Computational Methods for Multilevel Modelling

Douglas M. Bates
Department of Statistics
University of Wisconsin – Madison *

José C. Pinheiro
Bell Laboratories
Lucent Technologies

Abstract

A multilevel mixed-effects model has random effects at each of several nested levels of grouping of the observed responses. We may use these, for example, when modelling observations taken over time on students who are grouped into classes that are grouped into schools that are grouped into districts. If each of the distributions of the random effects is Gaussian and if the disturbance term at the lowest level of grouping is also Gaussian it is straightforward to define a likelihood for the fixed effects and the parameters defining the random effects distribution. We show that by expressing the random effects distribution in terms of relative precision factors and using matrix decompositions, this likelihood can be profiled and can be compactly expressed. The same decompositions provide rapid evaluation of the profiled log-restricted-likelihood for REML estimation.

The conditional distribution of the random effects given the data can be derived from the decomposed matrices. From this a compact and rapidly evaluated expression for the EM iterations can be derived. Reasonable starting estimates for the relative precision factors can be derived from the design alone. These starting estimates, refined by a moderate number of EM iterations, provide excellent starting values for a Newton-Raphson or quasi-Newton optimization of the log-likelihood or the log-restricted-likelihood. The methods we describe extend easily to models with non-spherical distributions for the within-group errors and to nonlinear multilevel models. *Key words and phrases:* mixed-effects models, EM algorithm,

maximum likelihood, restricted maximum likelihood

1 Introduction

We consider computational methods for Gaussian multilevel mixed-effects models as described, for example, in Longford (1993) or Goldstein (1995). These models are used with data where the individual observations are grouped at one or more hierarchical levels. For example, we may wish to model observations on students who are grouped into classes that are grouped into schools that are grouped into school districts. If, at

*This research was supported by the National Science Foundation through grant DMS-9704349.

each level in the grouping hierarchy, the experimental units that are observed constitute a sample from the population about which we wish to make inferences, we model the effect of each unit at each level as a random effect. A model including random effects for the individual units as well as overall fixed effects is called a mixed-effects model.

Our computational methods are suitable for maximum likelihood or maximum restricted likelihood estimation of the parameters in linear mixed-effects models with Gaussian distributions for the random effects at each level. These techniques can also form the basis for parameter estimation methods for general linear mixed-effects models or for nonlinear mixed-effects models.

1.1 Description of the model

For a single level of grouping, the linear mixed-effects model described by Laird and Ware (1982) expresses the n_i-dimensional response vector $\boldsymbol{y}_i$ for the ith unit as

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, M \\ \boldsymbol{b}_i &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \end{aligned} \tag{1}$$

where $\boldsymbol{\beta}$ is the p-dimensional vector of *fixed effects*, $\boldsymbol{b}_i$ is the q-dimensional vector of *random effects*, $\boldsymbol{X}_i$ (of size $n_i \times p$) and $\boldsymbol{Z}_i$ (of size $n_i \times q$) are known fixed-effects and random-effects regressor matrices, and $\boldsymbol{\epsilon}_i$ is the n_i-dimensional *within-group error* vector with a spherical Gaussian distribution. The assumption $\mathrm{Var}(\boldsymbol{\epsilon}_i) = \sigma^2\boldsymbol{I}$ can be relaxed as shown in §5.1.

The random effects for the ith unit, $\boldsymbol{b}_i$, are also assumed to have a Gaussian distribution but with a general positive-definite variance-covariance matrix $\boldsymbol{\Sigma}$.

In some of the multilevel modelling literature, notably in Goldstein (1995), the model (1) is called a "two-level model" because there are two levels of random variation: $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$. In other references this model would be described as having one level of random effects in the model. We will adopt the latter convention and count the "levels" in a multilevel model as the number of levels of nested random effects.

In a multilevel model with two levels of random effects the n_{ij}-dimensional vector of responses for the jth level-2 unit nested within the ith level-1 unit is written

$$\begin{aligned} \boldsymbol{y}_{ij} &= \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{i,j}\boldsymbol{b}_i + \boldsymbol{Z}_{ij}\boldsymbol{b}_{ij} + \boldsymbol{\epsilon}_{ij} \quad i = 1, \ldots, M \quad j = 1, \ldots, M_i \\ \boldsymbol{b}_i &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_1), \quad \boldsymbol{b}_{ij} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_2), \quad \boldsymbol{\epsilon}_{ij} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \end{aligned} \tag{2}$$

The regressor matrices $\boldsymbol{X}_{ij}$ (of size $n_{ij} \times p$), $\boldsymbol{Z}_{i,j}$ (of size $n_{ij} \times q_1$) and $\boldsymbol{Z}_{ij}$ (of size $n_{ij} \times q_2$) correspond to the fixed effects $\boldsymbol{\beta}$ and the first- and second-level random effects $\boldsymbol{b}_i$ and $\boldsymbol{b}_{ij}$.

Extensions to an arbitrary number of levels of random effects follow the same general pattern. For example, with three levels of random effects the response for the kth level-3 unit within the jth level-2 unit within the ith level-1 unit will be written

$$\begin{gathered} \boldsymbol{y}_{ijk} = \boldsymbol{X}_{ijk}\boldsymbol{\beta} + \boldsymbol{Z}_{i,jk}\boldsymbol{b}_i + \boldsymbol{Z}_{ij,k}\boldsymbol{b}_{ij} + \boldsymbol{Z}_{ijk}\boldsymbol{b}_{ijk} + \boldsymbol{\epsilon}_{ijk} \\ i = 1, \ldots, M \quad j = 1, \ldots, M_i \quad k = 1, \ldots, M_{ij} \\ \boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_1),\ \boldsymbol{b}_{ij} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_2),\ \boldsymbol{b}_{ijk} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_3),\ \boldsymbol{\epsilon}_{ijk} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \end{gathered}$$

Note that the distinction between, say, the kth horizontal section of the regressor matrix for the level-2 random effect $\boldsymbol{b}_{ij}$, written $\boldsymbol{Z}_{ij,k}$, and the jkth horizontal section of the regressor matrix for the level-1 random effect $\boldsymbol{b}_i$, written $\boldsymbol{Z}_{i,jk}$, is the position of the comma in the subscripts.

1.2 The variance-covariance of the random effects

In a model with Q nested levels of random effects, the variance-covariance matrices $\boldsymbol{\Sigma}_q,\ q = 1, \ldots, Q$ of the random effects must be symmetric and at least positive semi-definite. We consider only positive definite $\boldsymbol{\Sigma}_q$ because any indefiniteness can be removed by re-expressing the model in terms of a random-effects vector of lower dimension.

Within the set of positive-definite $\boldsymbol{\Sigma}_q$ we may impose further constraints such as requiring that $\boldsymbol{\Sigma}_q$ be diagonal or that it be a multiple of the identity or that it have a particular structure such as compound symmetry.

Whatever structure we assume for $\boldsymbol{\Sigma}_q$ we will always express $\boldsymbol{\Sigma}_q$ as a function of an unconstrained parameter vector $\boldsymbol{\theta}_q$ as described in Pinheiro and Bates (1996). By doing this we can ensure that the optimization of the likelihood can be expressed as an unconstrained optimization.

We extend the techniques of Pinheiro and Bates (1996) by expressing the conversion from $\boldsymbol{\theta}_q$ to $\boldsymbol{\Sigma}_q$ is two stages. First, we write

$$\boldsymbol{\Sigma}_q = \sigma^2 \boldsymbol{D}_q$$

where σ^2 is the variance of the components of $\boldsymbol{\epsilon}_i$. The matrix $\boldsymbol{D}_q$ is thus a relative or scaled variance-covariance matrix for the random effects. This re-expression of $\boldsymbol{\Sigma}_q$ does not change the form of the model but, as shown in §2, it does provide a more convenient expression of the likelihood or restricted likelihood for the model.

The next stage is to use a "square-root" factor of the relative precision matrix $\boldsymbol{D}_q^{-1}$. This *relative precision factor*, $\boldsymbol{\Delta}_q$, is any matrix such that

$$\boldsymbol{D}_q^{-1} = \boldsymbol{\Delta}_q' \boldsymbol{\Delta}_q$$

Such a $\boldsymbol{\Delta}_q$ always exists but does not have to be unique. We could, for example, use the Cholesky factor (Thisted, 1988, §3.3) of $\boldsymbol{D}_q^{-1}$ or the transpose of the inverse of the Cholesky factor of $\boldsymbol{D}_q$.

The parameter vector $\boldsymbol{\theta}_q$ defines $\boldsymbol{\Delta}_q$ through some unconstrained parameterization from which $\boldsymbol{D}_q$ and $\boldsymbol{\Sigma}_q$ are subsequently defined. Especially in deriving the EM iterations (§3.4) we can work with $\boldsymbol{\Delta}_q$ directly so the parameterization used is not an issue. When we do need to work in terms of $\boldsymbol{\theta}_q$ one of the parameterizations considered in Pinheiro and Bates (1996) could be used to define $\boldsymbol{\Delta}_q$ from $\boldsymbol{\theta}_q$.

1.3 An example: variability in IC manufacturing

To illustrate the multilevel model described in this section and the methodology presented in the following sections, we present an example from integrated circuit (IC) manufacturing.

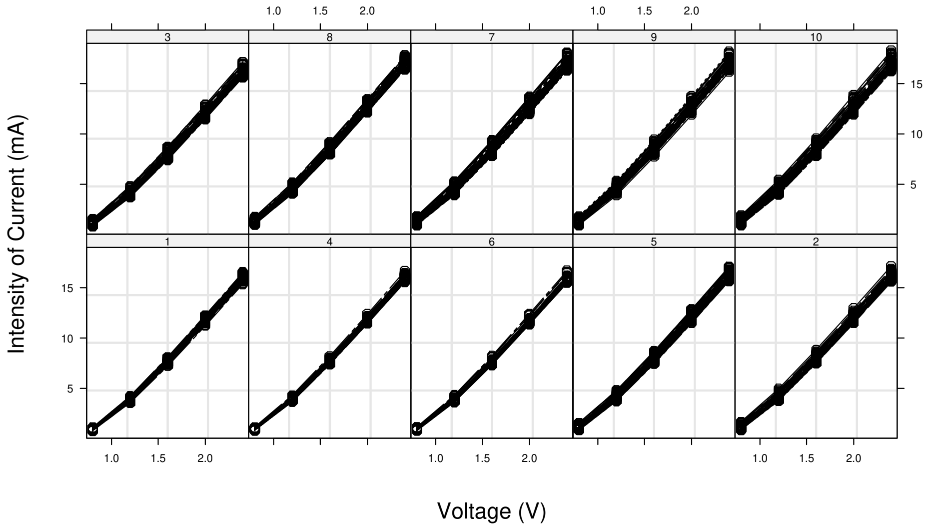

Figure 1: Current versus voltage curves for each of 67 sites within 10 wafers. Each panel presents data for the wafer whose number is given above the panel. The panels are ordered starting from the lower left by increasing maximum current intensity. There are 67 lines, one for each site, on each panel.

In an experiment conducted at the Microelectronics Division of Lucent Technologies to study different sources of variability in the manufacturing of analog MOS circuits, the intensity of current (in milli-Amperes) at 0.8, 1.2, 1.6, 2.0, and 2.4 Volts was measured on $80\mu\text{m}\times 0.6\mu\text{m}$ n-channel devices. Measurements were made on 10 wafers, each subdivided into 67 sites containing one device. The data are presented on Figure 1, where each panel represents a different wafer and each curve on a panel represents a different site.

Two levels of nesting are present in these data: *wafer* and *site within wafer*. The main objective of the experiment was to construct an empirical model for simulating the behavior of similar circuits.

In Figure 1 it appears that current could be modelled as a quadratic function of voltage. Preliminary analyses indicate that random effects are needed to account for the wafer-to-wafer variability of the intercept and the linear terms as well as for the site-to-site variability of the intercept. The corresponding multilevel model for the intensities of current in the jth site within the ith wafer is expressed, for $i = 1, \ldots, 10$

and $j = 1, \ldots, 67$, as

$$\underbrace{\begin{bmatrix} y_{ij_1} \\ y_{ij_2} \\ y_{ij_3} \\ y_{ij_4} \\ y_{ij_5} \end{bmatrix}}_{\boldsymbol{y}_{ij}} = \underbrace{\begin{bmatrix} 1 & 0.8 & 0.8^2 \\ 1 & 1.2 & 1.2^2 \\ 1 & 1.6 & 1.6^2 \\ 1 & 2.0 & 2.0^2 \\ 1 & 2.4 & 2.4^2 \end{bmatrix}}_{\boldsymbol{X}_{ij}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} 1 & 0.8 \\ 1 & 1.2 \\ 1 & 1.6 \\ 1 & 2.0 \\ 1 & 2.4 \end{bmatrix}}_{\boldsymbol{Z}_{i,j}} \underbrace{\begin{bmatrix} b_{i_1} \\ b_{i_2} \end{bmatrix}}_{\boldsymbol{b}_i} + \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}}_{\boldsymbol{Z}_{ij}} \underbrace{\begin{bmatrix} b_{ij_1} \end{bmatrix}}_{\boldsymbol{b}_{ij}} + \underbrace{\begin{bmatrix} \epsilon_{ij_1} \\ \epsilon_{ij_2} \\ \epsilon_{ij_3} \\ \epsilon_{ij_4} \\ \epsilon_{ij_5} \end{bmatrix}}_{\boldsymbol{\epsilon}_{ij}}$$

$$\begin{bmatrix} b_{i_1} \\ b_{i_2} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{1_{11}} & \Sigma_{1_{12}} \\ \Sigma_{1_{12}} & \Sigma_{1_{22}} \end{bmatrix} \right), \quad b_{ij_1} \sim \mathcal{N}(0, \Sigma_2), \quad \epsilon_{ij_k} \sim \mathcal{N}(0, \sigma^2)$$

In this example $Q = 2$, $M = 10$, $M_i = 67$, $n_{ij} = 5$, $q_1 = 2$, $q_2 = 1$, $\boldsymbol{\theta}_1$ is of dimension 3, and $\boldsymbol{\theta}_2$ is of dimension 1.

2 Expressing the likelihood

2.1 Single level of random effects

The likelihood function for model (1) can be written

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2 | \boldsymbol{y}) &= \prod_{i=1}^{M} p(\boldsymbol{y}_i | \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) \\ &= \prod_{i=1}^{M} \int p(\boldsymbol{y}_i | \boldsymbol{b}_i, \boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{b}_i | \boldsymbol{\theta}, \sigma^2)\, d\boldsymbol{b}_i \\ &= \prod_{i=1}^{M} \frac{1}{\sqrt{(2\pi\sigma^2)^{n_i} |\boldsymbol{D}|}} \times \\ &\qquad \int \frac{\exp\left[\frac{-1}{2\sigma^2}\left(\|\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i\|^2 + \boldsymbol{b}_i'\boldsymbol{D}^{-1}\boldsymbol{b}_i\right)\right]}{(2\pi\sigma^2)^{q/2}}\, d\boldsymbol{b}_i \end{aligned} \tag{3}$$

The expression $\|\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i\|^2 + \boldsymbol{b}_i'\boldsymbol{D}^{-1}\boldsymbol{b}_i$ in the exponent within the integral has the form of a penalized residual sum-of-squares. The term $\|\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i\|^2$ is exactly the residual sum of squares for the ith unit, and the additional term, $\boldsymbol{b}_i'\boldsymbol{D}^{-1}\boldsymbol{b}_i$, can be viewed as a "penalty" that inhibits the size of the random-effects vector $\boldsymbol{b}_i$. Using a relative precision factor $\boldsymbol{\Delta}$ we can write the penalty term in the form of a residual sum-of-squares as

$$\boldsymbol{b}_i'\boldsymbol{D}^{-1}\boldsymbol{b}_i = \|\boldsymbol{\Delta}\boldsymbol{b}_i\|^2 = \|\boldsymbol{0} - \boldsymbol{0}\boldsymbol{\beta} - \boldsymbol{\Delta}\boldsymbol{b}_i\|^2$$

This re-expression of the penalty is sometimes called a "pseudo-data" representation because it is equivalent to augmenting the observations and the regressor matrices with q new rows that look like additional observations. Writing

$$\tilde{\boldsymbol{y}}_i = \begin{bmatrix} \boldsymbol{y}_i \\ \boldsymbol{0} \end{bmatrix} \quad \tilde{\boldsymbol{X}}_i = \begin{bmatrix} \boldsymbol{X}_i \\ \boldsymbol{0} \end{bmatrix} \quad \tilde{\boldsymbol{Z}}_i = \begin{bmatrix} \boldsymbol{Z}_i \\ \boldsymbol{\Delta} \end{bmatrix}$$

the expression in the exponent becomes

$$\begin{aligned}\|\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i\|^2 + \boldsymbol{b}_i'\boldsymbol{D}^{-1}\boldsymbol{b}_i &= \|\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i\|^2 + \|\boldsymbol{0} - \boldsymbol{0}\boldsymbol{\beta} - \boldsymbol{\Delta}\boldsymbol{b}_i\|^2 \\ &= \|\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta} - \tilde{\boldsymbol{Z}}_i\boldsymbol{b}_i\|^2\end{aligned}$$

Having made the exponent expression look like a residual sum-of-squares we can employ standard numerical techniques for least squares problems. If we form the orthogonal-triangular decomposition (Thisted, 1988, §3.1)

$$\tilde{\boldsymbol{Z}}_i = \boldsymbol{Q}_{(i)} \begin{bmatrix}\boldsymbol{R}_{11(i)} \\ \boldsymbol{0}\end{bmatrix}$$

where $\boldsymbol{Q}_{(i)}$ is a $(n_i + q) \times (n_i + q)$ orthogonal matrix and $\boldsymbol{R}_{11(i)}$ is an upper-triangular $q \times q$ matrix, then the properties of orthogonal matrices ensure that

$$\begin{aligned}\left\|\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta} - \tilde{\boldsymbol{Z}}_i\boldsymbol{b}_i\right\|^2 &= \left\|\boldsymbol{Q}_{(i)}'\left(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta} - \tilde{\boldsymbol{Z}}_i\boldsymbol{b}_i\right)\right\|^2 \\ &= \left\|\boldsymbol{c}_{1(i)} - \boldsymbol{R}_{10(i)}\boldsymbol{\beta} - \boldsymbol{R}_{11(i)}\boldsymbol{b}_i\right\|^2 + \left\|\boldsymbol{c}_{0(i)} - \boldsymbol{R}_{00(i)}\boldsymbol{\beta}\right\|^2\end{aligned}$$

where the $q \times p$ matrix $\boldsymbol{R}_{10(i)}$, the $n_i \times p$ matrix $\boldsymbol{R}_{00(i)}$, the q-vector $\boldsymbol{c}_{1(i)}$ and n_i-vector $\boldsymbol{c}_{0(i)}$ are defined by

$$\begin{bmatrix}\boldsymbol{R}_{10(i)} \\ \boldsymbol{R}_{00(i)}\end{bmatrix} = \boldsymbol{Q}_{(i)}'\tilde{\boldsymbol{X}}_i \quad \text{and} \quad \begin{bmatrix}\boldsymbol{c}_{1(i)} \\ \boldsymbol{c}_{0(i)}\end{bmatrix} = \boldsymbol{Q}_{(i)}'\tilde{\boldsymbol{y}}_i$$

Furthermore, if $\boldsymbol{\Sigma}$ is positive definite, as we require, then $\boldsymbol{\Delta}$ is non-singular and hence $\boldsymbol{R}_{11(i)}$ is also non-singular.

Another way of thinking of this decomposition is as the orthogonal-triangular (QR) decomposition of an augmented matrix

$$\begin{bmatrix}\boldsymbol{Z}_i & \boldsymbol{X}_i & \boldsymbol{y}_i \\ \boldsymbol{\Delta} & \boldsymbol{0} & \boldsymbol{0}\end{bmatrix} = \boldsymbol{Q}_{(i)} \begin{bmatrix}\boldsymbol{R}_{11(i)} & \boldsymbol{R}_{10(i)} & \boldsymbol{c}_{1(i)} \\ \boldsymbol{0} & \boldsymbol{R}_{00(i)} & \boldsymbol{c}_{0(i)}\end{bmatrix} \tag{4}$$

where the reduction to triangular form is halted after the first q columns. (The peculiar numbering scheme for the submatrices and subvectors is designed to allow easy extension to more than one level of random effects as seen in §2.4 and §2.5.)

The calculation of the decomposition in (4) is straightforward, efficient, and numerically stable. Standard software such as Linpack (Dongarra, Bunch, Moler and Stewart, 1979) or LAPACK (Anderson, Bai, Bischoff, Demmel, Dongarra, DuCroz, Greenbaum, Hammarling, McKenney, Ostrouchov and Sorensen, 1994) can be used.

Returning to the integral in (3) we can now remove a constant factor and reduce it to

$$\begin{aligned}&\int \frac{\exp\left[\frac{-1}{2\sigma^2}\left(\|\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i\|^2 + \boldsymbol{b}_i'\boldsymbol{D}^{-1}\boldsymbol{b}_i\right)\right]}{(2\pi\sigma^2)^{q/2}}\, d\boldsymbol{b}_i \\ &= \exp\left[\frac{\|\boldsymbol{c}_{0(i)} - \boldsymbol{R}_{00(i)}\boldsymbol{\beta}\|^2}{-2\sigma^2}\right] \int \frac{\exp\left[\frac{-1}{2\sigma^2}\left(\|\boldsymbol{c}_{1(i)} - \boldsymbol{R}_{10(i)}\boldsymbol{\beta} - \boldsymbol{R}_{11(i)}\boldsymbol{b}_i\|^2\right)\right]}{(2\pi\sigma^2)^{q/2}}\, d\boldsymbol{b}_i\end{aligned} \tag{5}$$

Because $\boldsymbol{R}_{11(i)}$ is non-singular we can perform a change of variable to $\boldsymbol{\phi}_i = (\boldsymbol{c}_{1(i)} - \boldsymbol{R}_{10(i)}\boldsymbol{\beta} - \boldsymbol{R}_{11(i)}\boldsymbol{b}_i)/\sigma$ with the differential $d\boldsymbol{\phi}_i = \sigma^{-q} \operatorname{abs}|\boldsymbol{R}_{11(i)}|\, d\boldsymbol{b}_i$ and write the integral as

$$\begin{aligned} \int \frac{\exp\left[\frac{-1}{2\sigma^2}\left(\|\boldsymbol{c}_{1(i)} - \boldsymbol{R}_{10(i)}\boldsymbol{\beta} - \boldsymbol{R}_{11(i)}\boldsymbol{b}_i\|^2\right)\right]}{(2\pi\sigma^2)^{q/2}}\, d\boldsymbol{b}_i & \\ = \frac{1}{\operatorname{abs}|\boldsymbol{R}_{11(i)}|} \int \frac{\exp\left(-\|\boldsymbol{\phi}_i\|^2/2\right)}{(2\pi)^{q/2}}\, d\boldsymbol{\phi}_i & \\ = \operatorname{abs}|\boldsymbol{R}_{11(i)}|^{-1} & \end{aligned} \tag{6}$$

Since, by construction, $\boldsymbol{R}_{11(i)}$ is upper-triangular, its determinant is simply the product of its diagonal elements.

Substituting (6) into (5) into (3) provides the likelihood as

$$L(\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2|\boldsymbol{y}) = \prod_{i=1}^{M} \frac{\exp\left(\frac{\|\boldsymbol{c}_{0(i)} - \boldsymbol{R}_{00(i)}\boldsymbol{\beta}\|^2}{-2\sigma^2}\right)}{\sqrt{(2\pi\sigma^2)^{n_i}\,|\boldsymbol{D}|}} \operatorname{abs}|\boldsymbol{R}_{11(i)}|^{-1}$$

which is now in the form of a regression model for the fixed effects $\boldsymbol{\beta}$. Forming another orthogonal-triangular decomposition

$$\begin{bmatrix} \boldsymbol{R}_{00(1)} & \boldsymbol{c}_{0(1)} \\ \vdots & \vdots \\ \boldsymbol{R}_{00(M)} & \boldsymbol{c}_{0(M)} \end{bmatrix} = \boldsymbol{Q}_0 \begin{bmatrix} \boldsymbol{R}_{00} & \boldsymbol{c}_0 \\ \boldsymbol{0} & \boldsymbol{c}_{-1} \end{bmatrix} \tag{7}$$

and noting that $1/\sqrt{|\boldsymbol{D}|} = \operatorname{abs}|\boldsymbol{\Delta}|$ reduces this to

$$\begin{aligned} L(&\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) \\ &= \left(2\pi\sigma^2\right)^{-N/2} \exp\left(\frac{\|\boldsymbol{c}_{-1}\|^2 + \|\boldsymbol{c}_0 - \boldsymbol{R}_{00}\boldsymbol{\beta}\|^2}{-2\sigma^2}\right) \prod_{i=1}^{M} \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}|}{|\boldsymbol{R}_{11(i)}|}\right) \end{aligned} \tag{8}$$

where $N = \sum_{i=1}^{M} n_i$ is the total number of observations.

2.2 The profiled log-likelihood for $\boldsymbol{\theta}$

With the likelihood expressed as (8) we can derive explicit expressions for the optimal values of $\boldsymbol{\beta}$ and σ^2 conditional on a value of $\boldsymbol{\theta}$. We assume $\boldsymbol{R}_{00}$ is non-singular, in which case the optimal value of $\boldsymbol{\beta}$ satisfies

$$\boldsymbol{R}_{00}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \boldsymbol{c}_0 \tag{9}$$

Substituting this value in (8) and taking the logarithm provides the profiled log-likelihood

$$\begin{aligned} \ell(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) &= \log L(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) \\ &= -\frac{N\log(2\pi\sigma^2)}{2} + \frac{\|\boldsymbol{c}_{-1}\|^2}{-2\sigma^2} + \sum_{i=1}^{M} \log\operatorname{abs}\left(\frac{|\boldsymbol{\Delta}|}{|\boldsymbol{R}_{11(i)}|}\right) \end{aligned}$$

which is maximized with respect to σ^2 by $\widehat{\sigma^2}(\boldsymbol{\theta}) = \|\boldsymbol{c}_{-1}\|^2/N$

We can now write the profiled log-likelihood as a function of $\boldsymbol{\theta}$ alone as

$$\begin{aligned}\ell(\boldsymbol{\theta}|\boldsymbol{y}) &= \log L(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}, \widehat{\sigma^2}(\boldsymbol{\theta})|\boldsymbol{y}) \\ &= \mathit{const} - N\log\|\boldsymbol{c}_{-1}\| + \sum_{i=1}^{M}\log\operatorname{abs}\left(\frac{|\boldsymbol{\Delta}|}{|\boldsymbol{R}_{11(i)}|}\right)\end{aligned} \tag{10}$$

2.3 Restricted log-likelihood as a function of $\boldsymbol{\theta}$ alone

The *restricted likelihood* (Harville, 1976) is often preferred to the likelihood when defining an estimator for $\boldsymbol{\theta}$. One way of writing the restricted likelihood is

$$L_R(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) = \int L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2|\boldsymbol{y})\, d\boldsymbol{\beta}$$

which reduces to

$$L_R(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) = \left(2\pi\sigma^2\right)^{(N-p)/2}\exp\left(\frac{\|\boldsymbol{c}_{-1}\|^2}{-2\sigma^2}\right)\operatorname{abs}|\boldsymbol{R}_{00}|^{-1}\prod_{i=1}^{M}\operatorname{abs}\left(\frac{|\boldsymbol{\Delta}|}{|\boldsymbol{R}_{11(i)}|}\right)$$

using (8) and the same change-of-variable technique used to obtain (6).

Converting to the log-restricted-likelihood

$$\begin{aligned}&\ell_R(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) = \\ &\quad -\frac{N-p}{2}\log(2\pi\sigma^2) - \frac{\|\boldsymbol{c}_{-1}\|^2}{2\sigma^2} - \log\operatorname{abs}|\boldsymbol{R}_{00}| + \sum_{i=1}^{M}\log\operatorname{abs}\left(\frac{|\boldsymbol{\Delta}|}{|\boldsymbol{R}_{11(i)}|}\right)\end{aligned}$$

provides the conditional estimate $\widehat{\sigma^2_R}(\boldsymbol{\theta}) = \|\boldsymbol{c}_{-1}\|^2/(N-p)$ for σ^2 from which we obtain the profiled log-restricted-likelihood

$$\begin{aligned}\ell_R(\boldsymbol{\theta}|\boldsymbol{y}) &= \ell_R(\boldsymbol{\theta}, \widehat{\sigma^2_R}(\boldsymbol{\theta})) \\ &= \mathit{const} - (N-p)\log\|\boldsymbol{c}_{-1}\| - \log\operatorname{abs}|\boldsymbol{R}_{00}| + \sum_{i=1}^{M}\log\operatorname{abs}\left(\frac{|\boldsymbol{\Delta}|}{|\boldsymbol{R}_{11(i)}|}\right)\end{aligned}$$

2.4 Two levels of random effects

The likelihood for a model with two levels of random effects is defined as in (3) but integrating over both levels of random effects

$$\begin{aligned}&L(\boldsymbol{\beta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma^2|\boldsymbol{y}) = \\ &\quad \prod_{i=1}^{M}\int\prod_{j=1}^{M_i}\left[\int p(\boldsymbol{y}_{ij}|\boldsymbol{b}_{ij}, \boldsymbol{b}_i, \boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{b}_{ij}|\boldsymbol{\theta}_2, \sigma^2)\, d\boldsymbol{b}_{ij}\right] p(\boldsymbol{b}_i|\boldsymbol{\theta}_1, \sigma^2)\, d\boldsymbol{b}_i\end{aligned} \tag{11}$$

As with the single level of random effects, we can simplify the integrals in (11) if we augment the $\boldsymbol{Z}_{ij}$ matrices with $\boldsymbol{\Delta}_2$ and form orthogonal-triangular decompositions of these augmented arrays. This allows us to evaluate the inner integrals. To evaluate the outer integrals we iterate this process.

That is, we first form and decompose the arrays

$$\begin{bmatrix} \boldsymbol{Z}_{ij} & \boldsymbol{Z}_{i,j} & \boldsymbol{X}_{ij} & \boldsymbol{y}_{ij} \\ \boldsymbol{\Delta}_2 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} = \boldsymbol{Q}_{(ij)} \begin{bmatrix} \boldsymbol{R}_{22(ij)} & \boldsymbol{R}_{21(ij)} & \boldsymbol{R}_{20(ij)} & \boldsymbol{c}_{2(ij)} \\ \boldsymbol{0} & \boldsymbol{R}_{11(ij)} & \boldsymbol{R}_{10(ij)} & \boldsymbol{c}_{1(ij)} \end{bmatrix} \qquad i = 1, \dots, M \quad j = 1, \dots, M_i \tag{12}$$

The matrix $\boldsymbol{R}_{22(ij)}$ will be an upper-triangular matrix of dimension $q_2 \times q_2$. The other arrays in the first row of the decomposition in (12) are used only if the conditional estimates of $\boldsymbol{\beta}$ or the Best Linear Unbiased Predictors (BLUPs) (see §3.5) for $\boldsymbol{b}_{ij}$ and $\boldsymbol{b}_i$ are required. The arrays in the second row of the decomposition: $\boldsymbol{R}_{11(ij)}$, $\boldsymbol{R}_{10ij}$, and $\boldsymbol{c}_{1(ij)}$ each have n_{ij} rows.

To evaluate the outer integral in (11) we again form and decompose an augmented array

$$\begin{bmatrix} \boldsymbol{R}_{11(i1)} & \boldsymbol{R}_{10(i1)} & \boldsymbol{c}_{1(i1)} \\ \vdots & \vdots & \vdots \\ \boldsymbol{R}_{11(iM_i)} & \boldsymbol{R}_{10(iM_i)} & \boldsymbol{c}_{1(iM_i)} \\ \boldsymbol{\Delta}_1 & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} = \boldsymbol{Q}_{(i)} \begin{bmatrix} \boldsymbol{R}_{11(i)} & \boldsymbol{R}_{10(i)} & \boldsymbol{c}_{1(i)} \\ \boldsymbol{0} & \boldsymbol{R}_{00(i)} & \boldsymbol{c}_{0(i)} \end{bmatrix} \qquad i = 1, \dots, M \tag{13}$$

The final decomposition to produce $\boldsymbol{R}_{00}$, $\boldsymbol{c}_0$ and $\boldsymbol{c}_{-1}$ is the same as that in (7).

Using the matrices and vectors produced in (12), (13), and (7) and following the same steps as for the single level of nesting we can express the profiled log-likelihood for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ as

$$\begin{aligned} \ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \boldsymbol{y}) &= \log L(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \widehat{\sigma^2}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) | \boldsymbol{y}) \\ &= \mathit{const} - N \log \|\boldsymbol{c}_{-1}\| + \sum_{i=1}^{M} \log \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}_1|}{|\boldsymbol{R}_{11(i)}|}\right) \\ &\quad + \sum_{i=1}^{M} \sum_{j=1}^{M_i} \log \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}_2|}{|\boldsymbol{R}_{22(ij)}|}\right) \end{aligned}$$

Similarly, the profiled log-restricted-likelihood is

$$\begin{aligned} \ell_R(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \boldsymbol{y}) &= \log L_R(\widehat{\boldsymbol{\beta}}_R(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \widehat{\sigma^2}_R(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) | \boldsymbol{y}) \\ &= \mathit{const} - (N - p) \log \|\boldsymbol{c}_{-1}\| - \log \operatorname{abs} |\boldsymbol{R}_{00}| \\ &\quad + \sum_{i=1}^{M} \log \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}_1|}{|\boldsymbol{R}_{11(i)}|}\right) + \sum_{i=1}^{M} \sum_{j=1}^{M_i} \log \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}_2|}{|\boldsymbol{R}_{22(ij)}|}\right) \end{aligned}$$

To illustrate the decomposition in (12) and the first part of the calculation of the likelihood, we consider the data from the first site within the first wafer in the IC manufacturing example from §1.3. As described in §2.6 we use 0.84 as an initial value for

$\boldsymbol{\Delta}_2$. The decomposition is then

$$
\left[\begin{array}{c|cc|ccc|c}
\multicolumn{1}{c}{\boldsymbol{Z}_{11}} & \multicolumn{2}{c}{\boldsymbol{Z}_{1,1}} & \multicolumn{3}{c}{\boldsymbol{X}_{11}} & \multicolumn{1}{c}{\boldsymbol{y}_{11}} \\
1 & 1 & 0.8 & 1 & 0.8 & 0.8^2 & 0.92 \\
1 & 1 & 1.2 & 1 & 1.2 & 1.2^2 & 3.91 \\
1 & 1 & 1.6 & 1 & 1.6 & 1.6^2 & 7.69 \\
1 & 1 & 2.0 & 1 & 2.0 & 2.0^2 & 11.80 \\
1 & 1 & 2.4 & 1 & 2.4 & 2.4^2 & 16.00 \\
\hline
\boldsymbol{\Delta}_2 = 0.84 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right] =
$$

$$
\boldsymbol{Q}_{(11)}
\left[\begin{array}{c|cc|ccc|c}
\multicolumn{1}{c}{\boldsymbol{R}_{22(11)}} & \multicolumn{2}{c}{\boldsymbol{R}_{21(11)}} & \multicolumn{3}{c}{\boldsymbol{R}_{20(11)}} & \multicolumn{1}{c}{\boldsymbol{c}_{211}} \\
-2.389 & -2.093 & -3.349 & -2.093 & -3.349 & -6.029 & -16.880 \\
\hline
0 & 0.087 & -0.024 & 0.087 & -0.024 & -0.528 & -1.343 \\
0 & 0.087 & 0.376 & 0.087 & 0.376 & 0.592 & 2.437 \\
0 & 0.087 & 0.776 & 0.087 & 0.776 & 2.032 & 6.547 \\
0 & 0.087 & 1.176 & 0.087 & 1.176 & 3.792 & 10.747 \\
0 & -0.767 & -1.029 & -0.767 & -1.029 & -1.653 & -4.412 \\
\multicolumn{1}{c}{\mathbf{0}} & \multicolumn{2}{c}{\boldsymbol{R}_{11(11)}} & \multicolumn{3}{c}{\boldsymbol{R}_{10(11)}} & \multicolumn{1}{c}{\boldsymbol{c}_{1(11)}}
\end{array}\right]
$$

The direct contribution to the log-likelihood from this first site within the first wafer is

$$\log\operatorname{abs}\left(\frac{|\boldsymbol{\Delta}_2|}{|\boldsymbol{R}_{22(11)}|}\right) = \log\left(0.84/2.389\right) = -1.045$$

Since this was a balanced experiment where each wafer has the same number of sites and each site is measured at the same set of voltages, this contribution will be the same for each of the second level groups. The total direct contribution will be

$$\sum_{i=1}^{10}\sum_{j=1}^{67}\log\left(0.84/2.389\right) = -700.2$$

There will also be an indirect contribution from each site within each wafer according to the way they together determine $\|\boldsymbol{c}_{-1}\|$ for a given value of $\boldsymbol{\Delta}_2$.

2.5 Three or more levels of random effects

For every level of random effects added to the hierarchical model we simply extend the number of stages in the decompositions of the augmented $\boldsymbol{Z}$ and $\boldsymbol{X}$ matrices. For example, with three levels of random effects the decompositions begin with

$$
\begin{bmatrix}
\boldsymbol{Z}_{ijk} & \boldsymbol{Z}_{ij,k} & \boldsymbol{Z}_{i,jk} & \boldsymbol{X}_{ijk} & \boldsymbol{y}_{ijk} \\
\boldsymbol{\Delta}_3 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}
\end{bmatrix} =
$$

$$
\boldsymbol{Q}_{(ijk)}
\begin{bmatrix}
\boldsymbol{R}_{33(ijk)} & \boldsymbol{R}_{32(ijk)} & \boldsymbol{R}_{31(ijk)} & \boldsymbol{R}_{30(ijk)} & \boldsymbol{c}_{3(ijk)} \\
\mathbf{0} & \boldsymbol{R}_{22(ijk)} & \boldsymbol{R}_{21(ijk)} & \boldsymbol{R}_{20(ijk)} & \boldsymbol{c}_{2(ijk)}
\end{bmatrix}
\qquad
\begin{array}{l}
i = 1, \ldots, M \\
j = 1, \ldots, M_i \\
k = 1, \ldots, M_{ij}
\end{array}
$$

to evaluate the integral associated with the third level of random effects. The next set of decompositions are

$$\begin{bmatrix} \boldsymbol{R}_{22(ij1)} & \boldsymbol{R}_{21(ij1)} & \boldsymbol{R}_{20(ij1)} & \boldsymbol{c}_{2(ij1)} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{R}_{22(ijM_{ij})} & \boldsymbol{R}_{21(ijM_{ij})} & \boldsymbol{R}_{20(ijM_{ij})} & \boldsymbol{c}_{2(ijM_{ij})} \\ \boldsymbol{\Delta}_2 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} = \boldsymbol{Q}_{(ij)} \begin{bmatrix} \boldsymbol{R}_{22(ij)} & \boldsymbol{R}_{21(ij)} & \boldsymbol{R}_{20(ij)} & \boldsymbol{c}_{2(ij)} \\ \boldsymbol{0} & \boldsymbol{R}_{11(ij)} & \boldsymbol{R}_{10(ij)} & \boldsymbol{c}_{1(ij)} \end{bmatrix} \qquad \begin{array}{l} i = 1, \dots, M \\ j = 1, \dots, M_i \\ k = 1, \dots, M_{ij} \end{array}$$

and from there we proceed as before with (13) and (7).

The profiled log-likelihood becomes

$$\begin{aligned} \ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3 | \boldsymbol{y}) &= \log L(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3), \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \widehat{\sigma^2}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3) | \boldsymbol{y}) \\ &= \mathit{const} - N \log \|\boldsymbol{c}_{-1}\| + \sum_{i=1}^{M} \log \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}_1|}{|\boldsymbol{R}_{11(i)}|}\right) \\ &+ \sum_{i=1}^{M} \sum_{j=1}^{M_i} \log \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}_2|}{|\boldsymbol{R}_{22(ij)}|}\right) + \sum_{i=1}^{M} \sum_{j=1}^{M_i} \sum_{k=1}^{M_{ij}} \log \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}_3|}{|\boldsymbol{R}_{33(ijk)}|}\right) \end{aligned}$$

and the profiled restricted log-likelihood becomes

$$\begin{aligned} \ell_R(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3 | \boldsymbol{y}) &= \log L_R(\widehat{\boldsymbol{\beta}}_R(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3), \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \widehat{\sigma^2}_R(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3) | \boldsymbol{y}) \\ &= \mathit{const} - (N - p) \log \|\boldsymbol{c}_{-1}\| - \log \operatorname{abs} |\boldsymbol{R}_{00}| + \sum_{i=1}^{M} \log \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}_1|}{|\boldsymbol{R}_{11(i)}|}\right) \\ &+ \sum_{i=1}^{M} \sum_{j=1}^{M_i} \log \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}_2|}{|\boldsymbol{R}_{22(ij)}|}\right) + \sum_{i=1}^{M} \sum_{j=1}^{M_i} \sum_{k=1}^{M_{ij}} \log \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}_3|}{|\boldsymbol{R}_{33(ijk)}|}\right) \end{aligned}$$

2.6 Starting values for the θ parameters

Because we can express both the profiled log-likelihood and the profiled log-restricted-likelihood as a function of the $\boldsymbol{\theta}$ parameters, we only need to formulate starting values for $\boldsymbol{\theta}$ when performing iterative optimization. From (10) and from the discussion of the example in §2.4, we can see that $\boldsymbol{\theta}$ influences the profiled log-likelihood indirectly by changing $\|\boldsymbol{c}_{-1}\|^2$ and directly through terms of the form

$$\begin{aligned} \log \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}|}{|\boldsymbol{R}_{11(i)}|}\right) &= -\frac{1}{2} \log\left(\frac{|\boldsymbol{Z}_i'\boldsymbol{Z}_i + \boldsymbol{\Delta}'\boldsymbol{\Delta}|}{|\boldsymbol{\Delta}'\boldsymbol{\Delta}|}\right) \\ &= \frac{-\log\left|\boldsymbol{I} + \left(\boldsymbol{\Delta}^{-1}\right)' \boldsymbol{Z}_i'\boldsymbol{Z}_i\boldsymbol{\Delta}^{-1}\right|}{2} \leq 0 \end{aligned} \tag{14}$$

The size of $\boldsymbol{\Delta}$ relative to each of the $\boldsymbol{Z}_i, i = 1, \ldots, M$ affects the two types of terms in opposite ways.

As $\boldsymbol{\Delta}$ gets larger it pulls the conditional least-squares values of $\boldsymbol{b}_i$ closer to $\mathbf{0}$ thus increasing the residual sum-of-squares term $\|\boldsymbol{c}_{-1}\|^2$. Conversely when $\boldsymbol{\Delta}$ is very small relative to the $\boldsymbol{Z}_i$ it has little effect on the conditional least-squares values for $\boldsymbol{b}_i$ and $\|\boldsymbol{c}_{-1}\|^2$ will be smaller. However, small values of $\boldsymbol{\Delta}$ produce large values of the terms (14). For example, we showed earlier that $\boldsymbol{\Delta}_2 = 0.84$ produces

$$\log \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}_2|}{|\boldsymbol{R}_{22(11)}|}\right) = \log\left(0.84/2.389\right) = -1.045$$

in the IC manufacturing example. Doing the same calculations with a very small value, say $\boldsymbol{\Delta}_2 = 0.0084$ would produce

$$\log \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}_2|}{|\boldsymbol{R}_{22(11)}|}\right) = \log\left(0.0084/2.361\right) = -5.584$$

For a large value, say $\boldsymbol{\Delta}_2 = 84$, we would get

$$\log \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}_2|}{|\boldsymbol{R}_{22(11)}|}\right) = \log\left(84/84.030\right) = -0.0003542$$

In the limit as $\boldsymbol{\Delta}$ becomes very large relative to the $\boldsymbol{Z}_i$, the coefficients $\boldsymbol{b}_i$ in the penalized regression model are forced to zero and the residual sum-of-squares $\|\boldsymbol{c}_{-1}\|^2$ is the same as that from a regression of all the data on the fixed effects $\boldsymbol{\beta}$ only. The terms in (14) approach $-\log|\boldsymbol{I}|/2 = 0$, which is their upper bound.

The net result is that extremely large values of $\boldsymbol{\Delta}$ or extremely small values of $\boldsymbol{\Delta}$ tend to produce small values of the likelihood for most data sets. The optimal values of $\boldsymbol{\theta}$ will usually correspond to a $\boldsymbol{\Delta}$ that is comparable in size to the $\boldsymbol{Z}_i$. We generate a simple starting value $\boldsymbol{\Delta}^{(0)}$ as a diagonal matrix where each diagonal element is some fraction f of root-mean-square length of the corresponding column of the $\boldsymbol{Z}_i$ matrices. That is, letting $\boldsymbol{Z}_i(k)$ denote the kth column of $\boldsymbol{Z}_i$, the initial value for the kth diagonal element of $\boldsymbol{\Delta}$ is $f \times \left(\sum_{i=1}^{M} \|\boldsymbol{Z}_i(k)\|^2/M\right)^{1/2}$. Some limited experimentation with this formulation indicated that fractions f between 1/4 and 1/2 worked well. We use $f = 0.375$.

In some patterned variance-covariance matrices, such as the compound symmetry matrices, a general diagonal matrix cannot be represented. For those cases we use as a starting estimate for $\boldsymbol{\theta}$ the parameter values that provide the patterned matrix that is closest to the diagonal $\boldsymbol{\Delta}$ calculated as described above.

This procedure generalizes easily to multiple levels of random effects. As an example, we consider the two-level model for the IC manufacturing data of §1.3. Because the design is balanced (i.e. all wafers have the same number of sites and the intensity of current measurements were made at the same voltages for every site), it suffices to calculate the column norms for, say, $\boldsymbol{Z}_{11}$ and $\boldsymbol{Z}_1$. Therefore, the initial estimates for the precision matrices are

$$\boldsymbol{\Delta}_1^{(0)} = 0.375 \begin{bmatrix} \|\boldsymbol{Z}_1(1)\| & 0 \\ 0 & \|\boldsymbol{Z}_1(2)\| \end{bmatrix} = 0.375 \begin{bmatrix} \sqrt{335} & 0 \\ 0 & \sqrt{964.8} \end{bmatrix} = \begin{bmatrix} 6.86 & 0 \\ 0 & 11.65 \end{bmatrix}$$

$$\boldsymbol{\Delta}_2^{(0)} = 0.375\|\boldsymbol{Z}_{11}\| = 0.375\sqrt{5} = 0.84$$

3 Conditional estimates and EM iterations

In the last section we derived a compact representation of the profiled log-likelihood or profiled log-restricted-likelihood for the single level or nested mixed-effects model. This representation used the relative precision factors $\boldsymbol{\Delta}_q,\ q = 1, \dots, Q$ and orthogonal-triangular decompositions of augmented data arrays. A general nonlinear optimization algorithm, combined with the starting estimates for the $\boldsymbol{\Delta}_q$ derived in that section, could be used to obtain the maximum likelihood estimates, $\widehat{\boldsymbol{\theta}}$, or the restricted maximum likelihood estimates, $\widehat{\boldsymbol{\theta}}_R$, and, through (9), $\widehat{\boldsymbol{\beta}}$.

Before considering details of the optimization we examine the general conditional estimates $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and the conditional distribution of the random effects $p\left(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2\right)$ more closely. We relate these conditional estimates of the fixed-effects and the conditional expectations of the random effects to the parameter estimates in a penalized linear least squares problem. This conditional distribution of the random effects can be used to define EM iterations which are very helpful in refining starting estimates for the $\boldsymbol{\Delta}$.

3.1 Conditional distribution of the random effects

We can see from (5) that the conditional distribution of the random effects in a single-level model is

$$\begin{aligned} p\left(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2\right) &\propto p\left(\boldsymbol{b}_i, \boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2\right) \\ &\propto \exp\left(\frac{-1}{2\sigma^2}\left\|\boldsymbol{c}_{1(i)} - \boldsymbol{R}_{10(i)}\boldsymbol{\beta} - \boldsymbol{R}_{11(i)}\boldsymbol{b}_i\right\|^2\right) \end{aligned}$$

or, writing all the random effects together as $\boldsymbol{b}' = \left(\boldsymbol{b}_1', \dots, \boldsymbol{b}_M'\right)'$,

$$p\left(\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2\right) \propto \exp\left(\frac{-1}{2\sigma^2}\left\|\boldsymbol{c}_{\boldsymbol{b}} - \boldsymbol{R}_{\boldsymbol{b}}\boldsymbol{b} - \boldsymbol{R}_{\boldsymbol{\beta}}\boldsymbol{\beta}\right\|^2\right)$$

where

$$\boldsymbol{R}_{\boldsymbol{b}} = \begin{bmatrix} \boldsymbol{R}_{11(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R}_{11(2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{R}_{11(M)} \end{bmatrix} \quad \boldsymbol{R}_{\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{R}_{10(i)} \\ \boldsymbol{R}_{10(2)} \\ \vdots \\ \boldsymbol{R}_{10(M)} \end{bmatrix} \quad \boldsymbol{c}_{\boldsymbol{b}} = \begin{bmatrix} \boldsymbol{c}_{1(1)} \\ \boldsymbol{c}_{1(2)} \\ \vdots \\ \boldsymbol{c}_{1(M)} \end{bmatrix}$$

It then follows that

$$\boldsymbol{b}|\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{R}_{\boldsymbol{b}}^{-1}\left(\boldsymbol{c}_{\boldsymbol{b}} - \boldsymbol{R}_{\boldsymbol{\beta}}\boldsymbol{\beta}\right), \sigma^2 \boldsymbol{R}_{\boldsymbol{b}}^{-1}\left(\boldsymbol{R}_{\boldsymbol{b}}^{-1}\right)'\right) \tag{15}$$

We note that $\boldsymbol{R}_{\boldsymbol{b}}^{-1}$ is a block diagonal matrix with its ith diagonal block given by the inverse of the $q \times q$ upper-triangular matrix $\boldsymbol{R}_{11(i)}$.

From (8) we can see that the likelihood for the fixed effects $\boldsymbol{\beta}$ conditional on $\boldsymbol{\theta}$ and σ^2 has a similar form

$$L(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\theta}, \sigma^2) \propto \exp\left(\frac{-\|\boldsymbol{c}_0 - \boldsymbol{R}_{00}\boldsymbol{\beta}\|^2}{2\sigma^2}\right)$$

The distribution of the conditional estimate $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and the conditional distribution of $\boldsymbol{b}$ are the same as the distribution of the estimate of $\boldsymbol{\beta}$ and the conditional estimates $\boldsymbol{b}|\boldsymbol{\beta}$ in a linear regression model

$$\boldsymbol{c} = \boldsymbol{R}\boldsymbol{b}^* + \boldsymbol{\epsilon}^* \quad \boldsymbol{\epsilon}^* \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$$

where

$$\boldsymbol{b}^* = \begin{bmatrix}\boldsymbol{b}\\ \boldsymbol{\beta}\end{bmatrix} \quad \boldsymbol{R} = \begin{bmatrix}\boldsymbol{R}_{\boldsymbol{b}} & \boldsymbol{R}_{\boldsymbol{\beta}}\\ \boldsymbol{0} & \boldsymbol{R}_{00}\end{bmatrix} \quad \boldsymbol{c} = \begin{bmatrix}\boldsymbol{c}_{\boldsymbol{b}}\\ \boldsymbol{c}_0\end{bmatrix} \tag{16}$$

or, equivalently, the penalized linear regression model

$$\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{b}^* + \boldsymbol{\epsilon}^{**} \quad \boldsymbol{\epsilon}^{**} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$$

where

$$\boldsymbol{y}^* = \begin{bmatrix}\boldsymbol{y}_1\\ \boldsymbol{0}\\ \boldsymbol{y}_2\\ \boldsymbol{0}\\ \vdots\\ \boldsymbol{y}_M\\ \boldsymbol{0}\end{bmatrix} \quad \boldsymbol{X}^* = \begin{bmatrix}\boldsymbol{Z}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{X}_1\\ \boldsymbol{\Delta} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{Z}_2 & \cdots & \boldsymbol{0} & \boldsymbol{X}_2\\ \boldsymbol{0} & \boldsymbol{\Delta} & \cdots & \boldsymbol{0} & \boldsymbol{0}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Z}_M & \boldsymbol{X}_M\\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Delta} & \boldsymbol{0}\end{bmatrix} \tag{17}$$

The conditional distribution of the random effects generalizes to multiple levels. For example, in a model with two levels of random effects,

$$p\left(\boldsymbol{b}_i, \boldsymbol{b}_{i1}, \dots, \boldsymbol{b}_{iM_i}|\boldsymbol{y}_{i1}, \dots, \boldsymbol{y}_{iM_i}, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2\right) \propto \exp\left[\frac{-1}{2\sigma^2}\left(\left\|\boldsymbol{c}_{1(i)} - \boldsymbol{R}_{11(i)}\boldsymbol{b}_i - \boldsymbol{R}_{10(i)}\boldsymbol{\beta}\right\|^2 + \sum_{j=1}^{M_i}\left\|\boldsymbol{c}_{2(ij)} - \boldsymbol{R}_{22(ij)}\boldsymbol{b}_{ij} - \boldsymbol{R}_{21(ij)}\boldsymbol{b}_i - \boldsymbol{R}_{20(ij)}\boldsymbol{\beta}\right\|^2\right)\right]$$

In a vector notation of $\boldsymbol{b}(i)' = \left(\boldsymbol{b}_{i1}', \dots, \boldsymbol{b}_{iM_i}', \boldsymbol{b}_i'\right)'$ for the two-level model and writing

$$\boldsymbol{R_b}(i) = \begin{bmatrix} \boldsymbol{R}_{22(i1)} & \cdots & \mathbf{0} & \boldsymbol{R}_{21(i1)} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{R}_{22(iM_i)} & \boldsymbol{R}_{21(iM_i)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{R}_{11(i)} \end{bmatrix}$$

$$\boldsymbol{R_\beta}(i) = \begin{bmatrix} \boldsymbol{R}_{20(i1)} \\ \vdots \\ \boldsymbol{R}_{20(iM_i)} \\ \boldsymbol{R}_{10(i)} \end{bmatrix} \qquad \boldsymbol{c_b}(i) = \begin{bmatrix} \boldsymbol{c}_{2(i1)} \\ \vdots \\ \boldsymbol{c}_{2(iM_i)} \\ \boldsymbol{c}_{1(i)} \end{bmatrix}$$

we obtain

$$\boldsymbol{b}(i)|\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{R_b}^{-1}(i)\left(\boldsymbol{c_b}(i) - \boldsymbol{R_\beta}(i)\boldsymbol{\beta}\right), \sigma^2 \boldsymbol{R_b}^{-1}(i)\left(\boldsymbol{R_b}^{-1}(i)\right)'\right) \tag{18}$$

The equivalence of the distribution of the conditional estimates $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and the the conditional distribution of the random effects to the distribution of the parameter estimates in a penalized linear regression model extends to multiple levels of random effects. For two levels of random effects the penalized regression model is $\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{b}^* + \boldsymbol{\epsilon}^{**}$ where $(\boldsymbol{b}^*)' = (\boldsymbol{b}(1)', \boldsymbol{b}(2)', \dots, \boldsymbol{b}(M)', \boldsymbol{\beta})'$ and the substitutions

$$\begin{bmatrix} \boldsymbol{Z}_i \\ \boldsymbol{\Delta} \end{bmatrix} \Rightarrow \begin{bmatrix} \boldsymbol{Z}_{i1} & \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{Z}_{i,1} \\ \boldsymbol{\Delta}_2 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}_{i2} & \cdots & \mathbf{0} & \boldsymbol{Z}_{i,2} \\ \mathbf{0} & \boldsymbol{\Delta}_2 & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{Z}_{iM_i} & \boldsymbol{Z}_{i,M_i} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\Delta}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\Delta}_1 \end{bmatrix} \qquad \begin{bmatrix} \boldsymbol{y}_i \\ \mathbf{0} \end{bmatrix} \Rightarrow \begin{bmatrix} \boldsymbol{y}_{i1} \\ \mathbf{0} \\ \boldsymbol{y}_{i2} \\ \mathbf{0} \\ \vdots \\ \boldsymbol{y}_{iM_i} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

are made in the form (17).

3.2 Conditional distributions and restricted likelihood

The expressions (15) and (18) and their generalizations to more levels of random effects allow the EM iterations on $\boldsymbol{\theta}$ for the likelihood criterion to be developed, as shown in §3.4. Similar expressions are used to develop EM iterations for the restricted likelihood criterion. As described in Laird and Ware (1982) the restricted likelihood can be regarded as the likelihood associated with model (1) but incorporating $\boldsymbol{\beta}$ as a random effect, say $\boldsymbol{b}_0$, with an associated relative variance matrix $\boldsymbol{D}_0$ that tends to infinity. We could, for example, set $\boldsymbol{D}_0 = k\boldsymbol{I},\ k \to \infty$.

The conditional distribution of this extended random effects vector for the restricted model is simply the distribution of the estimates of $\boldsymbol{b}^*$ in the corresponding regression models (16) or (17). That is,

$$\boldsymbol{b}^*|\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{R}^{-1}\boldsymbol{c}, \sigma^2\boldsymbol{R}^{-1}(\boldsymbol{R}^{-1})'\right)$$

for a single level of random effects. This scales in the obvious way to multiple levels.

For both the original model and the restricted model the conditional distribution of the random effects corresponds to the distribution of the parameter estimates in a penalized linear regression model. The only difference is that in the original model it corresponds to the distribution of the estimates of $\boldsymbol{b}$ conditional on $\boldsymbol{\beta}$ and in the restricted model it corresponds to the marginal distribution of the estimates of $\boldsymbol{b}$.

3.3 Using matrix sparsity

We have expressed the conditional distributions of the random effects in terms of the inverses of the upper triangular matrices $\boldsymbol{R}$ and $\boldsymbol{R_b}$. These are large matrices but they are also patterned and sparse. We can take advantage of the patterns when forming the inverses and when doing further calculations.

For a single level of random effects we need only store

$$\begin{bmatrix} \boldsymbol{R}_{11(1)} & \boldsymbol{R}_{10(1)} & \boldsymbol{c}_{1(1)} \\ \boldsymbol{R}_{11(2)} & \boldsymbol{R}_{10(2)} & \boldsymbol{c}_{1(2)} \\ \vdots & \vdots & \vdots \\ \boldsymbol{R}_{11(M)} & \boldsymbol{R}_{10(M)} & \boldsymbol{c}_{1(M)} \\ \boldsymbol{0} & \boldsymbol{R}_{00} & \boldsymbol{c}_0 \end{bmatrix} \tag{19}$$

In calculating $\boldsymbol{R}^{-1}$ we can take advantage of the fact that the pattern of zeroes in $\boldsymbol{R}$ is repeated in

$$\boldsymbol{R}^{-1} = \begin{bmatrix} \boldsymbol{R}_{11(1)}^{-1} & \boldsymbol{0} & \cdots & \boldsymbol{0} & -\boldsymbol{R}_{11(1)}^{-1}\boldsymbol{R}_{10(1)}\boldsymbol{R}_{00}^{-1} \\ \boldsymbol{0} & \boldsymbol{R}_{11(2)}^{-1} & \cdots & \boldsymbol{0} & -\boldsymbol{R}_{11(2)}^{-1}\boldsymbol{R}_{10(2)}\boldsymbol{R}_{00}^{-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{R}_{11(M)}^{-1} & -\boldsymbol{R}_{11(M)}^{-1}\boldsymbol{R}_{10(M)}\boldsymbol{R}_{00}^{-1} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{R}_{00}^{-1} \end{bmatrix} \tag{20}$$

and perform the operations in place on an array stored as in (19). That is, there is no need to expand the array $\boldsymbol{R}$ out to its full size when calculating its inverse. We also note that each of the components such as $\boldsymbol{R}_{00}$ and $\boldsymbol{R}_{11(i)}, i = 1, \ldots, M$ that must be inverted to form (20) is a relatively small, triangular, non-singular matrix whose inverse is readily calculated.

One advantage of storing the intermediate results in this form is that the expressions for the conditional estimates $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and $\widehat{\boldsymbol{b}}_i(\boldsymbol{\theta}) = \mathrm{E}[\boldsymbol{b}_i|\boldsymbol{y}]$ evaluated at $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ have the form

$$\begin{aligned} \boldsymbol{R}_{00}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) &= \boldsymbol{c}_0 & \\ \boldsymbol{R}_{11(i)}\widehat{\boldsymbol{b}}_i(\boldsymbol{\theta}) &= \boldsymbol{c}_{1(i)} - \boldsymbol{R}_{10(i)}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) & i &= 1, \ldots, M \\ \boldsymbol{R}_{22(ij)}\widehat{\boldsymbol{b}}_{ij}(\boldsymbol{\theta}) &= \boldsymbol{c}_{2(ij)} - \boldsymbol{R}_{20(ij)}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) - \boldsymbol{R}_{21(ij)}\widehat{\boldsymbol{b}}_i(\boldsymbol{\theta}) & j &= 1, \ldots, M_i \\ &\cdots & &\cdots \end{aligned}$$

Once $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ has been calculated, the evaluation of $\boldsymbol{c}_{1(i)} - \boldsymbol{R}_{10(i)}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), i = 1, \dots, M$ can be performed as a single matrix multiplication. The structure corresponding to (19) for two levels of random effects has the form

$$\begin{bmatrix} \boldsymbol{R}_{22(11)} & \boldsymbol{R}_{21(11)} & \boldsymbol{R}_{20(11)} & \boldsymbol{c}_{2(11)} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{R}_{22(1M_1)} & \boldsymbol{R}_{21(1M_1)} & \boldsymbol{R}_{20(1M_1)} & \boldsymbol{c}_{2(1M_1)} \\ \mathbf{0} & \boldsymbol{R}_{11(1)} & \boldsymbol{R}_{10(1)} & \boldsymbol{c}_{1(1)} \\ \boldsymbol{R}_{22(21)} & \boldsymbol{R}_{21(21)} & \boldsymbol{R}_{20(21)} & \boldsymbol{c}_{2(21)} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{R}_{22(MM_M)} & \boldsymbol{R}_{21(MM_M)} & \boldsymbol{R}_{20(MM_M)} & \boldsymbol{c}_{2(MM_M)} \\ \mathbf{0} & \boldsymbol{R}_{11(M)} & \boldsymbol{R}_{10(M)} & \boldsymbol{c}_{1(M)} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{R}_{00} & \boldsymbol{c}_0 \end{bmatrix} \tag{21}$$

Again, this form facilitates calculation of expressions such as $\boldsymbol{c}_{1(i)} - \boldsymbol{R}_{10(i)}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and $\boldsymbol{c}_{2(ij)} - \boldsymbol{R}_{21(i)}\widehat{\boldsymbol{b}}_i(\boldsymbol{\theta}) - \boldsymbol{R}_{20(ij)}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$.

The corresponding block in $\boldsymbol{R}^{-1}$ is

$$\begin{bmatrix} \boldsymbol{R}^{-1}_{22(i1)} & \mathbf{0} & \cdots & \mathbf{0} & -\boldsymbol{R}^{-1}_{22(i1)}\boldsymbol{R}_{21(i1)}\boldsymbol{R}^{-1}_{i1(1)} \\ \mathbf{0} & \boldsymbol{R}^{-1}_{22(i2)} & \cdots & \mathbf{0} & -\boldsymbol{R}^{-1}_{22(12)}\boldsymbol{R}_{21(12)}\boldsymbol{R}^{-1}_{11(1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{R}^{-1}_{22(iM_i)} & -\boldsymbol{R}^{-1}_{22(iM_i)}\boldsymbol{R}_{21(iM_i)}\boldsymbol{R}^{-1}_{11(i)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{R}^{-1}_{11(i)} \end{bmatrix} \tag{22}$$

We can see that the pattern of (20) is repeated here. As before, the block does not need to be expanded to its full size to calculate the components of the inverse—the condensed array of the form (21) can be manipulated in place.

3.4 EM iterations

Optimization of the profiled log-likelihood or restricted log-likelihood is usually accomplished through EM iterations or through Newton-Raphson iterations (Laird and Ware, 1982; Lindstrom and Bates, 1988; Longford, 1993). The EM iterations have the advantage that the individual iterations are easy to compute and the initial iterations approach the optimum quite quickly. However, close to the optimum the EM iterations often proceed very slowly. It can be difficult to decide if the EM iterations have converged when they end up taking an exceedingly large number of very small steps toward the optimum. Also, taking so many iterations is itself expensive even if each iteration is relatively simple and fast.

The Newton-Raphson iterations, on the other hand, are individually more computationally intensive than the EM iterations and they can be quite unstable when far from the optimum. However, close to the optimum they converge very quickly.

We therefore recommend a hybrid approach of forming a initial $\boldsymbol{\theta}^{(0)}$ as described in section 2.6, performing a moderate number of EM iterations, then switching to

Newton-Raphson iterations. Essentially the EM iterations can be regarded as refining the starting estimates before beginning the more general optimization routine.

The EM iterations are based on regarding the random effects, such as the $\boldsymbol{b}_i, i = 1, \ldots, M$, as unobserved data. At iteration t we use the current variance-covariance parameter vector, $\boldsymbol{\theta}^{(t)}$, to evaluate the distribution of $\boldsymbol{b}|\boldsymbol{y}$ and derive the expectation of the log-likelihood for a new value of $\boldsymbol{\theta}$ given this conditional distribution. Because we are taking an expectation, this step is called the E step.

The M step consists of maximizing this expectation with respect to $\boldsymbol{\theta}$ to produce $\boldsymbol{\theta}^{(t+1)}$.

We can perform the expectation and maximization steps with respect to $\boldsymbol{\theta}$ only because we have explicit forms for the profiled likelihood as a function of $\boldsymbol{\theta}$. At iteration t we derive $\boldsymbol{\theta}^{(t+1)}$ from $\boldsymbol{\theta}^{(t)}$ fixing $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}^{(t)})$ and $\sigma^2 = \widehat{\sigma^2}(\boldsymbol{\theta}^{(t)})$. The EM algorithm ensures that the log-likelihood will increase from $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1)}$. If we then adjust $\boldsymbol{\beta}$ and σ^2 to the optimal values given $\boldsymbol{\theta}^{(t+1)}$ we can only increase the log-likelihood further.

For a single level of random effects we can see from (3) that the likelihood for the "full data"—both the unobserved $\boldsymbol{b}$ and the observed $\boldsymbol{y}$—depends on $\boldsymbol{\theta}$ only through $|\boldsymbol{D}|$ and quadratic forms in $(\boldsymbol{b}_i/\sigma)'\boldsymbol{D}^{-1}\boldsymbol{b}_i/\sigma$. To obtain the expected value for a new value of $\boldsymbol{\theta}$ we use standard results on the distribution of quadratic forms plus the expressions in (15) to derive

$$\begin{aligned}
\mathrm{E}[(\boldsymbol{b}_i/\sigma)'\boldsymbol{D}_1^{-1}\boldsymbol{b}_i/\sigma|\boldsymbol{y}] &= \mathrm{E}[\boldsymbol{b}_i/\sigma|\boldsymbol{y}]'\boldsymbol{D}_1^{-1}\mathrm{E}[\boldsymbol{b}_i/\sigma|\boldsymbol{y}] + \operatorname{tr}\left(\boldsymbol{D}^{-1}\mathrm{Var}(\boldsymbol{b}_i|\boldsymbol{y})\right) \\
&= \|\boldsymbol{\Delta}_1\mathrm{E}[\boldsymbol{b}_i|\boldsymbol{y}]/\sigma\|^2 + \left\| \left(\boldsymbol{R}_{11(i)}^{-1}\right)'\boldsymbol{\Delta}_1'\right\|_2^2 \\
&= \left\| \begin{bmatrix} \mathrm{E}[\boldsymbol{b}_i|\boldsymbol{y}]'/\sigma \\ \left(\boldsymbol{R}_{11(i)}^{-1}\right)' \end{bmatrix} \boldsymbol{\Delta}_1' \right\|_2^2
\end{aligned}$$

where the 2-norm of a matrix, written $\|\boldsymbol{A}\|_2$, is the square root of the sum of squares of the entries in the matrix. In this expression $\boldsymbol{R}_{11(i)}$ is evaluated using the data and $\boldsymbol{\theta}^{(t)}$ because it is a characteristic of the conditional distribution but $\boldsymbol{\Delta}_1$ is evaluated at the general $\boldsymbol{\theta}$ over which we will be optimizing the expected likelihood given the conditional distribution so as to produce $\boldsymbol{\theta}^{(t+1)}$.

The sum of the contributions for the level-1 random effects can be expressed as

$$\left\| \begin{bmatrix} \mathrm{E}[\boldsymbol{b}_1|\boldsymbol{y}]'/\sigma \\ \left(\boldsymbol{R}_{11(1)}^{-1}\right)' \\ \vdots \\ \mathrm{E}[\boldsymbol{b}_M|\boldsymbol{y}]'/\sigma \\ \left(\boldsymbol{R}_{11(M)}^{-1}\right)' \end{bmatrix} \boldsymbol{\Delta}_1' \right\|_2^2 = \|\boldsymbol{A}_1\boldsymbol{\Delta}_1'\|_2^2 \tag{23}$$

where $\boldsymbol{U}_1\boldsymbol{A}_1$ is an orthogonal-triangular decomposition of the stacked matrix on the left side of (23). Thus $\boldsymbol{U}_1$ is a $M(q_1 + 1) \times q_1$ matrix with orthonormal columns and $\boldsymbol{A}_1$ is a $q_1 \times q_1$ upper-triangular, non-singular matrix.

As described in §3.1, the conditional distribution of $\boldsymbol{b}|\boldsymbol{y}$ for the restricted model is also multivariate normal with the same mean but with a different variance-covariance matrix. For this case the non-zero blocks of the form $-\boldsymbol{R}_{11(1)}^{-1}\boldsymbol{R}_{10(1)}\boldsymbol{R}_{00}^{-1}$ in the $\boldsymbol{R}^{-1}$ matrix shown in (20) must be incorporated in the calculation of $\boldsymbol{A}_{R1}$. The decomposition has the form

$$\begin{bmatrix} \mathrm{E}[\boldsymbol{b}_1|\boldsymbol{y}]'/\sigma \\ \left(\boldsymbol{R}_{11(1)}^{-1}\right)' \\ -\left(\boldsymbol{R}_{11(1)}^{-1}\boldsymbol{R}_{10(1)}\boldsymbol{R}_{00}^{-1}\right)' \\ \vdots \\ \mathrm{E}[\boldsymbol{b}_M|\boldsymbol{y}]'/\sigma \\ \left(\boldsymbol{R}_{11(M)}^{-1}\right)' \\ -\left(\boldsymbol{R}_{11(M)}\boldsymbol{R}_{10(M)}\boldsymbol{R}_{00}^{-1}\right)' \end{bmatrix} = \boldsymbol{U}_{R1}\boldsymbol{A}_{R1}$$

With two levels of random effects, the matrix $\boldsymbol{A}_1$ is calculated as in (23) and the matrix $\boldsymbol{A}_2$ is calculated from a decomposition of a matrix formed from the scaled estimates of the random effects at that level and their estimated variances calculated according to (22).

$$\begin{bmatrix} \mathrm{E}[\boldsymbol{b}_{11}|\boldsymbol{y}]'/\sigma \\ \left(\boldsymbol{R}_{22(11)}^{-1}\right)' \\ -\left(\boldsymbol{R}_{22(11)}^{-1}\boldsymbol{R}_{21(11)}\boldsymbol{R}_{11(1)}^{-1}\right)' \\ \vdots \\ \mathrm{E}[\boldsymbol{b}_{MM_M}|\boldsymbol{y}]'/\sigma \\ \left(\boldsymbol{R}_{22(MM_M)}^{-1}\right)' \\ -\left(\boldsymbol{R}_{22(MM_M)}^{-1}\boldsymbol{R}_{21(MM_M)}\boldsymbol{R}_{11(M)}^{-1}\right)' \end{bmatrix} = \boldsymbol{U}_2\boldsymbol{A}_2$$

In general, $\boldsymbol{A}_j$ is calculated by computing the expected values of the conditional distribution of the level j random effects and scaling them by $1/\sigma$. The non-zero blocks in $\boldsymbol{R}_{\boldsymbol{b}}^{-1}$ are calculated and transposed. All those blocks corresponding to the level-j random effects are stacked, along with the scaled expected random effects. The resulting matrix of q_j columns is decomposed to give the upper-triangular matrix $\boldsymbol{A}_j$. When computing $\boldsymbol{A}_{Rj}$ for iterations based on the restricted likelihood, the matrix $\boldsymbol{R}_{\boldsymbol{b}}^{-1}$ is replaced by $\boldsymbol{R}^{-1}$.

Once $\boldsymbol{A}_1$ (or $\boldsymbol{A}_{R1}$) is available, the maximization problem to determine $\boldsymbol{\theta}_1^{(t+1)}$ reduces to

$$\max_{\boldsymbol{\theta}_1}\left\{M\log|\boldsymbol{D}_1^{-1}| - \operatorname{tr}(\boldsymbol{A}_1'\boldsymbol{D}_1^{-1}\boldsymbol{A}_1)\right\} \tag{24}$$

If $\boldsymbol{D}_1$ can be a general positive-definite matrix, the maximizer of (24) produces $\boldsymbol{D}_1^{-1} =$

$\boldsymbol{\Delta}_1'\boldsymbol{\Delta}_1$ where

$$\boldsymbol{\Delta}_1^{(t+1)} = \frac{\left(\boldsymbol{A}_1^{-1}\right)'}{\sqrt{M}} \tag{25}$$

Since only the factor $\boldsymbol{\Delta}_1$ and not the corresponding $\boldsymbol{\theta}$ is needed in the calculation of the next iteration, the iteration loop is complete. For two levels of random effects, the optimization problem becomes

$$\max_{\boldsymbol{\theta}_2} \left\{ \left(\sum_{i=1}^{M} M_i \right) \log |\boldsymbol{D}_2^{-1}| - \operatorname{tr}(\boldsymbol{A}_2' \boldsymbol{D}_2^{-1} \boldsymbol{A}_2) \right\}$$

with a solution, for general positive-definite $\boldsymbol{D}_2$ of

$$\boldsymbol{\Delta}_2^{(t+1)} = \frac{\left(\boldsymbol{A}_2^{-1}\right)'}{\sqrt{\sum_{i=1}^{M} M_i}} \tag{26}$$

This scales in the obvious way.

The only change required to perform EM iterations for the restricted likelihood criterion is to replace $\boldsymbol{A}_j$ by $\boldsymbol{A}_{Rj}$ in (25) or (26) or their extended versions for more levels.

In the IC manufacturing example of §1.3, starting with the initial estimates $\boldsymbol{\Delta}_1^{(0)}$ and $\boldsymbol{\Delta}_2^{(0)}$ given in §2.6, we obtain after 20 EM iterations

$$\boldsymbol{\Delta}_1^{(20)} = \begin{bmatrix} 2.498 & 0 \\ 1.097 & 0.681 \end{bmatrix} \quad \boldsymbol{\Delta}_2^{(20)} = 0.477$$

The maximum likelihood estimates for this example are

$$\widehat{\boldsymbol{\Delta}}_1 = \begin{bmatrix} 2.501 & 0 \\ 1.091 & 0.681 \end{bmatrix} \quad \widehat{\boldsymbol{\Delta}}_2 = 0.477$$

indicating that the EM algorithm quickly placed the estimates of $\boldsymbol{\theta}$ in the neighborhood of the MLEs.

3.5 General optimization and obtaining BLUPs

After refining the starting estimates for $\boldsymbol{\theta}$ with a moderate number of EM iterations, the parameters $\boldsymbol{\theta}$ are usually close to their optimal values. A general unconstrained optimization algorithm such as Newton-Raphson or a quasi-Newton algorithm can then be used to produce the final parameter estimates $\widehat{\boldsymbol{\theta}}$. The maximum likelihood estimates for $\boldsymbol{\beta}$ can be evaluated from $\boldsymbol{R}_{00}$ and $\boldsymbol{c}_0$ at $\widehat{\boldsymbol{\theta}}$ and (7). Furthermore, the matrices $\boldsymbol{R}_{11(i)}$ and $\boldsymbol{R}_{10(i)}$ and the vectors $\boldsymbol{c}_{1(i)}$ can be used with $\widehat{\boldsymbol{\beta}}$ to solve for $\mathrm{E}[\boldsymbol{b}_i|\boldsymbol{y}]$ at the parameter estimates. These are called the Best Linear Unbiased Predictors or BLUPs for the random effects.

When performing the optimization it is often helpful if the gradient of the objective function and possibly the Hessian of the objective function can be evaluated. It may be possible to use techniques similar to those in Golub and Pereyra (1973) or Kaufman (1975) and in Bates (1983) to derive an expression for the gradient using the penalized linear regression representation of the model. This will be difficult to generalize to several levels of nested random effects. Because the profiled log-likelihood or profiled log-restricted-likelihood can be evaluated efficiently for different values of $\boldsymbol{\theta}$, we prefer to use numerical values of the gradient and Hessian calculated through finite differences.

4 Computational considerations

4.1 Organizing the decompositions

Although several different matrices are being calculated and decomposed in the process of evaluating the profiled log-likelihood for $\boldsymbol{\theta}$, most of these matrices do not need to be stored unless $\widehat{\boldsymbol{\beta}}$ or the BLUPs for the $\boldsymbol{b}_i$ are to be calculated. In fact, the total amount of storage required for the calculation is essentially the same as that required to represent the original data.

The decompositions can be performed "in place" with a slightly augmented version of the original data array or they can be performed by copying horizontal slices of the original data array to temporary storage, augmenting them, decomposing the result, then copying back the pieces that are to be used later.

To demonstrate these techniques we begin with the data for a model with a single level of random effects, organized as

$$\begin{bmatrix} \boldsymbol{Z}_1 & \boldsymbol{X}_1 & \boldsymbol{y}_1 \\ \vdots & \vdots & \vdots \\ \boldsymbol{Z}_M & \boldsymbol{X}_M & \boldsymbol{y}_M \end{bmatrix}$$

For a given value of $\boldsymbol{\theta}$ we evaluate $\boldsymbol{\Delta}$ then, for each of the M horizontal slices in the array, form and decompose

$$\begin{bmatrix} \boldsymbol{Z}_i & \boldsymbol{X}_i & \boldsymbol{y}_i \\ \boldsymbol{\Delta} & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} = \boldsymbol{Q}_{(i)} \begin{bmatrix} \boldsymbol{R}_{11(i)} & \boldsymbol{R}_{10(i)} & \boldsymbol{c}_{1(i)} \\ \boldsymbol{0} & \boldsymbol{R}_{00(i)} & \boldsymbol{c}_{0(i)} \end{bmatrix}$$

exactly as in (4). The determinant of $\boldsymbol{R}_{11(i)}$ is calculated as the product of its diagonal elements and stored. After that none of $\boldsymbol{R}_{11(i)}$, $\boldsymbol{R}_{10(i)}$, or $\boldsymbol{c}_{1(i)}$ are required for the evaluation of the log-likelihood. They are only needed if the BLUP for $\boldsymbol{b}_i$ is to be calculated as the solution to $\boldsymbol{R}_{11(i)}\mathrm{E}[\boldsymbol{b}_i|\boldsymbol{y}] = \boldsymbol{c}_{1(i)} - \boldsymbol{R}_{10(i)}\widehat{\boldsymbol{\beta}}$. After the decomposition, the matrix $\boldsymbol{R}_{00(i)}$ and the vector $\boldsymbol{c}_{0(i)}$ can be copied back into the storage previously occupied by $\boldsymbol{X}_i$ and $\boldsymbol{y}_i$. When we have finished doing this for $i = 1, \ldots, M$ the last $p + 1$ columns of the original data array will be in exactly the form needed to evaluate the decomposition (7). This provides $\boldsymbol{c}_{-1}$ and $\boldsymbol{R}_{00}$. All the information to evaluate the profiled log-likelihood or the profiled log-restricted-likelihood is then available.

The operations of copying slices of the original data array to temporary storage then copying some of the results back into the data array can be avoided, if desired.

Beginning with an augmented form of the data array as

$$\begin{bmatrix} \boldsymbol{Z}_1 & \boldsymbol{X}_1 & \boldsymbol{y}_1 \\ \vdots & \vdots & \vdots \\ \boldsymbol{Z}_M & \boldsymbol{X}_M & \boldsymbol{y}_M \\ \boldsymbol{\Delta} & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$$

we can decompose the augmented array for the Mth case in place and evaluate $\operatorname{abs}|\boldsymbol{R}_{11(M)}|$. The rows containing $\boldsymbol{R}_{11(M)}$, $\boldsymbol{R}_{10(M)}$, and $\boldsymbol{c}_{1(M)}$ can then be overwritten with $\boldsymbol{\Delta}$ in the $\boldsymbol{Z}$ columns and zeroes in the $\boldsymbol{X}$ and $\boldsymbol{y}$ columns. This results in the augmented arrays for the $(M-1)$'st case being available to be decomposed. We continue in this fashion working from the M'th case up to the first case. The arrays needed to calculate the decomposition (7) are then in place in the last $p+1$ columns but beginning at row $q+1$, not at the first row.

In our experience the cost of copying the arrays to temporary storage before decomposing and copying some results back after decomposing is negligible compared to the cost of decomposing the arrays. Thus we prefer to work on copies. Furthermore, even though the "in place" calculation does not require copying the data it may at times take longer than the "copying" calculation. This is because the memory references are widely scattered in the memory space when decomposing a horizontal slice of a matrix with a very large number of rows. When working with extremely large data sets that cannot fit into the available physical memory of the computer such access patterns will degrade the performance of virtual memory systems. In the approach based on copying the widely scattered memory locations are only used twice; once as the source for the copying and once as the destination of the copying. When actually doing the decomposition on the copy, the memory references are more localized.

With two levels of random effects we begin with the data array in the form

$$\begin{bmatrix} \boldsymbol{Z}_{11} & \boldsymbol{Z}_{1,1} & \boldsymbol{X}_{11} & \boldsymbol{y}_{11} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{Z}_{1M_1} & \boldsymbol{Z}_{1,M_1} & \boldsymbol{X}_{1M_1} & \boldsymbol{y}_{1M_1} \\ \boldsymbol{Z}_{21} & \boldsymbol{Z}_{2,1} & \boldsymbol{X}_{21} & \boldsymbol{y}_{21} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{Z}_{MM_M} & \boldsymbol{Z}_{M,M_M} & \boldsymbol{X}_{MM_M} & \boldsymbol{y}_{MM_M} \end{bmatrix}$$

and decompose the augmented matrices from each horizontal slice corresponding to the level-2 random effects while accumulating $\operatorname{abs}|\boldsymbol{R}_{22(ij)}|, i=1,\ldots,M, j=1,\ldots,M_i$. The matrices $\boldsymbol{Z}_{i,j}$, $\boldsymbol{X}_{ij}$, and the vector $\boldsymbol{y}_{ij}$ are overwritten by $\boldsymbol{R}_{11(ij)}$, $\boldsymbol{R}_{10(ij)}$, and $\boldsymbol{c}_{1(ij)}$ produced in this decomposition. The modified data array is then in the form needed to perform the decompositions for the level-1 random effects (ignoring the first q_2 columns).

4.2 Pre-decomposition of the original data array

When the number of observations per lowest-level group is large compared to the number of random effects for the group, some time can be saved by decomposing the original data arrays and saving only the triangular parts of the decomposition. This technique is similar to the "two-stage orthogonal factorization" described by Golub and Pereyra (1973).

With one level of random effects the initial decomposition is similar to the reductions in (4) and (7). First we decompose

$$\begin{bmatrix}\boldsymbol{Z}_i & \boldsymbol{X}_i & \boldsymbol{y}_i\end{bmatrix} = \boldsymbol{S}_{(i)} \begin{bmatrix}\boldsymbol{T}_{11(i)} & \boldsymbol{T}_{10(i)} & \boldsymbol{d}_{1(i)} \\ \boldsymbol{0} & \boldsymbol{T}_{00(i)} & \boldsymbol{d}_{0(i)}\end{bmatrix} \quad i = 1, \ldots, M$$

then we accumulate and decompose all the $\boldsymbol{T}_{00(i)}$ matrices and $\boldsymbol{c}_{0(i)}$ vectors as

$$\begin{bmatrix}\boldsymbol{T}_{00(1)} & \boldsymbol{d}_{0(1)} \\ \vdots & \vdots \\ \boldsymbol{T}_{00(M)} & \boldsymbol{d}_{0(M)}\end{bmatrix} = \boldsymbol{S}_0 \begin{bmatrix}\boldsymbol{T}_{00} & \boldsymbol{d}_0 \\ \boldsymbol{0} & d_{-1} \\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix} \tag{27}$$

where d_{-1} is a scalar. The decomposition (27) is the one occasion where we carry out the orthogonal-triangular decomposition completely to reduce the matrix to triangular form.

The information in

$$\begin{bmatrix}\boldsymbol{T}_{11(1)} & \boldsymbol{T}_{10(1)} & \boldsymbol{d}_{1(1)} \\ \vdots & \vdots & \vdots \\ \boldsymbol{T}_{11(M)} & \boldsymbol{T}_{10(M)} & \boldsymbol{d}_{1(M)} \\ \boldsymbol{0} & \boldsymbol{T}_{00} & \boldsymbol{d}_0 \\ \boldsymbol{0} & \boldsymbol{0} & d_{-1}\end{bmatrix}$$

can now be used in place of the original data arrays. The number of rows in each group is reduced to q except for the M'th group where we include the rows with $\boldsymbol{T}_{00}$, $\boldsymbol{d}_0$, and d_{-1}. These rows must be included in the overall data array but could be included with any of the groups. We include them with the last group for convenience and to emulate the pattern in (19).

With two levels of random effects the initial decompositions are of the form

$$\begin{bmatrix}\boldsymbol{Z}_{ij} & \boldsymbol{Z}_{i,j} & \boldsymbol{X}_{ij} & \boldsymbol{y}_{ij}\end{bmatrix} = \boldsymbol{S}_{(ij)} \begin{bmatrix}\boldsymbol{T}_{22(ij)} & \boldsymbol{T}_{21(ij)} & \boldsymbol{T}_{20(ij)} & \boldsymbol{d}_{2(ij)} \\ \boldsymbol{0} & \boldsymbol{T}_{11(ij)} & \boldsymbol{T}_{10(ij)} & \boldsymbol{d}_{1(ij)}\end{bmatrix} \quad \begin{matrix} i = 1, \ldots, M \\ j = 1, \ldots, M_i \end{matrix}$$

followed by

$$\begin{bmatrix}\boldsymbol{T}_{11(i1)} & \boldsymbol{T}_{10(i1)} & \boldsymbol{d}_{1(i1)} \\ \vdots & \vdots & \vdots \\ \boldsymbol{T}_{11(iM_i)} & \boldsymbol{T}_{10(iM_i)} & \boldsymbol{d}_{1(iM_i)}\end{bmatrix} = \boldsymbol{S}_{(i)} \begin{bmatrix}\boldsymbol{T}_{11(i)} & \boldsymbol{T}_{10(i)} & \boldsymbol{d}_{1(i)} \\ \boldsymbol{0} & \boldsymbol{T}_{00(i)} & \boldsymbol{d}_{0(i)}\end{bmatrix} \quad i = 1, \ldots, M$$

and finally (27). The information required for later calculations is stored in the same form as that in (21) with a $\boldsymbol{R}$ matrix replaced by the corresponding $\boldsymbol{T}$ matrix and a $\boldsymbol{c}$ vector replaced by the corresponding $\boldsymbol{d}$ vector.

Generalization to more than two levels of random effects follows this type of pattern. The similarity between the decompositions needed for the evaluation of the log-likelihood or restricted log-likelihood and those needed for the pre-decomposition step can be exploited when writing software. In our code we use a single function or method for the decomposition step with an additional argument for the matrix $\mathbf{\Delta}$ that should be appended as a new set of rows before decomposing. When doing the pre-decomposition the number of rows to be appended is set to zero. Because the information to be stored for calculation of $\widehat{\boldsymbol{\beta}}$ and the BLUPs for the random effects or for the pre-decomposition is of a similar shape, we use the same function for both storage operations.

4.3 Parallelization

An important consideration in computational methods for modern computer systems is determining which parts of the calculation can be performed in parallel. The structure of the methods we have described divides neatly into separately evaluated pieces.

Within each level of random effects, the reductions and other computations related to the different groups are distinct. Thus they can be performed in parallel. Because all results from a given level of random effects must be available before proceeding to the next level, the extent of possible parallelization is clearly defined.

5 Conclusions and extensions to other models.

Computational methods for maximum likelihood or restricted maximum likelihood estimation of the parameters in a linear multilevel mixed-effects model are greatly enhanced by expressing the variance-covariance matrix of the random effects at each level in terms of a square root of the inverse of the relative variance matrix. These are the matrices that we have written as $\mathbf{\Delta}_q, q = 1, \dots, Q$.

Using this formulation and taking matrix decompositions, the profiled log-likelihood or profiled restricted log-likelihood can be compactly expressed and calculated. These expressions also give an indication of suitable starting values for the variance-covariance parameters.

The calculation of conditional estimates of the fixed effects or BLUPs for the random effects is also expressed compactly using matrix decomposition techniques. Using these expressions an EM algorithm for parameter estimation can be readily derived. The combination of starting estimates calculated from the original design matrices and a moderate number of EM iterations usually puts the parameter values very close to the final parameter estimates. The optimization can be finished with a few Newton-Raphson iterations in a suitable parameterization. Again, we can take advantage of the profiling of the log-likelihood over the values of the fixed-effects parameters.

The computational methods described in the previous sections can also be applied to extensions of the basic Gaussian linear multilevel model of §1. Two such extensions will be considered in this section: Gaussian linear multilevel models with *non-spherical* distributions for the within-group errors and Gaussian *nonlinear* multilevel models.

5.1 General linear multilevel model

The basic Gaussian linear multilevel model assumes that the within-group errors $\boldsymbol{\epsilon}_i$ are distributed as $\mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{I}\right)$. In many applications, especially when longitudinal or spatial data are collected, it is reasonable to allow for correlation among the within-group errors. The assumption of equal variances for the within-group errors also is frequently violated in practice. A more general formulation of the Gaussian multilevel model allows non-spherical Gaussian distributions for the within-group error.

In the single-level linear mixed-effects model (1), the general formulation of the model allows $\boldsymbol{\epsilon}_i \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}_i\right)$, where the scaled variance-covariance matrix $\boldsymbol{\Lambda}_i$ depends on i only through its dimensions and is generally parametrized by a small, fixed set of parameters $\boldsymbol{\rho}$. We assume that the $\boldsymbol{\Lambda}_i, \ i = 1, \ldots, M$ are positive definite. This formulation allows both heteroscedastic (e.g. variance increasing with a power of the expected response) and correlated (e.g. autoregressive-moving average structures) within-group errors. For the general linear two-level model (2) we have $\boldsymbol{\epsilon}_{ij} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}_{ij}\right)$ and this is similarly extended for more levels of nesting.

Let $\boldsymbol{\Lambda}_i^{1/2}$ denote a square-root factor of $\boldsymbol{\Lambda}_i$ (that is, $\boldsymbol{\Lambda}_i = \left(\boldsymbol{\Lambda}_i^{1/2}\right)' \boldsymbol{\Lambda}_i^{1/2}$) and define

$$\boldsymbol{y}_i^\star = \left(\boldsymbol{\Lambda}_i^{-1/2}\right)' \boldsymbol{y}_i \quad \boldsymbol{X}_i^\star = \left(\boldsymbol{\Lambda}_i^{-1/2}\right)' \boldsymbol{X}_i \quad \boldsymbol{Z}_i^\star = \left(\boldsymbol{\Lambda}_i^{-1/2}\right)' \boldsymbol{Z}_i \quad \boldsymbol{\epsilon}_i^\star = \left(\boldsymbol{\Lambda}_i^{-1/2}\right)' \boldsymbol{\epsilon}_i$$

It then follows from elementary properties of the multivariate normal distribution that

$$\begin{gathered} \boldsymbol{y}_i^\star = \boldsymbol{X}_i^\star \boldsymbol{\beta} + \boldsymbol{Z}_i^\star \boldsymbol{b}_i + \boldsymbol{\epsilon}_i^\star, \quad i = 1, \ldots, M \\ \boldsymbol{b}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\epsilon}_i^\star \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}) \end{gathered} \tag{28}$$

That is, $\boldsymbol{y}_i^\star$ follows the basic single-level linear mixed-effects (1). The Jacobian of the linear transformation $\boldsymbol{y}_i \to \boldsymbol{y}_i^\star$ is $1/\operatorname{abs}\left|\boldsymbol{\Lambda}^{1/2}\right|$ and by (10) and (28)

$$\begin{aligned} \ell(\boldsymbol{\theta}, \boldsymbol{\rho}|\boldsymbol{y}) &= \mathit{const} - N \log \|\boldsymbol{c}_{-1}^\star\| + \sum_{i=1}^{M} \log \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}|}{|\boldsymbol{R}_{11(i)}^\star|}\right) - \sum_{i=1}^{M} \log \operatorname{abs} |\boldsymbol{\Lambda}_i^{1/2}| \\ &= \ell\left(\boldsymbol{\theta}|\boldsymbol{y}^\star\right) - \sum_{i=1}^{M} \log \operatorname{abs} |\boldsymbol{\Lambda}_i^{1/2}| \end{aligned}$$

where $\boldsymbol{c}_{-1}^\star$ and $\boldsymbol{R}_{11(i)}^\star$ are the equivalent of $\boldsymbol{c}_{-1}$ and $\boldsymbol{R}_{11(i)}$ for model (28). The restricted log-likelihood is given by

$$\ell_R\left(\boldsymbol{\theta}, \boldsymbol{\rho}|\boldsymbol{y}\right) = \ell_R\left(\boldsymbol{\theta}|\boldsymbol{y}^\star\right) - \sum_{i=1}^{M} \log \operatorname{abs} |\boldsymbol{\Lambda}_i^{1/2}|$$

Similarly, the log-likelihood for the general linear two-level model is

$$\ell\left(\boldsymbol{\theta}, \boldsymbol{\rho}|\boldsymbol{y}\right) = \ell\left(\boldsymbol{\theta}|\boldsymbol{y}^\star\right) - \sum_{i=1}^{M} \sum_{j=1}^{M_i} \log \operatorname{abs} |\boldsymbol{\Lambda}_{ij}^{1/2}|$$

with an equivalent expression for the restricted log-likelihood. This extends to an arbitrary number of levels in the obvious way.

The computational methods described in §4 can, for the most part, be easily extended to the general linear multilevel model by replacing $\boldsymbol{y}, \boldsymbol{X}$, and $\boldsymbol{Z}$ with $\boldsymbol{y}^\star$, $\boldsymbol{X}^\star$, and $\boldsymbol{Z}^\star$. For example, given an initial value $\boldsymbol{\rho}^{(0)}$ for the within-group covariance structure parameters, initial values for $\boldsymbol{\theta}$ can be obtained using the methodology of §2.6 (e.g. $\left[\boldsymbol{\Delta}^{(0)}\right]_{kk} = 0.375\sqrt{\sum_{i=1}^{M} \|\boldsymbol{Z}_i^\star(k)\|^2/M}$ in the single-level case).

The derivation of $\widehat{\boldsymbol{\beta}}$ and the BLUPs for the general linear multilevel model is done as in §3.5, with the obvious substitutions. Pre-decompositions of the data, as described in §4.2, are not meaningful for the general linear multilevel model, as a new decomposition would have to be obtained for each new value of $\boldsymbol{\rho}$ in the optimization process. The results on the distributions of $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and $\boldsymbol{b}|\boldsymbol{y}$ in §3.1 extend in the obvious way to the general linear multilevel model.

Finally, the EM algorithm described in §3.4 has to be modified for the general linear multilevel model. For a given value of $\boldsymbol{\rho} = \boldsymbol{\rho}^{(t)}$, the methodology of §3.4 can be applied to the corresponding $\boldsymbol{y}^\star$ "observations" to obtain updated estimates $\boldsymbol{\theta}^{(t)}$, $\boldsymbol{\beta}^{(t)}$, and $\sigma^{2(t)}$. Assuming these fixed, updated estimates of $\boldsymbol{\rho}$ are obtained by maximizing $\ell\left(\boldsymbol{\rho}|\boldsymbol{y}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\beta}^{(t)}, \sigma^{2(t)}\right)$. This alternating optimization scheme is an example of the ECME algorithm proposed by Liu and Rubin (1994) and shares with the EM algorithm the property of monotone convergence.

5.2 Nonlinear multilevel model

A one-level nonlinear mixed effects model is similar in form to the linear mixed effects models (1) except that the expression $\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$, which is linear in both the fixed effects $\boldsymbol{\beta}$ and the random effects $\boldsymbol{b}_i$, is replaced by a nonlinear expression $\boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)$ where the components of $\boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)$ are given by $\{\boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i)\}_j = f(\boldsymbol{\phi}_{ij}, \boldsymbol{x}_{ij})$. Here f is a nonlinear model function and the subject-specific model parameter for group i at the jth observation is

$$\boldsymbol{\phi}_{ij} = \boldsymbol{A}_{ij}\boldsymbol{\beta} + \boldsymbol{B}_{ij}\boldsymbol{b}_i \qquad .$$

The matrices $\boldsymbol{A}_{ij}$ and $\boldsymbol{B}_{ij}$ are of appropriate dimension and depend on the group and possibly on the values of some covariates at the jth observation. This model is a slight generalization of that described in Lindstrom and Bates (1990) or Davidian and Giltinan (1995) in that $\boldsymbol{A}_{ij}$ and $\boldsymbol{B}_{ij}$ can depend on j. This generalization allows the incorporation of "time-varying" covariates with the fixed effects or the random effects in the model.

A nonlinear model with two levels of random effects can be written

$$\boldsymbol{y}_{ij} = \boldsymbol{f}_{ij}(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{b}_{ij}) + \boldsymbol{\epsilon}_{ij} \quad i = 1, \ldots, M \quad j = 1, \ldots, M_i$$

where $\left\{\boldsymbol{f}_{ij}(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{b}_{ij})\right\}_k = f(\boldsymbol{\phi}_{ijk}, \boldsymbol{x}_{ijk})$ and $\boldsymbol{\phi}_{ijk} = \boldsymbol{A}_{ijk}\boldsymbol{\beta} + \boldsymbol{B}_{i,jk}\boldsymbol{b}_i + \boldsymbol{B}_{ij,k}\boldsymbol{b}_{ij}$. The reason for the somewhat awkward notation is, as before, to allow "time-varying" covariates to be used with the fixed effects or the random effects. Often in practice the matrices $\boldsymbol{A}_{ijk}$, $\boldsymbol{B}_{i,jk}$, and $\boldsymbol{B}_{ij,k}$ are identity matrices or subsets of the columns of an identity matrix.

Extensions to more than two levels of random effects follow the obvious patterns.

Because $\boldsymbol{f}$ can be nonlinear in the fixed or random effects, the integrals required to express the log-likelihood do not have the succinct representations of those in the linear case. Also the conditional estimates of the fixed effects and the conditional expectations of the random effects do not have the analytical solutions they do in the case of the linear model.

As shown in §2 parameter estimation for the Gaussian linear mixed-effects model can be re-expressed as a penalized least-squares problem. It is natural also to re-phrase Gaussian nonlinear mixed-effects models as penalized nonlinear least squares problems approached through linear approximation to the nonlinear model.

For a purely fixed-effects nonlinear regression model a common iterative parameter estimation method is the Gauss-Newton method (Bates and Watts, 1988, §2.2) wherein the nonlinear model $\boldsymbol{f}(\boldsymbol{\beta})$ is replaced by a first-order Taylor series approximation about current estimates $\boldsymbol{\beta}^{(k)}$ as

$$\boldsymbol{f}(\boldsymbol{\beta}) \approx \boldsymbol{f}(\boldsymbol{\beta}^{(k)}) + \left.\frac{d\boldsymbol{f}}{d\boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}} (\boldsymbol{\beta} - \boldsymbol{\beta}^{(k)})$$

The parameter increment $\boldsymbol{\delta}^{(k)} = \boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}$ for the kth iteration is calculated as the least squares solution of

$$\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\beta}^{(k)}) = \left.\frac{d\boldsymbol{f}}{d\boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}} (\boldsymbol{\beta} - \boldsymbol{\beta}^{(k)}) + \boldsymbol{\epsilon}^{(k)}$$

A similar iterative scheme can be used to determine the conditional estimates $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and the conditional modes of the distribution of the random effects given a value of $\boldsymbol{\theta}$ in a nonlinear mixed effects models. Some care must be taken when incorporating the penalty terms such as $\boldsymbol{\Delta}\boldsymbol{b}_i$ into the nonlinear problem because the penalty term is a linear function of $\boldsymbol{b}_i$, not a linear function of the increment $\boldsymbol{b}_i^{(k+1)} - \boldsymbol{b}_i^{(k)}$. Two possible formulations of the penalized nonlinear least squares problem are

$$\begin{bmatrix} \boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\beta}^{(k)}, \boldsymbol{b}_i^{(k)}) \\ -\boldsymbol{\Delta}\boldsymbol{b}_i^{(k)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_i \\ \boldsymbol{0} \end{bmatrix} (\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}) + \begin{bmatrix} \boldsymbol{Z}_i \\ \boldsymbol{\Delta} \end{bmatrix} \left(\boldsymbol{b}_i^{(k+1)} - \boldsymbol{b}_i^{(k)}\right) \quad i = 1, \ldots, M$$

or

$$\begin{bmatrix} \boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\beta}^{(k)}, \boldsymbol{b}_i^{(k)}) + \boldsymbol{X}_i\boldsymbol{\beta}^{(k)} + \boldsymbol{Z}_i\boldsymbol{b}_i^{(k)} \\ \boldsymbol{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_i \\ \boldsymbol{0} \end{bmatrix} \boldsymbol{\beta}^{(k+1)} + \begin{bmatrix} \boldsymbol{Z}_i \\ \boldsymbol{\Delta} \end{bmatrix} \boldsymbol{b}_i^{(k+1)} \qquad (29)$$

where

$$\boldsymbol{X}_i = \frac{d\boldsymbol{f}}{d\boldsymbol{\beta}'} \quad \text{and} \quad \boldsymbol{Z}_i = \frac{d\boldsymbol{f}}{d\boldsymbol{b}_i'}$$

We prefer the second formulation (29) as the calculation of the increment more closely follows the form of the solution of the penalized least-squares problem in the linear case. This is essentially the form of the increment used in Lindstrom and Bates (1990) although this form allows generalization to a multiple levels of random effects.

Once updated values of the fixed-effects parameters $\boldsymbol{\beta}$ and the conditional modes of the random effects are available, the derivative matrices $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are re-evaluated and the variance-covariance parameters $\boldsymbol{\theta}$ are updated by several EM iterations. Only the EM iterations are used at this point because the $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ matrices are going to be recalculated the next time $\boldsymbol{\beta}$ and the conditional modes of the random effects are updated. It is not worthwhile determining the exact optimal $\boldsymbol{\theta}$ for the approximate problem that will subsequently be modified. Once the process of updating $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ separately stabilizes, they can be optimized jointly with higher order approximations to the log-likelihood for the nonlinear mixed-effects model as described in Pinheiro and Bates (1995). The Laplacian approximation would be a good choice of a higher-order approximation.

The fact that the profiled log-likelihood or profiled log-restricted-likelihood for a linear mixed-effects model can be quickly evaluated and also that EM iterations are very fast makes the early optimization of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ for a nonlinear mixed-effects model faster and more stable. Also, the current formulation of the linear mixed-effects model allows extension of nonlinear mixed-effects models to multiple nested levels of random effects which was not previously available.

5.3 Implementation

The computational methods for multilevel Gaussian mixed-effects models, general linear mixed-effects models and nonlinear nested mixed-effects models described in this paper are implemented in version 3.0 of the NLME library for S, S-PLUS, and R. Documentation and source code for this library is available at

http://franz.stat.wisc.edu/pub/NLME/

References

Anderson, E., Bai, Z., Bischoff, C., Demmel, J., Dongarra, J., DuCroz, J., Greenbaum, A., Hammarling, S., McKenney, A., Ostrouchov, S. and Sorensen, D. (1994). *LAPACK Users' Guide, 2nd ed.*, SIAM.

Bates, D. M. (1983). The derivative of $|X'X|$ and its uses, *Technometrics* **25**: 373–376.

Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*, Wiley, New York.

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*, 1st edn, Chapman & Hall, London.

Dongarra, J. J., Bunch, J. R., Moler, C. B. and Stewart, G. W. (1979). *Linpack Users' Guide*, SIAM, Philadelphia.

Goldstein, H. (1995). *Multilevel Statistical Models*, Halstead Press, New York.

Golub, G. H. and Pereyra, V. (1973). The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate, *J. SIAM* **10**: 413–432.

Harville, D. A. (1976). Extension of the Gauss-Markov theorem to include the estimation of random effects, *The Annals of Statistics* **4**: 384–395.

Kaufman, L. (1975). A variable projection method for solving separable nonlinear least squares problems, *BIT* **15**: 49–57.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data, *Biometrics* **38**: 963–974.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data, *Journal of the American Statistical Association* **83**: 1014–1022.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data, *Biometrics* **46**: 673–687.

Liu, C. and Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence, *Biometrika* **81**: 633–648.

Longford, N. T. (1993). *Random Coefficient Models*, Oxford University Press.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model, *Journal of Computational and Graphical Statistics* **4**(1): 12–35.

Pinheiro, J. C. and Bates, D. M. (1996). Unconstrained parameterizations for variance-covariance matrices, *Statistics and Computing* **6**: 289–296.

Thisted, R. A. (1988). *Elements of Statistical Computing*, Chapman & Hall.