Scientific Research Publishing

Open Journal of Statistics, 2021, 11, 558-605
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Nonparametric Estimation in Linear Mixed Models with Uncorrelated Homoscedastic Errors

Eugène-Patrice Ndong Nguéma, Betrand Fesuh Nono*, Henri Gwét

Department of Mathematics, Ecole Polytechnique, Yaoundé, Cameroon
Email: *fesuhbe@yahoo.co.uk

How to cite this paper: Ndong Nguéma, E.-P., Fesuh Nono, B. and Gwét, H. (2021) Nonparametric Estimation in Linear Mixed Models with Uncorrelated Homoscedastic Errors. *Open Journal of Statistics*, **11**, 558-605.
https://doi.org/10.4236/ojs.2021.114035

Received: July 2, 2021
Accepted: August 27, 2021
Published: August 30, 2021

Abstract

Today, Linear Mixed Models (LMMs) are fitted, mostly, by assuming that random effects and errors have Gaussian distributions, therefore using Maximum Likelihood (ML) or REML estimation. However, for many data sets, that double assumption is unlikely to hold, particularly for the random effects, a crucial component in which assessment of magnitude is key in such modeling. Alternative fitting methods not relying on that assumption (as ANOVA ones and Rao's MINQUE) apply, quite often, only to the very constrained class of variance components models. In this paper, a new computationally feasible estimation methodology is designed, first for the widely used class of 2-level (or longitudinal) LMMs with only assumption (beyond the usual basic ones) that residual errors are uncorrelated and homoscedastic, with no distributional assumption imposed on the random effects. A major asset of this new approach is that it yields nonnegative variance estimates and covariance matrices estimates which are symmetric and, at least, positive semi-definite. Furthermore, it is shown that when the LMM is, indeed, Gaussian, this new methodology differs from ML just through a slight variation in the denominator of the residual variance estimate. The new methodology actually generalizes to LMMs a well known nonparametric fitting procedure for standard Linear Models. Finally, the methodology is also extended to ANOVA LMMs, generalizing an old method by Henderson for ML estimation in such models under normality.

Keywords

Clustered Data, Linear Mixed Model, Fixed Effect, Uncorrelated Homoscedastic Error, Random Effects Predictor

1. Introduction

The Linear Mixed Model (LMM) is an extension of the classical Linear Model

(LM) aimed at modeling a continuous scalar response Y in terms of observed covariates some of which (say $X_1,\cdots,X_p$) have fixed effects as in the LM, being the same for all individuals in the population under study, and others (say $Z_1,\cdots,Z_r$) have random effects, thus possibly varying between some well identified subgroups in that population. Mixed models have been used to analyze data sets with clustered, longitudinal or multilevel structure in a variety of fields, such as medicine [1] [2], agriculture [3], animal breeding [4], small area estimation [5] [6], genetics [7] [8], growth modeling [9] [10], etc. Detailed presentations of LMMs can be found in [11] or Demidenko [12], with a more practical emphasis in West *et al*. [13]. The subclass of variance components (or ANOVA) LMMs is thoroughly examined in Searle *et al*. [14].

At the outset of this field, that latter class of LMMs was almost exclusively the only one considered to fit real world data sets, with ANOVA methods at the forefront. But things started to turn around by the end of the 1970s, with the advent of widespread powerful computational tools (both in hardware and software) which have now sufficiently matured and are widely accessible. As a consequence, nowadays by far the most popular approach to fitting an LMM to a given data set is to use a Gaussian LMM, *i.e*. assuming that both the random effects and the model residual errors have Gaussian distributions, then using either the Maximum Likelihood (ML) or the Restricted ML (REML) estimation procedures. This is not easy task, however, since the corresponding (restricted) likelihood equations are quite involved, even to simply be derived, and have no closed form solutions. Thus in the last 4 - 5 decades, considerable endeavor, both theoretical and computational, has been devoted to deriving and numerically solving these equations as efficiently as possible [11] [15] [16] [17] [18] [19]. The numerical solution of the nonlinear system of (restricted) likelihood equations is done, most often, through using one of two standard iterative methods: the Newton-Raphson (NR) or Fisher scoring (FS, a more statistical variant of NR) and the Expectation-Maximization (EM) algorithm in various forms [20] [21] [22].

Despite that unrivaled popularity, there is an obvious major issue with fitting Gaussian LMMs: the assumption of normality of both random effects and errors is dubious in many practical settings, especially for the former. Lange and Ryan [23], for instance, present practical cases of nonnormal random effects. In the last decade, faced with bigger and bigger data sets, both in size and dimensionality, a growing interest has been focused on how to analyze LMMs for such data. For instance, a quasi-likelihood approach for estimation and inference in linear mixed-effects models with high-dimensional fixed effects and possibly large or unbalanced cluster sizes has been recently proposed by Li *et al*. [24].

Now, prior to the widespread adoption of ML/REML in LMMs fitting, some methods not relying on Gaussian assumptions had been developed and used such as ANOVA methods [14] [25] [26] [27] [28], Henderson's Methods I, II and III [29], Rao's Minimum Norm Quadratic Unbiased Estimation (MINQUE) method [30] [31] [32] [33], iterative MINQUE [34]. But these are tailored only

to a very specific and limited class of LMMs, namely ANOVA ones (or variance components LMMs). Moreover, these methods share the common significant drawback of not ensuring nonnegative estimates of variance components. A qualified discussion in Searle *et al*. [14] seldom recommended them for general usage. Alternatively, a quasi-likelihood approach is presented by Jiang [35] using the REML equations in non-Gaussian ANOVA LMMs, showing consistency and asymptotic normality of the variance components estimators under some additional assumptions. Heyde [36] does the same, but emphasizing the relationship between quasi-likelihood and estimating functions. Iterative weighted least squares and iterative generalized least squares estimation methods for LMMs are presented, respectively, in Jiang *et al*. [37] and Goldstein [38].

Being based on quite restrictive assumptions, the aforementioned methods developed for non-Gaussian LMMs have a limited range of LMMs for which they are theoretically valid fitting methods. For instance, the overly simplifying assumptions on covariance matrices structure (like the ones in ANOVA and MINQUE methods) even exclude most Gaussian LMMs from that validity range, which helps explain why the ML and REML approaches have superseded them. In contrast, our goal in this work is to design an estimation methodology for LMMs in which the fixed effects vector of parameters and the random effects covariance matrix are estimated based on assumptions as weak as possible. We will first restrict attention to the very popular subclass of 2-level (also called *longitudinal*) LMMs pioneered by Laird and Ware [22], before extending the approach to fit ANOVA LMMs as well.

To achieve that goal, we take the practical standpoint here that, in most situations, except when some additional information is available about the data (such as serial correlation in errors for some longitudinal data), one has no other choice than to assume that the residual errors in the LMM to fit are uncorrelated and homoscedastic (u.ho.). This is probably the default option for error modeling (with the Gaussian assumptions when using ML or REML) in most statistical software packages tailored for parameters estimation in an LMM. In our modeling of 2-level LMMs here, based on adding only that assumption to the basic ones of zero mean and finite covariances for random effects and errors, we devise a new approach for estimating the fixed effects parameters vector β, the cluster random effects covariance matrix $\boldsymbol{D}$ and the residual errors variance σ_ε^2. Thus, we do not impose any assumption on the type of clusters random effects distribution. The approach can be viewed as an adaptation to 2-level LMMs of the well known 2-step procedure for fitting the LM to a given data set, under that same assumption, where first β is estimated by Ordinary Least Squares (OLS), then the variance σ_ε^2 of the errors by a carefully designed unbiased estimator. On closer scrutiny, it also turns out to generalize a little popularized estimating procedure for ML computation in Gaussian variance components models credited to Henderson by Harville [16] and detailed in Searle *et al*. ([14], pages 278-279).

As for the organization of the material presented in this article, Section 2 briefly reviews LMMs, up to the random effects prediction issue. Section 3 highlights some key preliminary results on LMMs which do not involve any Gaussian distributional assumptions, which will serve as backbone for our new estimation methodology to be designed. In Section 4, we construct that new estimation methodology for 2-level LMMs with u.ho. errors. Section 5 shows that our estimation methodology can be adapted to both 2-level LMMs with u.ho. errors and a diagonal cluster random effects covariance matrix and ANOVA LMMs. In Section 6, our new estimation approach is compared with the traditional Gaussian ML through a simulation study, an application to a classical data set with cluster structure, and a longitudinal data, with implementations done in the R software [39]. Section 7 draws our conclusion about the work presented. The Appendix contains the most lengthy proof of a result presented in the running text. A supplementary material document is also provided, which gathers some known results on LMMs which we have used, but scattered here and there in the literature, and some new results of our own, as well as important implementation details about our presented iterative methods for fitting LMMs. To make a distinction with the article, the numbering of sections, results and equations in that document is prefixed by the letter "S".

Before we proceed, please note that in this paper, all vectors are columns. Moreover, A^{T}, $\mathrm{tr}(A)$, $\|u\| = \sqrt{u^{\mathrm{T}}u}$ and $\|u\|_M = \sqrt{u^{\mathrm{T}}Mu}$ respectively denote transpose of matrix A, trace of square matrix A, Euclidean norm and norm w.r.t. the SPD matrix M of vector u, where "SPD" stands for *symmetric and positive definite* while "SPSD" means *symmetric and positive semi-definite* (matrix). $\boldsymbol{I}_n$ is the identity matrix of order n. $\mathscr{M}_{n,p}(\mathbb{R})$ and $\mathscr{M}_n(\mathbb{R})$ are the spaces of matrices with real elements, respectively $n\times p$ and $n\times n$. $\mathbb{E}(X)$ and $\mathbb{M}\mathrm{cov}(X)$ denote the mean and covariance matrix of random vector (or variable) X. $\mathscr{N}_q(\boldsymbol{m},\Sigma)$ is the q-dimensional Gaussian distribution with mean vector $\boldsymbol{m}$ and covariance matrix Σ, $\overset{\mathscr{D}}{\sim}$ is for "has probability distribution", while *estimating equation* will be abbreviated *EE*.

2. Linear Mixed Models: An Overview

2.1. The General Form of an LMM

The general form of an LMM is [11]:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{Z}U + \varepsilon, \tag{1}$$

where $\boldsymbol{Y} \in \mathbb{R}^n$ is the observed response, $\beta \in \mathbb{R}^p$ is the unknown vector of fixed effects parameters, $U \in \mathbb{R}^q$ is the vector of unobserved random effects, $\varepsilon \in \mathbb{R}^n$ comprises the unknown residual errors, while $\boldsymbol{X} \in \mathscr{M}_{n,p}(\mathbb{R})$ and $\boldsymbol{Z} \in \mathscr{M}_{n,q}(\mathbb{R})$ are given design matrices. The usual basic assumptions are [11] [12] [22]:

Assumption $\mathscr{A}1_g$. *U and ε are two independent zero mean random vectors with respective positive definite covariance matrices* $\mathbb{M}\mathrm{cov}(U) = \boldsymbol{G} \in \mathscr{M}_q(\mathbb{R})$

and $\mathbb{M}\mathrm{cov}(\varepsilon) = \boldsymbol{R} \in \mathscr{M}_n(\mathbb{R})$.

With these assumptions, the covariance matrix of $\boldsymbol{Y}$ is $\boldsymbol{V} = \mathbb{M}\mathrm{cov}(\boldsymbol{Y}) = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}} + \boldsymbol{R}$. We will also use the following assumption:

Assumption $\mathscr{A}2_g$. *The* $n \times p$ *design matrix* $\boldsymbol{X}$ *in* (1) *has full column rank*.

In (1), the main objective is to estimate β and the covariance matrices $\boldsymbol{R}$ and $\boldsymbol{G}$. LMMs are classified into two main groups, based on Gaussian distributional assumptions made or not on the random effects and errors in devising estimation procedures. Model (1) is termed Gaussian if, in addition to Assumption $\mathscr{A}1_g$, it holds that:

Assumption $\mathscr{A}3_g$. $U \overset{\mathscr{L}}{\sim} \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{G})$ *and* $\varepsilon \overset{\mathscr{L}}{\sim} \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{R})$.

If the random effects and/or errors are not assumed to be normal, but Assumption $\mathscr{A}1_g$ holds, then model (1) is usually said to be (rather abusively) a non-Gaussian LMM. The main goal in this work is to devise estimation procedures for LMMs which do not use Assumption $\mathscr{A}3_g$. The only key added assumption will rather be:

Assumption $\mathscr{A}4_g$. $\boldsymbol{R} = \sigma_\varepsilon^2 \boldsymbol{I}_n$ $(\sigma_\varepsilon^2 > 0)$, *i.e. the residual errors in* (1) *are uncorrelated and homoscedastic*.

2.2. The 2-Level (or Longitudinal) LMM

We will first design our new estimation methodology for one of the types of LMM most used in statistical analyzes. Since Laird and Ware [22], this type of LMM is usually qualified as *longitudinal*, but we feel it more encompassing to label them as 2-level. In this article, we will use the *units-cluster* terminology for 2-level LMMs. At the cluster level, such a model can be written by grouping the scalar responses for all units in each cluster *j* as:

$$\boldsymbol{Y}_j = \boldsymbol{X}_j \beta + \boldsymbol{Z}_j U_j + \varepsilon_j, \quad j = 1, \cdots, m, \tag{2}$$

where $\boldsymbol{Y}_j = \left(Y_{1j}, \cdots, Y_{n_j,j}\right)^{\mathrm{T}} \in \mathbb{R}^{n_j}$ is the response vector; $\varepsilon_j = \left(\varepsilon_{1j}, \cdots, \varepsilon_{n_j,j}\right)^{\mathrm{T}} \in \mathbb{R}^{n_j}$ is the vector of the errors; $\boldsymbol{X}_j \in \mathscr{M}_{n_j,p}(\mathbb{R})$ is the fixed effects design matrix; $\boldsymbol{Z}_j \in \mathscr{M}_{n_j,r}(\mathbb{R})$ is the random effects design matrix. The unknown parts of the model are $\beta = \left(\beta_1, \beta_2, \cdots, \beta_p\right)^{\mathrm{T}} \in \mathbb{R}^p$, the vector of fixed effects parameters, and $U_j \in \mathbb{R}^r$, the unobserved vector of random effects for cluster *j*. We will also consider the following usual basic assumptions for such models:

Assumption $\mathscr{A}1$. *The* U_j*'s are independent and identically distributed in* $\mathbb{R}^r$.

Assumption $\mathscr{A}2$. *The* ε_j*'s are independent*.

Assumption $\mathscr{A}3$. *The set of* U_j*'s and set of* ε_j*'s are independent from each other*.

Assumption $\mathscr{A}4$. $\mathbb{E}\left(U_j\right) = \boldsymbol{0} \in \mathbb{R}^r$, $\mathbb{E}\left(\varepsilon_j\right) = \boldsymbol{0} \in \mathbb{R}^{n_j}$, $\mathbb{M}\mathrm{cov}\left(U_j\right) = \boldsymbol{D} \in \mathscr{M}_r(\mathbb{R})$, $\mathbb{M}\mathrm{cov}\left(\varepsilon_j\right) = \boldsymbol{R}_j \in \mathscr{M}_{n_j}(\mathbb{R})$. *Here,* $\boldsymbol{D}$ *and* $\boldsymbol{R}_j$ *are, respectively, the covariance matrices of* U_j *and* ε_j *and are assumed finite and positive definite*.

Model (2) can be written in the form (1), but with the covariance matrices having special diagonal block structures: $\boldsymbol{G}=\operatorname{diag}(\boldsymbol{D},\cdots,\boldsymbol{D})$, $\boldsymbol{R}=\operatorname{diag}(\boldsymbol{R}_1,\cdots,\boldsymbol{R}_m)$, $\boldsymbol{V}=\operatorname{diag}(\boldsymbol{V}_1,\cdots,\boldsymbol{V}_m)$, with $\boldsymbol{V}_j=\boldsymbol{Z}_j\boldsymbol{D}\boldsymbol{Z}_j^{\mathrm{T}}+\boldsymbol{R}_j$, $j=1,\cdots,m$.

2.3. Variance Components Models (or ANOVA LMMs)

We will subsequently show that our methodology can also be adapted to the popular ANOVA LMMs, *i.e.* LMMs (1) in which the term $\boldsymbol{Z}U$ can be decomposed as:

$$\boldsymbol{Z}U=\boldsymbol{Z}_1U_1+\cdots+\boldsymbol{Z}_mU_m, \tag{3}$$

where $\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_m$ are given design matrices, with $\boldsymbol{Z}_j\in\mathscr{M}_{n,q_j}(\mathbb{R})$, $j=1,\cdots,m$, and $U_1,\cdots,U_m$ now become the random effects vectors of respective dimensions $q_1,\cdots,q_m$ ($q_1+\cdots+q_m=q$). Taking $\boldsymbol{Z}_{m+1}=\boldsymbol{I}_n$ and $U_{m+1}=\varepsilon$ (so $q_{m+1}=n$), the ANOVA LMM (1) with random effects term (3) is more neatly written:

$$\boldsymbol{Y}=\boldsymbol{X}\beta+\boldsymbol{Z}\tilde{U}, \tag{4}$$

with $\boldsymbol{Z}=(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_{m+1})$, $\tilde{U}=\left(U_1^{\mathrm{T}},\cdots,U_{m+1}^{\mathrm{T}}\right)^{\mathrm{T}}$ and default assumptions:

M1. $U_1,\cdots,U_{m+1}$ are independent;

M2. $\forall j=1,\cdots,m+1$, $\mathbb{E}(U_j)=\boldsymbol{0}$ and $\mathbb{M}\mathrm{cov}(U_j)=\sigma_j^2\boldsymbol{I}_{q_j}$, with $\sigma_j^2>0$ finite.

Then (4) implies $\boldsymbol{G}=\operatorname{diag}\left(\sigma_1^2\boldsymbol{I}_{q_1},\cdots,\sigma_m^2\boldsymbol{I}_{q_m}\right)$, and the only unknown parameters in (4), besides β, are the vector of components variances $\theta=\left(\sigma_1^2,\cdots,\sigma_{m+1}^2\right)^{\mathrm{T}}$. That is why (4) is called, under **M1** and **M2**, a *variance components model*.

2.4. Prediction of Random Effects: The BLP and the BLUP of U

Besides being sometimes an important target in LMM modeling, it turns out that having an effective and computable (given parameters values) predictor for the random effects vector U in the LMM (1) is a crucial component in our newly designed estimation methodology. In that respect, a first candidate is the best linear predictor (BLP) of U given $\boldsymbol{Y}$ (see [14], Chapter 7):

$$\mathrm{BLP}(U\mid\boldsymbol{Y})=\bar{U}=\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\beta). \tag{5}$$

But a far more popular predictor is the best linear unbiased predictor (BLUP) of $U\mid\boldsymbol{Y}$ given, under Assumptions $\mathscr{A}1_g$ - $\mathscr{A}2_g$, by Searle [40]:

$$\mathrm{BLUP}(U\mid\boldsymbol{Y})=\tilde{U}^*=\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\left(\boldsymbol{Y}-\boldsymbol{X}\tilde{\beta}^*\right), \tag{6}$$

where $\tilde{\beta}^*$ is the Best Linear Unbiased Estimator (BLUE) of β,

$$\tilde{\beta}^*=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{Y}. \tag{7}$$

But, in our new methodology (as in previous ones), trying to use either the BLP or the BLUP requires to first estimate $\boldsymbol{G}$ and $\boldsymbol{R}$. Replacing them in (6) by estimators $\hat{\boldsymbol{G}}$ and $\hat{\boldsymbol{R}}$ yields $\hat{U}^*=\mathrm{EBLUP}(U\mid\boldsymbol{Y})=\hat{\boldsymbol{G}}\boldsymbol{Z}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1}\left(\boldsymbol{Y}-\boldsymbol{X}\hat{\beta}^*\right)$, a so called Empirical BLUP (EBLUP) of $U\mid\boldsymbol{Y}$, with $\hat{\boldsymbol{V}}=\boldsymbol{Z}\hat{\boldsymbol{G}}\boldsymbol{Z}^{\mathrm{T}}+\hat{\boldsymbol{R}}$, and an Empirical BLUE (EBLUE) $\hat{\beta}^*=\left(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1}\boldsymbol{Y}$ of β. However, importantly for designing our new fitting methodology for LMMs, we stress that

the $\mathrm{BLUP}(U|\boldsymbol{Y})$ can also be viewed as a preliminary estimator of $\mathrm{BLP}(U|\boldsymbol{Y})$. Consequently, any $\mathrm{EBLUP}(U|\boldsymbol{Y})$ can also be taken as a practical estimator of $\bar{U}=\mathrm{BLP}(U|\boldsymbol{Y})$, thus qualifying also to be an Empirical BLP (EBLP) of $U|\boldsymbol{Y}$. Indeed, that is how we will estimate $\bar{U}$, using Henderson's mixed model equations briefly reviewed next.

2.5. Henderson's Mixed Model Equations

For given $\boldsymbol{G}$ and $\boldsymbol{R}$ in a Gaussian LMM (1), to simultaneously get an estimator for β and a predictor for U, Henderson [41] [42] maximized, w.r.t. β and U (the latter also viewed as a parameter), the joint density of the random couple $(\boldsymbol{Y},U)$. That maximization yields the so called *Henderson's mixed model equations* (HMMEs) given in matrix form as:

$$\begin{pmatrix}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{Z}\\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1}\end{pmatrix}\begin{pmatrix}\tilde{\beta}\\ \tilde{U}\end{pmatrix}=\begin{pmatrix}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{Y}\\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{Y}\end{pmatrix}. \tag{8}$$

From the outset, there has been debate over whether this was a valid way of trying to estimate β and predict U. Henderson himself acknowledged (see [43], page 16) that (even given an observed $\boldsymbol{Y}=\boldsymbol{y}\in\mathbb{R}^n$) the maximized function is not a true likelihood for the couple (β,U) because U is not a fixed unknown parameter but rather an unobserved random variable. So, strictly speaking, this does not qualify as an ML method. Nonetheless, from our standpoint, that debate is mostly peripheral. The only two relevant practical questions are: 1) What are the solutions $\tilde{\beta}$ and $\tilde{U}$ of the linear system (8)? 2) What properties do they have if taken as respective estimator of β and predictor of U? The answer to both questions is provided by the following result, credited by Harville [16] to Henderson in an unpublished 1963 report:

Theorem 1 ([16] [44]). *If* $\boldsymbol{X}$ *is full column rank while* $\boldsymbol{G}$ *and* $\boldsymbol{R}$ *are SPD matrices, then the solutions of the linear system* (8) *are* $\tilde{\beta}=\tilde{\beta}^{*}$ *and* $\tilde{U}=\tilde{U}^{*}$ *given by* (6)-(7), *with* $\boldsymbol{V}=\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}+\boldsymbol{R}$.

So, in the general LMM (1), given $\boldsymbol{G}$, $\boldsymbol{R}$ and the observed response vector $\boldsymbol{Y}$, the BLUE $\tilde{\beta}^{*}$ of β and the BLUP $\tilde{U}^{*}$ of $U|\boldsymbol{Y}$ are the unique solutions of the HMMEs.

3. More about Henderson's Mixed Model Equations and LMMs without Gaussian Assumptions

As announced in the introduction, we intend to devise a new estimation methodology, first for 2-level LMMs, which, contrary to usage outside of the Gaussian case, simultaneously estimates unknown parameters and predicts random effects values under Assumptions $\mathscr{A}1$ - $\mathscr{A}4$ and $\mathscr{A}4_g$. In that perspective, we give, in this section, and, especially, in Section 3.2, a series of new results about LMMs which do not use any Gaussian assumption, be it on the random effects or the residual errors. Those results will serve as basis for the design of our new fitting methodology for LMMs in Section 4.

3.1. An Important Preliminary: The BLUE, the BLUP and the HMMEs Do Not Need Gaussian Assumptions

Though Henderson, as reported above, famously discovered his eponymous mixed model Equations (8) coming from the Gaussian route, Theorem 1, which solves them, uses no Gaussian distributional assumptions for that. It actually uses no probabilistic, nor statistical framework or reasoning, whatsoever. It is rather a pure Matrix Algebra result asserting that the linear system (8) has as unique solutions the BLUE $\tilde{\beta}^*$ (7) of β and the BLUP $\tilde{U}^*$ (6) of $U|\boldsymbol{Y}$. Now, one can derive both $\tilde{\beta}^*$ and $\tilde{U}^*$ by optimizing, for each, a specific minimum variance criterion, but without relying on any Gaussian assumption or any parametric one for that matter [40]. That double observation makes it that the HMMEs are not attached to Gaussian distributional assumptions, contrary to what people always feel compelled to set before they use them.

Remark 1. *For the latter reason, we will use the HMMEs as our first two estimating equations when devising the first approach* (*coded* 3*S*-*A*1-*V*1) *of our new estimation methodology for LMMs in Section* 4.1.

3.2. More about the HMMEs Solutions

Here, we present a series of new Matrix Algebra results and their consequences for LMMs with u.ho. errors, without imposing any Gaussian assumption, much in line with Theorem 1. The discovery of those results triggered the design of our new fitting methodology for that class of LMMs. They derive from the very peculiar structure of the first HMME which we start by stressing.

3.2.1. The First HMME as a Weighted Least Squares Problem

In (8), the first equation can be rewritten: $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{X}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{R}^{-1}\left(\boldsymbol{Y}-\boldsymbol{Z}\tilde{U}\right)$. One then recognises the normal equations of the Weighted Least Squares (WLS) problem:

$$\min_{\beta\in\mathbb{R}^p}\left\|\left(\boldsymbol{Y}-\boldsymbol{Z}\tilde{U}\right)-\boldsymbol{X}\beta\right\|_{\boldsymbol{R}^{-1}}. \tag{9}$$

Combining that observation with Theorem 1 yields the remarkable result:

Theorem 2. *In the LMM* (1) *with Assumptions* $\mathscr{A}1_g$-$\mathscr{A}2_g$, *and given the BLUP* $\tilde{U}^*$ *of* $U|\boldsymbol{Y}$, *the BLUE* $\tilde{\beta}^*$ *of* β *is the WLS estimate of* β *in the LM*: $\boldsymbol{Y}-\boldsymbol{Z}\tilde{U}^*=\boldsymbol{X}\beta+\tilde{\varepsilon}$, *with response* $\boldsymbol{Y}-\boldsymbol{Z}\tilde{U}^*$, *design matrix* $\boldsymbol{X}$, *error* $\tilde{\varepsilon}$ *and weighting matrix* $\boldsymbol{R}^{-1}$.

But Theorem 2 will be more useful to us in designing our new methodology for LMM fitting in Section 4 when one adds the assumption that the LMM has uncorrelated and homoscedastic errors. We detail hereafter why.

3.2.2. More about the BLUE of β and the BLUP of U in an LMM with u.ho. Errors

Here, we focus attention on LMMs with u.ho. residual errors, *i.e.* Assumption $\mathscr{A}4_g$ holds, in addition to $\mathscr{A}1_g$-$\mathscr{A}2_g$. But again, no Gaussian assumption will be relied upon: only the u.ho. errors and their impact on the geometry of the

HMMEs solutions play a role here. Furthermore, we emphasize that our results to be presented hereafter are valid for any LMM (1) with u.ho. errors, and not only for 2-level LMMs with such errors. In that respect, first note that in the u.ho. residual errors scenario, the HMMEs (8) simplify significantly. Indeed, if $\boldsymbol{R}=\sigma_\varepsilon^2\boldsymbol{I}_n$ with $\sigma_\varepsilon^2>0$, then (8) reduces to:

$$\begin{pmatrix}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}\\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}+\sigma_\varepsilon^2\boldsymbol{G}^{-1}\end{pmatrix}\begin{pmatrix}\tilde{\beta}\\ \tilde{U}\end{pmatrix}=\begin{pmatrix}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Y}\end{pmatrix}. \tag{10}$$

The results to be derived hereafter highlight some peculiar properties about the solutions of the linear system (10), thus yielding some remarkable intertwined relationships, under Assumption $\mathscr{A}4_g$, between its unique solutions (Theorem 1), the BLUE $\tilde{\beta}=\tilde{\beta}^*$ of β and $\tilde{U}=\tilde{U}^*$, the BLUP of $U\,|\,\boldsymbol{Y}$.

We start with an identity which is a trivial consequence of the structure of $\boldsymbol{V}$ in an LMM (1) under u.ho. errors. Indeed, then $\boldsymbol{V}=\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}+\sigma_\varepsilon^2\boldsymbol{I}_n$, so:

$$\boldsymbol{I}_n-\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}=\sigma_\varepsilon^2\boldsymbol{V}^{-1}. \tag{11}$$

This is instrumental first for:

Lemma 3. *Consider the LMM* (1) *with Assumptions* $\mathscr{A}1_g$ *and* $\mathscr{A}4_g$. *Let* $\left(\tilde{\beta},\tilde{U}\right)\in\mathbb{R}^p\times\mathbb{R}^q$ *such that* $\tilde{U}=\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\left(\boldsymbol{Y}-\boldsymbol{X}\right)$. *Then one has*:

$$\boldsymbol{Y}-\boldsymbol{X}\tilde{\beta}-\boldsymbol{Z}\tilde{U}=\sigma_\varepsilon^2\boldsymbol{V}^{-1}\left(\boldsymbol{Y}-\boldsymbol{X}\tilde{\beta}\right). \tag{12a}$$

Moreover, if Assumption $\mathscr{A}2_g$ *also holds, then the following equivalence is true*:

$$\tilde{\beta}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{Y}\Leftrightarrow\tilde{\beta}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{Y}-\boldsymbol{Z}\tilde{U}\right). \tag{12b}$$

Proof. Let Assumptions $\mathscr{A}1_g$ and $\mathscr{A}4_g$ hold, $\boldsymbol{R}=\sigma_\varepsilon^2\boldsymbol{I}_n$ and $\tilde{U}=\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\tilde{\boldsymbol{Y}}$, with $\tilde{\boldsymbol{Y}}=\boldsymbol{Y}-\boldsymbol{X}$, for an arbitrary $\tilde{\beta}\in\mathbb{R}^p$. Then $\boldsymbol{Y}-\boldsymbol{X}\tilde{\beta}-\boldsymbol{Z}\tilde{U}=\tilde{\boldsymbol{Y}}-\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\tilde{\boldsymbol{Y}}=\left(\boldsymbol{I}_n-\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\right)\tilde{\boldsymbol{Y}}=\sigma_\varepsilon^2\boldsymbol{V}^{-1}\tilde{\boldsymbol{Y}}$, the last equality thanks to (11), which proves (12a).

Now, if Assumption $\mathscr{A}2_g$ is also true, then the square matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is nonsingular; so, given that $\boldsymbol{Y}-\boldsymbol{Z}\tilde{U}=\boldsymbol{X}\tilde{\beta}+\sigma_\varepsilon^2\boldsymbol{V}^{-1}\tilde{\boldsymbol{Y}}$ from (12a), then $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{Y}-\boldsymbol{Z}\tilde{U}\right)=\tilde{\beta}+\sigma_\varepsilon^2\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\left(\boldsymbol{Y}-\boldsymbol{X}\tilde{\beta}\right)$. Hence, since $\sigma_\varepsilon^2>0$, $\tilde{\beta}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{Y}-\boldsymbol{Z}\tilde{U}\right)\Leftrightarrow\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\left(\boldsymbol{Y}-\boldsymbol{X}\tilde{\beta}\right)=\boldsymbol{0}$. Then (12b) follows because Assumption $\mathscr{A}2_g$ implies that $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X}$ is an SPD matrix, thus nonsingular. □

Theorem 1 and Lemma 3 imply the following for the solutions of (10):

Theorem 4. *In the LMM* (1), *if* $P=\boldsymbol{V}^{-1}-\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}$ *and Assumptions* $\mathscr{A}1_g$ - $\mathscr{A}2_g$, $\mathscr{A}4_g$ *hold, then*:

$$\boldsymbol{Y}-\boldsymbol{X}\tilde{\beta}^*-\boldsymbol{Z}\tilde{U}^*=\sigma_\varepsilon^2\boldsymbol{V}^{-1}\left(\boldsymbol{Y}-\boldsymbol{X}\tilde{\beta}^*\right)=\sigma_\varepsilon^2P\boldsymbol{Y}, \tag{13a}$$

$$\tilde{\beta}^*=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{Y}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{Y}-\boldsymbol{Z}\tilde{U}^*\right). \tag{13b}$$

Proof. Thanks to Lemma 3, the respective definitions (7) and (6) of the BLUE $\tilde{\beta}^*$ and the BLUP $\tilde{U}^*$, and the standard identity $\boldsymbol{V}^{-1}\left(\boldsymbol{Y}-\boldsymbol{X}\tilde{\beta}^*\right)=P\boldsymbol{Y}$. □

Remark 2. *The usefulness of the second expression of* $\tilde{\beta}^*$ *in* (13b) *stems*

from the fact that one recognizes the OLS, under Assumption $\mathscr{A}2_g$, *of* β *in the LM* $\tilde{\boldsymbol{Y}} = \boldsymbol{X}\beta + \tilde{\varepsilon}$, *where* $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{Z}\tilde{U}^*$. *This clearly suggests how, in an iterative algorithm for LMM fitting, one may update the estimate of* β *given the current estimate of the BLUP of U. That observation is precisely what motivated the construction of the second approach* (*coded* 3*S*-*A*1-*V*2) *in our new estimation methodology for LMMs in Section* 4.1.

We then have the following corollary:

Corollary 1. *If* $A = \boldsymbol{Y}^{\mathrm{T}}\left(\boldsymbol{Y} - \boldsymbol{X}\tilde{\beta}^* - \boldsymbol{Z}\tilde{U}^*\right)$ *and* $B = \left\|\boldsymbol{Y} - \boldsymbol{X}\tilde{\beta}^* - \boldsymbol{Z}\tilde{U}^*\right\|^2$ *in the LMM* (1) *with Assumptions* $\mathscr{A}1_g$ - $\mathscr{A}2_g$ *and* $\mathscr{A}4_g$, *then*

$$A - B = \sigma_\varepsilon^2 \cdot \left\|\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\left(\boldsymbol{Y} - \boldsymbol{X}\tilde{\beta}^*\right)\right\|_{\boldsymbol{G}}^2 = \sigma_\varepsilon^2 \cdot \left\|\tilde{U}^*\right\|_{\boldsymbol{G}^{-1}}^2. \tag{14}$$

Therefore, the real random variable $\boldsymbol{Y}^{\mathrm{T}}\left(\boldsymbol{Y} - \boldsymbol{X}\tilde{\beta}^* - \boldsymbol{Z}\tilde{U}^*\right)$ *is always nonnegative.*

Proof. Let Assumptions $\mathscr{A}1_g$ - $\mathscr{A}2_g$ and $\mathscr{A}4_g$ be true and $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\tilde{\beta}^*$. First,

$$A - B = \left(\boldsymbol{X}\tilde{\beta}^* + \boldsymbol{Z}\tilde{U}^*\right)^{\mathrm{T}}\left(\tilde{\boldsymbol{Y}} - \boldsymbol{Z}\tilde{U}^*\right) = \tilde{U}^{*\mathrm{T}}\boldsymbol{Z}^{\mathrm{T}}\left(\tilde{\boldsymbol{Y}} - \boldsymbol{Z}\tilde{U}^*\right), \tag{15}$$

where we have used the fact that Theorem 4 and Remark 2 imply: $\boldsymbol{X}\tilde{\beta}^* \perp \tilde{\boldsymbol{Y}} - \boldsymbol{Z}\tilde{U}^*$. Now, thanks to (13a) in Theorem 4, $\tilde{\boldsymbol{Y}} - \boldsymbol{Z}\tilde{U}^* = \sigma_\varepsilon^2\boldsymbol{V}^{-1}\tilde{\boldsymbol{Y}}$. The latter, inserted in (15) alongside the expression (6) of $\tilde{U}^*$, implies:

$$A - B = \sigma_\varepsilon^2\tilde{\boldsymbol{Y}}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\tilde{\boldsymbol{Y}} = \sigma_\varepsilon^2\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\tilde{\boldsymbol{Y}}\right)^{\mathrm{T}}\boldsymbol{G}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\tilde{\boldsymbol{Y}}\right) = \sigma_\varepsilon^2\left\|\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\tilde{\boldsymbol{Y}}\right\|_{\boldsymbol{G}}^2.$$

Or, $A - B = \sigma_\varepsilon^2\left(\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\tilde{\boldsymbol{Y}}\right)^{\mathrm{T}}\boldsymbol{G}^{-1}\left(\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\tilde{\boldsymbol{Y}}\right) = \sigma_\varepsilon^2\tilde{U}^{*\mathrm{T}}\boldsymbol{G}^{-1}\tilde{U}^* = \sigma_\varepsilon^2\left\|\tilde{U}^*\right\|_{\boldsymbol{G}^{-1}}^2$. □

We end this series of results about the geometry of the HMMEs solutions with one about two remarkable expectations related to the u.ho. residual errors variance σ_ε^2 in the LMM (1):

Theorem 5. *In the LMM* (1) *with Assumptions* $\mathscr{A}1_g$ - $\mathscr{A}2_g$ *and* $\mathscr{A}4_g$, *one has, with P as in Theorem* 4:

$$\mathbb{E}\left[\boldsymbol{Y}^{\mathrm{T}}\left(\boldsymbol{Y} - \boldsymbol{X}\tilde{\beta}^* - \boldsymbol{Z}\tilde{U}^*\right)\right] = \sigma_\varepsilon^2 \cdot (n - p), \tag{16a}$$

$$\mathbb{E}\left\|\boldsymbol{Y} - \boldsymbol{X}\tilde{\beta}^* - \boldsymbol{Z}\tilde{U}^*\right\|^2 = \sigma_\varepsilon^4 \cdot \mathrm{tr}(P). \tag{16b}$$

Proof. First, (13a) in Theorem 4 implies that $\boldsymbol{Y}^{\mathrm{T}}\left(\boldsymbol{Y} - \boldsymbol{X}\tilde{\beta}^* - \boldsymbol{Z}\tilde{U}^*\right) = \sigma_\varepsilon^2\boldsymbol{Y}^{\mathrm{T}}P\boldsymbol{Y}$. Hence, with $m_{\boldsymbol{Y}} = \mathbb{E}(\boldsymbol{Y}) = \boldsymbol{X}\beta$ and $\boldsymbol{V} = \mathbb{M}\mathrm{cov}(\boldsymbol{Y})$,

$$\mathbb{E}\left[\boldsymbol{Y}^{\mathrm{T}}\left(\boldsymbol{Y} - \boldsymbol{X}\tilde{\beta}^* - \boldsymbol{Z}\tilde{U}^*\right)\right] = \sigma_\varepsilon^2\mathbb{E}\left(\boldsymbol{Y}^{\mathrm{T}}P\boldsymbol{Y}\right) = \sigma_\varepsilon^2\left[m_{\boldsymbol{Y}}^{\mathrm{T}}Pm_{\boldsymbol{Y}} + \mathrm{tr}\left(P\boldsymbol{V}\right)\right]. \tag{17a}$$

But $Pm_{\boldsymbol{Y}} = P\boldsymbol{X}\beta = \boldsymbol{V}^{-1}\boldsymbol{X}\beta - \boldsymbol{V}^{-1}\boldsymbol{X}\beta = \boldsymbol{0}$, so $m_{\boldsymbol{Y}}^{\mathrm{T}}Pm_{\boldsymbol{Y}} = 0$. Furthermore, since $P\boldsymbol{V} = \boldsymbol{I}_n - \boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}$, then $\mathrm{tr}\left(P\boldsymbol{V}\right) = n - \mathrm{tr}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X}\right] = n - p$. Substituting these results in (17a) yields (16a).

Similarly, from (13a), we get $\left\|\boldsymbol{Y} - \boldsymbol{X}\tilde{\beta}^* - \boldsymbol{Z}\tilde{U}^*\right\|^2 = \sigma_\varepsilon^4\boldsymbol{Y}^{\mathrm{T}}P^2\boldsymbol{Y}$, since P is symmetric. Hence

$$\mathbb{E}\left\|\boldsymbol{Y} - \boldsymbol{X}\tilde{\beta}^* - \boldsymbol{Z}\tilde{U}^*\right\|^2 = \sigma_\varepsilon^4\mathbb{E}\left(\boldsymbol{Y}^{\mathrm{T}}P^2\boldsymbol{Y}\right) = \sigma_\varepsilon^4 \cdot \left[m_{\boldsymbol{Y}}^{\mathrm{T}}P^2m_{\boldsymbol{Y}} + \mathrm{tr}\left(P^2\boldsymbol{V}\right)\right].$$

Now, $m_Y^{\mathrm{T}} P^2 m_Y = 0$ as $Pm_Y = \mathbf{0}$. Also, since $PVP = P$, $\mathrm{tr}\left(P^2 \boldsymbol{V}\right) = \mathrm{tr}\left(PVP\right) = \mathrm{tr}\left(P\right)$, hence (16b). □

Remark 3. *Identity* (16a) *is remarkable and seems familiar. Indeed, it is the analogue, for LMMs with u.ho. errors, of the identity which yields the unbiased estimator of the residual variance in an LM with that same type of errors. And we will use it in much the same way when deriving our new estimation methodology for LMMs with u.ho. errors. Indeed, in conjunction with the last sentence of Corollary* 1, (16a) *easily suggests how to compute, in an LMM fitting algorithm, a nonnegative update to estimate* σ_ε^2 *given the current estimates of* β *and the BLUP of* U.

3.3. More about the BLP of $U \mid Y$ in an LMM

The BLP $\bar{U} = \mathrm{BLP}\left(U \mid \boldsymbol{Y}\right)$, given by (5), will play a key role in our new estimation methodology for LMMs with u.ho. errors. That is because, first of all, its derivation does not require any Gaussian distributional assumption. Secondly, its computation and that of its covariance matrix are cheaper than those of the BLUP, especially under this scenario of u.ho. errors. See the needed formulas in the Supplementary material. In the latter, we also derive new expectations formulas specifically based on the BLP (whereas the ones in the previous section, such as (16a), were based on the BLUP) which can be used to obtain EEs for the residual variance in an LMM.

In our new methodology to be presented from Section 4 onwards, the results in Section 3.2 above and the Supplementary material will provide the tools to derive EEs for the β, σ_ε^2 and the $\mathrm{BLP}\left(U \mid \boldsymbol{Y}\right)$. On the other hand, to seek an EE for $\boldsymbol{D}$, we will start from the fact that given $\bar{U}_j = \mathrm{BLP}\left(U_j \mid \boldsymbol{Y}_j\right)$ and $\boldsymbol{V}_j^* = \boldsymbol{D} - \mathbb{M}\,\mathrm{cov}\left(U_j\right)$ ($j = 1, \cdots, m$), then, under Assumptions $\mathscr{A}1_g$ - $\mathscr{A}4_g$ and $\mathscr{A}2_g$, we get, as an unbiased estimator of $\boldsymbol{D}$, the $r \times r$ matrix

$$\tilde{\boldsymbol{D}}_1 = \frac{1}{m}\sum_{j=1}^{m}\left(\boldsymbol{V}_j^* + \bar{U}_j \bar{U}_j^{\mathrm{T}}\right). \tag{18}$$

4. Estimation in 2-Level LMMs with u.ho. Errors: A New Approach

We first present our approach for estimating the 3 main parameters of interest in a 2-level LMM (2) under Assumptions $\mathscr{A}1$ - $\mathscr{A}4$, $\mathscr{A}2_g$ and $\mathscr{A}4_g$: $\beta \in \mathbb{R}^p$, $\sigma_\varepsilon^2 > 0$ and $\boldsymbol{D} \in \mathscr{M}_r\left(\mathbb{R}\right)$, alongside obtaining a prediction for $U = \left(U_1^{\mathrm{T}}, \cdots, U_m^{\mathrm{T}}\right)^{\mathrm{T}} \in \mathbb{R}^q$. For that latter aspect, let, for $j = 1, \cdots, m$, $\bar{\boldsymbol{u}}_j = \mathrm{BLP}\left(U_j \mid \boldsymbol{Y}_j = \boldsymbol{y}_j\right) = \sigma_\varepsilon^{-2} \boldsymbol{V}_j^* \boldsymbol{Z}_j^{\mathrm{T}}\left(\boldsymbol{y}_j - \boldsymbol{X}_j \beta\right)$, the BLP of U_j given response $\boldsymbol{Y}_j = \boldsymbol{y}_j \in \mathbb{R}^{n_j}$ in cluster *j*. We will predict U through estimating

$$\bar{\boldsymbol{u}} = \mathrm{BLP}\left(U \mid \boldsymbol{Y} = \boldsymbol{y}\right) = \left(\bar{\boldsymbol{u}}_1^{\mathrm{T}}, \cdots, \bar{\boldsymbol{u}}_m^{\mathrm{T}}\right)^{\mathrm{T}} = \sigma_\varepsilon^{-2} \boldsymbol{V}^* \boldsymbol{Z}^{\mathrm{T}}\left(\boldsymbol{y} - \boldsymbol{X}\beta\right), \tag{19}$$

the BLP of U given response $\boldsymbol{Y} = \boldsymbol{y} \in \mathbb{R}^n$ for the whole sample.

Our approach is based on iteratively solving appropriate EEs for β, σ_ε^2, $\boldsymbol{D}$, $\bar{\boldsymbol{u}}$ derived using only sound nonparametric estimation principles. Devising

nonparametric EEs from given data to solve a statistical problem is an old idea in Statistics. The method of moments, whenever applicable, is in that category. In the field of LMMs, one may even say that that's where all started with the ANOVA methods for fitting variance components models. But it has turned especially hard to derive EEs applicable to fit a broad class of LMMs, as unconstrained as possible, but not relying on Gaussian assumptions. Attempts in that direction include the iterative weighted least squares and iterative generalized least squares estimation methods for LMMs presented, respectively, in Jiang *et al.* [37] and Goldstein [38]. But they are still not yet sufficiently general. What we propose hereafter is a 3-step construction to get closer to such a goal.

4.1. Two 3-Step Sequences for Estimation in 2-Level LMMs with u.ho. Errors

The key point in our approach is the fact that the BLP is computationally cheaper to handle than the BLUP (including at the covariance matrix level) while at the same time, as observed at the end of Section 2.4, an empirical BLUP can also be viewed as an empirical BLP. So, to devise an estimation procedure for β, σ_ε^2, $\boldsymbol{D}$ and prediction for U from $\boldsymbol{Y}=\boldsymbol{y}$, and given the preliminary results in Section 0, our starting ideas are described in what follows (we provide 2 different versions). In them, when $\hat{\boldsymbol{D}}$ is an estimate of $\boldsymbol{D}$, it is understood that $\hat{\boldsymbol{G}}=\operatorname{diag}\left(\hat{\boldsymbol{D}},\cdots,\hat{\boldsymbol{D}}\right)$ (m times) is the corresponding estimate of $\boldsymbol{G}$.

4.1.1. Starting Ideas: *Version* 1

- ***Step* 1: Estimating β and $\bar{\boldsymbol{u}}$, given estimates of σ_ε^2 and $\boldsymbol{D}$**

  Given preliminary estimates $\hat{\sigma}_\varepsilon^2$ and $\hat{\boldsymbol{D}}$ of σ_ε^2 and $\boldsymbol{D}$, one can obtain estimates $\hat{\beta}$ of β, and $\hat{\boldsymbol{u}}$ of $\bar{\boldsymbol{u}}$ by solving the system of HMMEs (10) using $\sigma_\varepsilon^2=\hat{\sigma}_\varepsilon^2$ and $\boldsymbol{G}=\hat{\boldsymbol{G}}$.

- ***Step* 2: Improved estimate of σ_ε^2, given $\hat{\beta}$, $\hat{\boldsymbol{u}}$ from Step 1**

  With $\hat{\beta}$ and $\hat{\boldsymbol{u}}$ obtained as above in Step 1, (16a) in Theorem 5 suggests to take

$$\hat{\sigma}_\varepsilon^2=\frac{\boldsymbol{y}^{\mathrm{T}}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\beta}-\boldsymbol{Z}\hat{\boldsymbol{u}}\right)}{n-p} \tag{20}$$

  as a hopefully improved estimate of σ_ε^2. Using Corollary 1, the way $\hat{\beta}$ and $\hat{\boldsymbol{u}}$ were obtained in Step 1 implies that $\hat{\sigma}_\varepsilon^2$ is a nonnegative real number.

- ***Step* 3: Improved estimate of $\boldsymbol{D}$ given preliminary estimates $\hat{\sigma}_\varepsilon^2$, $\tilde{\boldsymbol{D}}$, $\hat{\boldsymbol{u}}$ of σ_ε^2, $\boldsymbol{D}$, $\bar{\boldsymbol{u}}$**

  Here, we propose an approach to get an improved estimate $\hat{\boldsymbol{D}}$ of $\boldsymbol{D}$ from respective preliminary estimates $\hat{\sigma}_\varepsilon^2$, $\tilde{\boldsymbol{D}}$, $\hat{\boldsymbol{u}}$ of σ_ε^2, $\boldsymbol{D}$, and $\bar{\boldsymbol{u}}$ in (19). We start with the unbiased estimator $\tilde{\boldsymbol{D}}_1$ of $\boldsymbol{D}$ given by (18), but is not computable from the available data since it still depends on the unknown parameters, including $\boldsymbol{D}$ itself. Nevertheless, with the preliminary estimates $\tilde{\boldsymbol{D}}$ of $\boldsymbol{D}$, and $\hat{\sigma}_\varepsilon^2$ of σ_ε^2, Equation (S:1.7c) in the Supplementary material suggests estimating $\boldsymbol{V}_j^*$ by $\hat{\boldsymbol{V}}_j^*=\hat{\sigma}_\varepsilon^2\left(\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{Z}_j+\hat{\sigma}_\varepsilon^2\tilde{\boldsymbol{D}}^{-1}\right)^{-1}$. Then, for $j=1,\cdots,m$, given the available

estimate $\hat{\boldsymbol{u}}_j$ of $\overline{\boldsymbol{u}}_j$, (18) suggests as a hopefully improved computable estimate of $\boldsymbol{D}$:

$$\hat{\boldsymbol{D}} = \frac{1}{m}\sum_{j=1}^{m}\left(\hat{\boldsymbol{V}}_j^{*} + \hat{\boldsymbol{u}}_j\hat{\boldsymbol{u}}_j^{\mathrm{T}}\right). \tag{21}$$

4.1.2. Starting Ideas: *Version* 2

Here, Step 3 is the same as in Version 1 above, but Steps 1 and 2 rather go as follows:

- ***Step* 1: Estimation of $\overline{\boldsymbol{u}}$ and β, given estimates of β, $\boldsymbol{D}$, σ_ε^2**

  Assume that we have respective preliminary estimates $\tilde{\beta}$, $\hat{\boldsymbol{D}}$, $\hat{\sigma}_\varepsilon^2$ of β, $\boldsymbol{D}$, σ_ε^2. Given $\boldsymbol{Y} = \boldsymbol{y}$ and (19), we estimate $\overline{\boldsymbol{u}}$ by

$$\hat{\boldsymbol{u}} = \hat{\sigma}_\varepsilon^{-2}\hat{\boldsymbol{V}}^{*}\boldsymbol{Z}^{\mathrm{T}}\left(\boldsymbol{y} - \boldsymbol{X}\tilde{\beta}\right), \tag{22}$$

with $\hat{\boldsymbol{V}}^{*} = \hat{\sigma}_\varepsilon^2\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z} + \hat{\sigma}_\varepsilon^2\hat{\boldsymbol{G}}^{-1}\right)^{-1}$. From Theorem 4, we hope to get an improved estimate of β through:

$$\hat{\beta} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{u}}\right). \tag{23}$$

Remark 4. *This is not the same as Step* 1 *of Version* 1 *because* $\tilde{\beta}$ *appears on the l.h.s. of* (22) *instead of* $\hat{\beta}$. *So here,* $\hat{\boldsymbol{u}}$ *and* $\hat{\beta}$ *are not solutions to the HMMEs given* $\boldsymbol{G} = \hat{\boldsymbol{G}}$ *and* $\boldsymbol{R} = \hat{\sigma}_\varepsilon^2\boldsymbol{I}_n$.

Remark 5. *In* (22), *we have* $\hat{\boldsymbol{u}} = \left(\hat{\boldsymbol{u}}_1^{\mathrm{T}}, \cdots, \hat{\boldsymbol{u}}_m^{\mathrm{T}}\right)^{\mathrm{T}}$, *where for* $j = 1, \cdots, m$, $\hat{\boldsymbol{u}}_j \in \mathbb{R}^r$ *is given by*:

$$\hat{\boldsymbol{u}}_j = \hat{\sigma}_\varepsilon^{-2}\hat{\boldsymbol{V}}_j^{*}\boldsymbol{Z}_j^{\mathrm{T}}\left(\boldsymbol{y}_j - \boldsymbol{X}_j\tilde{\beta}\right), \quad \hat{\boldsymbol{V}}_j^{*} = \hat{\sigma}_\varepsilon^2\left(\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{Z}_j + \hat{\sigma}_\varepsilon^2\hat{\boldsymbol{D}}^{-1}\right)^{-1}. \tag{24}$$

Remark 6. *The vector* $\hat{\beta}$ *given by* (23) *is the OLS, under Assumption* $\mathscr{A}2_g$, *of* β *in the LM*: $\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{u}} = \boldsymbol{X} + \tilde{\varepsilon}$, *with* $\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{u}}$ *as response and* $\tilde{\varepsilon}$ *as vector of residual errors.*

- ***Step* 2: Improved estimate of σ_ε^2, given $\hat{\beta}$, $\hat{\boldsymbol{u}}$ from Step 1, and preliminary estimates of σ_ε^2, $\boldsymbol{D}$**

  With Remark 4 above, there is no guarantee that the numerator of (20) is a nonnegative number given $\hat{\boldsymbol{u}}$ and $\hat{\beta}$ from Step 1 in this Version 2. To ensure a nonnegative estimate of σ_ε^2, we combine Corollary 1 and Theorem 5 to obtain, from the available estimates $\tilde{\sigma}_\varepsilon^2$ and $\hat{\boldsymbol{D}}$ of σ_ε^2 and $\boldsymbol{D}$, the hopefully improved estimate of σ_ε^2:

$$\hat{\sigma}_\varepsilon^2 = \frac{\left\|\boldsymbol{y} - \boldsymbol{X}\hat{\beta} - \boldsymbol{Z}\hat{\boldsymbol{u}}\right\|^2 + \tilde{\sigma}_\varepsilon^2\left\|\hat{\boldsymbol{u}}\right\|_{\hat{\boldsymbol{G}}^{-1}}^2}{n - p}. \tag{25}$$

Remark 7. *From Corollary* 1, *Theorem* 5 *and Propositions S*1.4-*S*1.5, *S*1.8-*S*1.9, *two alternative somewhat more sophisticated computable nonnegative estimates of* σ_ε^2 *in this context are*:

$$\hat{\sigma}_\varepsilon^2 = \frac{\left\|\boldsymbol{y} - \boldsymbol{X}\hat{\beta} - \boldsymbol{Z}\hat{\boldsymbol{u}}\right\|^2 + q\tilde{\sigma}_\varepsilon^2}{n - p + \mathrm{tr}\left(\hat{\boldsymbol{G}}^{-1}\hat{\boldsymbol{V}}^{*}\right)} = \frac{\left\|\boldsymbol{y} - \boldsymbol{X}\hat{\beta} - \boldsymbol{Z}\hat{\boldsymbol{u}}\right\|^2 + q\tilde{\sigma}_\varepsilon^2}{n - p + \mathrm{tr}\left(\hat{\boldsymbol{D}}^{-1}\sum_{j=1}^{m}\hat{\boldsymbol{V}}_j^{*}\right)}, \tag{26a}$$

$$\hat{\sigma}_{\varepsilon}^{2}=\frac{\operatorname{tr}\left(\hat{\boldsymbol{V}}^{*}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right)}{\left\|\hat{\boldsymbol{u}}\right\|_{\hat{\boldsymbol{G}}^{-1}}^{2}}=\sum_{j=1}^{m}\operatorname{tr}\left(\hat{\boldsymbol{V}}_{j}^{*}\boldsymbol{Z}_{j}^{\mathrm{T}}\boldsymbol{Z}_{j}\right)\Big/\sum_{j=1}^{m}\left\|\hat{\boldsymbol{u}}_{j}\right\|_{\hat{\boldsymbol{B}}^{-1}}^{2}. \tag{26b}$$

4.2. Targeted Estimates for β, σ_{ε}^{2}, D and Prediction of U

4.2.1. The Estimating Equations

The reasoning in Steps 1-2-3 of Version 1 above leads us to seek, given the data vector $\boldsymbol{Y}=\boldsymbol{y}$, respective estimates $\hat{\beta}$, $\hat{\sigma}_{\varepsilon}^{2}$, $\hat{\boldsymbol{D}}$, of parameters β, σ_{ε}^{2}, $\boldsymbol{D}$, and prediction $\hat{\boldsymbol{u}}$ of U satisfying the system of equations:

$$\begin{pmatrix}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}\\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}+\hat{\sigma}_{\varepsilon}^{2}\hat{\boldsymbol{G}}^{-1}\end{pmatrix}\begin{pmatrix}\hat{\beta}\\ \hat{\boldsymbol{u}}\end{pmatrix}=\begin{pmatrix}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}\\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{y}\end{pmatrix}, \tag{27a}$$

$$\hat{\sigma}_{\varepsilon}^{2}=\frac{\boldsymbol{y}^{\mathrm{T}}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\beta}-\boldsymbol{Z}\hat{\boldsymbol{u}}\right)}{n-p}, \tag{27b}$$

$$\hat{\boldsymbol{D}}=\frac{1}{m}\sum_{j=1}^{m}\left(\hat{\boldsymbol{V}}_{j}^{*}+\hat{\boldsymbol{u}}_{j}\hat{\boldsymbol{u}}_{j}^{\mathrm{T}}\right). \tag{27c}$$

Steps 1-2-3 of Version 2 suggest the following system, with $\hat{\boldsymbol{D}}$ as in (27c):

$$\hat{\boldsymbol{u}}=\hat{\sigma}_{\varepsilon}^{-2}\hat{\boldsymbol{V}}^{*}\boldsymbol{Z}^{\mathrm{T}}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\beta}\right), \tag{28a}$$

$$\hat{\beta}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{y}-\boldsymbol{Z}\hat{\boldsymbol{u}}\right), \tag{28b}$$

$$\hat{\sigma}_{\varepsilon}^{2}=\frac{\left\|\boldsymbol{y}-\boldsymbol{X}\hat{\beta}-\boldsymbol{Z}\hat{\boldsymbol{u}}\right\|^{2}+\hat{\sigma}_{\varepsilon}^{2}\left\|\hat{\boldsymbol{u}}\right\|_{\hat{\boldsymbol{G}}^{-1}}^{2}}{n-p}. \tag{28c}$$

We so obtain two distinct sets of fixed points EEs for β, σ_{ε}^{2}, $\boldsymbol{D}$ and $\bar{\boldsymbol{u}}$. Our estimation approach for the 2-level LMM (2) under Assumptions $\mathscr{A}1$-$\mathscr{A}4$ and $\mathscr{A}2_{g}$, $\mathscr{A}4_{g}$ consists in solving either of them through successive approximations iterative procedure. However, we stress that, from Theorem 4 and Corollary 1, if $\hat{\boldsymbol{D}}$ is SPD, $\hat{\sigma}_{\varepsilon}^{2}>0$ and Assumption $\mathscr{A}2_{g}$ holds, then one has the following logical equivalence:

$$\text{(27a) – (27b)} \quad \Leftrightarrow \quad \text{(28a) – (28b) – (28c)}.$$

So the above two systems of equations are actually equivalent, and, thus, have the same solutions if any. However, they do suggest two different successive approximations algorithms to try to calculate these 4 unknowns.

4.2.2. 3*S*: A Code Name for the LMMs New Estimation Methodology

Our proposed two estimating algorithms are presented in the next section. Since each of them is based on a 3-step sequence construction, we shall use the code name 3*S* for this new methodology for fitting LMMs. In the future, we will present other variants of this approach. So the current one is coded 3*S*-A1, and the two algorithms we present hereafter are coded 3*S*-A1-V1 and 3*S*-A1-V2.

4.2.3. Two Iterative Estimating Procedures for 2-Level LMMs with u.ho. Errors

Given the response vector $\boldsymbol{y}=\left(\boldsymbol{y}_{1}^{\mathrm{T}},\cdots,\boldsymbol{y}_{m}^{\mathrm{T}}\right)^{\mathrm{T}}\in\mathbb{R}^{n}$, and the EEs (27a)-(27c), our

first estimating algorithm for a 2-level LMM with u.ho. errors is:

Algorithm 3*S*-A1-V1. ***Estimating*** β, σ_ε^2, $\boldsymbol{D}$ ***and predicting*** $U_1,\cdots,U_m$ ***in a 2-level LMM with u.ho. errors: Version*** 1

1) ***Initialization***: *At iteration* 0, *we estimate* β, σ_ε^2, $\boldsymbol{D}$ *and predict* $U_1,\cdots,U_m$, *as follows*:

a) $\hat{\beta}^{(0)}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$, *the OLS estimate of* β *in the linear model* $\boldsymbol{y}=\boldsymbol{X}\beta+\varepsilon$;

b) $\hat{\sigma}_\varepsilon^{2(0)}=\left\|\boldsymbol{y}-\boldsymbol{X}\hat{\beta}^{(0)}\right\|^2\Big/\left(n-p\right)$, *the corresponding unbiased estimate of the residual variance*;

c) $\forall j=1,\cdots,m$, $\hat{\boldsymbol{u}}_j^{(0)}=\left(\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{Z}_j\right)^{+}\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{y}_j^{*(0)}$, *the OLS estimate of* U_j *in the linear model* $\boldsymbol{y}_j^{*(0)}=\boldsymbol{Z}_jU_j+\varepsilon_j$, *with* U_j *considered as fixed parameter,* $\boldsymbol{Z}_j$ *the design matrix,* $\boldsymbol{y}_j^{*(0)}=\boldsymbol{y}_j-\boldsymbol{X}_j\hat{\beta}^{(0)}$ *as the response, and* $\left(\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{Z}_j\right)^{+}$ *being the Moore-Penrose pseudo-inverse of* $\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{Z}_j$;

d) $\hat{\boldsymbol{D}}^{(0)}=\dfrac{1}{m}\sum_{j=1}^{m}\hat{\boldsymbol{u}}_j^{(0)}\hat{\boldsymbol{u}}_j^{(0)\mathrm{T}}$;

2) ***The iterative process***: *Given* $\hat{\beta}^{(t)}$, $\hat{\sigma}_\varepsilon^{2(t)}$, $\hat{\boldsymbol{u}}^{(t)}=\left(\hat{\boldsymbol{u}}_1^{(t)\mathrm{T}},\cdots,\hat{\boldsymbol{u}}_m^{(t)\mathrm{T}}\right)^{\mathrm{T}}$, *and* $\hat{\boldsymbol{D}}^{(t)}$ *from iteration* t, *compute estimates and predictions at iteration* $t+1$ *as follows*:

a) $\hat{A}_j^{(t+1)}=\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{Z}_j+\hat{\sigma}_\varepsilon^{2(t)}\left[\hat{\boldsymbol{D}}^{(t)}\right]^{-1}$, $j=1,\cdots,m$;

b) *Solve for* $\hat{\beta}^{(t+1)}$ *and* $\hat{\boldsymbol{u}}^{(t+1)}$ *in the linear system, with* $\hat{A}^{(t+1)}=\mathrm{diag}\left(\hat{A}_1^{(t+1)},\cdots,\hat{A}_m^{(t+1)}\right)$:

$$\begin{pmatrix}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}\\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}+\hat{A}^{(t+1)}\end{pmatrix}\begin{pmatrix}\hat{\beta}^{(t+1)}\\ \hat{\boldsymbol{u}}^{(t+1)}\end{pmatrix}=\begin{pmatrix}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}\\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{y}\end{pmatrix};\tag{29}$$

c) $\hat{\sigma}_\varepsilon^{2(t+1)}=\boldsymbol{y}^{\mathrm{T}}\left(\boldsymbol{y}-\boldsymbol{X}^{(t+1)}-\boldsymbol{Z}\hat{\boldsymbol{u}}^{(t+1)}\right)\Big/\left(n-p\right)$;

d) $\hat{\boldsymbol{V}}_j^{*(t+1)}=\hat{\sigma}_\varepsilon^{2(t+1)}\left[\hat{A}_j^{(t+1)}\right]^{-1}$, $j=1,\cdots,m$;

e) $\hat{\boldsymbol{D}}^{(t+1)}=\dfrac{1}{m}\sum_{j=1}^{m}\left[\hat{\boldsymbol{V}}_j^{*(t+1)}+\hat{\boldsymbol{u}}_j^{(t+1)}\hat{\boldsymbol{u}}_j^{(t+1)\mathrm{T}}\right]$;

3) ***Stopping criterion***: *Assume convergence when* $\left\|\hat{\boldsymbol{D}}^{(t+1)}-\hat{\boldsymbol{D}}^{(t)}\right\|^{(M)}\le\delta_1\left\|\hat{\boldsymbol{D}}^{(t+1)}\right\|^{(M)}$, $\left\|\hat{\beta}^{(t+1)}-\hat{\beta}^{(t)}\right\|\le\delta_2\left\|\hat{\beta}^{(t+1)}\right\|$, $\left|\hat{\sigma}_\varepsilon^{2(t+1)}-\hat{\sigma}_\varepsilon^{2(t)}\right|\le\delta_3\hat{\sigma}_\varepsilon^{2(t+1)}$ *are all satisfied, where* $\delta_1,\delta_2,\delta_3$ *are relative tolerance levels set in* $(0,1)$, *and* $\|\cdot\|^{(M)}$ *is a chosen matrix norm. Otherwise,* $t\leftarrow t+1$ *and repeat Step* 2.

4) ***Extracting estimates***: *At convergence, take* $\hat{\beta}=\hat{\beta}^{(t+1)}$, $\hat{\sigma}_\varepsilon^2=\hat{\sigma}_\varepsilon^{2(t+1)}$, $\hat{\boldsymbol{D}}=\hat{\boldsymbol{D}}^{(t+1)}$ *as estimates of* β, σ_ε^2, $\boldsymbol{D}$. *Also, take* $\hat{\boldsymbol{u}}_1^{(t+1)},\cdots,\hat{\boldsymbol{u}}_m^{(t+1)}$ *as predictions of* $U_1,\cdots,U_m$.

Remark 8. *In that algorithm, whereas stabilization of the* $\boldsymbol{D}$ *and* σ_ε^2 *iterates is enough to ensure that the values of all the other estimates have stabilized as well, the stopping criterion of the algorithm is based on also monitoring the* β *iterates for more security.*

Given the alternative system of EEs (28a)-(28c), our second algorithm for 2-level LMMs with u.ho. errors goes as follows:

Algorithm 3*S*-A1-V2. ***Estimating*** β, σ_ε^2, $\boldsymbol{D}$ ***and predicting*** $U_1,\cdots,U_m$ ***in a 2-level LMM with u.ho. errors: Version*** 2

1) ***Initialization***: *The same as in Algorithm* 3*S-A*1-*V*1;

2) ***The iterative process***: *Given* $\hat{\beta}^{(t)}$, $\hat{\sigma}_\varepsilon^{2(t)}$, $\hat{\boldsymbol{u}}_1^{(t)},\cdots,\hat{\boldsymbol{u}}_m^{(t)}$, *hence* $\hat{\boldsymbol{u}}^{(t)}=\left(\hat{\boldsymbol{u}}_1^{(t)\mathrm{T}},\cdots,\hat{\boldsymbol{u}}_m^{(t)\mathrm{T}}\right)^{\mathrm{T}}$, *and* $\hat{\boldsymbol{D}}^{(t)}$ *from iteration* t, *compute estimates and predictions at iteration* $t+1$ *as follows*:

a) $\hat{B}_j^{(t+1)}=\left(\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{Z}_j+\hat{\sigma}_\varepsilon^{2(t)}\left[\hat{\boldsymbol{D}}^{(t)}\right]^{-1}\right)^{-1}$, $j=1,\cdots,m$;

b) $\hat{\beta}^{(t+1)}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{y}-\boldsymbol{Z}\hat{\boldsymbol{u}}^{(t)}\right)$;

c) $\hat{\boldsymbol{u}}_j^{(t+1)}=\hat{B}_j^{(t+1)}\boldsymbol{Z}_j^{\mathrm{T}}\left(\boldsymbol{y}_j-\boldsymbol{X}_j\hat{\beta}^{(t+1)}\right)$, $j=1,\cdots,m$;

d) $\hat{\sigma}_\varepsilon^{2(t+1)}=\dfrac{1}{n-p}\left(\left\|\boldsymbol{y}-\boldsymbol{Z}\hat{\boldsymbol{u}}^{(t+1)}-\boldsymbol{X}\hat{\beta}^{(t+1)}\right\|^2+\hat{\sigma}_\varepsilon^{2(t)}\left\|\hat{\boldsymbol{u}}^{(t+1)}\right\|_{\hat{\boldsymbol{G}}^{(t)-1}}^2\right)$,

$\hat{\boldsymbol{G}}^{(t)}=\mathrm{diag}\left(\hat{\boldsymbol{D}}^{(t)},\cdots,\hat{\boldsymbol{D}}^{(t)}\right)$;

e) $\hat{\boldsymbol{V}}_j^{*(t+1)}=\hat{\sigma}_\varepsilon^{2(t+1)}\hat{B}_j^{(t+1)}$, $j=1,\cdots,m$;

f) $\hat{\boldsymbol{D}}^{(t+1)}=\dfrac{1}{m}\sum_{j=1}^{m}\left[\hat{\boldsymbol{V}}_j^{*(t+1)}+\hat{\boldsymbol{u}}_j^{(t+1)}\hat{\boldsymbol{u}}_j^{(t+1)\mathrm{T}}\right]$;

3) ***Stopping criterion and extracting estimates***: *As in Algorithm* 3*S-A*1-*V*1.

Remark 9. *Important elements about the practical implementation of the above two designed algorithms are detailed further in Section S*2 *of the accompanying Supplementary material document. In particular, it is explained there how we monitor a possible rank deficiency of* $\boldsymbol{D}$ *and even whether or not random effects are really there for a given data set, in the first place. For the latter aspect, we introduce and motivate*

$$\rho=\left\|\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\right\|^{(M)}\Big/\sigma_\varepsilon^2, \tag{30}$$

a quantity called random effects ratio, through formula (S:2.6b) *to try to assess the likelihood of having significant random effects or not in the data.*

4.3. First Properties of the Estimates $\hat{\beta}$, $\hat{\sigma}_\varepsilon^2$, $\hat{D}$ and $\hat{u}$

The convergence of the above two algorithms has not been investigated yet. However, the fact that their convergence can only occur to estimates $\hat{\beta}$, $\hat{\sigma}_\varepsilon^2$, $\hat{\boldsymbol{u}}$, $\hat{\boldsymbol{D}}$ (of β, σ_ε^2, $\boldsymbol{D}$, $\bar{\boldsymbol{u}}$) satisfying the two equivalent systems of EEs of Section 0.0.5 allows to identify some first properties of those estimates. Indeed, assume that any of the above two algorithms has converged to $\hat{\beta}$, $\hat{\sigma}_\varepsilon^2$, $\hat{\boldsymbol{D}}$, $\hat{\boldsymbol{u}}$, with $\hat{\boldsymbol{D}}$ SPD, $\hat{\sigma}_\varepsilon^2>0$ and Assumption $\mathscr{A}2_g$ true. We highlight two important properties of these estimates.

4.3.1. Relationship with the HMMEs

Firstly, Theorems 1 & 4 and Corollary 1 imply the following about $\hat{\beta}$, $\hat{\sigma}_\varepsilon^2$, $\hat{\boldsymbol{D}}$, $\hat{\boldsymbol{u}}$:

Theorem 6. *If one takes* $\boldsymbol{G}=\hat{\boldsymbol{G}}$ *and* $\sigma_\varepsilon^2=\hat{\sigma}_\varepsilon^2$ *in the HMMEs* (10), *then* $\tilde{\beta}=\hat{\beta}$ *and* $\tilde{U}=\hat{\boldsymbol{u}}$ *are the solutions to* (10). *Therefore, the estimates* $\hat{\beta}$, $\hat{\sigma}_\varepsilon^2$, $\hat{\boldsymbol{D}}$ *and prediction* $\hat{\boldsymbol{u}}$ *satisfy*:

$$\hat{\beta}=\mathrm{EBLUE}\left(\beta\,|\,\boldsymbol{y},\hat{\boldsymbol{G}},\hat{\sigma}_\varepsilon^2\right)=\left(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}, \tag{31}$$

$$\hat{\boldsymbol{u}} = \text{EBLUP}\left(U \mid \boldsymbol{y}, \hat{\boldsymbol{G}}, \hat{\sigma}_\varepsilon^2\right) = \hat{\boldsymbol{G}}\boldsymbol{Z}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\right), \tag{32}$$

where $\hat{\boldsymbol{V}} = \boldsymbol{Z}\hat{\boldsymbol{G}}\boldsymbol{Z}^{\mathrm{T}} + \hat{\sigma}_\varepsilon^2\boldsymbol{I}_n$. *That is* $\hat{\beta}$ *and* $\hat{\boldsymbol{u}}$ *are, respectively, an empirical BLUE of* β *and an empirical BLUP of* U *when* $\boldsymbol{G}$ *is estimated by* $\hat{\boldsymbol{G}}$ *and* σ_ε^2 *by* $\hat{\sigma}_\varepsilon^2$.

Remark 10. *The main difference between Algorithms* 3*S*-*A*1-*V*1 *and* 3*S*-*A*1-*V*2 *is that the former enforces satisfaction of the HMMEs for the computed estimates at each iteration* $t \geq 1$, *while the latter only guarantees that for the final estimates if convergence is achieved.*

4.3.2. Relationship with the 2-Level Gaussian LMM Likelihood Equations under i.i.d. Errors

The EEs (27b) and (28b) indicate that our 2 estimation procedures above are extensions to 2-level LMMs with u.ho. errors of the well known 2-step method to fit an LM with such errors whereby β is first estimated by OLS and, then, that estimator is used to get an unbiased estimator for σ_ε^2, and this without using any parametric distributional assumption. Now, it is trivial that for the LM with i.i.d. Gaussian errors, that 2-step estimation method is asymptotically equivalent to the ML estimation of β and σ_ε^2. We are going to show that the same property holds for our 2 estimation procedures when the 2-level LMM is, indeed, a Gaussian one with i.i.d. errors, at least in terms of the solutions $\hat{\beta}$, $\hat{\sigma}_\varepsilon^2$, $\hat{\boldsymbol{D}}$ satisfying the *likelihood equations*. So, we assume a 2-level LMM satisfying Assumptions $\mathscr{A}1$-$\mathscr{A}4$ and, for $j = 1, \cdots, m$:

Assumption $\mathscr{A}5$. $U_j \overset{\mathscr{D}}{\sim} \mathscr{N}_r(\boldsymbol{0}, \boldsymbol{D})$ *and* $\varepsilon_j \overset{\mathscr{D}}{\sim} \mathscr{N}_{n_j}\left(\boldsymbol{0}, \sigma_\varepsilon^2\boldsymbol{I}_{n_j}\right)$.

Estimating the parameters β, σ_ε^2, $\boldsymbol{D}$ of that 2-level Gaussian LMM through ML usually starts by trying to solve the likelihood equations:

$$\frac{\partial}{\partial\beta}\ell\left(\beta, \sigma_\varepsilon^2, \boldsymbol{D} \mid \boldsymbol{y}\right) = \boldsymbol{0}, \tag{33a}$$

$$\frac{\partial}{\partial\boldsymbol{D}}\ell\left(\beta, \sigma_\varepsilon^2, \boldsymbol{D} \mid \boldsymbol{y}\right) = \boldsymbol{0}, \tag{33b}$$

$$\frac{\partial}{\partial\sigma_\varepsilon^2}\ell\left(\beta, \sigma_\varepsilon^2, \boldsymbol{D} \mid \boldsymbol{y}\right) = 0, \tag{33c}$$

where $\ell\left(\beta, \sigma_\varepsilon^2, \boldsymbol{D} \mid \boldsymbol{y}\right)$ is the log-likelihood of the model with observed response $\boldsymbol{Y} = \boldsymbol{y}$.

First, it is well known that, given our assumptions:

$$(33a) \Leftrightarrow \beta = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{y}. \tag{34}$$

Thus, thanks to (31), the triplet $\left(\beta, \sigma_\varepsilon^2, \boldsymbol{D}\right) = \left(\hat{\beta}, \hat{\sigma}_\varepsilon^2, \hat{\boldsymbol{D}}\right)$ satisfies the first likelihood Equation (33a). For the other two likelihood equations, generalizing computations in Searle *et al.* ([14], pages 278-279) for the ANOVA model (4), one has (Proof in the **Appendix**):

Theorem 7. *Let* $\tilde{\beta}$ *and* $\tilde{U}$ *satisfy the HMMEs* (10) *given* σ_ε^2, $\boldsymbol{D}$ *for the* 2*-level LMM* (2), *with Assumptions* $\mathscr{A}1$-$\mathscr{A}4$, $\mathscr{A}2_g$ *and* $\mathscr{A}5$.

1) *If* (33a) *holds, then*:

$$(33b) \quad \Leftrightarrow \boldsymbol{D} = \frac{1}{m}\sum_{j=1}^{m}\left(\boldsymbol{V}_j^{*} + \tilde{U}_j\tilde{U}_j^{\mathrm{T}}\right).$$

2) *If* (33a) *and* (33b) *both hold, then*:

$$(33c) \quad \Leftrightarrow \sigma_\varepsilon^2 = \boldsymbol{y}^{\mathrm{T}}\left(\boldsymbol{y} - \boldsymbol{X}\tilde{\beta} - \boldsymbol{Z}\tilde{U}\right)\Big/ n.$$

Given the EE (27c), the first equivalence in that theorem shows that the second likelihood Equation (33b) is satisfied by the quadruplet $\left(\beta, \sigma_\varepsilon^2, \boldsymbol{D}, U\right) = \left(\hat{\beta}, \hat{\sigma}_\varepsilon^2, \hat{\boldsymbol{D}}, \hat{\boldsymbol{u}}\right)$. But, given our EE (27b), we see that the triplet $\left(\beta, \sigma_\varepsilon^2, U\right) = \left(\hat{\beta}, \hat{\sigma}_\varepsilon^2, \hat{\boldsymbol{u}}\right)$ falls short to satisfying (33c) only through having a denominator $n - p$ instead of n, exactly like in the LM case. So, when the 2-level LMM is a Gaussian one with i.i.d. errors, our 2 estimation procedures above are asymptotically equivalent to the ML method. However, as for the LM case, we stick here with our denominator $n - p$, inspired by the unbiasedness property (16a) in Theorem 5.

Remark 11. *An important by-product of Theorem* 7 *is that it implies that by replacing* $n - p$ *by* n *in the denominator of the formula updating the residual variance estimate during the iterations in Algorithms* 3*S*-*A*1-*V*1 *and* 3*S*-*A*1-*V*2, *we get two new algorithms for computing the Gaussian ML estimates in a* 2*-level LMM. They may be coded* 3*S*-*ML*-*A*1-*V*1 *and* 3*S*-*ML*-*A*1-*V*2 *and can be viewed as alternatives to existing algorithms based on Newton-Raphson, Fisher scoring or EM* (*Expectation-Maximization*).

4.4. 3*S* LMM Fitting: What about the Accuracy of the Estimates?

Obviously, once a 3*S* algorithm has converged to the targeted estimates of parameters for fitting an LMM to a given data set, the next question is, as usual in statistical inference: *what about the accuracy of those estimates*? In a strictly parametric modeling with Gaussian ML or REML estimation, the answer is customarily built through estimating the inverse of Fisher's Information Matrix for the estimated parameters. But we cannot go that route here because we precisely aimed here at a nonparametric modeling with the LMM for the provided data set. Instead, to assess the accuracy of the estimates computed by a 3*S* algorithm, we use the bootstrap. But bootstrapping mixed models necessitates significantly more care than is usually the case when using routine i.i.d. data. Methods and up-to-date discussions about this can be found for instance in Carpenter *et al*. [45], Van der Leeden *et al*. [46], Chambers and Chandra [47], Thai *et al*. [48], Modugno and Giannerini [49].

5. 3*S* Fitting of LMMs: Beyond 2-Level LMMs

As stated from the outset, the first goal in this work was to design a methodology for parameters estimation in a 2-level LMM where the only assumption added to the basic ones ($\mathscr{A}1$ to $\mathscr{A}4$) would be to have u.ho. errors (Assumption $\mathscr{A}4_g$). Nonetheless, analyzing the 3-step sequence upon which each of our devised 3*S* iterative algorithms was constructed in Section 4, one is struck by the

fact that the 2-level element only intervenes in Step 3, *i.e.* when deriving an EE for $\boldsymbol{D}$. In the construction of both algorithms, what is done in Steps 1-2 is based on results derived in Section 3.2.2 and classical formulas of the BLP in an LMM and its covariance matrix, thus only uses Assumptions $\mathscr{A}1_g$, $\mathscr{A}2_g$ and $\mathscr{A}4_g$ of an arbitrary LMM (1). It comes that for any LMM satisfying the latter 3 assumptions, if one can reliably devise an EE for $\boldsymbol{G}$, then one would immediately get respective adaptations of Algorithms 3*S*-A1-V1 and 3*S*-A1-V2 for fitting it. In that order, one can start from the fact that (analogous to (18) for the 2-level case)

$$\tilde{\boldsymbol{G}}_1 = \boldsymbol{V}^* + \bar{U}\bar{U}^{\mathrm{T}} \tag{35}$$

is an unbiased estimator of $\boldsymbol{G}$, where $\boldsymbol{V}^* = \boldsymbol{G} - \mathbb{M}\mathrm{cov}\left(\bar{U}\right)$.

The most obvious case is when $\boldsymbol{G}$ is a diagonal matrix, but with some diagonal elements being equal by design. We examine two classical such situations hereafter: the 2-level LMM with $\boldsymbol{D}$ diagonal and the ANOVA LMM (4).

5.1. 3*S* Fitting of 2-Level LMMs with u.ho. Errors and D Diagonal

In a 2-level LMM (2), when the dimension r of the vectors $U_1, \cdots, U_m$ is big while the dimension n of the response vector is small to moderate, the number $p + mr + r\left(r+1\right)/2$ of scalar parameters to estimate in β, $\boldsymbol{D}$ and the random effects predictors might become too big for any LMM fitting method. In that case, imposing a diagonal structure for $\boldsymbol{D}$ may be a convenient way to achieve a reasonable estimation process for β, $\boldsymbol{D}$ and σ_ε^2 from the observed response vector $\boldsymbol{Y} = \boldsymbol{y} \in \mathbb{R}^n$. This motivates assuming then that $\boldsymbol{D} = \mathrm{diag}\left(\sigma_1^2, \cdots, \sigma_r^2\right)$, with $\sigma_k^2 > 0$, $\forall k = 1, \cdots, r$.

With that, the unbiased estimator $\tilde{\boldsymbol{D}}_1$ of $\boldsymbol{D}$ given by (18) implies that for each $k = 1, \cdots, r$, an unbiased preliminary estimator of the variance component σ_k^2 is

$$\tilde{\sigma}_k^2 = \frac{1}{m}\sum_{j=1}^{m}\left[\left(\boldsymbol{V}_j^*\right)_{kk} + \left(\bar{U}_j\right)_k^2\right], \tag{36}$$

where $\left(\boldsymbol{V}_j^*\right)_{kk}$ and $\left(\bar{U}_j\right)_k$ are the k^{th} diagonal element of matrix $\boldsymbol{V}_j^*$ and k^{th} component of vector $\bar{U}_j$. Thus the EE (21) for $\boldsymbol{D}$ is simplified here to:

$$\hat{\boldsymbol{D}} = \mathrm{diag}\left(\hat{\sigma}_1^2, \cdots, \hat{\sigma}_r^2\right), \quad \hat{\sigma}_k^2 = \frac{1}{m}\sum_{j=1}^{m}\left[\left(\hat{\boldsymbol{V}}_j^*\right)_{kk} + \left(\hat{\boldsymbol{u}}_j\right)_k^2\right]. \tag{37}$$

Then one can easily adapt Algorithms 3*S*-A1-V1 and 3*S*-A1-V2 for the 2-level LMM with u.ho. errors and diagonal $\boldsymbol{D}$, to get, say, Algorithms 3*S*-A1-V1-diag and 3*S*-A1-V2-diag. The latter are presented in Section S4.1 of the Supplementary material document.

5.2. 3*S* Fitting of an ANOVA LMM

In the ANOVA LMM (4), we can partition $\bar{U} = \mathrm{BLP}\left(U \mid \boldsymbol{Y}\right)$ as in (S:1.6a), but where each $\bar{U}_j$ is rather a q_j-vector ($j = 1, \cdots, m$). Now, let: $s_0 = 0$, and

$\forall j = 1, \cdots, m$, $s_j = q_1 + \cdots + q_j = s_{j-1} + q_j$. Then each vector $\bar{U}_j$ is made up of the $\left(s_{j-1}+1\right)^{\text{th}}$ to $\left(s_j\right)^{\text{th}}$ components of $\bar{U}$.

With the structure of $\boldsymbol{G}$ in this case, the estimator $\tilde{\boldsymbol{G}}_1$ in (35) implies that an unbiased preliminary estimator of each variance component σ_j^2 is

$$\tilde{\sigma}_j^2 = \frac{1}{q_j}\left[\sum_{k=s_{j-1}+1}^{s_j}\left(\boldsymbol{V}^*\right)_{kk} + \left\|\bar{U}_j\right\|^2\right] \tag{38}$$

for $j = 1, \cdots, m$. Hence, we get an EE for $\boldsymbol{G}$ here as:

$$\hat{\boldsymbol{G}} = \operatorname{diag}\left(\hat{\sigma}_1^2, \cdots, \hat{\sigma}_m^2\right), \quad \hat{\sigma}_j^2 = \frac{1}{q_j}\left[\sum_{k=s_{j-1}+1}^{s_j}\left(\hat{\boldsymbol{V}}^*\right)_{kk} + \left\|\hat{\boldsymbol{u}}_j\right\|^2\right]. \tag{39}$$

From that, Algorithms 3*S*-A1-V1 and 3*S*-A1-V2 can be adapted to fit an ANOVA LMM to get, say, Algorithms 3*S*-A1-V1-ANOVA and 3*S*-A1-V2-ANOVA. They are presented in Section S4.2 of the Supplementary material. The first one is nearly the same as one of those derived by Henderson for Gaussian ANOVA LMMs [14] [16]. The only difference is again the $n-p$ in the denominator in the r.h.s. of the EE for σ_ε^2 in (27b).

6. Numerical Examples

In this section, we carry out some numerical experiments on simulated data sets (Section 6.1) and two real world data sets (Section 6.2). For both situations, we compared Algorithm 3*S*-A1-V1 vs. Gaussian ML. For the latter, and in light of Theorem 7, we could have used Algorithm 3*S*-ML-A1-V1 in Remark 11. But that would have looked like a self comparison. So we instead chose to use the implementation of Gaussian ML by the lmer function in the reference lme4 package [50] of the R software. Hereafter, we will denote the latter by lmer-ML. The reason why Algorithm 3*S*-A1-V2 is not included in the study is that, as expected, it tends to give the same results as Algorithm 3*S*-A1-V1.

6.1. A Simulation Study

Here, to investigate the performance of Algorithm 3*S*-A1-V1 vs. lmer-ML, we fitted both to 500 simulated data sets, each of size $n = 200$, under various distributional scenarios. First, we assumed we have a population Ω comprising $m = 10$ clusters in equal proportions of 1/10 and that in each cluster *j*, both the vector *X* of fixed effects covariates and *Z*, that of random effects covariates, follow Gaussian distributions with given parameters, respectively in $\mathbb{R}^4$ and $\mathbb{R}^3$. To get a unit in a simulated data set, we first sample its cluster among the 10 in Ω, then its covariates by sampling from the Gaussian distributions of *X* and *Z* in that cluster. The distributions parameters used for that and the simulated distributions described below are given in Section S5 of the Supplementary material document, alongside what we used as vector of fixed effects parameters $\beta \in \mathbb{R}^4$.

To get one data set, after simulating the 200 units in it with their fixed and random effects covariates as just described,

- we first simulate a vector of random effects $U_j \in \mathbb{R}^3$ for each cluster

$j \in \{1,2,\cdots,10\}$ according to a chosen distribution;

- then, for each unit i sampled in cluster j and included in the data set, with simulated fixed effects covariates $X_{ij} \in \mathbb{R}^4$ and random effects covariates $Z_{ij} \in \mathbb{R}^3$, and a residual error $\varepsilon_{ij} \in \mathbb{R}$ according to a chosen distribution. Then its simulated exact response is calculated as $Y_{ij} = X_{ij}^{\mathrm{T}}\beta + Z_{ij}^{\mathrm{T}}U_j + \varepsilon_{ij}$.

For the distribution of the residual errors, we examined three options: a Gaussian, a mixture of 4 Gaussians with given mixing proportions, and a discrete distribution with 4 mass points. On the other hand, the covariance matrix $\boldsymbol{D}$ of each U_j is 3×3. For the distribution of the U_j's in $\mathbb{R}^3$, we examined five options: Gaussian with $\boldsymbol{D}$ SPD, Gaussian with $\operatorname{rank}(\boldsymbol{D}) = 1$, Gaussian with $\operatorname{rank}(\boldsymbol{D}) = 2$, mixture of 3 Gaussians with given mixing proportions, and a discrete distribution with 3 mass points. Thus, we have $3\times 5 = 15$ possible scenarios. Hence, for each of the algorithms 1 and lmer-ML, in each of the 15 scenarios, we simulated 500 200-sized data sets and fitted the 2-level LMM (expressed in the R software notations):

$$\mathrm{Y} \sim \mathrm{fe}(-1+\mathrm{X1}+\mathrm{X2}+\mathrm{X3}+\mathrm{X4}) + \mathrm{re}(-1+\mathrm{Z1}+\mathrm{Z2}+\mathrm{Z3}) + \mathrm{Gr}(\mathrm{cluster}) \qquad (40)$$

where "−1" signifies no intercept included, be it on the fixed or the random effects.

The fit of the LMM (40) to each simulated data set yields an estimate $\hat{\theta} \in \left\{\hat{\beta}, \hat{\boldsymbol{D}}, \hat{\sigma}^2\right\}$ of each parameter $\theta \in \left\{\beta, \boldsymbol{D}, \sigma^2\right\}$, and a prediction $\hat{U} = \left(\hat{U}_j^{\mathrm{T}}\right)_{j=1,\cdots,10}^{\mathrm{T}}$ of the random effects vector $U = \left(U_j^{\mathrm{T}}\right)_{j=1,\cdots,10}^{\mathrm{T}} \in \mathbb{R}^{30}$.

Furthermore, to estimate the Mean Squared Prediction Error (MSPE) of the unit response Y by its prediction $\hat{Y}$ from a fitted model, for each simulated data, we independently simulated an additional data set of size 100 with simulated response vector $\boldsymbol{Y} \in \mathbb{R}^{100}$ and computed $\hat{\boldsymbol{Y}}$, the response vector predicted by the fitted LMM. For each $\theta \in \left\{\beta, \boldsymbol{D}, \sigma^2\right\}$, we then used its simulated replicates $\hat{\theta}_1, \cdots, \hat{\theta}_{500}$ to estimate

$$b_R\left(\hat{\theta}\right) = \left\|\mathbb{E}\left(\hat{\theta}\right) - \theta\right\| \Big/ \|\theta\| \text{ and } \operatorname{RMSE}\left(\hat{\theta}\right) = \mathbb{E}\left\|\hat{\theta} - \theta\right\|^2 \Big/ \|\theta\|^2, \qquad (41)$$

respectively the Euclidean norm of the Relative Bias and the Relative Mean Squared Error of $\hat{\theta}$. Each expectation was estimated by the corresponding Monte Carlo estimate. In that calculation, for the symmetric 3×3 matrices $\theta = \boldsymbol{D}$ and $\hat{\theta} = \hat{\boldsymbol{D}}$, we identified, each, to the vector of elements in its upper triangular part (comprising the diagonal). Likewise, for each predictor $\hat{P} \in \left\{\hat{U}, \hat{Y}\right\}$ of $P \in \{U, Y\}$, we used its simulated replicates to estimate the norm of its bias and its MSPE.

$$b\left(\hat{P}\right) = \left\|\mathbb{E}\left(\hat{P}\right) - P\right\| \text{ and } \operatorname{MSPE}\left(\hat{P}\right) = \mathbb{E}\left\|\hat{P} - P\right\|^2. \qquad (42)$$

For both algorithms, the results about the estimation of β, $\boldsymbol{D}$, σ_ε^2 and the prediction of random effects and responses are reported in Table 1. Before we specifically comment on them, note, however, that when running an iterative algorithm on simulated data sets, the algorithm might fail to converge on some

Table 1. Estimates of the bias and MSE of estimates of parameters, predicted random effects and responses in the LMM (40) applied to simulated data sets with different distributions for the random effects and residual errors.

Distribution of random effects and errors	$b_R(\widehat{\beta})$	$\sqrt{\text{RMSE}(\widehat{\beta})}$	$b_R(\widehat{\sigma}^2_\varepsilon)$	$\sqrt{\text{RMSE}(\widehat{\sigma}^2_\varepsilon)}$	$b_R(\widehat{\mathbf{D}})$	$\sqrt{\text{RMSE}(\widehat{\mathbf{D}})}$	$b(\widehat{U})$	$\sqrt{\text{MSPE}(\widehat{U})}$	$b(\widehat{Y})$	$\sqrt{\text{MSPE}(\widehat{Y})}$
Gaussian errors										
Gaussian RE ($\mathbf{D}$ *SPD*)	0.00355	0.0689	0.00216	0.112	0.532	0.958	0.0165	0.662	$3.80e-05$	$5.96e-03$
	(0.000861)	*(0.0678)*	*(0.0188)*	*(0.108)*	*(0.521)*	*(0.951)*	*(0.00782)*	*(0.673)*	*(9.46e−05)*	*(5.87e−03)*
Gaussian RE ($\mathrm{R}(\mathbf{D})=1$)	0.00327	0.0535	0.0172	0.103	0.878	1.02	0.00693	0.358	$4.55e-06$	$5.17e-03$
	(0.00351)	*(0.0540)*	*(0.0410)*	*(0.105)*	*(0.918)*	*(2.39)*	*(0.00202)*	*(0.363)*	*(2.18e−05)*	*(5.28e−03)*
Gaussian RE ($\mathrm{R}(\mathbf{D})=2$)	0.00449	0.0625	0.0164	0.108	0.74	1.26	0.00350	0.533	$4.20e-05$	$5.76e-03$
	(0.00177)	*(0.0624)*	*(0.0401)*	*(0.115)*	*(0.765)*	*(1.25)*	*(0.00291)*	*(0.551)*	*(7.57e−05)*	*(5.66e−03)*
Gaussian mixture RE	0.0022	0.0676	0.0055	0.109	0.425	1.15	0.0116	0.684	$1.45e-04$	$4.43e-03$
	(0.00152)	*(0.0665)*	*(0.0155)*	*(0.104)*	*(0.521)*	*(0.737)*	*(0.00709)*	*(0.674)*	*(6.95e−05)*	*(4.43e−03)*
Discrete RE	0.00224	0.0606	0.0219	0.105	0.818	1.84	0.00411	0.513	$2.91e-05$	$3.57e-03$
	(0.00249)	*(0.0612)*	*(0.0380)*	*(0.111)*	*(0.718)*	*(1.21)*	*(0.00553)*	*(0.516)*	*(6.56e−05)*	*(3.63e−03)*
Gaussian mixture errors										
Gaussian RE ($\mathbf{D}$ *SPD*)	0.00109	0.0641	0.364	0.406	0.724	1.44	0.00819	0.598	$9.23e-05$	$5.84e-03$
	(0.00254)	*(0.0644)*	*(0.416)*	*(0.446)*	*(0.567)*	*(1.12)*	*(0.00515)*	*(0.598)*	*(8.87e−05)*	*(5.88e−03)*
Gaussian RE ($\mathrm{R}(\mathbf{D})=1$)	0.00165	0.0622	0.404	0.441	0.935	2.20	0.00377	0.438	$2.21e-05$	$5.42e-03$
	(0.00421)	*(0.0621)*	*(0.457)*	*(0.483)*	*(2.87)*	*(7.43)*	*(0.00484)*	*(0.461)*	*(5.48e−05)*	*(5.16e−03)*
Gaussian RE ($\mathrm{R}(\mathbf{D})=2$)	0.00368	0.0628	0.402	0.441	0.670	0.991	0.00212	0.519	$2.80e-05$	$5.72e-03$
	(0.00300)	*(0.0622)*	*(0.417)*	*(0.452)*	*(0.816)*	*(1.01)*	*(0.00764)*	*(0.532)*	*(1.48e−04)*	*(5.65e−03)*
Gaussian mixture RE	0.00375	0.0626	0.395	0.433	0.510	0.898	0.00826	0.609	$1.14e-06$	$4.45e-03$
	(0.00287)	*(0.0646)*	*(0.404)*	*(0.438)*	*(0.722)*	*(0.786)*	*(0.00707)*	*(0.596)*	*(2.58e−05)*	*(4.43e−03)*
Discrete RE	0.00331	0.0641	0.384	0.424	0.728	1.18	0.00756	0.512	$1.87e-05$	$3.69e-03$
	(0.00310)	*(0.0632)*	*(0.426)*	*(0.457)*	*(0.638)*	*(1.25)*	*(0.00637)*	*(0.509)*	*(2.75e−06)*	*(3.75e−03)*
Discrete errors										
Gaussian RE ($\mathbf{D}$ *SPD*)	0.00134	0.0577	0.774	0.778	0.643	1.57	0.00341	0.507	$6.21e-05$	$5.99e-03$
	(0.00195)	*(0.0552)*	*(0.784)*	*(0.788)*	*(0.589)*	*(1.37)*	*(0.00724)*	*(0.504)*	*(2.56e−05)*	*(5.80e−03)*
Gaussian RE ($\mathrm{R}(\mathbf{D})=1$)	0.00132	0.0597	0.778	0.782	0.723	1.43	0.00364	0.472	$1.23e-05$	$5.20e-03$
	(0.00301)	*(0.0608)*	*(0.792)*	*(0.796)*	*(0.737)*	*(1.73)*	*(0.00410)*	*(0.479)*	*(7.19e−05)*	*(5.19e−03)*
Gaussian RE ($\mathrm{R}(\mathbf{D})=2$)	0.00221	0.0586	0.773	0.778	1.10	2.55	0.00726	0.486	$5.39e-07$	$5.73e-03$
	(0.00266)	*(0.0595)*	*(0.783)*	*(0.787)*	*(1.51)*	*(3.27)*	*(0.00719)*	*(0.490)*	*(3.26e−05)*	*(5.76e−03)*
Gaussian mixture RE	0.00174	0.0570	0.778	0.783	0.384	0.955	0.00119	0.500	$7.66e-05$	$4.53e-03$
	(0.00266)	*(0.0560)*	*(0.777)*	*(0.781)*	*(0.614)*	*(1.39)*	*(0.00358)*	*(0.510)*	*(1.20e−06)*	*(4.53e−03)*
Discrete RE	0.00188	0.0599	0.772	0.777	0.848	1.40	0.00786	0.485	$2.97e-05$	$3.69e-03$
	(0.00168)	*(0.0598)*	*(0.786)*	*(0.790)*	*(0.633)*	*(1.35)*	*(0.00666)*	*(0.489)*	*(1.60e−05)*	*(3.62e−03)*

$\boldsymbol{D}$ is the cluster random effects (RE) 3 × 3 covariance matrix; R($\boldsymbol{D}$) = rank($\boldsymbol{D}$) is the rank of $\boldsymbol{D}$. In each cell of two numbers in the table, the one on top is the result from the Algorithm 3*S*-A1-V1, while the one below it, in parentheses, is the result from Gaussian ML.

of them. In our case, the strategy was to discard such data sets, but, in each scenario, run the simulation until 500 data sets have been successfully fitted by the considered algorithm. Nonetheless, while Algorithm 3*S*-A1-V1 never failed during this set of simulations, the situation appeared particularly tricky when running Algorithm lmer-ML. Indeed, with its default settings, its inner workings make that it signaled a failure of convergence on a massive number of simulated data sets, so that we decided actually not to reject them all, rejecting only the cases where the algorithm signaled a difficulty to converge due to a suspected singular fit, *i.e.* true model parameters likely on or close to the boundary of their space of values. Even with that, the only scenarios where we did not witness any failure of lmer-ML was under the Gaussian mixture distribution for the cluster random effects in $\mathbb{R}^3$, whereas one would have only expected failure when the covariance matrix $\boldsymbol{D}$ is rank deficient.

As might be expected, the results of both algorithms are comparable. Nevertheless, while there is no aspect in both tables in which the lmer-ML algorithm uniformly or almost uniformly dominates 3*S*-A1-V1, a dominance of the latter over the former can be observed, in almost all scenarios, on a lower bias of the estimate of σ^2_ε, and not far from so on a lower MSE of that same estimate. Such an almost uniform dominance of the 3*S*-A1-V1 algorithm can also be seen in the

prediction of the response in terms of bias. As for comparing the two algorithms in various individual scenarios, the main striking observation is that 1 almost uniformly dominates lmer-ML when the cluster random effects covariance matrix $\boldsymbol{D}$ is rank deficient, more decisively so the smaller that rank.

6.2. Two Real World Data Sets

In this section, we apply Algorithm 3*S*-A1-V1 on two classical data sets, cake [51] and Blackmore (from the car R package) data, and compare it to the results from Gaussian ML. The results of estimating β, σ_ε^2, $\boldsymbol{D}$ and predicting $U_1,\cdots,U_m$ by Algorithm 3*S*-A1-V1 will be presented. Each of these estimates is given with variability indicators such as estimates of Bias, Mean Squared Error (MSE) and *t*-statistics, *p*-values for significance and 95% two-sided confidence intervals (CI), all computed using the nonparametric residuals bootstrap approach outlined in Carpenter *et al*. [45]. In each LMM fit to a data set, we will also provide an estimate of the random effects ratio ρ in the data, given by (30).

6.3. Application to the Cake Data

The cake data set consists of observations of the *breakage angle* of chocolate cakes made with 3 different recipes and baked at 6 different temperatures, with 15 replicates for each combination of a recipe and a temperature. Hence, there are $3\times 6\times 15 = 270$ sample units (the baked cakes) for each of which there is a recorded value of the variables: angle (numeric), recipe (factor with levels A, B, C), temperature (ordered factor with levels 175 < 185 < 195 < 205 < 215 < 225), replicate (factor with levels 1 to 15), temp (with the same values as temperature, but viewed as numeric). In the R software notations [39], we fitted to these data the two LMMs:

$$\text{angle} \sim \text{fe}(-1+\text{temp}) + \text{re}(1) + \text{Gr}(\text{recipe}:\text{replicate}) \tag{43a}$$

$$\text{angle} \sim \text{fe}(-1+\text{temperature}) + \text{re}(1) + \text{Gr}(\text{recipe}:\text{replicate}) \tag{43b}$$

In the LMM (43a), the response is angle, the fixed effects variable is temp, we have a random intercept as only random effects variable while the clustering variable is obtained by crossing the categories of the two factors recipe and replicate, thus yielding $3\times 15 = 45$ clusters. The difference between the LMMs (43a) and (43b) is that the latter uses instead, as lone fixed effects variable, temperature viewed as a 6-level factor. The reason the intercept has been excluded in the fixed effects part of both models is that a preliminary fitting of the LMM (43a) with an intercept among the fixed effects revealed it to be insignificant through the bootstrap procedure. The interest in fitting the two models is that it allows to assess whether the true values of the temperature really matter in the quality of the fit.

The results for models (43a) and (43b) are respectively presented in **Table 2** and **Table 3**. For a given parameter estimate, the *t*-statistic is the estimate value divided by the square root of its MSE and the *p*-value is calculated assuming that *t*-statistic has, approximately, a standard Gaussian distribution when the true

Table 2. Results of algorithm 3*S*-A1-V1 fitting the LMM (43a) to the cake data.

parameters	*est.*	$[2.5\% - 97.5\%]$	*bias*	*std.dev*	$\sqrt{\text{MSE}}$	*t-stat*	*p–value*
$\beta_{\texttt{temp}}$	0.1604	[0.1519, 0.1705]	0.0002258	0.004798	0.004803	33.400	$1.545e-244$
$\sigma^2_{\texttt{intercept}}$	39.28	[20.77, 60.00]	−0.1352	9.850	9.851	3.988	$6.675e-05$
σ^2_ε	20.71	[17.01, 24.64]	−0.1051	1.946	1.949	10.62	$2.299e-26$
ρ	1.897	[1.003, 2.978]	0.02226	0.5226	0.5231	3.626	$2.876e-04$

β_{temp}: coeff. of fixed effects temp; $\sigma^2_{\text{intercept}}$: random intercept variance; σ^2_ε: residual variance; ρ: random effects ratio; *est.*: estimate; [2.5% - 97.5%], bias, std.dev, $\sqrt{\text{MSE}}$, *t*-stat, *p*-value: respectively estimated two-sided 95% percentile confidence interval, bias, standard deviation, square root of the MSE, *t*-statistic and *p*-value of the *t*-test for a zero value, all computed by a bootstrap simulation of 1000 replicates of the model fit using the nonparametric residuals bootstrap approach described in Carpenter *et al.* [45]. These variability estimates are computed from the 1000 bootstrap replicates using the standard formulas introduced by Efron (see [52]).

Table 3. Results of Algorithm 3*S*-A1-V1 fitting the LMM (43b) to the cake data.

parameters	*est.*	$[2.5\% - 97.5\%]$	*bias*	*std.dev*	$\sqrt{\text{MSE}}$	*t-stat*	*p–value*
$\beta_{\texttt{temperature.175}}$	27.980	[25.89, 30.28]	0.01359	1.141	1.141	24.52	$8.184e-133$
$\beta_{\texttt{temperature.185}}$	29.96	[27.82, 32.37]	−0.001430	1.176	1.176	25.47	$3.735e-143$
$\beta_{\texttt{temperature.195}}$	31.42	[29.30, 33.92]	0.03457	1.191	1.192	26.36	$3.469e-153$
$\beta_{\texttt{temperature.205}}$	32.18	[29.90, 34.51]	0.009740	1.194	1.194	26.94	$7.656e-160$
$\beta_{\texttt{temperature.215}}$	35.84	[33.76, 38.43]	0.04854	1.189	1.190	30.13	$1.970e-199$
$\beta_{\texttt{temperature.225}}$	35.36	[33.16, 37.83]	0.0009604	1.161	1.161	30.45	$1.121e-203$
$\sigma^2_{\texttt{intercept}}$	39.30	[20.720, 60.00]	−0.2556	9.807	9.811	4.006	$6.168e-05$
σ^2_ε	20.57	[16.66, 24.66]	−0.1119	2.042	2.045	10.06	$8.470e-24$
ρ	1.911	[1.006, 3.040]	0.01894	0.5280	0.5284	3.616	$2.987e-04$

Meaning of parameters as in **Table 2**.

value of the parameter is 0. The latter is rarely a bad approximation in this context. Overall, the two model fits are quite similar in terms of random effects variance, residual variance and MSPE estimate. Moreover, the relative difference in terms of fitted response values between the two fits ranges between −0.047 and 0.030. Finally, the *p*-value of the random effects ratio ρ is roughly 3×10^{-4}, strongly suggesting the presence of random effects on the intercept w.r.t. the recipe:replicate clustering considered for the baked cakes.

For comparison, we also fitted each of the two LMMs (43a) and (43b) by the Gaussian ML method. The results are given in **Table 4** and, as expected, the parameters estimates are almost identical to those in the two previous tables.

6.4. Application to the Blackmore Longitudinal Data

To illustrate our methods on longitudinal LMMs, we consider the Blackmore data described in detail in Davis *et al.* [53]. It is a longitudinal, retrospective, self-reported data from a case-control study on the physical exercise histories of 231 teenage girls, with 138 who are eating disorder (*anorexia nervosa*) patients recruited from a 4 years inpatient Eating Disorder Program at the Toronto hospital for sick children. The other 93 comparable "control" subjects with no history of a psychiatric disorder (determined by asking relevant clinical questions) were recruited from informational letters through school boards to parents, inviting the teenage daughter (the letter bearer) to take part in the study. Retrospective recall of leisure-time sport and exercise activities at target ages 8, 10, 12, 14 years (if applicable in relation to the girl's current age), and the 12 months

Table 4. LMM Gaussian maximum likelihood fitting to the cake data.

LMM (43a)			
parameters	*est.*	*std.dev*	*t-stat*
$\beta_{\texttt{temp}}$	0.1604	0.004666	34.38
$\sigma^2_{\texttt{intercept}}$	39.30		
σ^2_{ε}	20.62		

LMM (43b)			
parameters	*est.*	*std.dev*	*t-stat*
$\beta_{\texttt{temperature.175}}$	27.98	1.149	24.35
$\beta_{\texttt{temperature.185}}$	29.96	1.149	26.07
$\beta_{\texttt{temperature.195}}$	31.42	1.149	27.35
$\beta_{\texttt{temperature.205}}$	32.18	1.149	28.00
$\beta_{\texttt{temperature.215}}$	35.84	1.149	31.19
$\beta_{\texttt{temperature.225}}$	35.36	1.149	30.77
$\sigma^2_{\texttt{intercept}}$	39.40		
σ^2_{ε}	20.02		

prior to the study, were obtained during a structured interview. Since the girls were recruited at different ages, the number of observations and the age at the last observation vary, respectively from 2 to 5 observations and from 11.58 to 17.92 years. The Blackmore data are given in the long format for a longitudinal data set, with the following variables: subject (an identification code for each girl); age (the girl's age in years at the time of observation); exercise (the variable of interest, the amount of physical activity in which the girl engaged, expressed as estimated hours per week) and group (a factor indicating whether the girl is a "patient" or a "control"). For modeling exercise in terms of age and group across the sample of girls, Davis *et al.* [53] first $\log_2$-transformed exercise to make its distribution for both groups more symmetric and linearised its relationship with age. Because there are some 0 values of exercise, 5 minutes (5/60 of an hour) were added to each value of exercise prior to taking logs.

For these data, we fit an LMM using as fixed effects: an intercept, age minus 8 (denoted age8), group and the interaction of the two latter. As for random effects in this study, they are attached to variations from girl to girl, so the clustering variable is subject. Now, a follow-up plot of $\log_2$ exercise by age for 20 randomly selected patients and 20 randomly selected control girls showed that the $\log_2$ exercise at the start of follow-up varies considerably, suggesting a random intercept in the LMM. Also, the evolution of $\log_2$ exercise with age differs from girl to girl, informing on the inclusion of a random slope for age8. The model to fit is therefore:

$$\begin{aligned}\log_2 \text{exercise} \sim\ & \text{fe}(1+\text{age8}+\text{group}+\text{age8}*\text{group}) \\ & +\text{re}(1+\text{age8})+\text{Gr}(\text{subject})\end{aligned} \tag{44}$$

when fitting that model, and as traditional in this type of study, in the numerical coding of the binary variable group, our reference category is "control". Moreover, the covariance matrix $\boldsymbol{D}$ is 2×2, with diagonal elements $\sigma^2_{\text{intercept}}$ and σ^2_{age8}, and both off-diagonal ones equal to $\text{cov}_{\text{intercept,age8}}$.

The results for the model using Algorithm 3*S*-A1-V1 versus ML are presented in Table 5 and Table 6, respectively. The interaction of age with group is highly significant in both methods, reflecting a steeper average trend in number of exercises with age in the patient group. The fixed intercept and group effects are not significant at a 5% level. Parameters estimates from Algorithm 3*S*-A1-V1 are

Table 5. Results of algorithm 3*S*-A1-V1 fitting the LMM (44) to the Blackmore data.

parameters	*est.*	[2.5% − 97.5%]	*bias*	*std.dev*	$\sqrt{\text{MSE}}$	*t-stat*	*p–value*
$\beta_{\texttt{intercept}}$	−0.2754	[−0.6371, 0.09391]	0.0007563	0.1855	0.1855	−1.485	$1.376e-01$
$\beta_{\texttt{age8}}$	0.06366	[0.005921, 0.1248]	0.001169	0.03069	0.03071	2.073	$3.818e-02$
$\beta_{\texttt{group:patient}}$	−0.3552	[−0.8579, 0.1008]	−0.0008574	0.2406	0.2406	−1.476	$1.399e-01$
$\beta_{\texttt{age8*[group:patient]}}$	0.2405	[0.1610, 0.3200]	−0.001338	0.03956	0.03958	6.077	$1.225e-09$
$\sigma^2_{\texttt{intercept}}$	2.097	[1.668, 2.651]	0.06293	0.2451	0.2531	8.288	$1.155e-16$
$\sigma^2_{\texttt{age8}}$	0.02954	[0.02174, 0.05089]	0.006141	0.007498	0.009692	3.048	$2.301e-03$
$\text{corr}_{\texttt{intercept,age8}}$	−0.3006	[−0.5171, −0.1187]	−0.03928	0.3193	0.3306	−2.751	$5.936e-03$
σ^2_ε	1.533	[1.289, 1.678]	−0.04473	0.09846	0.1081	14.17	$1.336e-45$
ρ	2.297	[1.692, 3.490]	0.2573	0.4720	0.5376	4.272	$1.937e-05$

$\text{corr}_{\text{intercept;age8}}$: correlation coefficient between random effects variables intercept and age8.

Table 6. Results of Gaussian ML fitting of the LMM (44) to the Blackmore data.

parameters	*est.*	*std.dev*	*t-stat*	*parameters*	*est.*
$\beta_{\texttt{intercept}}$	−0.2762	0.1816	−1.521	$\sigma^2_{\texttt{intercept}}$	2.058
$\beta_{\texttt{age8}}$	0.06412	0.03122	2.054	$\sigma^2_{\texttt{age8}}$	0.02644
$\beta_{\texttt{group:patient}}$	−0.3536	0.2343	−1.510	$\text{corr}_{\texttt{intercept,age8}}$	−0.28
$\beta_{\texttt{age8*[group:patient]}}$	0.2396	0.03922	6.110	σ^2_ε	1.548

very close to those from ML. The *p*-value for the random effects ratio estimate is less than 0.001, confirming the presence, w.r.t. the girls, of random effects on the intercept and/or age8, and probably both since the estimates of their variances $\sigma^2_{\text{intercept}}$ and σ^2_{age8} are significantly greater than zero. It also appears that there is a significant negative correlation between those two random effects.

In the Supplementary material, we fitted another LMM to this data set by adding a random slope for age8*group in (44). The results using Algorithm 3*S*-A1-V1 show that, in addition to the parameters already significant in (44), the same is true of the variance of the interaction age8*group, implying a probable random effect on it also. We remark that Gaussian ML fitting of this model failed to converge with the default settings in the R function lmer.

7. Concluding Remarks

Till today, it has been difficult to routinely fit LMMs without assuming both random effects and residual errors to have Gaussian distributions (the default in almost all statistical software packages designed for that purpose). Yet, for many data sets, that assumption may be debatable, especially for the random effects. This is disturbing since modeling of random effects behavior is one of the main goals of LMM fitting in the first place. Therefore, there has been an implicit need, for long now, to develop fitting methods for LMMs not requiring Gaussian assumptions, while being applicable to as wide a range of LMMs as possible. Being restricted to variance components models, the venerable ANOVA methods and Rao's MINQUE largely fall short in that respect. In the work presented here, we were able to devise a new iterative fitting methodology for 2-level (or longitudinal) LMMs with only added assumption (to the basic ones) that the residual errors were uncorrelated and homoscedastic. Each variant of that estimation methodology iterates through a small set of estimating equations and, when

convergent, yields nonnegative estimates of variances and SPSD estimates of covariance matrices. Though no Gaussian assumption is involved in the derivation of these EEs, we, however, showed that if the 2-level LMM is, indeed, Gaussian with i.i.d. errors, then these EEs are equivalent to the likelihood ones, safe for a denominator *n-p* in lieu of *n* in the EE for the residual variance. Furthermore, the newly developed estimation methodology for LMMs, nicknamed 3*S*, is not exclusive to 2-level ones. The same ideas can be used to fit some other classes of LMMs as well, as we showed for variance components (or ANOVA) LMMs, generalizing an old method of Henderson for ML estimation in such models of Gaussian type. An interesting by-product we got is also, actually, an extension of that Henderson method to Gaussian ML fitting of 2-level LMMs with u.ho. errors.

Proof of Theorem 7

In addition to Lemmas S3.1 and S3.2, we shall need the following one:

Lemma 8. *In the* 2*-level LMM* (2), *if the clusters random effects matrix* $\boldsymbol{D}$ *satisfies*:

$$\boldsymbol{D} = \frac{1}{m}\sum_{j=1}^{m}\left(\boldsymbol{V}_j^* + \tilde{U}_j\tilde{U}_j^{\mathrm{T}}\right), \tag{45}$$

with $\tilde{U}_1,\cdots,\tilde{U}_m \in \mathbb{R}^m$, *then, letting* $\tilde{U} = \left(\tilde{U}_1^{\mathrm{T}},\cdots,\tilde{U}_m^{\mathrm{T}}\right)^{\mathrm{T}} \in \mathbb{R}^q$,

$$\mathrm{tr}\left(\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\right) = \left\|\tilde{U}\right\|_{\boldsymbol{G}^{-1}}^2. \tag{46}$$

Proof. From the proof of Proposition S1.4, we already deduce that

$$\mathrm{tr}\left(\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\right) = \mathrm{tr}\left[\left(\boldsymbol{G}-\boldsymbol{V}^*\right)\boldsymbol{G}^{-1}\right]. \tag{47a}$$

Now, in the 2-level LMM (2), we know that $\boldsymbol{G} = \mathrm{diag}\left(\boldsymbol{D},\cdots,\boldsymbol{D}\right)$, and $\boldsymbol{V}^* = \mathrm{diag}\left(\boldsymbol{V}_1^*,\cdots,\boldsymbol{V}_m^*\right)$, which are two conformal $q\times q$ block-diagonal matrices. Hence, one also has:

$$\boldsymbol{G}-\boldsymbol{V}^* = \mathrm{diag}\left(\boldsymbol{D}-\boldsymbol{V}_1^*,\cdots,\boldsymbol{D}-\boldsymbol{V}_m^*\right) \text{ and } \boldsymbol{G}^{-1} = \mathrm{diag}\left(\boldsymbol{D}^{-1},\cdots,\boldsymbol{D}^{-1}\right),$$

which implies that $\left(\boldsymbol{G}-\boldsymbol{V}^*\right)\boldsymbol{G}^{-1} = \mathrm{diag}\left(\left(\boldsymbol{D}-\boldsymbol{V}_1^*\right)\boldsymbol{D}^{-1},\cdots,\left(\boldsymbol{D}-\boldsymbol{V}_m^*\right)\boldsymbol{D}^{-1}\right)$. Therefore,

$$\mathrm{tr}\left[\left(\boldsymbol{G}-\boldsymbol{V}^*\right)\boldsymbol{G}^{-1}\right] = \sum_{j=1}^{m}\mathrm{tr}\left[\left(\boldsymbol{D}-\boldsymbol{V}_j^*\right)\boldsymbol{D}^{-1}\right] = \mathrm{tr}\left[\sum_{j=1}^{m}\left(\boldsymbol{D}-\boldsymbol{V}_j^*\right)\cdot\boldsymbol{D}^{-1}\right]. \tag{47b}$$

Notice then that (45) is equivalent to $\sum_{j=1}^{m}\left(\boldsymbol{D}-\boldsymbol{V}_j^*\right) = \sum_{j=1}^{m}\tilde{U}_j\tilde{U}_j^T$, which, inserted in (47b), yields:

$$\mathrm{tr}\left[\left(\boldsymbol{G}-\boldsymbol{V}^*\right)\boldsymbol{G}^{-1}\right] = \mathrm{tr}\left[\sum_{j=1}^{m}\tilde{U}_j\tilde{U}_j^{\mathrm{T}}\cdot\boldsymbol{D}^{-1}\right] = \sum_{j=1}^{m}\mathrm{tr}\left(\tilde{U}_j\tilde{U}_j^{\mathrm{T}}\boldsymbol{D}^{-1}\right). \tag{47c}$$

Now, for $j=1,\cdots,m$, $\mathrm{tr}\left(\tilde{U}_j\tilde{U}_j^{\mathrm{T}}\boldsymbol{D}^{-1}\right) = \mathrm{tr}\left(\tilde{U}_j^{\mathrm{T}}\boldsymbol{D}^{-1}\tilde{U}_j\right) = \tilde{U}_j^{\mathrm{T}}\boldsymbol{D}^{-1}\tilde{U}_j \in \mathbb{R}$. Hence, (47c) implies: $\mathrm{tr}\left[\left(\boldsymbol{G}-\boldsymbol{V}^*\right)\boldsymbol{G}^{-1}\right] = \sum_{j=1}^{m}\tilde{U}_j^{\mathrm{T}}\boldsymbol{D}^{-1}\tilde{U}_j = \tilde{U}^{\mathrm{T}}\boldsymbol{G}^{-1}\tilde{U} = \left\|\tilde{U}\right\|_{\boldsymbol{G}^{-1}}^2$, which, with (47a), yields (46). □

- We now proceed properly to prove Theorem 7:

Proof. Let $\tilde{\beta}$ and $\tilde{U}$ satisfy the HMMEs (10) given σ_ε^2, $\boldsymbol{D}$ and $\boldsymbol{Y}=\boldsymbol{y}$ in the 2-level LMM (2). Then, thanks to Theorem 1,

$$\tilde{\beta}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{y} \text{ and } \tilde{U}=\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\tilde{\beta}\right), \tag{48a}$$

the latter being equivalent here to:

$$\tilde{U}_j=\boldsymbol{D}\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{V}_j^{-1}\left(\boldsymbol{y}_j-\boldsymbol{X}_j\tilde{\beta}\right), \quad \text{for } j=1,\cdots,m. \tag{48b}$$

Now, let Assumptions $\mathscr{A}1$ - $\mathscr{A}5$ be true. Then, for $j=1,\cdots,m$, $\boldsymbol{Y}_j \overset{\mathscr{L}}{\sim} \mathscr{N}_{n_j}\left(\boldsymbol{X}_j\beta,\boldsymbol{V}_j\right)$, with density:

$$f_{\boldsymbol{Y}_j}\left(\boldsymbol{y}_j \mid \beta,\sigma_\varepsilon^2,\boldsymbol{D}\right)=\left(2\pi\right)^{-n_j/2}\left|\boldsymbol{V}_j\right|^{-1/2}\exp\left[-\frac{1}{2}\left\|\boldsymbol{y}_j-\boldsymbol{X}_j\beta\right\|_{\boldsymbol{V}_j^{-1}}^2\right],$$

The log-likelihood over the data from all the m clusters is thus:

$$\ell\left(\beta,\sigma_\varepsilon^2,\boldsymbol{D}\mid\boldsymbol{y}\right)=c-\frac{1}{2}\sum_{j=1}^{m}\left[\log\left|\boldsymbol{V}_j\right|+\left(\boldsymbol{y}_j-\boldsymbol{X}_j\beta\right)^{\mathrm{T}}\boldsymbol{V}_j^{-1}\left(\boldsymbol{y}_j-\boldsymbol{X}_j\beta\right)\right]. \tag{48c}$$

1) Now, assume also that (33a) holds, hence $\beta=\tilde{\beta}$, thanks to (34). On the other hand, using (S:3.1a) and (S:3.1b) in Lemma S3.2 to differentiate (48c) w.r.t. $\boldsymbol{D}$ gives:

$$-2\frac{\partial\ell}{\partial\boldsymbol{D}}=\sum_{j=1}^{m}\left[\frac{\partial\log\left|\boldsymbol{V}_j\right|}{\partial\boldsymbol{D}}+\frac{\partial\left(\boldsymbol{y}_j-\boldsymbol{X}_j\beta\right)^{\mathrm{T}}\boldsymbol{V}_j^{-1}\left(\boldsymbol{y}_j-\boldsymbol{X}_j\beta\right)}{\partial\boldsymbol{D}}\right]=\operatorname{diag}\left(S\right)-2S,$$

where $S=\sum_{j=1}^{m}S_j$, with $S_j=W_j-B_j$, $B_j=\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{V}_j^{-1}\boldsymbol{Z}_j$, $W_j=\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{V}_j^{-1}\left(\boldsymbol{y}_j-\boldsymbol{X}_j\right)\left(\boldsymbol{y}_j-\boldsymbol{X}_j\right)^{\mathrm{T}}\boldsymbol{V}_j^{-1}\boldsymbol{Z}_j$. With Lemma S3.1, the nonsingularity of $\boldsymbol{D}$, the fact that $\beta=\tilde{\beta}$ and using (48b), then (S:1.6c),

$$\begin{aligned}&\frac{\partial\ell}{\partial\boldsymbol{D}}=0\Leftrightarrow S=0\Leftrightarrow\sum_{j=1}^{m}W_j=\sum_{j=1}^{m}B_j\Leftrightarrow\boldsymbol{D}\cdot\sum_{j=1}^{m}W_j\cdot\boldsymbol{D}=\boldsymbol{D}\cdot\sum_{j=1}^{m}B_j\cdot\boldsymbol{D}\\&\Leftrightarrow\sum_{j=1}^{m}\tilde{U}_j\tilde{U}_j^{\mathrm{T}}=\sum_{j=1}^{m}\left(\boldsymbol{D}-\boldsymbol{V}_j^{*}\right)\Leftrightarrow\boldsymbol{D}=\frac{1}{m}\sum_{j=1}^{m}\left(\boldsymbol{V}_j^{*}+\tilde{U}_j\tilde{U}_j^{\mathrm{T}}\right)=\tilde{\boldsymbol{D}}.\end{aligned} \tag{49a}$$

2) Now, let also (33a) and (33b) both hold. Hence $\tilde{\beta}=\beta$ and (49a) are both true. Furthermore, (49a) implies that $\boldsymbol{G}=\tilde{\boldsymbol{G}}=\operatorname{diag}\left(\tilde{\boldsymbol{D}},\cdots,\tilde{\boldsymbol{D}}\right)$. Using (S:3.1c) and (S:3.1d) in Lemma S3.2 to differentiate (48c) w.r.t. σ_ε^2,

$$-2\frac{\partial\ell}{\partial\sigma_\varepsilon^2}=\operatorname{tr}\left(\boldsymbol{V}^{-1}\right)-\left\|\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\beta\right)\right\|^2. \tag{49b}$$

Now, given that $\beta=\tilde{\beta}$ and (48a), we get, from (12a) in Lemma 3:

$$\sigma_\varepsilon^2\left\|\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\tilde{\beta}\right)\right\|=\left\|\boldsymbol{y}-\boldsymbol{X}\tilde{\beta}-\boldsymbol{Z}\tilde{U}\right\|. \tag{49c}$$

Also, using (11), then Lemma 8 (given that (49a) is true, by assumption),

$$\sigma_\varepsilon^2\operatorname{tr}\left(\boldsymbol{V}^{-1}\right)=\operatorname{tr}\left(\boldsymbol{I}_n\right)-\operatorname{tr}\left(\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\right)=n-\left\|\tilde{U}\right\|_{\tilde{\boldsymbol{G}}^{-1}}^2. \tag{49d}$$

Inserting (49c) and (49d) in (49b), we get:

$$-2\frac{\partial\ell}{\partial\sigma_\varepsilon^2}=\frac{n-\left\|\tilde{U}\right\|_{\tilde{\boldsymbol{G}}^{-1}}^2}{\sigma_\varepsilon^2}-\frac{1}{\sigma_\varepsilon^4}\left\|\boldsymbol{y}-\boldsymbol{X}\tilde{\beta}-\boldsymbol{Z}\tilde{U}\right\|^2.$$

Consequently, from Corollary 1,

$$\frac{\partial \ell}{\partial \sigma_\varepsilon^2} = 0 \Leftrightarrow n\sigma_\varepsilon^2 = \left\| \boldsymbol{y} - \boldsymbol{X}\tilde{\beta} - \boldsymbol{Z}\tilde{U} \right\|^2 + \sigma_\varepsilon^2 \left\| \tilde{U} \right\|_{\tilde{G}^{-1}}^2 = \boldsymbol{y}^{\mathrm{T}} \left(\boldsymbol{y} - \boldsymbol{X}\tilde{\beta} - \boldsymbol{Z}\tilde{U} \right). \quad \square$$

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Gibbons, R.D. and Hedeker, D. (2000) Applications of Mixed-Effects Models in Biostatistics. *Sankhy*, **62**, 70-103.

[2] Yarkiner, Z., Hunter, G., O'Neil, R. and de Lusignan, S. (2013) Applications of Mixed Models for Investigating Progression of Chronic Disease in a Longitudinal Dataset of Patient Records from General Practice. *Journal of Biometrics & Biostatistics*, **S9**, Article No. 001. https://doi.org/10.4172/2155-6180.S9-001

[3] Van der Merwe, A.J. and Pretorius, A.L. (2003) An Application of the Mixed Linear Model and the Dirichlet Process Prior in Veterinary Medicine Research. *Journal of Agricultural, Biological, and Environmental Statistics*, **8**, Article No. 328. https://doi.org/10.1198/1085711032192

[4] Milkevych, V., Madsen, P., Gao, H. and Jensen, J. (2021) The Relative Effect of Genomic Information on Efficiency of Bayesian Analysis of the Mixed Linear Model with Unknown Variance. *Journal of Animal Breeding and Genetics*, **138**, 14-22.

[5] Torkashvand, E., Jozani, M.J. and Torabi, M. (2017) Clustering in Small Area Estimation with Area Level Linear Mixed Models. *Journal of the Royal Statistical Society Series A* (*Statistics in Society*), **180**, 1253-1279. https://doi.org/10.1111/rssa.12308

[6] Maiti, T. (2001) Robust Generalized Linear Mixed Models for Small Area Estimation. *Journal of Statistical Planning and Inference*, **98**, 225-238. https://doi.org/10.1016/S0378-3758(00)00302-5

[7] Kizilkaya, K., Garrick, D.J., Fernando, R.L., Mestav, B. and Yildiz, M.A. (2010) Use of Linear Mixed Models for Genetic Evaluation of Gestation Length and Birth Weight Allowing for Heavy-Tailed Residual Effects. *Genetics Selection Evolution*, **42**, Article No. 26. https://doi.org/10.1186/1297-9686-42-26

[8] Dandine-Roulland, C. and Perdry, H. (2015) The Use of the Linear Mixed Model in Human Genetics. *Human Heredity*, **80**, 196-206. https://doi.org/10.1159/000447634

[9] Lussetti, D., Kuljus, K., Ranneby, B., Ilstedt, U., Falck, J. and Karlsson, A. (2019) Using Linear Mixed Models to Evaluate Stand Level Growth Rates for Dipterocarps and *Macaranga* Species Following Two Selective Logging Methods in Sabah, Borneo. *Forest Ecology and Management*, **437**, 372-379. https://doi.org/10.1016/j.foreco.2019.01.044

[10] Zamudio, F., Yanez, M., Guerra, F., Fuentes, D. and Gonzalez, A. (2020) Comparative Analysis of SNP Data and Hybrid Taxa Information by Using a Classificatory Linear Mixed Model to Study the Genetic Variation and Heritability of Initial Height Growth in Selected Poplar Hybrids. *Tree Genetics & Genomes* **16**, Article No. 69. https://doi.org/10.1007/s11295-020-01435-1

[11] Jiang, J. (2007) Linear and Generalized Linear Mixed Models and Their Applications. Springer, New York.

[12] Demidenko, E. (2013) Mixed Models: Theory and Applications with R. 2nd Edition,

John Wiley & Sons, Inc., Hoboken.

[13] West, B.T., Welch, K.B. and Galecki, A.T. (2007) Linear Mixed Models: A Practical Guide Using Statistical Software. Chapman & Hall/CRC, London.

[14] Searle, S.R., Casella, G. and McCulloch, C.E. (2006) Variance Components. Wiley, New York.

[15] Hartley, H.O. and Rao, J.N.K. (1967) Maximum Likelihood Estimation for the Mixed Analysis of Variance Model. *Biometrika*, **54**, 93-108. https://doi.org/10.1093/biomet/54.1-2.93

[16] Harville, D.A. (1977) Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, **72**, 320-338. https://doi.org/10.1080/01621459.1977.10480998

[17] 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.

[18] Pinheiro, J.C. and Bates, D.M. (2000) Mixed-Effects Models in S and S-PLUS. Springer, NY.

[19] Gumedze, F.N. and Dunne, T.T. (2011) Parameter Estimation and Inference in the Linear Mixed Model. *Linear Algebra and its Applications*, **435**, 1920-1944. https://doi.org/10.1016/j.laa.2011.04.015

[20] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society Series B*, **39**, 1-38.

[21] Laird, N.M. (1982) Computation of Variance Components Using the EM Algorithm. *Journal of Statistical Computation and Simulation*, **14**, 295-303. https://doi.org/10.1080/00949658208810550

[22] Laird, N.M. and Ware, J.M. (1982) Random Effects Models for Longitudinal Data. *Biometrics*, **38**, 963-974. https://doi.org/10.2307/2529876

[23] Lange, N. and Ryan, L. (1989) Assessing Normality in Random Effects Models. *Annals of Statistics*, **17**, 624-642. https://doi.org/10.1214/aos/1176347130

[24] Li, S., Cai, T.T. and Li, H. (2021) Inference for High-Dimensional Linear Mixed-Effects Models: A Quasi-Likelihood Approach. *Journal of the American Statistical Association*. (In Press) https://doi.org/10.1080/01621459.2021.1888740

[25] Eisenhart, C. (1947) The Assumptions Underlying the Analysis of Variance. *Biometrics*, **3**, 1-21. https://doi.org/10.2307/3001534

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

[27] Montgomery, D.C. (2004) Design and Analysis of Experiments. Wiley, New York.

[28] Oehlert, G.W. (2010) A First Course in Design and Analysis of Experiments. Library of Congress Cataloging-in-Publication Data, USA.

[29] Henderson, C.R. (1953) Estimation of Variance and Covariance Components. *Biometrics*, **9**, 226-252. https://doi.org/10.2307/3001853

[30] Rao, C.R. (1970) Estimation of Heteroscedastic Variances in Linear Models. *Journal of the American Statistical Association*, **65**, 161-172. https://doi.org/10.1080/01621459.1970.10481070

[31] Rao, C.R. (1971) Estimation of Variance and Covariance Components: MINQUE Theory. *Journal of Multivariate Analysis*, **1**, 257-275. https://doi.org/10.1016/0047-259X(71)90001-7

[32] Rao, C.R. (1971) Minimum Variance Quadratic Unbiased Estimation of Variance Components. *Journal of Multivariate Analysis*, **1**, 445-456. https://doi.org/10.1016/0047-259X(71)90019-4

[33] Rao, C.R. (1972) Estimation of Variance and Covariance Components in Linear Models. *Journal of the American Statistical Association*, **67**, 112-115. https://doi.org/10.1080/01621459.1972.10481212

[34] Rao, C.R. and Kleffe, J. (1988) Estimation of Variance Components and Applications. North-Holland, Amsterdam.

[35] Jiang, J. (1996) REML Estimation: Asymptotic Behavior and Related Topics. *Annals of Statistics*, **24**, 255-286. https://doi.org/10.1214/aos/1033066209

[36] Heyde, C.C. (1994) A Quasi-Likelihood Approach to the REML Estimating Equations. *Statistics & Probability Letters*, **21**, 381-384. https://doi.org/10.1016/0167-7152(94)00035-2

[37] Jiang, J., Luan, Y. and Wang, Y. (2007) Iterative Estimating Equations: Linear Convergence and Asymptotic Properties. *The Annals of Statistics*, **35**, 2233-2260. https://doi.org/10.1214/009053607000000208

[38] Goldstein, H. (1986) Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares. *Biometrika*, **73**, 43-56. https://doi.org/10.1093/biomet/73.1.43

[39] R Core Team (2019) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/

[40] Searle, S.R. (1995) The Matrix Handling of BLUE and BLUP in the Mixed Linear Model. BU-1275-MA, Biometrics Unit, Cornell University, Ithaca.

[41] Henderson, C.R. (1950) Estimation of Genetic Parameters (Abstract). *Annals of Mathematical Statistics*, **21**, 309-310.

[42] Henderson, C.R., Kempthorne, O., Searle, S.R. and von Krosigk, C.N. (1959) The Estimation of Environmental and Genetic Trends from Records Subject to Culling. *Biometrics*, **15**, 192-218. https://doi.org/10.2307/2527669

[43] Henderson, C.R. (1973) Sire Evaluation and Genetic Trends. *Journal of Animal Science*, 10-41. https://doi.org/10.1093/ansci/1973.Symposium.10

[44] Henderson, C.R. (1963) Selection Index and Expected Genetic Advance. In: *Statistical Genetics and Plant Breeding*, *National Academy of Sciences*, No. 982, National Research Council Publication, Washington DC, 141-163.

[45] Carpenter, J.R., Goldstein, H. and Rasbash, J. (2003) A Novel Bootstrap Procedure for Assessing the Relationship between Class Size and Achievement. *Journal of the Royal Statistical Society*: *Series C*, **52**, 431-443. https://doi.org/10.1111/1467-9876.00415

[46] Van der Leeden, R., Meijer, E. and Busing, F. (2008) Chapter 11. Resampling Multilevel Models. In: de Leeuw, J. and Meijer, E., Eds., *Handbook of Multilevel Analysis*, Springer, New York, 401-433.

[47] Chambers, R. and Chandra, H. (2013) A Random Effect Block Bootstrap for Clustered Data. *Journal of Computational and Graphical Statistics*, **22**, 452-470. https://doi.org/10.1080/10618600.2012.681216

[48] Thai, H.-T., Mentré, F., Holford, N., Veyrat-Follet, C. and Comets, E. (2013) A Comparison of Bootstrap Approaches for Estimating Uncertainty of Parameters in Linear Mixed-Effects Models. *Pharmaceutical Statistics*, **12**, 129-140. https://doi.org/10.1002/pst.1561

[49] Modugno, L. and Giannerini, T. (2015) The Wild Bootstrap for Multilevel Models. *Communications in Statistics—Theory and Methods*, **44**, 4812-4825.

https://doi.org/10.1080/03610926.2013.802807

[50] Bates, D., Maechler, M., Bolker, B. and Walker, S. (2015) Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, **67**, 1-48. https://doi.org/10.18637/jss.v067.i01

[51] Cook, F.E. (1938) Chocolate Cake, I. Optimum Baking Temperature. Master's Thesis, Iowa State College, Iowa.

[52] Efron, B. and Tibsharani, J. (1993) An Introduction to the Bootstrap. Chapman and Hall, London.

[53] Davis, C., Blackmore, E., Katzman, D.K. and Fox, J. (2005) Female Adolescents with Anorexia Nervosa and Their Parents: A Case-Control Study of Exercise Attitudes and Behaviours. *Psychological Medicine*, **35**, 377-386. https://doi.org/10.1017/S0033291704003447

Supplementary Material

Supplementary Information for the article: Nonparametric Estimation in Linear Mixed Models with Uncorrelated Homoscedastic Errors

Introduction

In this document, we present supplementary materials useful for the understanding of the article. For that, the first section gathers some known results on LMMs scattered here and there in the literature, and some new results of our own. Section S2 presents some important implementation details about our presented iterative methods for fitting LMMs, while Section S3 presents two lemmas useful for proving Theorem 7 in the article. In Section S4, 3*S* fitting algorithms for 2-level LMMs with u.ho. errors and diagonal covariance matrix for the cluster random effects, and ANOVA LMMs are presented. More analysis of the Blackmore data is presented in the last section.

S1. LMMs without Gaussian Assumptions: Some Useful Results

S1.1. An Equivalent Formulation of Assumptions $\mathcal{A}1$ - $\mathcal{A}3$

It is useful to note the following equivalent formulation of Assumptions $\mathcal{A}1$ - $\mathcal{A}3$:

Proposition S1.1. *Assumptions* $\mathcal{A}1$ - $\mathcal{A}3$ *are equivalent to the following set of* 3 *assumptions*:

a1) *The* U_j*'s are identically distributed in* $\mathbb{R}^r$.

a2) $\forall j$, *the two random vectors* U_j *and* ε_j *are independent.*

a3) *The random couples* $(U_1,\varepsilon_1),\cdots,(U_m,\varepsilon_m)$ *are mutually independent.*

S1.2. More about the BLP of $U \mid Y$ in an LMM

The best linear predictor $\bar{U} = \mathrm{BLP}(U \mid \boldsymbol{Y})$, given by (5), plays a key role in our new estimation methodology for LMMs with u.ho. errors. That is why we expand on it here. And, again, no Gaussian distributional assumption is involved.

S1.2.1. Covariance Matrix and Another Expression of the BLP of $U \mid Y$ in an LMM

The relationship between the covariance matrix of $\bar{U} = \mathrm{BLP}(U \mid \boldsymbol{Y})$ and that of the true random effects vector U will be instrumental in estimating the latter in our methodology, starting with:

$$\mathbb{M}\mathrm{cov}(\bar{U}) = \boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathbb{M}\mathrm{cov}(\boldsymbol{Y})\,\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G} = \boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G} = \boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}, \quad \text{(S:1.1a)}$$

thanks to (5). Hence,

$$\mathbb{M}\mathrm{cov}(\bar{U}) = \mathbb{E}(\bar{U}\bar{U}^{\mathrm{T}}) = \boldsymbol{G} - \boldsymbol{V}^{*}, \quad \text{(S:1.1b)}$$

where $\boldsymbol{V}^{*}$ is the symmetric matrix:

$$\boldsymbol{V}^{*} = \boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}. \quad \text{(S:1.1c)}$$

Introducing the apparently complicated matrix $\boldsymbol{V}^*$ is motivated by the fact that it is only $q\times q$ (and not $n\times n$ like $\boldsymbol{V}$) and turns out instrumental in devising an alternative useful way of rewriting the BLP (5) in the LMM (1), more so with u.ho. residual errors:

Lemma S1.2. *In the LMM* (1) *with Assumption* $\mathcal{A}1_g$, *the following identities hold*:

$$\boldsymbol{V}^* = \left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right)^{-1}, \tag{S:1.2a}$$

$$\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1} = \boldsymbol{V}^*\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}^{-1}. \tag{S:1.2b}$$

Moreover, *if Assumption* $\mathcal{A}4_g$ *also holds*, *then these identities simplify to*:

$$\boldsymbol{V}^* = \left(\sigma_\varepsilon^{-2}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right)^{-1} = \sigma_\varepsilon^2\cdot\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z} + \sigma_{\mathrm{e}}^2\boldsymbol{G}^{-1}\right)^{-1}, \tag{S:1.2c}$$

$$\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1} = \sigma_\varepsilon^{-2}\boldsymbol{V}^*\boldsymbol{Z}^{\mathrm{T}}. \tag{S:1.2d}$$

The advantage of (S:1.2c)-(S:1.2d) is that they give respective expressions of $\boldsymbol{V}^*$ and $\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}$ using only $q\times q$ and $n\times q$ matrices, and no $n\times n$ one. These imply a much computationally cheaper formula for the BLP of $U\,|\,\boldsymbol{Y}$ in an LMM with u.ho. errors, when compared with (5):

Proposition S1.3. *In the LMM* (1) *with Assumption* $\mathcal{A}1_g$, *one has*:

$$\bar{U} = \mathrm{BLP}\left(U\,|\,\boldsymbol{Y}\right) = \boldsymbol{V}^*\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}^{-1}\left(\boldsymbol{Y} - \boldsymbol{X}\beta\right). \tag{S:1.3a}$$

Moreover, *if Assumption* $\mathcal{A}4_g$ *also holds*, *then this simplifies to*:

$$\bar{U} = \mathrm{BLP}\left(U\,|\,\boldsymbol{Y}\right) = \sigma_\varepsilon^{-2}\boldsymbol{V}^*\boldsymbol{Z}^{\mathrm{T}}\left(\boldsymbol{Y} - \boldsymbol{X}\beta\right). \tag{S:1.3b}$$

It is also worth noting:

Proposition S1.4. *In the LMM* (1) *with Assumption* $\mathcal{A}1_g$, *one has*:

$$\mathbb{E}\left(\left\|\bar{U}\right\|_{\boldsymbol{G}^{-1}}^2\right) = q - \mathrm{tr}\left(\boldsymbol{G}^{-1}\boldsymbol{V}^*\right). \tag{S:1.4a}$$

Moreover, *if Assumption* $\mathcal{A}4_g$ *also holds*, *then we also have*:

$$\mathbb{E}\left(\left\|\bar{U}\right\|_{\boldsymbol{G}^{-1}}^2\right) = n - \sigma_\varepsilon^2\mathrm{tr}\left(\boldsymbol{V}^{-1}\right) = \sigma_\varepsilon^{-2}\mathrm{tr}\left(\boldsymbol{V}^*\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right). \tag{S:1.4b}$$

Proof. *Given that* $\left\|\bar{U}\right\|_{\boldsymbol{G}^{-1}}^2 = \bar{U}^{\mathrm{T}}\boldsymbol{G}^{-1}\bar{U}$, $\mathbb{E}\left(\bar{U}\right) = \boldsymbol{0}$ *and* (S:1.1b),

$$\mathbb{E}\left(\left\|\bar{U}\right\|_{\boldsymbol{G}^{-1}}^2\right) = \mathbb{E}\left(\bar{U}\right)^{\mathrm{T}}\boldsymbol{G}^{-1}\mathbb{E}\left(\bar{U}\right) + \mathrm{tr}\left[\boldsymbol{G}^{-1}\mathbb{M}\mathrm{cov}\left(\bar{U}\right)\right] = \mathrm{tr}\left[\boldsymbol{G}^{-1}\mathbb{M}\mathrm{cov}\left(\bar{U}\right)\right]$$
$$= \mathrm{tr}\left[\boldsymbol{G}^{-1}\left(\boldsymbol{G} - \boldsymbol{V}^*\right)\right] = \mathrm{tr}\left(\boldsymbol{I}_q - \boldsymbol{G}^{-1}\boldsymbol{V}^*\right) = q - \mathrm{tr}\left(\boldsymbol{G}^{-1}\boldsymbol{V}^*\right).$$

Secondly, *from* (S:1.1a), $\mathrm{tr}\left[\boldsymbol{G}^{-1}\mathbb{M}\mathrm{cov}\left(\bar{U}\right)\right] = \mathrm{tr}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\right) = \mathrm{tr}\left(\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\right)$. *Now*, *if Assumption* $\mathcal{A}4_g$ *also holds*, *then*

$$\mathrm{tr}\left(\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\right) = \mathrm{tr}\left[\boldsymbol{V}^{-1}\left(\boldsymbol{V} - \sigma_\varepsilon^2\boldsymbol{I}_n\right)\right] = \mathrm{tr}\left(\boldsymbol{I}_n\right) - \sigma_\varepsilon^2\mathrm{tr}\left(\boldsymbol{V}^{-1}\right),$$

which entails the first equality in (S:1.4b). *For the second one*, *using* (S:1.2d),

$$\mathrm{tr}\left(\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\right) = \mathrm{tr}\left(\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}^{-1}\right) = \sigma_\varepsilon^{-2}\mathrm{tr}\left(\boldsymbol{Z}\boldsymbol{V}^*\boldsymbol{Z}^{\mathrm{T}}\right) = \sigma_\varepsilon^{-2}\mathrm{tr}\left(\boldsymbol{V}^*\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right).$$

The last proposition implies some additional striking matrix identities in an LMM with u.ho. errors:

Proposition S1.5. *In the LMM* (1) *with Assumptions* $\mathcal{A}1_g$ *and* $\mathcal{A}4_g$,

$$\sigma_\varepsilon^2 \operatorname{tr}\left(\boldsymbol{V}^{-1}\right) + \sigma_\varepsilon^{-2} \operatorname{tr}\left(\boldsymbol{V}^* \boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z}\right) = n, \tag{S:1.5a}$$

$$\operatorname{tr}\left(\boldsymbol{G}^{-1} \boldsymbol{V}^*\right) + \sigma_\varepsilon^{-2} \operatorname{tr}\left(\boldsymbol{V}^* \boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z}\right) = q. \tag{S:1.5b}$$

S1.2.2. Cluster Partitioning of the $\mathrm{BLP}(U \mid Y)$ in a 2-Level LMM

In the 2-level LMM (2), we can partition $\bar{U} = \mathrm{BLP}(U \mid \boldsymbol{Y})$ cluster-wise:

$$\bar{U} = \begin{pmatrix} \bar{U}_1 \\ \vdots \\ \bar{U}_m \end{pmatrix}, \quad \text{with each } \bar{U}_j \text{ a random } r\text{-vector } (j = 1, \cdots, m). \tag{S:1.6a}$$

Combining that with (5), the assumptions and the partitioned structures of the various design and covariance matrices $\boldsymbol{G}$, $\boldsymbol{R}$ and $\boldsymbol{V}$, we get:

$$\bar{U}_j = \mathrm{BLP}\left(U_j \mid \boldsymbol{Y}_j\right) = \boldsymbol{D}\boldsymbol{Z}_j^{\mathrm{T}} \boldsymbol{V}_j^{-1} \left(\boldsymbol{Y}_j - \boldsymbol{X}_j \beta\right) \quad (j = 1, \cdots, m). \tag{S:1.6b}$$

Similarly, from (S:1.1c), one sees that $\boldsymbol{V}^* = \operatorname{diag}\left(\boldsymbol{V}_1^*, \cdots, \boldsymbol{V}_m^*\right)$, where

$$\boldsymbol{V}_j^* = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{Z}_j^{\mathrm{T}} \boldsymbol{V}_j^{-1} \boldsymbol{Z}_j \boldsymbol{D} \in \mathcal{M}_r\left(\mathbb{R}\right) \quad (j = 1, \cdots, m). \tag{S:1.6c}$$

Furthermore, from (S:1.6b), calculations similar to (S:1.1a) give:

$$\mathbb{M}\mathrm{cov}\left(\bar{U}_j\right) = \mathbb{E}\left(\bar{U}_j \bar{U}_j^{\mathrm{T}}\right) = \boldsymbol{D}\boldsymbol{Z}_j^{\mathrm{T}} \boldsymbol{V}_j^{-1} \boldsymbol{Z}_j \boldsymbol{D} = \boldsymbol{D} - \boldsymbol{V}_j^* \quad (j = 1, \cdots, m). \tag{S:1.6d}$$

So, here, $\mathbb{M}\mathrm{cov}\left(\bar{U}\right) = \operatorname{diag}\left(\mathbb{M}\mathrm{cov}\left(\bar{U}_1\right), \cdots, \mathbb{M}\mathrm{cov}\left(\bar{U}_m\right)\right)$, thus yielding

$$\mathbb{M}\mathrm{cov}\left(\bar{U}\right) = \operatorname{diag}\left(\boldsymbol{D} - \boldsymbol{V}_1^*, \cdots, \boldsymbol{D} - \boldsymbol{V}_m^*\right) = \boldsymbol{G} - \boldsymbol{V}^*. \tag{S:1.6e}$$

Finally, the cluster-wise versions of Lemma S1.2 and Propositions S1.3-S1.5 for the 2-level LMM (2) are:

Lemma S1.6. *In the* 2*-level LMM* (2), *the following identities hold, for* $j = 1, \cdots, m$:

$$\boldsymbol{V}_j^* = \left(\boldsymbol{Z}_j^{\mathrm{T}} \boldsymbol{R}_j^{-1} \boldsymbol{Z}_j + \boldsymbol{D}^{-1}\right)^{-1}, \tag{S:1.7a}$$

$$\boldsymbol{D}\boldsymbol{Z}_j^{\mathrm{T}} \boldsymbol{V}_j^{-1} = \boldsymbol{V}_j^* \boldsymbol{Z}_j^{\mathrm{T}} \boldsymbol{R}_j^{-1}. \tag{S:1.7b}$$

Moreover, if Assumption $\mathcal{A}4_g$ *also holds, then these identities simplify to*:

$$\boldsymbol{V}_j^* = \left(\sigma_\varepsilon^{-2} \boldsymbol{Z}_j^{\mathrm{T}} \boldsymbol{Z}_j + \boldsymbol{D}^{-1}\right)^{-1} = \sigma_\varepsilon^2 \left(\boldsymbol{Z}_j^{\mathrm{T}} \boldsymbol{Z}_j + \sigma_\varepsilon^2 \boldsymbol{D}^{-1}\right)^{-1}, \tag{S:1.7c}$$

$$\boldsymbol{D}\boldsymbol{Z}_j^{\mathrm{T}} \boldsymbol{V}_j^{-1} = \sigma_\varepsilon^{-2} \boldsymbol{V}_j^* \boldsymbol{Z}_j^{\mathrm{T}}. \tag{S:1.7d}$$

Proposition S1.7. *In the* 2*-level LMM* (2), *one has, for* $j = 1, \cdots, m$:

$$\bar{U}_j = \mathrm{BLP}\left(U_j \mid \boldsymbol{Y}_j\right) = \boldsymbol{V}_j^* \boldsymbol{Z}_j^{\mathrm{T}} \boldsymbol{R}_j^{-1} \left(\boldsymbol{Y}_j - \boldsymbol{X}_j \beta\right). \tag{S:1.8a}$$

Moreover, if Assumption $\mathcal{A}4_g$ *also holds, then this simplifies to*:

$$\bar{U}_j = \mathrm{BLP}\left(U_j \mid \boldsymbol{Y}_j\right) = \sigma_\varepsilon^{-2} \boldsymbol{V}_j^* \boldsymbol{Z}_j^{\mathrm{T}} \left(\boldsymbol{Y}_j - \boldsymbol{X}_j \beta\right). \tag{S:1.8b}$$

Proposition S1.8. *In the* 2*-level LMM* (2), *one has, for* $j = 1, \cdots, m$:

$$\mathbb{E}\left(\left\|\bar{U}_j\right\|_{\boldsymbol{D}^{-1}}^2\right) = r - \operatorname{tr}\left(\boldsymbol{D}^{-1} \boldsymbol{V}_j^*\right). \tag{S:1.9}$$

Moreover, if Assumption $\mathcal{A}4_g$ *also holds, then we also have*:

$$\mathbb{E}\left(\left\|\bar{U}_j\right\|_{\boldsymbol{D}^{-1}}^2\right)=n_j-\sigma_\varepsilon^2\operatorname{tr}\left(\boldsymbol{V}_j^{-1}\right)=\sigma_\varepsilon^{-2}\operatorname{tr}\left(\boldsymbol{V}_j^*\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{Z}_j\right). \tag{S:1.10}$$

Proposition S1.9. *In the 2-level LMM* (2) *with Assumptions* $\mathcal{A}1_g$ *and* $\mathcal{A}4_g$, *for* $j=1,\cdots,m$:

$$\sigma_\varepsilon^2\operatorname{tr}\left(\boldsymbol{V}_j^{-1}\right)+\sigma_\varepsilon^{-2}\operatorname{tr}\left(\boldsymbol{V}_j^*\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{Z}_j\right)=n_j, \tag{S:1.11a}$$

$$\operatorname{tr}\left(\boldsymbol{D}^{-1}\boldsymbol{V}_j^*\right)+\sigma_\varepsilon^{-2}tr\left(\boldsymbol{V}_j^*\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{Z}_j\right)=r. \tag{S:1.11b}$$

S2. 3S Iterative Algorithms: Some Important Implementation Details

We give here some useful precisions to effectively program Algorithms 3*S*-A1-V1 and 3*S*-A1-V2.

S2.1. Solving the HMMEs in Algorithm 3*S*-A1-V1

Our way of solving a system of the form (29) is based on the following result which we admit:

Theorem S2.1. (***Solving a*** 2×2 ***block nonsingular system***) *Let* $A=\begin{pmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{pmatrix}\in\mathcal{M}_n(\mathbb{R})$ *be nonsingular, with* $n=n_1+n_2$ *and* $n_1,n_2\in\mathbb{N}^*$. *If* $A_{11}\in\mathcal{M}_{n_1}(\mathbb{R})$ *and is nonsingular, then*:

1) $B_{22}=A_{22}-A_{21}A_{11}^{-1}A_{12}\in\mathcal{M}_{n_2}(\mathbb{R})$ *and is nonsingular.*

2) *The unique solution of the linear system* $AX=Y$ *in* $\mathbb{R}^n$ *is given by*:

$$\begin{cases}X_2=B_{22}^{-1}\left(Y_2-A_{21}A_{11}^{-1}Y_1\right)\in\mathbb{R}^{n_2}\\ X_1=A_{11}^{-1}\left(Y_1-A_{12}X_2\right)\in\mathbb{R}^{n_1},\end{cases}$$

with $X_1,Y_1\in\mathbb{R}^{n_1}$, $X_2,Y_2\in\mathbb{R}^{n_2}$ *such that* $X=\begin{pmatrix}X_1\\ X_2\end{pmatrix}$ *and* $Y=\begin{pmatrix}Y_1\\ Y_2\end{pmatrix}$.

Obviously, Theorem S2.1 has a twin version whereby it is rather the lower diagonal block $A_{22}\in\mathcal{M}_{n_2}(\mathbb{R})$ which is assumed nonsingular. To solve the linear system (29) at the beginning of each new iteration $t+1$ in Algorithm 3*S*-A1-V1, we use Theorem S2.1 with:

$$A_{11}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\in\mathcal{M}_p(\mathbb{R}),\ A_{12}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}\in\mathcal{M}_{p,q}(\mathbb{R}),$$

$$A_{21}=\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X}=A_{12}^{\mathrm{T}}\in\mathcal{M}_{q,p}(\mathbb{R}),\ A_{22}=\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}+\hat{\sigma}_\varepsilon^{2(t)}\hat{\boldsymbol{G}}^{(t)-1}\in\mathcal{M}_q(\mathbb{R}),$$

$$Y_1=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}\in\mathbb{R}^p,\ Y_2=\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{y}\in\mathbb{R}^q.$$

Remark S2.1. *The product matrices* A_{11}, A_{12}, A_{21}, $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$ *and vectors* Y_1, Y_2 *are all independent from the iteration index t. Hence, they can be computed before the iterations start, which is a major computational advantage for the running time of Algorithm* 3*S*-*A*1-*V*1.

S2.2. Choice of the Matrix Norm $\|\cdot\|^{(M)}$ in *Instruction* 3 of the 3*S* Algorithms

In the two iterative estimating algorithms for 2-level LMMs presented in Section

4.2.3, all the matrix iterates $\hat{\boldsymbol{D}}^{(t)}$ are at least SPSD. Based on that, we choose an appropriate matrix norm $\|\cdot\|^{(M)}$ to use in *Instruction* 3 of these algorithms to cheaply compute $\left\|\hat{\boldsymbol{D}}^{(t+1)}\right\|^{(M)}$ at any new iteration $t+1$. Indeed, if $A=\left(a_{ij}\right)\in\mathcal{M}_r\left(\mathbb{R}\right)$ is SPSD, then it is known that:

$$\max_{k,l}\left|a_{kl}\right|=\max_{k}a_{kk}.$$

Consequently, if we take as matrix norm:

$$\|A\|^{(M)}=\max_{k,l}\left|a_{kl}\right|, \tag{S:2.1}$$

then $\|A\|^{(M)}=\max\left[\operatorname{diag}\left(A\right)\right]$, where $\operatorname{diag}\left(A\right)$ is the vector of diagonal elements in matrix A. So using the matrix norm given by (S:2.1), $\left\|\hat{\boldsymbol{D}}^{(t+1)}\right\|^{(M)}$ is very cheap to compute and the first control test inequality of the stopping criterion in *Instruction* 3 then simplifies to:

$$\max_{k,l}\left|\left(\hat{\boldsymbol{D}}^{(t+1)}-\hat{\boldsymbol{D}}^{(t)}\right)_{kl}\right|\le\delta_1\cdot\max\left[\operatorname{diag}\left(\hat{\boldsymbol{D}}^{(t+1)}\right)\right].$$

S2.3. Terminating a 3*S* Iterative Algorithm: *The Various Scenarios*

An iterative algorithm aimed at equations solving (or optimization), however theoretically well crafted, almost always runs the risk of *premature termination* in some cases, i.e. the algorithm stalls before its expected convergence has been achieved. Algorithms 3*S*-A1-V1 and 3*S*-A1-V2 are not immune to that threat: each of them may fail on a given data set because the latter is far from satisfying one of the LMM assumptions upon which the algorithm is built. Therefore, it is necessary to include, in the inner workings of our 3*S* algorithms, capabilities to detect, as early as possible, the most plausible scenarios of abnormal termination or behavior.

So, as is traditional in equations solving and optimization iterative algorithms, we use a *categorical flag variable* code.Stop in each of our algorithms to record the cause of the algorithm termination during a particular run, with success code:

code.Stop = OK:1 *convergence*, *with satisfaction of the intended stopping criterion*.

Then we handle the following cases of possible premature stoppage, in which we also include a code. Warning flag variable giving rather a specific warning about the iterations (even if they ended apparently well).

S2.3.1. code.Stop = KO:X: *Fixed Effects Design Matrix* X *Not of Full Column Rank*

As developed till today, a key assumption used in our 3*S* methodology for estimation in LMMs is $\mathcal{A}2_g$, i.e. the fixed effects design matrix $\boldsymbol{X}$ is of full column rank. If not, then some of our derivations no longer hold. Although the methodology might be extended in the future to bypass the need of that assumption, at present we stop any of our algorithms if $\mathcal{A}2_g$ is not satisfied at the outset.

1) Checking if the fixed effects design matrix $\boldsymbol{X}$ is full column rank in Algorithm 3*S*-A1-V1

For Algorithm 3*S*-A1-V1, the checking is based on the fact that computing and storing the $p \times p$ product matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is a necessary requirement in that algorithm, to form the coefficient matrix of the linear system of HMMEs (29). Now, $\boldsymbol{X}$ and $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ always have the same rank, so $\boldsymbol{X}$ has full column rank *if, and only if*, $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is nonsingular. But the latter holds *if, and only if*, the symmetric matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ (which is at least SPSD) is SPD, which is equivalent to having a *Cholesky factorization*, *i.e.* $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = U^{\mathrm{T}}U$, where U is a square upper triangular matrix with positive diagonal elements.

Hence, at the beginning of Algorithm 3*S*-A1-V1, to check for Assumption $\mathcal{A}2_g$, once $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ has been obtained, we compute its *pivoted Cholesky factorization*, using an available software program. Modern numerical software such as the LAPACK routines (Anderson et al., 1999), called by the R statistical software (our programming tool) functions handling Numerical Linear Algebra (NLA) calculations, can compute that factorization and check whether an input symmetric matrix is even SPSD in the first place, and if yes, then check if it is likely, or not, SPD, at least up to the numerical instabilities due to the computer rounding errors. They can also output an estimate of the rank k of an SPSD matrix and indicate its k columns most likely linearly uncorrelated. Applying that to $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, we can decide whether it is clearly SPD, or likely close to a singular SPSD matrix.

In the latter case, we set code.Stop = KO:X, and do not launch Algorithm 3*S*-A1-V1 at all, informing the user why. But we also output the estimated rank k of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ and the indices of its k columns most likely uncorrelated. Since these two pieces of information are the same for $\boldsymbol{X}$, the user can then, if he/she wishes, re-launch the 3*S* iterative algorithm, but with that fixed effects design matrix changed by suppressing its other $p-k$ columns.

2) Checking if the fixed effects design matrix $\boldsymbol{X}$ is full column rank in Algorithm 3*S*-A1-V2

Algorithm 3*S*-A1-V2, instead of the HMMEs, needs to solve, at each iteration, an OLS problem which design matrix is $\boldsymbol{X}$. To achieve that, rather than first computing the product matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, numerical analysts recommend to compute the QR factorization (Demmel, 1997) of $\boldsymbol{X}$. Interestingly, modern numerical software such as the LINPACK routines (Dongarra et al., 1978) called by the R NLA functions can do so while also giving an estimate of the rank of $\boldsymbol{X}$. So for Algorithm 3*S*-A1-V2, our checking of Assumption $\mathcal{A}2_g$ will be based on the results of the *QR factorization with pivoting* of $\boldsymbol{X}$, with the conclusion drawn as for Algorithm 3*S*-A1-V1 above.

S2.3.2. A Vital Issue for LMM Fitting: *Random Effects or Not*?

People trying to fit an LMM rather than an LM to their data most often do not worry whether the incurred excess in mathematical modeling and computational time is worth the price in the first place. They do not ask themselves the obvious

question:

Are there really random effects in the data with the provided clustering?

But that's a mistake, noticeably because it turns out that the answer to that question directly impacts the way an LMM fitting iterative procedure really behaves on a given data set. In the programming of a 3*S* algorithm, we anticipate three possible scenarios related to that issue.

3) code.Stop = KO:1: *RE covariance matrix likely* (*or close to*) *singular*

Even when the cluster random effects are there, they may have been poorly parameterized, so that their scalar components are correlated, resulting in a rank deficient (or singular) covariance matrix $\boldsymbol{D}$. If so, this will affect adversely any of our 3*S* algorithms since they all assume $\boldsymbol{D}$ to be SPD.

To deal with that potential difficulty in a 3*S* algorithm, since $\boldsymbol{D}$ is not known (we are trying to estimate it among other parameters...), our strategy is that after computing each new iterate $\hat{\boldsymbol{D}}^{(t)}$ ($t \geq 1$), we test whether it is of full rank or not. Since in our computations, $\hat{\boldsymbol{D}}^{(t)}$ is always at least SPSD, we numerically check for its rank by attempting a pivoted Cholesky factorization on it along the same lines as described above for the matrix $\boldsymbol{X}^T\boldsymbol{X}$ in the code.Stop = KO:X scenario. If the Cholesky software routine declares the newly computed $\hat{\boldsymbol{D}}^{(t)}$ at an iteration t of not being full rank, the iterative algorithm is stopped, informing the user why: *the true cluster random effects covariance matrix* $\boldsymbol{D}$ *is either singular or close to such a matrix*.

After that premature termination (and as for the case of singular $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ previously), the algorithm also returns the estimated rank ℓ of $\hat{\boldsymbol{D}}^{(t)}$ and the indices of its ℓ most uncorrelated columns. Using these pieces of information as estimates of the corresponding features of the true (but unknown) $\boldsymbol{D}$, the user can re-launch the 3*S* iterative algorithm, but with cluster random effects vectors $U_1,\cdots,U_m$ of dimension reduced from r to ℓ, only keeping the ℓ scalar components associated to the ℓ columns most uncorrelated in $\boldsymbol{D}$. In the data, this amounts to keeping just the corresponding ℓ columns in each of the cluster random effects design matrices $\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_m$.

Note, however, that this does not work when $r=1$, i.e. each cluster random effect is actually a scalar rather than a vector because, then, the cluster random effects covariance matrix $\boldsymbol{D}$ also reduces to a (nonnegative) scalar and is, therefore, rank deficient *if, and only if*, that scalar is zero. This will also be true of its $\hat{\boldsymbol{D}}^{(t)}$ iterates. But the latter matrices will unlikely be numerically close to zero up to computer rounding errors, so there is little chance that any of them would be detected as rank deficient based on the Choleski Factorization. So we need to handle other abnormal possible termination scenarios including that subcase.

4) code.Stop = KO:2: *Slow convergence, likely because of no cluster RE*

Recall that a 3*S* algorithm has converged when the 3 inequalities

$$\left\|\hat{\boldsymbol{D}}^{(t+1)}-\hat{\boldsymbol{D}}^{(t)}\right\|^{(M)} \leq \delta_1 \cdot \left\|\hat{\boldsymbol{D}}^{(t+1)}\right\|^{(M)}, \qquad \text{(S:2.2a)}$$

$$\left\|\hat{\beta}^{(t+1)}-\hat{\beta}^{(t)}\right\| \le \delta_2 \cdot \left\|\hat{\beta}^{(t+1)}\right\|, \tag{S:2.2b}$$

$$\left|\hat{\sigma}_{\varepsilon}^{2(t+1)}-\hat{\sigma}_{\varepsilon}^{2(t)}\right| \le \delta_3 \cdot \hat{\sigma}_{\varepsilon}^{2(t+1)}, \tag{S:2.2c}$$

are simultaneously satisfied, where $\delta_1, \delta_2, \delta_3$ are relative tolerance levels set in $(0,1)$, and $\|\cdot\|^{(M)}$ is a chosen matrix norm. Now, when a 3*S* algorithm appears to be converging slowly, an inspection of the evolution of the iterations intermediate results reveals that this is, more often than not, entirely due to the slow convergence of the sequence of $\hat{\boldsymbol{D}}^{(t)}$ iterates. This manifests itself through the difficulty of having the relative error control inequality (S:2.2a) satisfied, whereas there is really a decrease in the sequence of absolute errors estimates $\left\|\hat{\boldsymbol{D}}^{(t+1)}-\hat{\boldsymbol{D}}^{(t)}\right\|^{(M)}$. From Numerical Analysis, we know that such a situation typically happens when the sought limit value of the $\hat{\boldsymbol{D}}^{(t)}$ iterates (*i.e.* our targeted estimate of $\boldsymbol{D}$) is either *zero* or *close to zero* (probably because the true $\boldsymbol{D}$ is so). A traditional way to anticipate such a situation usually consists in replacing (S:2.2a) with a mix of both relative and absolute error controls:

$$\left\|\hat{\boldsymbol{D}}^{(t+1)}-\hat{\boldsymbol{D}}^{(t)}\right\|^{(M)} \le \delta_1 \cdot \left\|\hat{\boldsymbol{D}}^{(t+1)}\right\|^{(M)} \quad \text{or} \quad \left\|\hat{\boldsymbol{D}}^{(t+1)}-\hat{\boldsymbol{D}}^{(t)}\right\|^{(M)} \le \delta_{11}, \tag{S:2.3}$$

where δ_{11} is an absolute error tolerance upper bound.

But because it is generally difficult to set an appropriate value for δ_{11} in a given context, here we instead use the following context dependent alternative to the second inequality in (S:2.3):

$$\left\|\hat{\boldsymbol{D}}^{(t+1)}-\hat{\boldsymbol{D}}^{(t)}\right\|^{(M)} \le 0.5 \cdot \left\|\hat{\boldsymbol{D}}^{(t+1)}\right\|^{(M)} \quad \text{and} \quad \left\|\boldsymbol{Z}\hat{\boldsymbol{G}}^{(t+1)}\boldsymbol{Z}^{\mathrm{T}}\right\|^{(M)} \le 0.01 \cdot \hat{\sigma}_{\varepsilon}^{2(t+1)}, \tag{S:2.4}$$

the latter inequality aiming at detecting a high probability of zero cluster random effects in the given data set. Hence, the final stopping criterion used in our 3*S* algorithms is rather:

$$((\text{S:2.2a}) \text{ or } (\text{S:2.4})) \text{ and } (\text{S:2.2b}) \text{ and } (\text{S:2.2c}). \tag{S:2.5}$$

The rightmost inequality in (S:2.4) tries to test whether the contribution to the response covariance matrix $\boldsymbol{V}$ of the random effects covariance matrix $\boldsymbol{G}$ is (at least) *two orders of magnitude smaller* than that of the residual errors variance σ_{ε}^2. This occurring gives an indication that these random effects either are not really there or are insignificant, which is slowing down the convergence of the $\hat{\boldsymbol{D}}^{(t)}$ iterates, and thus the algorithm. So when (S:2.4), (S:2.2b) and (S:2.2c) are satisfied, but not (S:2.2a), we stop with code.Stop = OK:2, inviting the user to view this as a *partial success* scenario.

5) code.Warning = 3: *RE might not be there, anyway...*

A 3*S* algorithm being outright successful (i.e. termination with code.Stop = OK:1) does not imply that there are, indeed, random effects in the provided data. So it is useful to try to detect their presence or absence anyway, convergence of the algorithm or not. Our way of doing so is to always test the rightmost inequality in (S:2.4) after exiting a 3*S* algorithm. If satisfied, a warning is issued to the user: *despite the observed convergence* (*if any*), *with the given clustering variable*(*s*), *random effects might actually not be there, or seem insignificant*. An ad-

vantage in that way of trying to detect the zero random effects scenario is that it works whatever the dimension r of these ones in each cluster.

In any event, we always include, in the output from the algorithm, the value of the last ratio:

$$\hat{\rho} = \frac{\left\| \boldsymbol{Z}\hat{\boldsymbol{G}}^{(t)}\boldsymbol{Z}^{\mathrm{T}} \right\|^{(M)}}{\hat{\sigma}_{\varepsilon}^{2(t)}}, \tag{S:2.6a}$$

estimating

$$\rho = \frac{\left\| \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}} \right\|^{(M)}}{\sigma_{\varepsilon}^{2}}. \tag{S:2.6b}$$

The latter can be called the *random effects ratio* in the considered LMM for the given data set, and can be roughly interpreted as the coefficient of random effects really present in the data:

1) the higher above the number 1 the ratio ρ is, the more effective such a presence is;

2) the smaller below 1 the ratio ρ is, the more doubt can be cast about that presence.

But, for a 2-level LMM (2), note that the block diagonal structure of $\boldsymbol{Z}$ implies:

$$\left\| \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}} \right\|^{(M)} = \max_{1\le j\le m} \left\| \boldsymbol{Z}_j \boldsymbol{D} \boldsymbol{Z}_j^{\mathrm{T}} \right\|^{(M)}, \quad \left\| \boldsymbol{Z}\hat{\boldsymbol{G}}^{(t)}\boldsymbol{Z}^{\mathrm{T}} \right\|^{(M)} = \max_{1\le j\le m} \left\| \boldsymbol{Z}_j \boldsymbol{D}^{(t)} \boldsymbol{Z}_j^{\mathrm{T}} \right\|^{(M)}. \tag{S:2.7}$$

Remark S2.2. *One may wonder: why not simply use a statistical test to decide whether there are cluster random effects or not in the given data? The answer is twofold. First, theoretically valid statistical tests are hard to develop for LMMs, more so in our context where we aim at fitting these models without imposing any parametric distributional assumptions on the data. Secondly, as seen above, the situation of no cluster random effects actually first impacts negatively on the convergence of a* 3*S algorithm* (*as well as other iterative algorithms for LMM fitting*), *often severely slowing it down or even flat out stalling it. Hence in that scenario, it is difficult to even compute reliable estimates on which to perform statistical tests. Here, we wanted to develop practical numerical tests aimed at detecting it during the run of the algorithm. And such a detection is of the utmost practical importance since, if well done, it may allow to decide whether LMM modeling even makes sense or not for a given data set. Nevertheless, upon exiting a* 3*S algorithm, we can perform a bootstrap test to decide whether the ratio* ρ *is, or not, significantly greater than zero for the given data set.*

S2.3.3. code.Warning = 4: *Almost Constant Sequence of* $\hat{\beta}^{(t)}$ *Iterates*

For many data sets, we observed that the sequence of $\hat{\beta}^{(t)}$ iterates produced by any of our 3*S* algorithms is nearly constant. This has to do with our way of initializing that sequence by the OLS in an LM fit of the data, i.e. an LMM with no RE. With that initialization, if the sequence ($\hat{\beta}^{(t)}$) exhibits a constant trend from

the outset, one may suspect that the LM is probably better suited for the data set at hand than a pure LMM, *i.e.* an LMM with nonzero RE. And this is likely so because either there are no cluster random effects or these are nonsignificant or would require a bigger sample size to be detected. Hence, whatever the stoppage condition (*convergence or not*) of the iterative algorithm, if for all iterations carried, the $\hat{\beta}^{(t)}$ iterates were all only negligibly different from their OLS starting point $\hat{\beta}^{(0)}$, we conclude that we might be in that *zero RE* scenario, meaning that *what was supposed to be cluster random effects actually behave, more or less, the same from one cluster to the other*.

More specifically, we always perform the following test after termination of any of our 3*S* algorithms (successfully or not):

$$\text{for } t = 1(1)T, \quad \left\| \hat{\beta}^{(t)} - \hat{\beta}^{(0)} \right\| \le \delta_2 \left\| \hat{\beta}^{(0)} \right\|, \tag{S:2.8}$$

where $t = T$ is the iteration at which the algorithm was stopped, and δ_2 is the same as in (S:2.2b). If (S:2.8) is satisfied, then the user is informed that: *may be the LM is better suited for her/his data set*. The recommendation then is either to use the LM, or redesign the clustering variable because it might have been poorly chosen in the first place.

S2.3.4. code.Stop = KO:5: *Authorized Maximum Number of Iterations Reached without Convergence*

As is customary for iterative algorithms expected to stop at satisfaction of a convergence criterion, we must also anticipate the possibility that, for the provided data set, a 3*S* algorithm either does not actually converge or does so slowly. For that, we set, as is usual, a maximum number max.Its (default = 200) of authorized iterations in the algorithm. If that number is reached without the convergence stopping criterion satisfied, the algorithm is terminated with code.Stop = 5, and we inform the user why.

S2.3.5. Complement: *Authorized Minimum Number of Iterations*

Somewhat untraditionally, but to limit the risk of an optimistic premature successful termination, we impose that a 3*S* algorithm always performs a minimum number min.Its (default = 10) before starting to check for a success in convergence.

S3. Two Lemmas Useful for Proving Theorem 7

We admit the following two lemmas (the second one can be proved using multivariable differential tools presented in Graham (1981)):

Lemma S3.1. *Let* $S \in \mathcal{M}_n(\mathbb{R})$. *Then, one has*: $2S - \operatorname{diag}(S) = 0 \Leftrightarrow S = 0$.

Lemma S3.2. *Let* $\boldsymbol{x} \in \mathbb{R}^n$ *and* $N \in \mathcal{M}_n(\mathbb{R})$, *SPD, such that*

$$N = ZQZ^{\mathrm{T}} + \alpha \boldsymbol{I}_n, \quad \text{with } Q \in \mathcal{M}_r(\mathbb{R}), \text{symmetric}, \alpha \in \mathbb{R} \text{ and } Z \in \mathcal{M}_{n,r}(\mathbb{R}).$$

If the upper triangular part of Q has functionally independent elements and which are independent from α, *whereas* $\boldsymbol{x}$ *and Z are independent from Q and* α, *then*:

$$\frac{\partial\left(\boldsymbol{x}^{\mathrm{T}}N^{-1}\boldsymbol{x}\right)}{\partial Q}=\operatorname{diag}\left(W\right)-2W, \tag{S:3.1a}$$

$$\frac{\partial\log\left|N\right|}{\partial Q}=2B-\operatorname{diag}\left(B\right), \tag{S:3.1b}$$

$$\frac{\partial\left(\boldsymbol{x}^{\mathrm{T}}N^{-1}\boldsymbol{x}\right)}{\partial\alpha}=-\left\|N^{-1}\boldsymbol{x}\right\|^{2}, \tag{S:3.1c}$$

$$\frac{\partial\log\left|N\right|}{\partial\alpha}=\operatorname{tr}\left(N^{-1}\right), \tag{S:3.1d}$$

where $W=Z^{\mathrm{T}}N^{-1}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}N^{-1}Z$, $B=Z^{\mathrm{T}}N^{-1}Z$, *and if* A *is a square matrix, then* $\operatorname{diag}\left(A\right)$ *is the diagonal matrix with same main diagonal as* A.

S4. 3*S* Fitting Algorithms for 2-Level LMMs with u.ho. Errors and Diagonal Covariance Matrix for the Cluster Random Effects, and ANOVA LMMs

S4.1. 3*S* Algorithms for 2-Level LMMs with u.ho. Errors Assuming a Diagonal Covariance Matrix for the Cluster Random Effects

We present, here, the equivalence of Algorithms 3*S*-A1-V1 and 3*S*-A1-V2 respectively for 2-level LMMs with u.ho. errors, assuming a diagonal covariance matrix for the cluster random effects vector, that is:

$$\boldsymbol{D}=\operatorname{diag}\left(\sigma_{1}^{2},\cdots,\sigma_{r}^{2}\right).$$

They carry respectively the code names 3*S*-A1-V1-d and 3*S*-A1-V2-d.

Algorithm 3*S*-A1-V1-d. ***Estimation of*** β, σ_{ε}^{2}, $\boldsymbol{D}=\operatorname{diag}\left(\sigma_{1}^{2},\cdots,\sigma_{r}^{2}\right)$ ***and prediction of*** $U_{1},\cdots,U_{m}$ ***in a 2-level LMM with u.ho. errors: Version 1***

1) ***Initialization***: *At iteration* 0, *we estimate* β, σ_{ε}^{2}, $\boldsymbol{D}$ *and predict* $U_{1},\cdots,U_{m}$, *as follows*:

a) *As in Algorithm* 3*S*-*A*1-*V*1;

b) $\hat{\sigma}_{k}^{2(0)}=\frac{1}{m}\sum_{j=1}^{m}\left(\hat{u}_{j,k}^{(0)}\right)^{2}$, $k=1,\cdots,r$, *with* $\hat{u}_{j,k}^{(0)}$ *the* k^{th} *element of the vector* $\hat{\boldsymbol{u}}_{j}^{(0)}$;

c) $\hat{\boldsymbol{D}}^{(0)}=\operatorname{diag}\left(\hat{\sigma}_{1}^{2(0)},\cdots,\hat{\sigma}_{r}^{2(0)}\right)$;

2) ***The iterative process***: *Given* $\hat{\beta}^{(t)}$, $\hat{\sigma}_{\varepsilon}^{2(t)}$, $\hat{\boldsymbol{u}}_{1}^{(t)},\cdots,\hat{\boldsymbol{u}}_{m}^{(t)}$, *hence* $\hat{\boldsymbol{u}}^{(t)}=\left(\hat{\boldsymbol{u}}_{1}^{(t)\mathrm{T}},\cdots,\hat{\boldsymbol{u}}_{m}^{(t)\mathrm{T}}\right)^{\mathrm{T}}$, *and* $\hat{\boldsymbol{D}}^{(t)}=\operatorname{diag}\left(\sigma_{1}^{2(t)},\cdots,\sigma_{r}^{2(t)}\right)$ *from iteration* t, *we obtain estimates and predictions at iteration* $t+1$ *as follows*:

a) *As in Algorithm* 3*S*-*A*1-*V*1;

b) $\hat{\sigma}_{k}^{2(t+1)}=\frac{1}{m}\sum_{j=1}^{m}\left[\left(\hat{\boldsymbol{V}}_{j}^{*(t+1)}\right)_{kk}+\left(\hat{u}_{j,k}^{(t+1)}\right)^{2}\right]$, $k=1,\cdots,r$;

c) $\hat{\boldsymbol{D}}^{(t+1)}=\operatorname{diag}\left(\hat{\sigma}_{1}^{2(t+1)},\cdots,\hat{\sigma}_{r}^{2(t+1)}\right)$;

3) ***Stopping criterion***: *We assume convergence when all the* 3 *inequalities*

$$\left\|\hat{\vartheta}^{(t+1)}-\hat{\vartheta}^{(t)}\right\|\le\delta_{1}\cdot\left\|\hat{\vartheta}^{(t+1)}\right\|, \tag{S:4.1a}$$

$$\left\|\hat{\beta}^{(t+1)}-\hat{\beta}^{(t)}\right\|\le\delta_{2}\cdot\left\|\hat{\beta}^{(t+1)}\right\|, \tag{S:4.1b}$$

$$\left|\hat{\sigma}_{\varepsilon}^{2(t+1)}-\hat{\sigma}_{\varepsilon}^{2(t)}\right| \le \delta_3 \cdot \hat{\sigma}_{\varepsilon}^{2(t+1)}, \quad \text{(S:4.1c)}$$

are satisfied, where $\delta_1, \delta_2, \delta_3$ *are relative tolerance levels set in* $(0,1)$, *and* $\hat{\vartheta} = \left(\hat{\sigma}_1^2, \cdots, \hat{\sigma}_r^2\right)^{\mathrm{T}}$. *Otherwise, repeat Step* 2 *with* $t \leftarrow t+1$.

4) ***Extracting estimates***: *At convergence, take* $\hat{\beta} = \hat{\beta}^{(t+1)}$, $\hat{\sigma}_{\varepsilon}^2 = \hat{\sigma}_{\varepsilon}^{2(t+1)}$, $\hat{\boldsymbol{D}} = \hat{\boldsymbol{D}}^{(t+1)}$ *as estimates of* β, σ_{ε}^2, $\boldsymbol{D}$. *Also, take* $\hat{\boldsymbol{u}}_1^{(t+1)}, \cdots, \hat{\boldsymbol{u}}_m^{(t+1)}$ *as predictions of* $U_1, \cdots, U_m$.

Algorithm 3*S*-A1-V2-d. ***Estimation of*** β, σ_{ε}^2, $\boldsymbol{D} = \mathrm{diag}\left(\sigma_1^2, \cdots, \sigma_r^2\right)$ ***and prediction of*** $U_1, \cdots, U_m$ ***in a 2-level LMM with u.ho. errors: Version 2***

1) ***Initialization***: *The same as in Algorithm* 3*S*-*A*1-*V*1-*d*;

2) ***The iterative process***: *Given* $\hat{\beta}^{(t)}$, $\hat{\sigma}_{\varepsilon}^{2(t)}$, $\hat{\boldsymbol{u}}_1^{(t)}, \cdots, \hat{\boldsymbol{u}}_m^{(t)}$, *hence* $\hat{\boldsymbol{u}}^{(t)} = \left(\hat{\boldsymbol{u}}_1^{(t)\mathrm{T}}, \cdots, \hat{\boldsymbol{u}}_m^{(t)\mathrm{T}}\right)^{\mathrm{T}}$, *and* $\hat{\boldsymbol{D}}^{(t)} = \mathrm{diag}\left(\sigma_1^{2(t)}, \cdots, \sigma_r^{2(t)}\right)$ *from iteration t, we obtain estimates and predictions at iteration* $t+1$ *as follows*:

a) *As in Algorithm* 3*S*-*A*1-*V*2;

b) *As in Algorithm* 3*S*-*A*1-*V*1-*d*;

3) ***Stopping criterion***: *As in Algorithm* 3*S*-*A*1-*V*1-*d*. *Otherwise, repeat Step* 2 *with* $t \leftarrow t+1$.

4) ***Extracting estimates***: *As in Algorithm* 3*S*-*A*1-*V*1-*d*.

S4.2. 3*S* Algorithms for ANOVA LMMs

We present the adaptation of Algorithms 3*S*-A1-V1 and 3*S*-A1-V2 respectively for ANOVA LMMs. That is, the case where

$$\boldsymbol{D} = \boldsymbol{G} = \mathrm{diag}\left(\sigma_1^2 \boldsymbol{I}_{m_1}, \cdots, \sigma_g^2 \boldsymbol{I}_{m_g}\right),$$

where, for $k = 1, \cdots, g$, each variance σ_k^2 corresponds to the *k*th categorical variable with m_k categories, with *g* the total number of categorical variables with random effects. They carry respectively the code names 3*S*-ANOVA-A1-V1 and 3*S*-ANOVA-A1-V2.

Algorithm 3*S*-ANOVA-A1-V1. ***Estimation of*** β, σ_{ε}^2, $\boldsymbol{D} = \boldsymbol{G}$ ***and prediction of U in a 2-level ANOVA LMM: Version 1***

1) ***Initialization***: *At iteration* 0, *we estimate* β, σ_{ε}^2, $\boldsymbol{G}$ *and predict U, as follows*:

a) *As in Algorithm* 3*S*-*A*1-*V*1;

b) $\hat{\boldsymbol{u}}^{(0)} = \left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^{\mathrm{T}}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}^{(0)}\right)$, *the OLS estimate of U in the linear model* $\boldsymbol{Y}^{*(0)} = \boldsymbol{Z}U + \varepsilon$, *with U considered as fixed parameter in the model,* $\boldsymbol{Z}$ *the design matrix, and* $\boldsymbol{Y}^{*(0)} = \boldsymbol{Y} - \boldsymbol{X}\hat{\beta}^{(0)}$ *as the response vector*;

c) $\hat{\sigma}_{\varepsilon}^{2(0)} = \left\|\boldsymbol{y} - \boldsymbol{X}\hat{\beta}^{(0)} - \boldsymbol{Z}\hat{\boldsymbol{u}}^{(0)}\right\|^2 \Big/ (n-p)$;

d) $\hat{\sigma}_k^{2(0)} = \dfrac{1}{m_k}\left\|\hat{\boldsymbol{u}}_k^{(0)}\right\|^2 = \dfrac{1}{m_k}\sum_{l=1}^{m_k}\left(\hat{u}_{k,l}^{(0)}\right)^2$, *for* $k = 1, \cdots, g$, *with* $\hat{\boldsymbol{u}}^{(0)} = \left(\hat{\boldsymbol{u}}_1^{(0)\mathrm{T}}, \cdots, \hat{\boldsymbol{u}}_g^{(0)\mathrm{T}}\right)^{\mathrm{T}}$ *split according to the g categorical variables that form the boolean matrix* $\boldsymbol{Z}$, *such that* $\hat{\boldsymbol{u}}_k^{(0)} = \left(\hat{u}_{k,1}^{(0)}, \cdots, \hat{u}_{k,m_k}^{(0)}\right)^{\mathrm{T}}$ *is the* m_k*-vector of random effects of the* k^{th} *categorical variable.*

e) $\hat{\boldsymbol{G}}^{(0)} = \mathrm{diag}\left(\hat{\sigma}_1^{2(0)}\boldsymbol{I}_{m_1}, \cdots, \hat{\sigma}_g^{2(0)}\boldsymbol{I}_{m_g}\right)$;

2) ***The iterative process***: *Given* $\hat{\beta}^{(t)}$, $\hat{\sigma}_{\varepsilon}^{2(t)}$, $\hat{\boldsymbol{u}}^{(t)}$, *and* $\hat{\boldsymbol{G}}^{(t)} = \mathrm{diag}\left(\hat{\sigma}_1^{2(t)}\boldsymbol{I}_{m_1}, \cdots, \hat{\sigma}_g^{2(t)}\boldsymbol{I}_{m_g}\right)$ *from iteration* t, *we obtain estimates and predictions at iteration* $t+1$ *as follows*:

a) *As in Algorithm* 3*S*-*A*1-*V*1;

b) $\hat{\boldsymbol{V}}^{*(t+1)} = \left(\hat{\sigma}_{\varepsilon}^{-2(t+1)}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z} + \hat{\boldsymbol{G}}^{(t)-1}\right)^{-1}$;

c) $\hat{\sigma}_k^{2(t+1)} = \dfrac{1}{m_k}\sum_{l=1}^{m_k}\left[\left(\hat{\boldsymbol{V}}_{kk}^{*(t+1)}\right)_{ll} + \left(\hat{u}_{k,l}^{(t+1)}\right)^2\right]$, $k = 1, \cdots, g$, *where* $\hat{\boldsymbol{V}}_{kk}^{*(t+1)}$ *is the* k^{th} *diagonal block of* $\hat{\boldsymbol{V}}^{*(t+1)}$ *corresponding to the* m_k *categories of the* k^{th} *categorical random effects variable*;

d) $\hat{\boldsymbol{G}}^{(t+1)} = \mathrm{diag}\left(\hat{\sigma}_1^{2(t+1)}\boldsymbol{I}_{m_1}, \cdots, \hat{\sigma}_g^{2(t+1)}\boldsymbol{I}_{m_g}\right)$;

3) ***Stopping criterion***: *We assume convergence when all the* 3 *inequalities*

$$\left\|\hat{\vartheta}^{(t+1)} - \hat{\vartheta}^{(t)}\right\| \le \delta_1 \cdot \left\|\hat{\vartheta}^{(t+1)}\right\|, \tag{S:4.2a}$$

$$\left\|\hat{\beta}^{(t+1)} - \hat{\beta}^{(t)}\right\| \le \delta_2 \cdot \left\|\hat{\beta}^{(t+1)}\right\|, \tag{S:4.2b}$$

$$\left|\hat{\sigma}_{\varepsilon}^{2(t+1)} - \hat{\sigma}_{\varepsilon}^{2(t)}\right| \le \delta_3 \cdot \hat{\sigma}_{\varepsilon}^{2(t+1)}, \tag{S:4.2c}$$

are satisfied, where $\delta_1, \delta_2, \delta_3$ *are relative tolerance levels set in* $(0,1)$, *and* $\hat{\vartheta} = \left(\hat{\sigma}_1^2, \cdots, \hat{\sigma}_g^2\right)^{\mathrm{T}}$. *Otherwise, repeat Step* 2 *with* $t \leftarrow t+1$.

4) ***Extracting estimates***: *At convergence, take* $\hat{\beta} = \hat{\beta}^{(t+1)}$, $\hat{\sigma}_{\varepsilon}^2 = \hat{\sigma}_{\varepsilon}^{2(t+1)}$, $\hat{\boldsymbol{G}} = \hat{\boldsymbol{G}}^{(t+1)}$ *as estimates of* β, σ_{ε}^2, $\boldsymbol{G}$. *Also, take* $\hat{\boldsymbol{u}}^{(t+1)}$ *as predictions of* U.

Algorithm 3*S*-ANOVA-A1-V2. ***Estimation of*** β, σ_{ε}^2, $\boldsymbol{D} = \boldsymbol{G}$ ***and prediction of*** U ***in a*** 2***-level ANOVA LMM: Version*** 2

1) ***Initialization***: *The same as in Algorithm* 3*S*-*ANOVA*-*A*1-*V*1;

2) ***The iterative process***: *Given* $\hat{\beta}^{(t)}$, $\hat{\sigma}_{\varepsilon}^{2(t)}$, $\hat{\boldsymbol{u}}^{(t)}$, *and* $\hat{\boldsymbol{G}}^{(t)} = \mathrm{diag}\left(\hat{\sigma}_1^{2(t)}\boldsymbol{I}_{m_1}, \cdots, \hat{\sigma}_g^{2(t)}\boldsymbol{I}_{m_g}\right)$ *from iteration* t, *we obtain estimates and predictions at iteration* $t+1$ *as follows*:

a) *As in Algorithm* 3*S*-*A*1-*V*2, *but removing the precision about* $\hat{\boldsymbol{G}}^{(t)}$;

b) $A^{(t+1)} = \left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z} + \hat{\sigma}_{\varepsilon}^{2(t+1)}\hat{\boldsymbol{G}}^{(t)-1}\right)^{-1}$;

c) $\hat{\boldsymbol{u}}^{(t+1)} = A^{(t+1)}\boldsymbol{Z}^{\mathrm{T}}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}^{(t+1)}\right)$;

d) $\hat{\boldsymbol{V}}^{*(t+1)} = \hat{\sigma}_{\varepsilon}^{2(t+1)}A^{(t+1)}$;

e) *As in Algorithm* 3*S*-*ANOVA*-*A*1-*V*1;

3) ***Stopping criterion***: *As in Algorithm* 3*S*-*ANOVA*-*A*1-*V*1. *Otherwise, repeat Step* 2 *with* $t \leftarrow t+1$.

4) ***Extracting estimates***: *As in* 3*S*-*ANOVA*-*A*1-*V*1.

S5. Parameters of the Distributions Used for the Simulations

In this section, we detail the distributions and parameters used for the simulations that yield the results in Section 6.1. As a general rule, the values of most of these parameters were themselves generated randomly before launching the simulations.

S5.1. Simulation of Sample Items, Clusters, Fixed and Random Effects Covariates

We assume that we have a population comprising $m = 10$ clusters in the same

proportions of 1/10. To get a sample item, we first sampled its cluster, then its fixed and random effects covariates, respectively in $\mathbb{R}^4$ and $\mathbb{R}^3$, assuming that in each cluster j, the vector X of fixed effects covariates followed $\mathcal{N}_4\left(\boldsymbol{m}_j^X, \boldsymbol{I}_4\right)$. Likewise, in each cluster j, the vector Z of random effects covariates followed $\mathcal{N}_3\left(\boldsymbol{m}_j^Z, 0.7\boldsymbol{I}_3\right)$. We used:

- $\boldsymbol{m}_1^X = (-2,1,-2,2)^{\mathrm{T}}$, $\boldsymbol{m}_2^X = (-4,-4,-4,1)^{\mathrm{T}}$, $\boldsymbol{m}_3^X = (4,-1,3,2)^{\mathrm{T}}$, $\boldsymbol{m}_4^X = (-1,3,0,-2)^{\mathrm{T}}$, $\boldsymbol{m}_5^X = (-3,3,2,-2)^{\mathrm{T}}$, $\boldsymbol{m}_6^X = (2,-1,3,-3)^{\mathrm{T}}$, $\boldsymbol{m}_7^X = (-3,-3,4,4)^{\mathrm{T}}$, $\boldsymbol{m}_8^X = (-4,1,4,-3)^{\mathrm{T}}$, $\boldsymbol{m}_9^X = (-1,2,0,-3)^{\mathrm{T}}$, $\boldsymbol{m}_{10}^X = (-1,0,0,4)^{\mathrm{T}}$;
- $\boldsymbol{m}_1^Z = (-2,0,-3)^{\mathrm{T}}$, $\boldsymbol{m}_2^Z = (3,1,1)^{\mathrm{T}}$, $\boldsymbol{m}_3^Z = (0,-1,-1)^{\mathrm{T}}$, $\boldsymbol{m}_4^Z = (2,0,-2)^{\mathrm{T}}$, $\boldsymbol{m}_5^Z = (-3,2,3)^{\mathrm{T}}$, $\boldsymbol{m}_6^Z = (3,-3,0)^{\mathrm{T}}$, $\boldsymbol{m}_7^Z = (-2,3,2)^{\mathrm{T}}$, $\boldsymbol{m}_8^Z = (2,3,-3)^{\mathrm{T}}$, $\boldsymbol{m}_9^Z = (0,3,1)^{\mathrm{T}}$, $\boldsymbol{m}_{10}^Z = (-2,2,0)^{\mathrm{T}}$.

In addition, no intercept was included, be it in the fixed part or the random part of the simulated LMM, while we took $\beta = (-1.5, 0.7, 0, 2.3)^{\mathrm{T}} \in \mathbb{R}^4$ as vector of fixed effects parameters.

S5.2. Parameters of the Distributions for Simulating Random Effects and Residual Errors

To simulate the cluster random effects in $\mathbb{R}^3$, we examined 5 distributions, while for the residual errors, we examined 3, hence, testing in total, $5 \times 3 = 15$ scenarios.

S5.2.1. Distributions Tested for the Random Effects

Case 1: Non-degenerate Gaussian distribution in $\mathbb{R}^3$ with mean vector and covariance matrix:

$$\boldsymbol{m} = \boldsymbol{0}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 32.784612 & 28.546477 & -9.726209 \\ 28.546477 & 44.499170 & -3.915344 \\ -9.726209 & -3.915344 & 27.435629 \end{pmatrix}.$$

Case 2: A degenerate Gaussian distribution in $\mathbb{R}^3$ with mean vector and a rank one covariance matrix:

$$\boldsymbol{m} = \boldsymbol{0}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 2.5820307 & -11.232197 & -0.46004207 \\ -11.2321974 & 48.861640 & 2.00124781 \\ -0.4600421 & 2.001248 & 0.08196599 \end{pmatrix}.$$

Case 3: As in Case 2, but with mean vector and a rank two covariance matrix:

$$\boldsymbol{m} = \boldsymbol{0}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 10.19483 & 6.661080 & -17.54595 \\ 6.66108 & 5.462537 & -10.91097 \\ -17.54595 & -10.910967 & 30.47329 \end{pmatrix}.$$

Case 4: Mixture of 3 Gaussian distributions in $\mathbb{R}^3$ with vector of mixing proportions $\pi = (1/6, 1/3, 1/2)^{\mathrm{T}}$, common covariance matrix $\boldsymbol{\Sigma} = 2\boldsymbol{I}_3$ and respective vector means:

$$\boldsymbol{m}_1 = (0.2815777, 1.8254139, -0.5799608)^{\mathrm{T}},$$

$$\boldsymbol{m}_2 = (-0.16064371, 0.01403464, -1.63159078)^{\mathrm{T}},$$

$$\boldsymbol{m}_3 = (0.01323658, -0.61782773, 1.28104746)^{\mathrm{T}}.$$

Case 5: A discrete distribution in $\mathbb{R}^3$ with vector of respective probabilities $\pi = (1/6, 1/3, 1/2)^{\mathrm{T}}$ on the 3 mass points:

$$\boldsymbol{u}_1 = \begin{pmatrix} 0.7043073 \\ 0.7533598 \\ -0.7370090 \end{pmatrix}, \quad \boldsymbol{u}_2 = \begin{pmatrix} 0.6410311 \\ 0.8449623 \\ -0.7769852 \end{pmatrix}, \quad \boldsymbol{u}_3 = \begin{pmatrix} -3.3851412 \\ 0.3501683 \\ 0.8949348 \end{pmatrix}.$$

S5.2.2. Distributions Tested for the Residual Errors

Case 1: Univariate Gaussian distribution $\mathcal{N}(0, 2.25)$.

Case 2: Mixture of 4 Gaussian distributions in $\mathbb{R}$ of variance $\sigma^2 = 0.9613748$ and respective means:

$$\boldsymbol{m}_1 = 1.293309, \quad \boldsymbol{m}_2 = -1.912166, \quad \boldsymbol{m}_3 = 0.4601711, \quad \boldsymbol{m}_4 = 0.690338,$$

and vector of mixing proportions $\pi = (1/8, 1/4, 1/2, 1/8)^{\mathrm{T}}$.

Case 3: A discrete distribution in $\mathbb{R}$ with the 4 mass points:

$$u_1 = 1.532882, \quad u_2 = 1.332857, \quad u_3 = -1.494380, \quad u_4 = 1.778924$$

of respective probabilities $\pi = (1/8, 1/4, 1/2, 1/8)^{\mathrm{T}}$.

S6. More Analysis of the Blackmore Data

One can also consider fitting an LMM to the Blackmore data where, in addition to having a random slope for age as in (44), we also have a random slope for the interaction age8*group, thus the following LMM:

$$\begin{aligned} \log_2 \text{exercise} \sim\ & \text{fe}(1 + \text{age8} + \text{group} + \text{age8} * \text{group}) \\ & + \text{re}(1 + \text{age8} + \text{age8} * \text{group}) + \text{Gr}(\text{subject}) \end{aligned} \tag{S:6.1}$$

The results for this model using Algorithm 3*S*-A1-V1 are presented in **Table S1**. In addition to the parameters which are already significant in the LMM (44), the estimate of the variance $\sigma^2_{\text{age8.group}}$ of the random effect age8*group in the LMM (S:6.1) is also significantly greater than zero, confirming that a random effect on age8*group is also probable. This suggests that there is a likely variation of the interaction between age and group across the girls population in this study.

Table S1. Results of algorithm 3*S*-A1-V1 fitting the LMM (S:6.1) to the Blackmore data.

parameters	*est.*	$[2.5\% - 97.5\%]$	*bias*	*std.dev*	$\sqrt{\text{MSE}}$	*t-stat*	*p–value*
β_{int}	−0.2759	[−0.6373, 0.09259]	0.0008132	0.1852	0.1852	−1.490	$1.363e-01$
β_{age8}	0.06383	[0.008403, 0.1242]	0.001310	0.03032	0.03035	2.103	$3.547e-02$
$\beta_{\text{gr:pat}}$	−0.3546	[−0.8601, 0.1041]	−0.0005366	0.2402	0.2402	−1.476	$1.399e-01$
$\beta_{\text{age8*[gr:pat]}}$	0.2402	[0.1585, 0.3151]	−0.002138	0.03931	0.03936	6.103	$1.041e-09$
σ^2_{int}	2.094	[1.693, 2.652]	0.07693	0.2418	0.2537	8.254	$1.532e-16$
σ^2_{age8}	0.02712	[0.01910, 0.04529]	0.003153	0.006638	0.007349	3.690	$2.239e-04$
$\sigma^2_{\text{age8*[gr:pat]}}$	0.01366	[0.01123, 0.01957]	0.001324	0.002206	0.002573	5.309	$1.101e-07$
$\text{corr}_{\text{int, age8}}$	−0.1995	[−0.4627, −0.04233]	−0.06773	0.3306	0.3586	−1.551	$1.209e-01$
$\text{corr}_{\text{int, age8*[gr:pat]}}$	−0.2141	[−0.4588, 0.01548]	−0.01022	0.3530	0.3536	−1.713	$8.677e-02$
$\text{corr}_{\text{age8, age8*[gr:pat]}}$	−0.2666	[−0.3945, 0.1250]	0.1443	0.3628	0.4419	−1.365	$1.722e-01$
σ^2_ε	1.530	[1.2860, 1.662]	−0.05381	0.09631	0.1103	13.870	$1.017e-43$
ρ	2.497	[1.841, 3.986]	0.3309	0.5463	0.6388	3.909	$9.252e-05$

int, gr, pat: respective shortcuts for intercept, group, patient.

In contrast, we signal that Gaussian ML fitting of the LMM (S:6.1) to the Blackmore data failed to converge with the default settings in the lmer function in the lme4 R package.

References

Anderson, E. and ten others (1999). *LAPACK Users' Guide*, *3rd Edition*, SIAM, Philadelphia, Available on-line at http://www.netlib.org/lapack/lug/lapack_lug.ht.

Demmel, J. W. (1997). *Applied Numerical Linear Algebra*, SIAM, Philadelphia.

Dongarra, J. J., Bunch, J. R., Moler, C. B. and Stewart, G. W. (1978). *LINPACK Users Guide*, SIAM, Philadelphia, PA, USA.

Graham, A. (1981). *Kronecker Products and Matrix Calculus*: *with Applications*, Ellis Horwood, Chichester, West Sussex, England.