon} \\ &= \boldsymbol{u_m} + \boldsymbol{u_\epsilon} \end{aligned} \tag{3}$$

where $\boldsymbol{M}$ is the $m \times r$ matrix of marker covariate data; $\boldsymbol{\alpha}$ is the associated $r \times 1$ vector of random marker effects (regression coefficents) and $\boldsymbol{u_\epsilon}$ is the $m \times 1$ vector of lack of fit effects for the marker regressions. The vector $\boldsymbol{u_m} = \boldsymbol{M}\boldsymbol{\alpha}$ represents the additive genetic effects due to the markers.

Thus the model in equation (1) can be written as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z_g}\boldsymbol{M}\boldsymbol{\alpha} + \boldsymbol{Z_g}\boldsymbol{u_\epsilon} + \boldsymbol{Z_g}\boldsymbol{u_e} + \boldsymbol{Z_p}\boldsymbol{u_p} + \boldsymbol{e} \\ &= \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z_g}\boldsymbol{u_m} + \boldsymbol{Z_g}\boldsymbol{u_\epsilon} + \boldsymbol{Z_g}\boldsymbol{u_e} + \boldsymbol{Z_p}\boldsymbol{u_p} + \boldsymbol{e} \end{aligned} \begin{matrix} (4) \\ (5) \end{matrix}$$

We assume that the variance matrices of random genetic effects are given by

$$
\begin{aligned}
\mathrm{var}\left(\boldsymbol{\alpha}\right) &= \sigma_m^2 \boldsymbol{D} \\
\mathrm{var}\left(\boldsymbol{u_m}\right) &= \sigma_m^2 \boldsymbol{MDM}^\top = \sigma_m^2 \boldsymbol{K} \\
\mathrm{var}\left(\boldsymbol{u_\epsilon}\right) &= \sigma_\epsilon^2 \boldsymbol{A} \\
\mathrm{var}\left(\boldsymbol{u_e}\right) &= \sigma_e^2 \boldsymbol{I_m}
\end{aligned}
$$

where $\boldsymbol{A}$ is the $m \times m$ block of the numerator relationship matrix that relates to the genotypes with marker data; $\boldsymbol{D}$ is an $r \times r$ matrix, often assumed to be the identity matrix $\boldsymbol{I_r}$, and σ_m^2, σ_ϵ^2 and σ_e^2 are the variances for marker effects, marker lack of fit effects and residual genetic effects, respectively. The matrix $\boldsymbol{K} = \boldsymbol{MDM}^\top$ is the $m \times m$ genomic relationship matrix.

The variance matrix for the (total) genetic effects, denoted $\boldsymbol{G_g}$, is therefore given by

$$
\boldsymbol{G_g} = \mathrm{var}\left(\boldsymbol{u_g}\right) = \sigma_m^2 \boldsymbol{K} + \sigma_\epsilon^2 \boldsymbol{A} + \sigma_e^2 \boldsymbol{I_m} \tag{6}
$$

We write $\boldsymbol{G_g} = \boldsymbol{G_g}(\sigma_m^2, \sigma_\epsilon^2, \sigma_e^2)$ to highlight that in the maximal genetic model in which both pedigree and marker information is included, it is a function of three unknown parameters.

2   Statistical model for a multi-environment trial

Here we extend the models for the analysis of a single trial to a series of trials, known as a multi-environment trial (MET). We now let $\boldsymbol{y}$ denote the $n \times 1$ combined vector of data across all trials in the MET. blah blah blah

We assume that the variance matrices of random genetic effects are given by

$$
\begin{aligned}
\mathrm{var}\left(\boldsymbol{\alpha}\right) &= \boldsymbol{\Sigma_m} \otimes \boldsymbol{D} \\
\mathrm{var}\left(\boldsymbol{u_\epsilon}\right) &= \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{A} \\
\mathrm{var}\left(\boldsymbol{u_e}\right) &= \boldsymbol{\Sigma_e} \otimes \boldsymbol{I_m}
\end{aligned}
$$

where the matrices $\boldsymbol{D}$ and $\boldsymbol{A}$ are as defined previously. The matrices $\boldsymbol{\Sigma_m}$, $\boldsymbol{\Sigma_\epsilon}$ and $\boldsymbol{\Sigma_e}$ are $t \times t$ symmetric positive (semi)-definite matrices ane will be referred to as the between environment marker, marker lack of fit and residual genetic variance matrices. Finally, the variance matrix, $\boldsymbol{G_g}$ for the total genetic effects is given by

$$
\mathrm{var}\left(\boldsymbol{u_g}\right) = \boldsymbol{\Sigma_m} \otimes \boldsymbol{K} + \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{A} + \boldsymbol{\Sigma_e} \otimes \boldsymbol{I_m} \tag{7}
$$

We write $\mathrm{var}\left(\boldsymbol{u_g}\right) = \boldsymbol{G_g} = \boldsymbol{G_g}(\boldsymbol{\Sigma_m}, \boldsymbol{\Sigma_\epsilon}, \boldsymbol{\Sigma_e})$ to highlight that in the maximal genetic model in which both pedigree and marker information is included, it is a function of three matrices of unknown parameters. We have found that the Factor Analytic form provides a useful form for the component matrices. In this case we write

$$
\boldsymbol{\Sigma_s} = \boldsymbol{\Lambda_s}\boldsymbol{\Lambda_s}^\top + \boldsymbol{\Psi_s} \tag{8}
$$

for $s \in (m, \epsilon, e)$.

3   Appendix I

We consider the case where there are m_d genotypes with phenotypic data, but there is pedigree information available on $m_p > m_d$ genotypes. Without loss of generality, we consider the analysis for a single site and we exclude the random peripheral (non-genetic) effects so write the linear mixed model for the $n \times 1$ data vector $\boldsymbol{y}$ as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z_g}\boldsymbol{u_g} + \boldsymbol{e} \tag{9}$$

where $\boldsymbol{\tau}$ is a vector of fixed effects with associated design matrix $\boldsymbol{X}$; $\boldsymbol{u_g}$ is the $m_p \times 1$ vector of genetic effects with associated $n \times m_p$ design matrix $\boldsymbol{Z_g}$ and $\boldsymbol{e}$ is the vector of residuals.

We write the genetic effects as $\boldsymbol{u_g} = (\boldsymbol{u_{g_1}^\top}, \boldsymbol{u_{g_2}^\top})^\top$ where $\boldsymbol{u_{g_1}}$ and $\boldsymbol{u_{g_2}}$ represent the genetic effects for genotypes without and with phenotypic data, respectively. The design matrix is therefore given by $\boldsymbol{Z_g} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{Z_{g_2}} \end{bmatrix}$ where $\boldsymbol{0}$ is an $n \times (m_p - m_d)$ matrix of zeros. The genetic variance matrix and its inverse are partitioned conformably as

$$\text{var}\,(\boldsymbol{u_g}) = \boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}_{11} & \boldsymbol{G}_{12} \\ \boldsymbol{G}_{21} & \boldsymbol{G}_{22} \end{bmatrix} \quad \text{with} \quad \boldsymbol{G}^{-1} = \begin{bmatrix} \boldsymbol{G}^{11} & \boldsymbol{G}^{12} \\ \boldsymbol{G}^{21} & \boldsymbol{G}^{22} \end{bmatrix} \tag{10}$$

The MME for the model in equation (9) are given by

$$\begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{R}^{-1} \boldsymbol{X} & \boldsymbol{0} & \boldsymbol{X}^\top \boldsymbol{R}^{-1} \boldsymbol{Z_{g_2}} \\ \boldsymbol{0} & \boldsymbol{G}^{11} & \boldsymbol{G}^{12} \\ \boldsymbol{Z_{g_2}^\top} \boldsymbol{R}^{-1} \boldsymbol{X} & \boldsymbol{G}^{21} & \boldsymbol{Z_{g_2}^\top} \boldsymbol{R}^{-1} \boldsymbol{Z_{g_2}} + \boldsymbol{G}^{22} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\tau}} \\ \tilde{\boldsymbol{u}}_{g_1} \\ \tilde{\boldsymbol{u}}_{g_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{R}^{-1} \boldsymbol{y} \\ \boldsymbol{0} \\ \boldsymbol{Z_{g_2}^\top} \boldsymbol{R}^{-1} \boldsymbol{y} \end{bmatrix} \tag{11}$$

From the second equation in (11) we have that

$$\tilde{\boldsymbol{u}}_{g_1} = -(\boldsymbol{G}^{11})^{-1} \boldsymbol{G}^{12} \tilde{\boldsymbol{u}}_{g_2} \tag{12}$$

and substituting this into the third equation in (11) yields the reduced set of MME given by

$$\begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{R}^{-1} \boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{R}^{-1} \boldsymbol{Z_{g_2}} \\ \boldsymbol{Z_{g_2}^\top} \boldsymbol{R}^{-1} \boldsymbol{X} & \boldsymbol{Z_{g_2}^\top} \boldsymbol{R}^{-1} \boldsymbol{Z_{g_2}} + \boldsymbol{G}_{22}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\tau}} \\ \tilde{\boldsymbol{u}}_{g_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{R}^{-1} \boldsymbol{y} \\ \boldsymbol{Z_{g_2}^\top} \boldsymbol{R}^{-1} \boldsymbol{y} \end{bmatrix} \tag{13}$$

Therefore, instead of working with the linear mixed model of equation (9), in which the vector of genetic effects, $\boldsymbol{u_g}$, is of length m_p and corresponds to all genotypes in the pedigree, we could use the model commensurate with the MME in equation (13), namely

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z_{g_2}}\boldsymbol{u_{g_2}} + \boldsymbol{e} \tag{14}$$

In this model the vector of genetic effects, $\boldsymbol{u_{g_2}}$, is of length m_d and corresponds only to those genotypes grown in the trial, that is, those genotypes with phenotypic data.

Then we would obtain the E-BLUPs of the genetic effects for genotypes with data via solution of the MME in equation (13) and the genetic effects for genotypes without data

using equation (12). Note that we propose the form of the model in equation (14) for ease of illustration of the concepts presented in this paper. When the variance matrix $\boldsymbol{G}$ involves the numerator relationship matrix, and when, as is typically the case, the majority of genotypes with data are non-parental genotypes, then it is computationally more efficient to use the model as in equation (9) with MME as in equation (11). This is due to the fact that the block of the inverse of the numerator relationship matrix that relates to non-parental genotypes is diagonal (see Cullis et al., 2014).

4   Appendix II

We consider the case where there are m_d genotypes with phenotypic data, but we are only interested in $m < m_d$ of these genotypes. For example, parental genotypes may have been grown in the field trial but may not be of interest, or, we may not have marker data for all of the genotypes grown in the trial. In order to preserve the spatial structure of the trial, we choose not to remove any phenotypic data but instead exclude effects from the genetic model. Without loss of generality, we consider the analysis for a single site and we exclude the random peripheral (non-genetic) effects so write the linear mixed model for the $n \times 1$ data vector $\boldsymbol{y}$ as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z_g}\boldsymbol{u_g} + \boldsymbol{e} \tag{15}$$

where $\boldsymbol{\tau}$ is a vector of fixed effects with associated design matrix $\boldsymbol{X}$; $\boldsymbol{u_g}$ is the $m_d \times 1$ vector of genetic effects with associated $n \times m_d$ design matrix $\boldsymbol{Z_g}$ and $\boldsymbol{e}$ is the vector of residuals.

We write the fixed effects as $\boldsymbol{\tau} = (\boldsymbol{\tau_0}^\top, \boldsymbol{\tau_g}^\top)^\top$ where $\boldsymbol{\tau_g}$ is the $(m_d - m) \times 1$ vector of fixed effects corresponding to the genotypes to be excluded and we let $\boldsymbol{X_g}$ denote the associated $n \times (m_d - m)$ design matrix. Thus $\boldsymbol{X} = [\boldsymbol{X_0} \ \ \boldsymbol{X_g}]$ where $\boldsymbol{X_0}$ is the design matrix associated with the (non-genetic) fixed effects $\boldsymbol{\tau_0}$.

In an analogous manner we write the genetic effects as $\boldsymbol{u_g} = (\boldsymbol{u_{g_1}}^\top, \boldsymbol{u_{g_2}}^\top)^\top$ where $\boldsymbol{u_{g_1}}$ is the $(m_d - m) \times 1$ vector of genetic effects corresponding to the genotypes to be excluded and $\boldsymbol{u_{g_2}}$ is the $m \times 1$ vector of genetic effects of interest. The design matrix is therefore given by $\boldsymbol{Z_g} = \begin{bmatrix} \boldsymbol{X_g} & \boldsymbol{Z_{g_2}} \end{bmatrix}$. The genetic variance matrix and its inverse are partitioned conformably as

$$\text{var}\,(\boldsymbol{u_g}) = \boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}_{11} & \boldsymbol{G}_{12} \\ \boldsymbol{G}_{21} & \boldsymbol{G}_{22} \end{bmatrix} \quad \text{with} \quad \boldsymbol{G}^{-1} = \begin{bmatrix} \boldsymbol{G}^{11} & \boldsymbol{G}^{12} \\ \boldsymbol{G}^{21} & \boldsymbol{G}^{22} \end{bmatrix} \tag{16}$$

5 Appendix III

The MME for the model in equation (15) are given by

$$
\begin{bmatrix}
\boldsymbol{X_0^\top R^{-1} X_0} & \boldsymbol{X_0^\top R^{-1} Z_{g_2}} & \boldsymbol{X_0^\top R^{-1} X_g} & \boldsymbol{X_0^\top R^{-1} X_g} \\
\boldsymbol{Z_{g_2}^\top R^{-1} X_0} & \boldsymbol{Z_{g_2}^\top R^{-1} Z_{g_2} + G^{22}} & \boldsymbol{Z_{g_2}^\top R^{-1} X_g + G^{21}} & \boldsymbol{Z_{g_2}^\top R^{-1} X_g} \\
\boldsymbol{X_g^\top R^{-1} X_0} & \boldsymbol{X_g^\top R^{-1} Z_{g_2} + G^{12}} & \boldsymbol{X_g^\top R^{-1} X_g + G^{11}} & \boldsymbol{X_g^\top R^{-1} X_g} \\
\boldsymbol{X_g^\top R^{-1} X_0} & \boldsymbol{X_g^\top R^{-1} Z_{g_2}} & \boldsymbol{X_g^\top R^{-1} X_g} & \boldsymbol{X_g^\top R^{-1} X_g}
\end{bmatrix}
\begin{bmatrix}
\boldsymbol{\hat{\tau}_0} \\
\boldsymbol{\tilde{u}_{g_2}} \\
\boldsymbol{\tilde{u}_{g_1}} \\
\boldsymbol{\hat{\tau}_g}
\end{bmatrix} =
$$

$$
\begin{bmatrix}
\boldsymbol{X_0^\top R^{-1} y} \\
\boldsymbol{Z_{g_2}^\top R^{-1} y} \\
\boldsymbol{X_g^\top R^{-1} y} \\
\boldsymbol{X_g^\top R^{-1} y}
\end{bmatrix}
\tag{17}
$$

Absorbing the equation for $\boldsymbol{\hat{\tau}_g}$ gives

$$
\begin{bmatrix}
\boldsymbol{X_0^\top S X_0} & \boldsymbol{X_0^\top S Z_{g_2}} & \boldsymbol{0} \\
\boldsymbol{Z_{g_2}^\top S X_0} & \boldsymbol{Z_{g_2}^\top S Z_{g_2} + G^{22}} & \boldsymbol{G^{21}} \\
\boldsymbol{0} & \boldsymbol{G^{12}} & \boldsymbol{G^{11}}
\end{bmatrix}
\begin{bmatrix}
\boldsymbol{\hat{\tau}_0} \\
\boldsymbol{\tilde{u}_{g_2}} \\
\boldsymbol{\tilde{u}_{g_1}}
\end{bmatrix} =
\begin{bmatrix}
\boldsymbol{X_0^\top S y} \\
\boldsymbol{Z_{g_2}^\top S y} \\
\boldsymbol{0}
\end{bmatrix}
\tag{18}
$$

where $\boldsymbol{S} = \boldsymbol{R^{-1}} - \boldsymbol{R^{-1} X_g} \left(\boldsymbol{X_g^\top R^{-1} X_g} \right)^{-1} \boldsymbol{X_g^\top R^{-1}}$. Thus, in a similar manner to Appendix I, the third equation in (18) gives

$$
\boldsymbol{\tilde{u}_{g_1}} = -(\boldsymbol{G^{11}})^{-1} \boldsymbol{G^{12}} \boldsymbol{\tilde{u}_{g_2}}
\tag{19}
$$

and substituting this into the second equation in (17) yields the reduced set of MME, after absorbing $\boldsymbol{\hat{\tau}_g}$, given by

$$
\begin{bmatrix}
\boldsymbol{X_0^\top S X_0} & \boldsymbol{X_0^\top S Z_{g_2}} \\
\boldsymbol{Z_{g_2}^\top S X_0} & \boldsymbol{Z_{g_2}^\top S Z_{g_2} + G_{22}^{-1}}
\end{bmatrix}
\begin{bmatrix}
\boldsymbol{\hat{\tau}_0} \\
\boldsymbol{\tilde{u}_{g_2}}
\end{bmatrix} =
\begin{bmatrix}
\boldsymbol{X_0^\top S y} \\
\boldsymbol{Z_{g_2}^\top S y}
\end{bmatrix}
\tag{20}
$$

Therefore, instead of working with the linear mixed model of equation (15), in which the vector of random genetic effects, $\boldsymbol{u_g}$, is of length m_d and corresponds to all genotypes grown in the trial, that is, all genotypes with phenotypic data, we could use the model commensurate with the MME in equation (20), namely

$$
\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Z_{g_2} u_{g_2}} + \boldsymbol{e}
\tag{21}
$$

In this model the vector of random genetic effects, $\boldsymbol{u_{g_2}}$, is of length m and corresponds only to those genotypes of interest, for example, those with marker data. Additionally, the model includes fixed effects, $\boldsymbol{\tau_g}$, corresponding to the genotypes to be excluded.

5 Appendix III

We consider the case where the number of markers is much larger than the number of genotypes with marker data. It is therefore computaionally efficient to fit the linear mixed model using the form given in equation (5) rather than equation (4). If we require E-BLUPs and associated PEVs for the marker effects $\boldsymbol{\alpha}$ it is convenient to expand equation (5) to include both $\boldsymbol{u_m}$ and $\boldsymbol{\alpha}$. We therefore write the model as

$$
\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Z_g^* u_m^*} + \boldsymbol{e^*}
\tag{22}
$$

5 Appendix III

where $\boldsymbol{Z_g}^* = [\boldsymbol{Z_g} \ \boldsymbol{0}]$ and $\boldsymbol{0}$ is an $n \times r$ matrix of zeros; $\boldsymbol{u_m}^* = (\boldsymbol{u_m}^\top, \ \boldsymbol{\alpha}^\top)^\top$ and $\boldsymbol{e}^* = \boldsymbol{Z_g} \boldsymbol{u_\epsilon} + \boldsymbol{Z_g} \boldsymbol{u_e} + \boldsymbol{Z_p} \boldsymbol{u_p} + \boldsymbol{e}$ with associated variance matrix $\boldsymbol{R}^* = \sigma_\epsilon^2 \boldsymbol{Z_g} \boldsymbol{A} \boldsymbol{Z_g}^\top + \sigma_e^2 \boldsymbol{Z_g} \boldsymbol{Z_g}^\top + \boldsymbol{Z_p} \boldsymbol{G_p} \boldsymbol{Z_p}^\top + \boldsymbol{R}$.

The variance matrix for $\boldsymbol{u_m}^*$ is given by

$$\mathrm{var} \begin{pmatrix} \boldsymbol{u_m} \\ \boldsymbol{\alpha} \end{pmatrix} = \boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}_{11} & \boldsymbol{G}_{12} \\ \boldsymbol{G}_{21} & \boldsymbol{G}_{22} \end{bmatrix} = \sigma_m^2 \begin{bmatrix} \boldsymbol{K} & \boldsymbol{M} \boldsymbol{D} \\ \boldsymbol{D} \boldsymbol{M}^\top & \boldsymbol{D} \end{bmatrix} \tag{23}$$

The inverse is partitioned conformably as

$$\boldsymbol{G}^{-1} = \begin{bmatrix} \boldsymbol{G}^{11} & \boldsymbol{G}^{12} \\ \boldsymbol{G}^{21} & \boldsymbol{G}^{22} \end{bmatrix} \tag{24}$$

The MME for the model in equation (22) are given by

$$\begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{R}^{*-1} \boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{R}^{*-1} \boldsymbol{Z_g} & \boldsymbol{0} \\ \boldsymbol{Z_g}^\top \boldsymbol{R}^{*-1} \boldsymbol{X} & \boldsymbol{Z_g}^\top \boldsymbol{R}^{*-1} \boldsymbol{Z_g} + \boldsymbol{G}^{11} & \boldsymbol{G}^{12} \\ \boldsymbol{0} & \boldsymbol{G}^{21} & \boldsymbol{G}^{22} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\tau}} \\ \tilde{\boldsymbol{u}}_m \\ \tilde{\boldsymbol{\alpha}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{R}^{*-1} \boldsymbol{y} \\ \boldsymbol{Z_g}^\top \boldsymbol{R}^{*-1} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} \tag{25}$$

Absorbing the equation for $\hat{\boldsymbol{\tau}}$ gives

$$\begin{bmatrix} \boldsymbol{Z_g}^\top \boldsymbol{S}^* \boldsymbol{Z_g} + \boldsymbol{G}^{11} & \boldsymbol{G}^{12} \\ \boldsymbol{G}^{21} & \boldsymbol{G}^{22} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{u}}_m \\ \tilde{\boldsymbol{\alpha}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z_g}^\top \boldsymbol{S}^* \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} \tag{26}$$

where $\boldsymbol{S}^* = \boldsymbol{R}^{*-1} - \boldsymbol{R}^{*-1} \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{R}^{*-1} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{R}^{*-1}$. Thus, in a similar manner to Appendix I, the second equation in (26) gives

$$\begin{aligned} \tilde{\boldsymbol{\alpha}} &= -(\boldsymbol{G}^{22})^{-1} \boldsymbol{G}^{21} \tilde{\boldsymbol{u}}_m \\ &= (\boldsymbol{G}^{22})^{-1} \boldsymbol{G}^{22} \boldsymbol{G}_{21} \boldsymbol{G}_{11}^{-1} \tilde{\boldsymbol{u}}_m \\ &= \boldsymbol{D} \boldsymbol{M}^\top \boldsymbol{K}^{-1} \tilde{\boldsymbol{u}}_m \end{aligned} \tag{27}$$

Also note that, substituting this into the first equation in (26) gives the reduced set of MME given by

$$\left(\boldsymbol{Z_g}^\top \boldsymbol{S}^* \boldsymbol{Z_g} + (\sigma_m^2 \boldsymbol{K})^{-1} \right) \tilde{\boldsymbol{u}}_m = \boldsymbol{Z_g}^\top \boldsymbol{S}^* \boldsymbol{y} \tag{28}$$

which is identical to the equation for $\tilde{\boldsymbol{u}}_m$ that would be achieved using the standard (non-expanded) form of the model in equation (5). Thus we can obtain E-BLUPs of $\boldsymbol{\alpha}$ by fitting the model as in equation (5) to obtain $\tilde{\boldsymbol{u}}_m$, then using equation (27).

In terms of PEVs, we let $\boldsymbol{C}$ denote the coefficient matrix of the MME in equation (26), and partition as for $\boldsymbol{G}$. The PEV for $\tilde{\boldsymbol{\alpha}}$ is then given by $\boldsymbol{C}^{22}$ where this is the partition of the inverse of $\boldsymbol{C}$ corresponding to $\tilde{\boldsymbol{\alpha}}$. Similarly, the PEV for $\tilde{\boldsymbol{u}}_m$ is given by $\boldsymbol{C}^{11}$. Using standard results for the inverse of partitioned matrices, we have that

$$\begin{aligned} \mathrm{var}\,(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) &= \boldsymbol{C}_{22}^{-1} + \boldsymbol{C}_{22}^{-1} \boldsymbol{C}_{21} \boldsymbol{C}^{11} \boldsymbol{C}_{12} \boldsymbol{C}_{22}^{-1} \\ &= (\boldsymbol{G}^{22})^{-1} + \boldsymbol{G}_{21} \boldsymbol{G}_{11}^{-1} \boldsymbol{C}^{11} \boldsymbol{G}_{11}^{-1} \boldsymbol{G}_{12} \\ &= \sigma_m^2 \left(\boldsymbol{D} - \boldsymbol{D} \boldsymbol{M}^\top \boldsymbol{K}^{-1} \boldsymbol{M} \boldsymbol{D} \right) + \\ &\qquad \boldsymbol{D} \boldsymbol{M}^\top \boldsymbol{K}^{-1} \mathrm{var}\,(\tilde{\boldsymbol{u}}_m - \boldsymbol{u}_m) \boldsymbol{K}^{-1} \boldsymbol{M} \boldsymbol{D} \end{aligned} \tag{29}$$

where $\operatorname{var}(\tilde{\boldsymbol{u}}_{\boldsymbol{m}} - \boldsymbol{u}_{\boldsymbol{m}}) = (\boldsymbol{Z}_{\boldsymbol{g}}^{\top}\boldsymbol{S}^{*}\boldsymbol{Z}_{\boldsymbol{g}} + (\sigma_m^2 \boldsymbol{K})^{-1})^{-1}$ is the PEV for $\tilde{\boldsymbol{u}}_{\boldsymbol{m}}$ as would be obtained using the standard (non-expanded) form of the model in equation (5). Thus we can obtain the PEV for $\tilde{\boldsymbol{\alpha}}$ by fitting the model as in equation (5) to obtain the PEV of $\tilde{\boldsymbol{u}}_{\boldsymbol{m}}$, then using equation (29).

References

CULLIS, B. R., JEFFERSON, P., THOMPSON, R., & SMITH, A. B. (2014). Factor analytic and reduced animal models for the investigation of additive genotype by environment interaction in outcrossing plant species with application to a *pinus radiata* breeding program. *Theoretical and Applied Genetics.* **127**, 2193–2210.