Published in final edited form as:

J Stat Comput Simul. 2013 ; 83(12): . doi:10.1080/00949655.2012.690763.

Adaptive Fitting of Linear Mixed-Effects Models with Correlated Random-effects

Guangxiang Zhang and John J. Chen*

Biostatistics & Data Management Core, University of Hawaii John A. Burns School of Medicine, Honolulu, Hawaii 96813, USA

Abstract

Linear mixed-effects model has been widely used in longitudinal data analyses. In practice, the fitting algorithm can fail to converge due to boundary issues of the estimated random-effects covariance matrix G , i.e., being near-singular, non-positive definite, or both. Current available algorithms are not computationally optimal because the condition number of matrix G is unnecessarily increased when the random-effects correlation estimate is not zero. We propose an adaptive fitting (AF) algorithm using an optimal linear transformation of the random-effects design matrix. It is a data-driven adaptive procedure, aiming at reducing subsequent random-effects correlation estimates down to zero in the optimal transformed estimation space. Simulations show that AF significantly improves the convergent properties, especially under small sample size, relative large noise and high correlation settings. One real data for Insulin-like Growth Factor (IGF) protein is used to illustrate the application of this algorithm implemented with software package R (nlme).

Keywords

Linear mixed-effects; Convergence rate; Condition number; Collinearity between random-effects; Centering; Optimal linear transformation; Random slope

1. Introduction

Linear mixed-effects model (LMM) has been widely used for the analysis of hierarchical and longitudinal data [1]. By incorporating random-effects into an ordinal regression model, LMM accommodates correlations among multiple observations made on the same unit (e.g., subject, group, cluster, classroom, center) and allows for unbalanced designs where all subjects do not require an equal number of observations and/or the same data collection occasions. It is also known as multilevel model [2], random coefficient model [3] or hierarchical linear model [4] in different substantive fields.

The maximum likelihood estimation of LMM parameters can be implemented by various numerical optimization algorithms, such as expectation-maximization (EM) (e.g., [1, 5–8]), Newton-Raphson (e.g., [9–12]), iterative generalized least squares (IGLS) [2] or Fisher scoring [13], and is available from many standard or specialized softwares, see [14] for details. However, an important practical problem of applying mixed-effects models is the slow or non-convergence issue during the nonlinear iterative maximizing likelihood process. When random-effects are highly correlated, it can sometimes be difficult to achieve convergence using these available packages, or alternatively, an algorithm can converge to a

*Corresponding author. Tel: +1 808 692 1814; Fax: +1 808 692 1985; john.chen@hawaii.edu.

random-effects covariance matrix that is near-singular or non-positive definite (PD). Several papers reported that the non-convergence rate can vary from a few percents to more than 50% in simulation studies [8, 15–18], depending on algorithms used, simulation scale and settings, e.g., sample size, relative size of random-effects variance to the residual variance, and correlation level. It has also been shown that the estimated covariance matrix for random-effects could be non-PD [8, 14, 16, 19, 20] and the estimated correlation between random-effects could be close to unity [8, 21–23]. The non-convergence problem has been shown to get worse as the correlation level increased, either between random-effects [16] or between fixed-effects [18].

During LMM fitting, the traditional centering technique, which transforms a predictor covariate around its grand mean, has been extensively used in practice [e.g., 16, 22, 24–27]. It is even built in as a default or optional setting for some softwares [24], because of its ability to facilitate parameter interpretation and the possibility of improving numerical stability. For independent data, centering can remove the non-essential ill-conditioning in ordinary regressions, e.g., the sample covariance between intercept and slope being zero after centering (e.g., [28, 29]; [30, chap. 16]). However, for hierarchical or correlated data modeled by LMM, the computational consequences of centering on random-effects are complicated and not well understood. The reported correlation estimate between random intercept and random slope might be partly reduced after centering, but centering did not necessarily eliminate the collinearity problems between random-effects (e.g., [21, 24]). After centering, both non-convergence and non-PD could still occur [16, 21]. Therefore, grand mean centering is not a computationally optimal linear transformation for mixed-effects models in general.

For a general linear transformation, the invariant nature of LMM inferences before and after transformation has been discussed [3, 26]. Longford [3] also identified an optimal linear transformation with zero correlation between random-effects in the search for minimum variance for response variable. Little work has been done on this optimal transformation with respect to its improving fitting of mixed-effects models. To the best of our knowledge, this optimal transformation has not been linked to the non-convergence issue during numerical estimation process nor been used as a replacement for the traditional centering technique.

The primary goal of this article is to utilize the optimal linear transformation of the predictor variable to improve the numerical stability of LMM fitting. Focusing on the potential computational benefits from using this optimal transformation, we formulate an adaptive fitting (AF) algorithm, aiming to improve the non-convergence problems during the fitting of linear mixed-effects models with highly correlated random-effects.

Our article is organized as follows. We first review the LMM and describe its linear transformation in section 2. In section 3, the AF algorithm is proposed, and its correlation and condition number reduction properties are provided. The performance of AF is studied through simulations (section 4) and its application is shown using a real life example (section 5). Discussions about this AF procedure are provided in section 6.

2. LMM and linear transformation

2.1. LMM and its estimation

LMM [1, 21] can be specified as:

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i, i=1, \dots, m \quad (1)$$

$$b_i \sim N_q(0, G), \varepsilon_i \sim N(0, \sigma_e^2 R_i), b_i \perp \varepsilon_i,$$

where y_i is the observed response vector for i th subject, with scalar components $y_{ij}, j = 1, \dots, n_i$; $X_i(n_i \times p)$ and $Z_i(n_i \times q)$ are design matrices of known covariates; β are the $p \times 1$ fixed-effects coefficients modeling the population-average effects; b_i are the $q \times 1$ random-effects modeling subject-specific effects and are assumed to be normally distributed with mean zero and a general covariance matrix G ; ε_i are the $n_i \times 1$ unexplained errors and $R_i = I_{n_i}$ is usually assumed; b_i and ε_i are assumed to be independent.

If a LMM has two random-effects $b_i^T = (b_{0i}, b_{1i})$ with the corresponding $Z_i = (\mathbf{1}, x_i)$, where $\mathbf{1}$ is a constant vector of one and x_i is the observed slope predictor vector with scalar components x_{ij} , then it is a random intercept and slope (RIS) model. If the random slope b_{1i} term is dropped from a RIS model, then it becomes a random intercept only (RI) model.

Following the notations of [31], let $\theta \triangleq (\beta, \phi)$, where $\phi \triangleq (G, \sigma_e^2)$ is the vector of covariance parameters. The vector θ can be estimated by maximizing the log-likelihood function of (1),

$$l(\theta) = -\frac{1}{2} \sum_{i=1}^m \{ \log(|V_i|) + (y_i - X_i \beta)^T V_i^{-1} (y_i - X_i \beta) + n_i \log(2\pi) \}, \quad (2)$$

where V_i is the marginal covariance matrix of response variable,

$$V_i = \text{Cov}(y_i) = Z_i G Z_i^T + \sigma_e^2 I. \quad (3)$$

The MLE of fixed-effects parameters β can be obtained after solving the score function $\partial l(\theta) / \partial \beta = 0$,

$$\hat{\beta}(\hat{\phi}) = \left(\sum_i X_i^T V_i(\hat{\phi})^{-1} X_i \right)^{-1} \sum_i X_i^T V_i(\hat{\phi})^{-1} y_i, \quad (4)$$

which is a generalized least squares estimator. However, the MLEs of covariance parameters

do not have a general closed form solution and must be estimated iteratively by maximizing the log-likelihood function after plugging $\hat{\beta}(\hat{\phi})$ in (4) into (2). The asymptotic covariance matrix of the MLEs is estimated by the inverse of the Hessian matrix at the optimum $-\partial^2 l(\hat{\theta}) / \partial \theta \partial \theta^T$. The MLEs of β and ϕ are asymptotically independent given that $E[\partial l(\hat{\theta}) / \partial \phi] = 0$. The MLEs of covariance parameters ϕ are biased and can be corrected by the restricted likelihood (REML) method, adjusting for the loss of degree of freedom due to estimating the fixed-effects in β . The corresponding REML log-likelihood differs from the full likelihood (ML) function in (2) only by a constant term plus

$$-0.5 \sum_i \log(|X_i^T V_i^{-1} X_i|).$$

There are several iterative optimization algorithms for fitting LMM. EM algorithm is known for its slow rate of convergence, but it can be used to provide starting values for other algorithms (e.g., in R, Stata and HLM). The Newton-Raphson algorithm and its variations are the most commonly used algorithms to fit LMM, where an observed Hessian matrix is needed. Iterative generalized least squares (IGLS) algorithm and the Fisher scoring

algorithm are mathematically equivalent under normality assumption [32]. The Fisher scoring algorithm uses the expected Hessian matrix and can be considered as a modification of Newton-Raphson algorithm.

The main computational difficulty in applying LMM is the estimation of covariance matrix [14, pp. 30-31]. It is well known that the Newton-Raphson algorithm does not guarantee convergence or can converge to a non-PD covariance matrix (e.g., [11, 15, 19], [14, chap. 6]).

2.2. Linear transformation

To improve the numerical stability of LMM fitting, traditional centering technique has been extensively used in practice, with the expectation of a computational benefit similar to that in ordinary regressions without random-effects for independent data (e.g., [25], pp. 600; [26], pp. 339). However, for clustered or hierarchical data, centering becomes more complicated, and the computational benefits in ordinal regression may no longer exist. In the context of RI model, there exists an optimal linear transformation of the response variable which can convert the correlated random vector (intercept and error) into an uncorrelated one to facilitate testing for independent normality [33]. For RIS model, an orthogonal linear transformation for random-effects has been identified but not for numerical optimization purpose [3]. On the other hand, reparameterization of the design matrix $[X_i, Z_i]$ by an orthogonal-triangular form has also been shown to reduce the complexity of computation and improve the LMM fitting [11], although it is not a linear transformation method for the design matrix.

The general linear transformation for LMM has been discussed [3, 26]. Let the original design matrices X_i and Z_i be transformed to $X_i^* = Z_i A_1$ and $X_i^* = X_i A_2$, respectively. Namely,

$$\begin{aligned} y_i &= X_i \beta + Z_i b_i + \varepsilon_i \\ &= X_i A_1 A_1^{-1} \beta + Z_i A_2 A_2^{-1} b_i + \varepsilon_i \quad (5) \\ &= X_i^* \beta^* + Z_i^* b_i^* + \varepsilon_i, \end{aligned}$$

where $\beta^* = A_1^{-1} \beta$, $b_i^* = A_2^{-1} b_i$, with both $A_1 (p \times p)$ and $A_2 (q \times q)$ assumed invertible. Morrell et al. [26] showed that the likelihood function in (2) is invariant under linear transformations and that the REML function is only affected by a constant related to the determinant of A_1 if the scaling of fixed-effects exists ($|A_1|^2 = 1$). Thus REML is also invariant for a location shift linear transformation, i.e., matrix A_1 being a unit upper triangular matrix with all diagonal elements of one. Random-effects design matrix Z_i is usually a subset of its fixed-effects counterpart X_i since the mean of random-effects $E(b_i)$ is assumed zero [21]. For convenience and simplicity, we assume here the non-singular linear transformation matrices for X_i and Z_i in the equation (5) $A_1 = A_2 = A$. Thus,

$$G^* = \text{Cov}(b_i^*) = \text{Cov}(A^{-1} b_i) = A^{-1} \text{Cov}(b_i) (A^{-1})^T = A^{-1} G (A^{-1})^T. \quad (6)$$

3. The proposed adaptive fitting (AF) algorithm and its properties

For a RIS model with random intercept b_0 and random slope b_1 , let $Z_i = (\mathbf{1}, x_i)$. Given a non-singular linear transformation (A) of observed data Z_i , we have $Z_i^* = Z_i A$. The corresponding implicit change in random-effects is $b_i \Rightarrow b_i^* = A^{-1} b_i$.

3.1. Notations

$A_\delta: A = A_\delta \triangleq \begin{pmatrix} 1 & \delta \\ 0 & 1 \end{pmatrix}$, where A is a general location shift transformation matrix specified by a scalar location shift index of δ , corresponding to the transformation, $x_{ij} \Rightarrow x_{ij}^* = x_{ij} + \delta$.

$G(\rho): G = G(\rho) = \text{Cov}(b_i) \triangleq \begin{pmatrix} \sigma_{b_0}^2 & \rho\sigma_{b_0}\sigma_{b_1} \\ \rho\sigma_{b_0}\sigma_{b_1} & \sigma_{b_1}^2 \end{pmatrix}$, $|\rho| < 1$, where G is the random-effects covariance matrix in the original space, with the corresponding correlation between random intercept and random slope.

$G_\delta^*(\rho^*): G^* = G_\delta^*(\rho^*) = \text{Cov}(b_i^*) = \text{Cov}(A_\delta^{-1}b_i)$, where G^* is the random-effects covariance matrix in a transformed space ($Z_i \Rightarrow Z_i^* = Z_i A_\delta$), with the corresponding new correlation ρ^* . If $\rho = 0$, then $G^* = G$ and $\rho^* = 0$.

3.2. Optimal linear transformation

Applying a general location shift transformation to a RIS model, by equation (6), we have,

Lemma 3.1—For a general initial covariance matrix $G = \text{Cov}(b_i)$ and a general transformation matrix $A = A_\delta$, the new covariance matrix in the transformed space is,

$$G^* = \text{Cov}(b_i^*) = A^{-1}G(A^{-1})^T = \begin{pmatrix} \sigma_{b_0}^2 - 2\delta\rho\sigma_{b_0}\sigma_{b_1} + \delta^2\sigma_{b_1}^2 & \rho\sigma_{b_0}\sigma_{b_1} - \delta\sigma_{b_1}^2 \\ \rho\sigma_{b_0}\sigma_{b_1} - \delta\sigma_{b_1}^2 & \sigma_{b_1}^2 \end{pmatrix}.$$

We define the optimal shift d to be a location shift such that the resulting matrix G^* above becomes diagonal in the transformed space. To diagonalize G^* , set $\rho\sigma_{b_0}\sigma_{b_1} - \delta\sigma_{b_1}^2 = 0$. and we get $\delta = \rho\sigma_{b_0}/\sigma_{b_1}$. Thus, the optimal shift d is the ratio of the covariance between random intercept and random slope divided by the variance of random slope, i.e.,

$$d \triangleq \rho \frac{\sigma_{b_0}}{\sigma_{b_1}} = \frac{\text{Cov}(b_0, b_1)}{\text{Var}(b_1)}. \quad (7)$$

The optimal covariance matrix G^* can be obtained after replacing the general ρ in Lemma 3.1 with the optimal shift d . Thus,

Theorem 3.2—After a location shift by A , the new covariance matrix in the transformed space G^* will become diagonal (denoted as G_d^* or Ω), i.e.,

$$G^* = \text{Cov}(b_i^*) = G_\delta^*(\rho^* = 0) = \begin{pmatrix} \sigma_{b_0}^2(1 - \rho^2) & 0 \\ 0 & \sigma_{b_1}^2 \end{pmatrix} \triangleq \Omega,$$

if and only if $\delta = d$.

In other words, correlated random-effects will become uncorrelated after the optimal linear transformation $Z_i = (1, x_i) \Rightarrow Z_i^* = Z_i A = (1, x_i + d \cdot 1)$, where matrix $A = A_d$ is specified by the d in equation (7). For the longitudinal data settings, this is a location shift transformation changing the origin of time variable. For example, a longitudinal study with predictor time

variable x_{ij} for i -th subject at j -th occasion, transforming of the slope covariate $x_{ij} \Rightarrow x_{ij}^* = x_{ij} + d$ will lead to a zero correlation among random intercept and slope.

3.3. AF algorithm

Inspired by the equation (7), we propose an adaptive fitting (AF) algorithm to reduce the subsequent estimated correlation numerically closer to zero in the transformed estimation space, by transforming the slope covariate using the optimal location shift d adaptively. This process can be repeated and the convergence can be improved iteratively until the current location shift d estimate is down to zero.

For a RIS model, the steps of an AF algorithm can be described as:

1. Fit the model using the untransformed initial covariate x_{ij} . Let iteration index $k = 0$, and denote the initial $x_{ij} = x_{ij}^0$ and the initial location shift $d_0 = 0$.
2. Obtain random-effects covariance matrix estimate $\hat{G}$ from current fitting outputs. For non-convergent cases, results from the last iteration can be used. Denote the initial $\hat{G} = \hat{G}_0$.
3. Calculate current observed optimal location shift d_{k+1} from current covariance matrix estimate $\hat{G}_k$,

$$d_{k+1} = \frac{\widehat{\text{Cov}}(b_0, b_1)}{\widehat{\text{Var}}(b_1)} = \hat{\rho}_k \frac{\widehat{\sigma}_{b_0, k}}{\widehat{\sigma}_{b_1, k}} \quad (8)$$

4. If d_{k+1} is not zero, go to the next step; otherwise stop and go to Step 7.
5. Adaptively fit the model using updated transformed covariate $x_{ij}^* = x_{ij}^{k+1}$,

$$x_{ij}^{k+1} = x_{ij}^k + d_{k+1},$$

and obtain updated parameter estimates in the transformed space, such as $\hat{G}^*$ and denote as $\hat{G}_{k+1}$.

6. Update $k = k + 1$ and go to Step 3.
7. Obtain the final estimates for (β, φ) in the original space by taking the one-to-one inverse transformation. Based on the fitted $\hat{G}^*$ at the last iteration, the corresponding random-effects covariance matrix estimate $\tilde{G}$ in the original space can be obtained by

$$\tilde{G} = A \hat{G}^* A^T, \quad (9)$$

where A is specified by the sum of cumulative shifts $\left(\sum_{i=0}^{k+1} d_i\right)$ relative to the original space.

In Step 4 above, whether another AF step is needed depends on whether the current observed location shift d_{k+1} is close enough to zero. Empirically, the stopping rule with a magnitude at 0.001 provides reasonable good results. Step 7 may not be needed for those parameters or measures which are invariant with a location shift, e.g., the determinant of a estimated random-effects covariance matrix.

3.4. AF algorithm properties

We first illustrate the general impact of a location shift on the random-effects covariance matrix in a transformed space by Lemma 3.3, then establish Theorems 3.4–3.7 for the optimal shift d . Theorem 3.4 addresses what kind of location shift transformation will reduce the correlation level and why traditional centering approach could actually increase the correlation level. Theorems 3.5 to 3.7 show whether and how the condition number of random-effects covariance matrix is reduced in the optimal transformed space. Please refer to the Appendices for the relevant proofs.

Recall that in the original estimation space with observed data (Y, X) , we have covariance matrix G and random-effects $b_i = (b_0, b_1)^T$. After a non-singular transformation of the original data X by A , the model is refitted in a transformed space using pseudo-new observed data $(Y, X^* = XA)$, with new matrix G^* and transformed random-effects $b_i^* = A^{-1}b_i$.

3.4.1. Impact of a general location shift

Lemma 3.3: Assume a diagonal initial covariance matrix,

$$G = \text{Cov}(b_i) = G(\rho_0 = 0) = \begin{pmatrix} \sigma_{b_0}^2 & 0 \\ 0 & \sigma_{b_1}^2 \end{pmatrix} \text{ and a general } A = A. \text{ Then}$$

$$G^* = \text{Cov}(b_i^*) + A^{-1}G(A^{-1})^T = \begin{pmatrix} \sigma_{b_0}^2 + \delta^2 \sigma_{b_1}^2 & -\delta \sigma_{b_1}^2 \\ -\delta \sigma_{b_1}^2 & \sigma_{b_1}^2 \end{pmatrix}, \text{ and}$$

- i. $\forall \delta \neq 0 \Rightarrow \rho^* = \text{Cor}(b_0^*, b_1^*) \neq 0$;
- ii. $|\rho^*| \leq 1$, as $|\rho| \leq 1$; $|\rho^*|$ is a monotonically increasing function of $|\delta|$.

Lemma 3.3 can be illustrated by plotting the new correlation between random-effects in a transformed space as a function of location shift across various variance component ratios of the initial diagonal matrix G (Fig. 1). After a location shift perturbation, the new correlation will move away from the initial zero correlation. The correlation can approach the boundaries (± 1.0) after a large enough location shift for different variance component combinations.

Lemma 3.3 also shows that the new random intercept variance in the transformed space also strictly increases with the increase in the location shift level, if the initial $G = G(\rho_0 = 0)$. It is possible that for a large amount of δ , not only the new correlation is going to unity, but also the new intercept variance $\sigma_{b_0}^{*2} = \sigma_{b_0}^2 + \delta^2 \sigma_{b_1}^2$ is larger than the range of y_{ij} . This can cause numerical instability. If the origin of x_{ij} is too far away from the zero-correlation position (i.e., $|\delta| \gg 0$), then the numerical stability to estimate such a random-effects covariance matrix may be poor, as indicated by both the random-effects correlation and the magnitude of variance of random intercept. A good location shift should change the origin of x_{ij} such that it reduces the correlation between random-effects.

3.4.2. Correlation reduction—We define a neighborhood of an optimal shift d to be δ_τ , a set of location shifts around the optimal shift, i.e.,

$$\delta_\tau \in \{0 \leq |\delta - d| < |\delta|; \forall d \neq 0\}. \quad (10)$$

Theorem 3.4: There exists a neighborhood of the optimal shift such that a location shift within this neighborhood will reduce the initial correlation level in the transformed space,

i.e., given the initial ρ_0 ,

$$\delta \in \delta_\tau \iff |\rho_*| < |\rho|.$$

Specifically,

- i. if the initial $\rho_0 > 0$ (thus $d > 0$), then

$$\delta_\tau \in \{0 < \delta < 2d\} \iff |\rho_*| < |\rho_0|;$$

- ii. if the initial $\rho_0 < 0$ (thus $d < 0$), then

$$\delta_\tau \in \{2d < \delta < 0\} \iff |\rho_*| < |\rho_0|;$$

- iii. beyond the optimal shift neighborhood, new correlation becomes larger in absolute value than the initial value,

$$|\delta - d| > |d| \iff |\rho_*| > |\rho_0|.$$

Fig. 2 illustrates the neighborhood of the optimal shift in the case of the initial correlation $\rho_0 > 0$. Shifting to the right but still within the neighborhood of the optimal shift will ensure the reduction of correlation level in the transformed space. For the case with a negative initial correlation, by Lemma 3.3, the shifting direction will be to the left (figure not shown).

Note that the traditional centering will not work if the origin after centering transformation does not fall in the neighborhood of the optimal shift. For example, if the initial $\rho_0 > 0$, then $d > 0$. Centering which shifts the origin in the opposite direction will always increase $|\rho|$ in the transformed space.

3.4.3. Condition number improvement—A common measure of numerical stability and singularity of a matrix is the condition number (CN) [34–36]. The CN of a matrix M is defined as the square root of the ratio of the maximal eigenvalue to the minimal eigenvalue of MM^T ,

$$CN(M) = \sqrt{\frac{\lambda_{\max}(MM^T)}{\lambda_{\min}(MM^T)}}. \quad (11)$$

Larger condition number corresponds to less numerical stability. If the condition number is one $CN(M) = 1$, M is said to be perfectly conditioned. If a matrix M is near-singular, $CN(M)$ can be very large and M is ill-conditioned. When the condition number is huge and the inverse of the condition number of a matrix becomes comparable to computer round-off error, one can expect the computing quality will be poor for a nontrivial matrix operation on this matrix, such as inverse operation. The rates of convergence of many iterative algorithms are strongly influenced by the size of condition number [37]. For example, the condition number of the correlation matrix of the discrete Fourier transform vector has been directly linked to the rate of converge of algorithm [38].

Theorem 3.5: Let $G = \begin{pmatrix} \sigma_{b_0}^2 & \rho\sigma_{b_0}\sigma_{b_1} \\ \rho\sigma_{b_0}\sigma_{b_1} & \sigma_{b_1}^2 \end{pmatrix}$, where $0 < |\rho| < 1$, with eigenvalues $\lambda_1 > \lambda_2 > 0$. Assume $A = A_d$ thus $G^* = G_d^* = \Omega = \begin{pmatrix} \sigma_{b_0}^2(1-\rho^2) & 0 \\ 0 & \sigma_{b_1}^2 \end{pmatrix}$ by Theorem 3.2, with eigenvalues $\lambda(G^*) : \lambda_1^* \geq \lambda_2^* > 0$. Compared to the initial G , G^* in the transformed space has

- i. more clustered eigenvalues, $\lambda_2 < \lambda_2^* \leq \lambda_1^* < \lambda_1$;
- ii. larger smallest eigenvalue, $\min(G^*) > \min(G)$;
- iii. smaller condition number, $CN(G^*) < CN(G)$.

To measure the extent of CN reduction, we define the condition number ratio as

$$CNR = CNR[G] \triangleq \frac{CN(G)}{CN(G^*)}. \quad (12)$$

If the CN is reduced in the transformed space, the CNR will be greater than one.

Theorem 3.6: Larger CN reduction will be obtained for a setting with higher initial correlation, given the same initial variance component ratio, i.e., if $|\rho_1| > |\rho_2|$, then

$$CNR \left[\begin{pmatrix} \sigma_{b_0}^2 & \rho_1\sigma_{b_0}\sigma_{b_1} \\ \rho_1\sigma_{b_0}\sigma_{b_1} & \sigma_{b_1}^2 \end{pmatrix} \right] > CNR \left[\begin{pmatrix} \sigma_{b_0}^2 & \rho_2\sigma_{b_0}\sigma_{b_1} \\ \rho_2\sigma_{b_0}\sigma_{b_1} & \sigma_{b_1}^2 \end{pmatrix} \right].$$

Theorem 3.7: Let two settings have the same initial correlation but reversed variance component ratios. The setting with larger variance component ratio will have larger CN reduction after an optimal transformation, i.e., given $0 < \rho < 1$ and $\sigma = \sigma_{b_0}/\sigma_{b_1} > 1$, we have

$$CNR \left[\begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right] > CNR \left[\begin{pmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{pmatrix} \right].$$

Thus, a large reduction of CN can be expected under challenging data fitting situations, e.g., with extreme correlation and/or large variance component ratio. Such challenging scenarios also produce a wide neighborhood of the optimal shift, where the proposed AF algorithm is expected to perform well.

4. Simulations

4.1. Convergence performance measures

We used non-convergence rate to measure the numerical convergence performance of the AF algorithm. The non-convergence rate was defined as the observed proportion of non-convergent runs, i.e., the number of failed runs divided by the total number runs for a simulation setting. Although a statistical software package may not signal any error messages related to the convergence status, a nominally convergent run may produce a non-positive definite covariance matrix estimate. The non-PD here was defined for random-effects covariance matrix rather than for the covariance matrix of outcome variable [16, 20]. As long as the estimated covariance matrix G had a negative eigenvalue, it was a non-PD

run. Due to the skewed distribution of observed CNs, we used the geometric mean instead of arithmetic mean to measure the condition number estimate for a simulation setting.

4.2. Softwares

The simulated data sets were fitted by routine *lme(nlme)* [39] in the *R* environment using REML estimation method. By package default, the fitting function first ran 25 steps of EM iterations to provide initial values before entering Newton-Raphson iterations. The default convergence criterion is the change of parameter estimate. The default convergence tolerance ($1e-6$) was used. The intermediate fitting results were always available at any iteration step by setting option *returnObject = TRUE* in the *lmeControl* list of the *lme* routine, no matter whether the convergence was obtained or not, or whether the covariance matrix estimate is PD or not. To simplify the convergence diagnosis, two default options in *lmeControl* list, the maximum number of Newton-Raphson iterations (*msMaxIter*, default 50) and the maximum number of evaluations of the objective function permitted for nlminb (*msMaxEval*, default 200), were also both increased to 500. As a result, the relevant warning messages for these two options did not occur in our simulations. The same *lme* routine was used both before and after AF.

All simulations and analyses were conducted on a 2.13 GHz Intel Core (TM) 2 CPU 6400 processor on the Windows XP Professional (version 2002) platform.

4.3. Simulation settings

We fitted a RIS model to series of datasets generated from the following LMM.

$$y_{ij} = \beta_0 + x_{ij}\beta_1 + b_{0i} + x_{ij}b_{1i} + \varepsilon_{ij}, i=1, \dots, m, j=1, \dots, n, \quad (13)$$

where y_{ij} being scalar, the j th observations within i th group; fixed-effects $\beta_0 = -1$ and $\beta_1 = 1$; both predictor variable x_{ij} and residual ε_{ij} were sampled from standard normal distribution, $N(0,1)$.

Sample sizes ($N = m \times n$) varied from 125 to 2000, with number of groups $m \in (25, 50, 100)$ and group size $n \in (5, 10, 20)$, a balanced design. Each random effect had three levels, b_{0i} (intercept) or b_{1i} (slope) $\in (1, 1/2, 1/4)$. The correlation between random-effects covered five levels, $\rho \in (0.99, 0.95, 0.80, 0.30, 0.00)$, indicating extremely high, very high, high, moderate and zero correlation, respectively. Thus, there were a total of 405 scenarios ($9 \text{ sample sizes} \times 9 \text{ variance combinations} \times 5 \text{ correlations}$).

The number of replication runs per simulation design scenario was set at 1000. For each simulated dataset, the RIS model was fitted both in the original space without AF and in the transformed space with AF.

4.4. Simulation results

4.4.1. Non-convergence rate—The AF procedure is found to improve the non-convergence rate significantly (Table 1). Across 405 scenarios, the average non-convergence rate is as high as 17.23% before AF, but is reduced to zero level (0.0005%) in the transformed space after AF. Correspondingly, the percentage of scenarios with non-convergence issues is 80.0% and 0.5%, before and after AF, respectively. The number of non-convergent runs out of 405,000 simulated runs is dropped from 69,787 to 165 and 2, before AF, after single and two AF iteration steps, respectively. The following discussions about the performance of AF algorithm are all based on the fitting results after two AF steps.

Before AF, the average non-convergence rate increases from 3.88%, 4.62%, 13.47%, 27.14% to 37.05% as the population correlation level increases from $\rho = 0.00, 0.30, 0.80, 0.95$ to 0.99 , respectively. However, the population correlation level has little impact on the efficiency of AF or the number of AF steps needed. After the first AF step, there are 165 non-convergent runs, evenly distributed in 52 scenarios, with 12, 12, 11, 10, and 7 scenarios against the population ρ from $\rho = 0.00, 0.30, 0.80, 0.95$ to 0.99 , respectively. After the second AF step, there are two failed runs, one from zero and the other from moderate correlation (0.30) scenarios. Therefore, the strong impact of high correlation on non-convergence in the original space before AF disappears in the optimal transformed space. The AF algorithm is generally applicable and especially effective when the correlation is extremely high.

Small random slope variance ($\sigma_{b_1}^2$) also has a significant impact on the non-convergence rate before AF (Fig. 3A and 3B). When the random slope variance is reduced (Fig. 3A, from left panels to right panels), the non-convergence rate increases across all three levels of random intercept variances ($\sigma_{b_0}^2$). Controlling random slope at the same level (e.g., $b_1 = 0.25$), however, when the random intercept variance becomes larger (Fig. 3A, from top to bottom), the non-convergence rate was reduces slightly. Fig. 3B also confirms that small random slope variance has a larger impact on the non-convergence rate than small random intercept variance across various sample size settings. On the other hand, the two factors of sample size, the number of groups (m) and group size (n), seem to have comparable impact on the non-convergence rate.

The non-convergence issue can become more severe as the variances combination decreases, namely, relative noise increases. For the 45 scenarios with the highest level of variance component ($b_0 = b_1 = 1$) considered, only 25 scenarios have all 1000 convergent runs, with a mean (range) non-convergence rate of 5.06% (0.0%, 35.9%) across the 45 scenarios. On the other hand, for the 45 scenarios with the smallest level of variance component ($b_0 = b_1 = 0.25$) examined, only 2 scenarios have a zero non-convergence rate, with a mean (range) non-convergence rate of 32.56% (0.0%, 60.3%) across the 45 scenarios.

It is not surprising that the non-convergence rate can be improved by increasing sample size. For the 45 scenarios with the smallest number of observations ($m \times n = 25 \times 5 = 125$) considered, none of the scenarios has a zero non-convergence rate before AF, with a mean (range) non-convergence rate of 30.06% (1.0%, 57.7%) across the 45 scenarios. On the other hand, for the 45 scenarios with the maximum number of observations ($100 \times 20 = 2000$) considered, only 22 scenarios have a zero non-convergence rate before AF, with a mean (range) non-convergence rate of 0.81% (0.0%, 44.9%) across the 45 scenarios.

4.4.2. Non-positive definite—Besides the non-convergence rate, the utility of AF can also be illustrated by the improvement in the fraction of positive definite estimates of random-effects covariance matrix (Table 2). Across all 405 scenarios, there is only one run with the non-PD issue after AF. However, before AF, there are 48,278 (11.92%) runs with non-PD issues, which are distributed in 321 out of 405 scenarios. Importantly, the 11.92% non-PD runs is composed of 8.56% non-convergent and 3.36% “nominally convergent” runs. In other words, there are on average more than 33 “implicit” non-PD runs per scenario, which were considered as convergent runs without signaling any error message by the software package. If stratified by the convergence status, the conditional non-PD rate per run is much higher within failed runs than that within convergent runs (49.67% vs. 4.06%). Therefore, non-PD issue might still be a numerical problem even for a nominally convergent run before AF but not in the optimal transformed space.

4.4.3. Change in correlation—The above overall improvement in the convergence rate and positive definite property can be further understood by observing the change in the estimated random-effects correlation before and after AF (Table 2). On average, the near-zero correlation ($|\rho| < 0.10$) rate improves from 9.6% before AF to 85.6% after AF. More importantly extreme correlation ($|\rho| \approx 0.99$) almost disappears in the optimal transformed space. The average extreme correlation rate drops from 26.4% before AF down to 0.0042% after AF. Our simulations show that the correlation estimate has a 99.3% empirical rate to become smaller after AF.

The near-zero correlation after AF can also be demonstrated by looking at the per scenario average of observed random-effects correlations in the optimal transformed space. The observed average correlation per scenario scatters around zero and has a tight range of $(-0.104, 0.166)$ across all simulated 405 scenarios. For those 81 scenarios with zero non-convergence rate, the observed average correlation in the optimal transformed space are very close to zero, with a range of $(-0.00021, 0.00041)$.

Therefore, to improve the non-convergence rate, the estimated correlation after AF is not required to be exactly zero, which is the ideal theoretical value after the proposed optimal transformation. The crucial factor is to move the estimated correlation away from the extreme boundary region.

4.4.4. Change in condition number—Theorem 3.5 predicts that the condition number will become smaller after AF as long as the original population correlation is not zero. Our simulations show that the condition number estimate has a 91.3% empirical rate to become smaller after AF across all 405,000 simulated runs. The condition number ratio

$\left(CNR = \frac{CN(G)}{CN(G^*)} \right)$ can also measure the extent of CN reduction. Larger CNR means larger CN improvement after AF. The theoretical CNR has a geometric mean (range) of 2.83 (1.00, 288.40) for all 405 scenarios here. Our simulations show that the empirical CNR has a geometric mean (range) of 2.61 (0.63, 142.17) for all scenarios, where 389 scenarios (96.0%) have $CNR > 1$. Therefore, the numerical optimization in the optimal transformed space is significantly improved in both the intuitive correlation measure and the more essential condition number measure.

5. An Example: IGF data

Davidian and Giltinan [40] described a data set obtained during quality control radioimmunoassays for radioactive tracer used to calibrate the Insulin-like Growth Factor (IGF-I) protein concentration measurements. The data contained 237 measurements from 10 lots during 1 to 50 days, each lot with 4 to 39 measurements. The median and mean of collection time points were 22 and 22.45 days, respectively.

This publicly available IGF data is also included in the R package *nlme*. It failed to converge using a RIS model either in the original or mean-centered space, where the correlation estimates at last iteration were -1.000 and 1.000 , respectively [21]. Table 3 summarizes the RIS model fitting results for IGF data in the original and centered spaces against those obtained in the optimal transformed space after AF. For this specific data, centering could not alleviate the random-effects collinearity problem. The reversing sign of correlation could be an indication of an over-shifting after centering. Compared to the original space, the centering results was slightly worse in terms of fitted log-likelihood, minimal eigenvalue and condition number.

Among the three estimation spaces, the proposed transformed space produced best fitting diagnosis results. For instance, the convergence was achieved with positive definite covariance matrix estimate, and the condition number was reduced more than 10,000 times under the optimal transformed space. After AF, the estimated correlation was relatively small (-0.187), the AIC, BIC and number of Newton-Raphson iteration steps needed were also improved slightly.

Theoretically, it is invariant for the parameters of fixed effect slope β_1 random noise and random slope variance after a non-singular location shift linear transformation of RIS model, but the intercept-related parameters (e.g., random intercept variance and covariance) could change under different spaces. This theoretical expectation could be observed by comparing these parameter estimates at last iteration across three estimation spaces, although some of them were from non-convergent fittings.

We note that the random-effects variance estimate was relatively small compared to the random noise, and that the minimal eigenvalue of random-effects covariance matrix was negative but very close to zero in both the original and centering spaces. Thus, the non-PD issue in a non-optimal space could be due to larger round-off error during the nonlinear iteration process.

6. Discussions

Linear mixed-effects models have been widely used to model correlated data from many research fields. In practice it is important to alleviate non-convergence issues during the iterative optimization process. Non-PD may also be an issue and can be implicitly masked as a “nominally convergent” run by the default criteria of a software package. The proposed AF algorithm provides a straightforward technique to achieve these goals by reducing the collinearity between random-effects of LMM.

Simulations show that both the non-convergence rate and the non-PD rate are significantly reduced to zero level after AF, using the optimal linear transformation of the random slope variable. The AF algorithm is generally effective across various settings, including those challenging scenarios with very high correlation, relatively large noise or small random-effects variance, and small sample size.

The computational advantages of AF algorithm are demonstrated by several measures. The reduction of random-effects correlation down to zero provides an intuitive measure of the benefits of AF. The smallest eigenvalue of random-effects covariance matrix in the optimal transformed space is increased and less susceptible to boundary issue. The condition number of random-effects covariance matrix is also generally reduced in the transformed space. Larger reduction of condition number coincides with a setting with high non-convergence rate, e.g., extremely high correlation and near-zero random slope variance.

The core idea of AF algorithm

The core idea of our proposed AF algorithm is to utilize the existence of an optimal linear transformation where the resulting zero correlation is the farthest away from the parameter boundary, i.e., -1 or $+1$. Such a strategy is expected to further reduce the possibility for correlation estimate to converge near or out of the parameter boundary than that of explicitly direct reparameterization of model parameters, e.g., using some nonlinear transformations, such as log or logit, to force variance and correlation estimates “within” their domains (e.g., [41] for linear and non-linear mixed-effects models which had been implemented in the *lme(nlme)* routine; [42] for generalized linear mixed-effects model in disease mapping setting). Such direct reparameterization methods can still encounter the convergence on the

boundary issue if the estimated random-effects are highly correlated. For example, our simulations show that there is only a 79.41% empirical rate to converge with PD covariance matrix across all simulation scenarios.

The feasibility of AF algorithm

The AF approach is feasible for several reasons. First, conceptually, the existence of a neighborhood of the optimal shift d provides a working window for AF algorithm to reduce the random-effects correlation. We can expect that the AF algorithm has more tolerance to a non-accurate estimate of d if the underlying neighborhood is wider. In our simulations, non-convergent runs showed very high observed correlations between random-effects before AF even when the initial population correlation was zero. Thus both d and its neighborhood were determined mainly by the variance ratio. Second, by Lemma 3.3, when random-effects are extremely correlated, the random intercept variance is also expected to be large, thus very likely to be larger than the random slope variance. In real data analysis using a RIS model, smaller random slope variance relative to random intercept one was usually observed (e.g., [16, 22, 24, 27, 31, 43]). Overall, a non-convergent run before AF will be strongly associated with a nonzero optimal shift, a wide neighborhood of the shift, and a large CN reduction potential after AF. Third, note that a location shift matrix A used in AF, even not optimal or estimated with some error, will not introduce extra computational error during the transformation of observed data, since $\det(A) = \det(A^{-1}) = 1$ and $CN(A) = CN(A^{-1}) = 1$ for any A . Lastly, the AF algorithm can be straightforwardly implemented in existing LMM routines without modifying the internal optimization algorithm. The only needed coding effort in using AF is to extract the covariance matrix estimate from the algorithm outputs and calculate the optimal shift d . For a non-convergent run, the estimate of optimal shift d can be available from the last iteration output. Even if a highly accurate estimate of d from previous challenging fitting may be not realistic, an approximate estimate can be adequate as long as the subsequent correlation estimate becomes smaller and is not near the parameter boundary.

AF algorithm compared to centering

The proposed AF procedure differs from traditional grand mean centering technique in several important aspects. First, the random-effects correlation and the relative size of random-effects variances are taken into account by AF while centering only uses first-moment information. Centering is not a numerically optimal linear transformation for LMM. If the new origin of a slope variable after centering is not within the proposed optimal shift neighborhood, centering may even increase the correlation level between random-effects. Second, the optimal location shift $d = b_0/b_1$ is a unit-free scalar. It cannot be directly calculated from the observed data itself before estimating a random-effects covariance matrix. Third, AF can be iterative while centering is a one step approach. Lastly, centering is a one-direction location shift method and the new origin is always within the observed range of slope covariate. However, as pointed out by [3], the direction of a location shift can be either positive or negative, and even beyond the range of the observed raw data. In real longitudinal data analysis, the calculated value of d may be close to the observed mean of time variable, e.g., for a fetal growth data [22], or the two may not be close to each other as in the IGF data in section 5. In our simulations, it is possible for the new origin of slope covariate to go beyond the range of the data during numerical optimization process, especially when the random intercept variance is larger than the random slope variance in high correlation settings. Since the AF algorithm focuses on the numerical optimization process rather than the inference of model parameters, this is acceptable. If a run can converge both before and after AF, the parameter estimates in the original space will be almost identical.

The implications of AF algorithm for RIS

models The numerical optimization of a RIS model can be much more challenging than that of a RI model. When a RIS model fails to converge due to highly correlated random-effects, it is a common practice to remove random slope from the RIS model and fit a RI model. However, the simple compound symmetric covariance structure and the constant correlation assumption of RI model may be unrealistic or not adequate for actual data modelings [44]. RIS model is probably the most standard tool to study changes over time in longitudinal growth curve modeling, as it allows to model non-stationary covariance structure and to investigate the growth velocity variations across different subjects using subject-specific random slope estimates. Several studies have showed that the maximum likelihood inference on fixed-effects is more robust to misspecification of the covariance structure under a RIS model, compared to a RI model [31, 45, 46]. Thus, the excellent convergence property of AF algorithm is useful not only for simulation studies but also for real data analyses.

The general application of AF algorithm

The AF procedure can be considered as a general purpose and efficient algorithm for mixed-effects models with correlated random-effects. We note that the linear transformation used by the AF algorithm does not impose any distribution assumption on the response variable. Although Newton-Raphson algorithm is the main iterative optimization tool used in our simulations, EM algorithm is also found to be faster after AF based on our limited simulations (results not shown). We recommend that the AF procedure should be considered as a routine collinearity diagnostic and sensitivity analysis tool during LMM fittings, especially when there are reasons to suspect that convergence on the boundary issue is present.

Limitations

There are several potential limitations for the current study. First, whether a non-convergent run achieves convergence after AF can be influenced by the magnitude of the calculated optimal shift d . If the automatically estimated shift is at zero level, the AF procedure cannot further improve the optimization and may result in non-convergence. In this type of situation, one can introduce some extent of manual perturbation which will help achieve numerical convergence. In our simulations, there were only two runs with non-convergence issues after AF. In both cases, convergences were achieved after manually introducing small perturbations (e.g., less than one) once or twice. Second, the performance of AF algorithm and convergence results may not be the same across different software packages when the covariance matrix estimate is near-singular or non-PD. As long as a package can provide the intermediate estimates of random-effects covariance matrix, the shift d can be calculated from the output. In these cases, the proposed AF algorithm will still be applicable and expected to improve the LMM fitting. Lastly, the idea of AF algorithm still needs to be further evaluated for various challenging settings for mixed models, such as unbalanced data structures, multiple dimensions of random-effects, and non-normal outcome data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the editor and anonymous referee for their constructive comments. The research was partly supported by Awards U54RR026136, G12RR003061 and P20RR016467 from the National Center for Research Resources (NCRR), National Institutes of Health (NIH). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NCRR or NIH.

Appendix A

Proof of Lemma 3.3

Proof

The first result is obvious. We just need to prove the second result in the Lemma 3.3. By

$$\text{definition, } \rho^* = -\frac{\delta\sigma_{b_1}}{\sqrt{\sigma_{b_0}^2 + \delta^2\sigma_{b_1}^2}} = -\frac{\delta}{\sqrt{\sigma^2 + \delta^2}} = -\frac{\delta}{h} = f(\delta, h), \text{ where } \sigma = \frac{\sigma_{b_0}}{\sigma_{b_1}} \text{ and } h = \sqrt{\sigma^2 + \delta^2}.$$

Taking first order partial derivative, we have

$$\frac{\partial f}{\partial \delta} = -\frac{h - \delta h'}{h^2} = -\frac{\delta^2/h - h}{h^2} = -\frac{\delta^2 - h^2}{h^3} = -\frac{\sigma^2}{h^3} < 0.$$

Thus, $|\rho^*| = \frac{|\delta|}{\sqrt{\sigma^2 + \delta^2}}$ monotonically goes to unity, as $|\delta| \rightarrow \infty$ for a fixed ratio $\sigma = \sigma_{b_0}/\sigma_{b_1}$.

Appendix B

Proof of Theorem 3.4

Proof

$$\text{By Lemma 3.1, in general, } G^* = \begin{pmatrix} \sigma_{b_0}^2 - 2\delta\rho_{b_0}\sigma_{b_1} + \delta^2\sigma_{b_1}^2 & \rho\sigma_{b_0}\sigma_{b_1} - \delta\sigma_{b_1}^2 \\ \rho\sigma_{b_0}\sigma_{b_1} - \delta\sigma_{b_1}^2 & \sigma_{b_1}^2 \end{pmatrix}.$$

First, it is straightforward to show that doubling the optimal shift will change the sign of correlation, by replacing general δ in above G^* with the value of $d = \rho\sigma_{b_0}/\sigma_{b_1}$. That is,

$$\delta = 2d \iff G^* = G_{2d}^* = A_{2d}^{-1}G(A_{2d}^{-1})^T = \begin{pmatrix} \sigma_{b_0}^2 & -\rho\sigma_{b_0}\sigma_{b_1} \\ -\rho\sigma_{b_0}\sigma_{b_1} & \sigma_{b_1}^2 \end{pmatrix} \quad (\text{B1})$$

By reparameterizing, $A = A_\delta = \begin{pmatrix} 1 & \delta \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \delta - d \\ 0 & 1 \end{pmatrix} = A_d A_{\delta-d}$. Obviously both A_d and A_{-d} are invertible location shift transformation matrices.

$$\begin{aligned} G^* &= \text{Cov}(b_i^*) \\ &= A^{-1}G(A^{-1})^T \\ &= (A_d A_{\delta-d})^{-1}G(A_{\delta-d}^{-1}A_d^{-1})^T \\ &= A_{\delta-d}^{-1}A_d^{-1}G(A_d^{-1})^T(A_{\delta-d}^{-1})^T \\ &= A_{\delta-d}^{-1}[A_d^{-1}G(A_d^{-1})^T](A_{\delta-d}^{-1})^T \\ &= A_{\delta-d}^{-1}[\Omega](A_{\delta-d}^{-1})^T \quad (\text{by Theorem 3.2}) \end{aligned}$$

Applying Lemma 3.3 to G^* relative to the diagonal matrix Λ , we have,

- i. $|\rho^*|$ monotonically increases with $|\delta - d|$;
- ii. G^* with $\rho^* = 0$, as long as no shift from $\delta = d$, e.g., $|\delta - d| = 0$ and $\delta = d$;

- iii. G^* with $|*| = |0| = 0$, as long as current shift, $|\delta - d| = |d| = |\rho_0 \sigma_{b_0} / \sigma_{b_1}|$, thus, $= 0$ or $2d$.

Theorem 3.4 is proved after combining (i), (ii) and (iii).

Appendix C

Proof of Theorem 3.5

The proof for the first result (i) in Theorem 3.5 will be provided after Lemma C.3. The result (ii) is directly from (i), and (iii) can also be derived from (i), since $CN(G^*) = \lambda_1^* / \lambda_2^*$ and $CN(G) = \lambda_1 / \lambda_2$.

Lemma C.1

For a square symmetric matrix $M_{m \times m}$ relative to a scalar $\alpha \neq 0$, the condition number of M is invariant, i.e., $CN(\alpha M) = CN(M)$; eigenvalues of M are not invariant and $\lambda(\alpha M) = \alpha \lambda(M)$.

Proof— $M^T = M$ eigenvalue $\lambda(M)$ $(M^T) = (M)$ $(MM^T) = [\lambda(M)]^2$. By the definition of CN in (11), for a square symmetric matrix M ,

$$CN(M) = \sqrt{\frac{\lambda_{\max}(MM^T)}{\lambda_{\min}(MM^T)}} = \frac{|\lambda_{\max}(M)|}{|\lambda_{\min}(M)|}.$$

Then,

$$\begin{aligned} \lambda = \text{eigen}(M) &\iff \det(M - \lambda I_n) = 0 \\ &\iff \alpha^n \det(M - \lambda I_n) = 0, \forall \alpha \neq 0 \\ &\iff \det(\alpha M - \alpha \lambda I_n) = 0 \\ &\iff \text{eigen}(\alpha M) = \alpha \lambda \end{aligned}$$

Also,

$$CN(\alpha M) = \frac{|\lambda_{\max}(\alpha M)|}{|\lambda_{\min}(\alpha M)|} = \frac{|\alpha \lambda_{\max}(M)|}{|\alpha \lambda_{\min}(M)|} = CN(M).$$

Lemma C.2

For a positive definite matrix $G = \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}$, where $\sigma > 0$ and $0 < |\rho| < 1$, denote diagonal elements $G_{11} = \sigma^2$, $G_{22} = 1$, and eigenvalues $\lambda(G)$, $\lambda_1 > \lambda_2 > 0$. Then the two eigenvalues are less clustered than the two diagonal elements, i.e.,

$$\lambda_1 > G_{22} > \lambda_2 > 0, \text{ and } \lambda_1 > G_{11} > \lambda_2 > 0, \text{ for } \forall \sigma > 0, \forall \theta < |\rho| < 1.$$

Proof—The characteristic equation is $\det(G - \lambda I_2) = 0$, then

$$\lambda^2 - (\sigma^2 + 1)\lambda + \sigma^2(1 - \rho^2) = 0,$$

then,

$$\Delta = (\sigma^2 + 1)^2 - 4\sigma^2(1 - \rho^2) = (\sigma^2 - 1)^2 + 4\rho^2\sigma^2 \geq 0, \quad (C1)$$

Thus,

$$\lambda_1 = \frac{\sigma^2 + 1 + \sqrt{\Delta}}{2}, \lambda_2 = \frac{\sigma^2 + 1 - \sqrt{\Delta}}{2}. \quad (C2)$$

Note that

$$\lambda_1 = \lambda_2 \iff \Delta = 0 \iff \sigma = 1 \text{ and } \rho = 0 \iff G = I_2 \iff \lambda_1 = \lambda_2 = 1 = G_{22}.$$

It means that non-zero correlation will result in unequal eigenvalues and that neither of two eigenvalues will be equal to a diagonal element G_{22} ,

$$\rho \neq 0 \iff G_{22} \neq \lambda_1 \neq \lambda_2 \neq G_{22}.$$

Thus, Lemma C.2 is obviously true if $\sigma = 1$.

If $\sigma > 1$, then $\lambda_1 > 1 = G_{22}$, by (C2). If $\sigma < 1$,

$$\begin{aligned} \lambda_1 &= \frac{\sigma^2 + 1 + \sqrt{\Delta}}{2} > G_{22} = 1; \forall 0 < \sigma < 1 \\ &\iff \sqrt{\Delta} > 1 - \sigma^2 \\ &\iff \sqrt{\Delta} > 1 - \sigma^2 \geq 0 \\ &\iff \Delta > (1 - \sigma^2)^2 \\ &\iff (\sigma^2 - 1)^2 + 4\rho^2\sigma^2 > (1 - \sigma^2)^2 \\ &\iff 4\rho^2\sigma^2 > 0; \forall \rho \neq 0, \forall \sigma \neq 0. \end{aligned}$$

If $\sigma < 1$, then $\lambda_2 < 1 = G_{22}$, by (C2). If $\sigma > 1$,

$$\begin{aligned} \lambda_2 &= \frac{\sigma^2 + 1 - \sqrt{\Delta}}{2} < G_{22} = 1; \forall \sigma > 1 \\ &\iff \sqrt{\Delta} > \sigma^2 - 1 \\ &\iff \sqrt{\Delta} > \sigma^2 - 1 \geq 0 \\ &\iff \Delta > (\sigma^2 - 1)^2 \\ &\iff (\sigma^2 - 1)^2 + 4\rho^2\sigma^2 > (\sigma^2 - 1)^2 \\ &\iff 4\rho^2\sigma^2 > 0; \forall \rho \neq 0, \forall \sigma \neq 0. \end{aligned}$$

Similarly, we can prove $\lambda_1 > G_{11} > \lambda_2$.

Based on above Lemma C.1 and Lemma C.2, we have

Lemma C.3—For a positive definite matrix $G = \begin{pmatrix} \sigma_{b_0}^2 & \rho\sigma_{b_0}\sigma_{b_1} \\ \rho\sigma_{b_0}\sigma_{b_1} & \sigma_{b_1}^2 \end{pmatrix}$, where $\sigma_{b_0} > 0$, $\sigma_{b_1} > 0$ and $0 < |\rho| < 1$, denote two eigenvalues $(\lambda_1, \lambda_2) > 0$. Then the two eigenvalues are more separated than the two diagonal elements, i.e., for $0 < |\rho| < 1$,

$$\lambda_1 > \sigma_{b_1}^2 > \lambda_2 > 0, \text{ and } \lambda_1 > \sigma_{b_0}^2 > \lambda_2 > 0$$

$$\iff \lambda_1 > \max(\sigma_{b_0}^2, \sigma_{b_1}^2) \text{ \& } \lambda_2 < \min(\sigma_{b_0}^2, \sigma_{b_1}^2).$$

Proof: Theorem 3.5 (i)

Recall $\lambda_1^* = \max(\sigma_{b_0}^2(1 - \rho^2), \sigma_{b_1}^2)$ and $\lambda_2^* = \min(\sigma_{b_0}^2(1 - \rho^2), \sigma_{b_1}^2)$ for diagonal G^* . If $\sigma_{b_1}^2 \leq \sigma_{b_0}^2(1 - \rho^2)$, then

$$\begin{aligned} & \sigma_{b_1}^2 \leq \sigma_{b_0}^2(1 - \rho^2); 0 < |\rho| < 1 \\ \iff & \lambda_2^* = \sigma_{b_1}^2 \leq \sigma_{b_0}^2(1 - \rho^2) = \lambda_1^* \\ \iff & \sigma_{b_1}^2 = \lambda_2^* \leq \sigma_{b_0}^2(1 - \rho^2) < \sigma_{b_0}^2 \\ \Rightarrow & \lambda_2 < \sigma_{b_1}^2 = \lambda_2^* \leq \lambda_1^* < \sigma_{b_0}^2 < \lambda_1 \text{ (by Lemma C.3)} \\ \iff & \lambda_2 < \lambda_2^* \leq \lambda_1^* < \lambda_1. \end{aligned}$$

If $\sigma_{b_1}^2 > \sigma_{b_0}^2(1 - \rho^2)$, then

$$\begin{aligned} & \sigma_{b_0}^2(1 - \rho^2) < \sigma_{b_1}^2; 0 < |\rho| < 1 \\ \iff & \lambda_2^* = \sigma_{b_0}^2(1 - \rho^2) < \sigma_{b_1}^2 = \lambda_1^* \\ \iff & \lambda_2^* < \lambda_1^* = \sigma_{b_1}^2 \\ \Rightarrow & \lambda_2^* < \lambda_1^* = \sigma_{b_1}^2 < \lambda_1 \text{ (by Lemma C.3)} \\ \iff & \lambda_2^* < \lambda_1^* < \lambda_1 \text{ (i)} \\ \Rightarrow & \lambda_1^* < \lambda_1 \\ \iff & \lambda_1^* \lambda_2 < \lambda_1 \lambda_2 = \lambda_1^* \lambda_2^*, (\det(G) = \det(G^*)) \\ \iff & \lambda_2 < \lambda_2^* \text{ (ii)} \\ \Rightarrow & \lambda_2 < \lambda_2^* < \lambda_1^* < \lambda_1 \text{ (i \& ii)}. \end{aligned}$$

Appendix D

Proof of Theorem 3.6

Theorem 3.6 will be proved after the following Lemma D.1.

Lemma D.1

For a positive definite matrix $G = \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}$, the condition number ratio $CNR[G]$ is either the square of the largest eigenvalue of G , or the inverse of the square of the smallest eigenvalue of G , i.e.,

$$CNR[G] \triangleq \frac{CN(G)}{CN(G^*)} = \begin{cases} [\lambda_{\max}(G)]^2, & \text{if } \sigma^2 \leq \frac{1}{1-\rho^2} \\ \frac{1}{[\lambda_{\min}(G)]^2}, & \text{if } \sigma^2 > \frac{1}{1-\rho^2} \end{cases}$$

Proof—Denote the eigenvalues of the matrix G to be (λ_1, λ_2) : $\lambda_1 > \lambda_2 > 0$, and the eigenvalues

of the corresponding matrix $G^* = \begin{pmatrix} \sigma^2(1-\rho^2) & 0 \\ 0 & 1 \end{pmatrix}$ in the optimal transformed space to be $\lambda(G^*)$: $\lambda_1^* \geq \lambda_2^* > 0$. By definition,

$$\begin{cases} \lambda_1^* = 1, & \text{if } \sigma^2 \leq \frac{1}{1-\rho^2} \\ \lambda_2^* = 1, & \text{if } \sigma^2 > \frac{1}{1-\rho^2} \end{cases} \quad (D1)$$

By definition, $\det(G^*) = \det(G)$, i.e., $\lambda_1 \lambda_2 = \lambda_1^* \lambda_2^*$.

Therefore,

$$CNR[G] \triangleq \frac{CN(G)}{CN(G^*)} = \frac{\lambda_1/\lambda_2}{\lambda_1^*/\lambda_2^*} = \left(\frac{\lambda_1}{\lambda_1^*}\right)^2 = \left(\frac{\lambda_2^*}{\lambda_2}\right)^2 \quad (D2)$$

Thus, Lemma D.1 is proved after combining equation D2 with D1. Combining Lemma D.1 with Lemma C.1 and Lemma C.2, we also have $CNR[G] > 1$ for a general covariance matrix G .

Proof: Theorem 3.6

Given matrix $G = \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}$ with eigenvalues: $\lambda_1 > \lambda_2 > 0$, by Lemma C.2, we have

$$\lambda_1 = \frac{\sigma^2 + 1 + \sqrt{\Delta}}{2}, \lambda_2 = \frac{\sigma^2 + 1 - \sqrt{\Delta}}{2}, \text{ where } \Delta = (\sigma^2 - 1)^2 + 4\rho^2\sigma^2 \geq 0.$$

For a fixed ρ , if $|\sigma|$ increases, then λ_1 strictly increases, and therefore λ_2 strictly decreases while λ_2 strictly decreases. By Lemma D.1, $CNR[G]$ will increase with the change in either λ_1 or λ_2 . Thus, given $|\lambda_1| > |\lambda_2|$, we have

$$CNR \left[\begin{pmatrix} \sigma^2 & \rho_1\sigma \\ \rho_1\sigma & 1 \end{pmatrix} \right] > CNR \left[\begin{pmatrix} \sigma^2 & \rho_2\sigma \\ \rho_2\sigma & 1 \end{pmatrix} \right].$$

Together with Lemma C.1, Theorem 3.6 is thus proved.

Appendix E

Proof of Theorem 3.7

Proof: Theorem 3.7

Denote covariance matrix $G_\sigma = \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}$ with eigenvalues: $\lambda_1 > \lambda_2 > 0$. Denote covariance matrix $G_{1/\sigma} = \begin{pmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{pmatrix}$. Then $CN(G_{1/\sigma}) = CN(G) = \lambda_1/\lambda_2$ by Lemma C.2.

We also have the corresponding new matrix $G_{1/\sigma}^* = \begin{pmatrix} 1 - \rho^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$ in the optimal transformed space. Obviously, $CN(G_{1/\sigma}^*) = \sigma^2 / (1 - \rho^2)$, given $\sigma^2 > 1$. Thus

$$CNR[G_{1/\sigma}] \triangleq \frac{CN(G_{1/\sigma})}{CN(G_{1/\sigma}^*)} = \frac{\lambda_1/\lambda_2}{\lambda_1^* \lambda_2^*} = \frac{\lambda_1/\lambda_2}{\sigma^2 / (1 - \rho^2)} \quad (E1)$$

By Lemma D.1,

$$CNR[G_\sigma] = \begin{cases} [\lambda_1]^2, & \text{if } \sigma^2 \leq \frac{1}{1 - \rho^2} \\ \frac{1}{[\lambda_2]^2}, & \text{if } \sigma^2 > \frac{1}{1 - \rho^2} \end{cases} \quad (E2)$$

Taking the ratio and noting that $\lambda_1 \lambda_2 = \sigma^2(1 - \rho^2)$, we have

$$\frac{CNR[G_\sigma]}{CNR[G_{1/\sigma}]} = \begin{cases} \lambda_1 \lambda_2 \times \frac{\sigma^2}{1 - \rho^2} = \sigma^4 > 1, & \text{if } \sigma^2 \leq \frac{1}{1 - \rho^2} \\ \frac{1}{\lambda_1 \lambda_2} \times \frac{\sigma^2}{1 - \rho^2} = \frac{1}{(1 - \rho^2)^2} > 1, & \text{if } \sigma^2 > \frac{1}{1 - \rho^2} \end{cases} \quad (E3)$$

Thus $CNR[G] > CNR[G_{1/\sigma}]$ is always true, regardless of the relative size of σ^2 and $\frac{1}{1 - \rho^2}$.

References

1. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics*. 1982; 38:963–974. [PubMed: 7168798]
2. Goldstein H. Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*. 1986; 73:43–56.
3. Longford, NT. Random Coefficient Models. Oxford: Clarendon Press; 1993.
4. Raudenbush, SW.; Bryk, AS. Hierarchical Linear Models: Applications and Data Analysis Methods. 2nd. Sage Publications; 2002.
5. Dempster AP, Selwyn MR, Patel CM, Roth AJ. Statistical and computational aspects of mixed model analysis. *Applied Statistics*. 1984; 33:203–214.
6. Laird N, Lange N, Stram D. Maximum-likelihood computations with repeated measures - application of the EM-algorithm. *Journal of the American Statistical Association*. 1987; 82:97–105.
7. Liu C, Rubin DB. The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*. 1994; 81:633–648.

8. Meng XL, van Dyk D. Fast EM-type implementations for mixed effects models. *Journal of the Royal Statistical Society Series B-Statistical Methodology*. 1998; 60:559–578.
9. Jennrich R, Schluchter M. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*. 1986; 42:805–820. [PubMed: 3814725]
10. Thompson R, Meyer K. Estimation of variance components: What is missing in the EM algorithm. *Journal of Statistical Computation and Simulation*. 1986; 24:215–230.
11. Lindstrom MJ, Bates DM. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*. 1988; 83:1014–1022.
12. Callanan TP, Harville DA. Some new algorithms for computing restricted maximum likelihood estimates of variance components. *Journal of Statistical Computation and Simulation*. 1991; 38:239–259.
13. Longford NT. A fast scoring algorithm for maximum-likelihood-estimation in unbalanced mixed models with nested random effects. *Biometrika*. 1987; 74:817–827.
14. West, BT.; Welch, KB.; Galecki, AT. *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman and Hall/CRC; Boca Raton, FL: 2006.
15. van Dyk DA. Fitting mixed-effects models using efficient EM-type algorithms. *Journal of Computational and Graphical Statistics*. 2000; 9:78–98.
16. Browne WJ, Draper D. Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*. 2000; 15:391–420.
17. Berkhof J, Snijders T. Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics*. 2001; 26:133–152.
18. Shieh Y, Fouladi R. The effect of multicollinearity on multilevel modeling parameter estimates and standard errors. *Educational and Psychological Measurement*. 2003; 63:951–985.
19. Mikulich SK, Zerbe GO, Jones RH, Crowley TJ. Relating the classical covariance adjustment techniques of multivariate growth curve models to modern univariate mixed effects models. *Biometrics*. 1999; 55:957–964. [PubMed: 11315035]
20. Pryseley A, Tchonlafi C, Verbeke G, Molenberghs G. Estimating negative variance components from Gaussian and non-Gaussian data: A mixed models approach. *Computational Statistics & Data Analysis*. 2011; 55:1071–1085.
21. Pinheiro, JC.; Bates, DM. *Mixed-Effects Models in S and S-PLUS*. Springer; New York: 2000.
22. Gurrin L, Blake K, Evans S, Newnham J. Statistical measures of foetal growth using linear mixed models applied to the foetal origins hypothesis. *Statistics in Medicine*. 2001; 20:3391–3409. [PubMed: 11746325]
23. Solaro N, Ferrari PA. Robustness of parameter estimation procedures in multilevel models when random effects are MEP distributed. *Statistical Methods & Applications*. 2007; 16:51–67.
24. Kreft IGG, Leeuwde J, Aiken L. The effect of different forms of centering in hierarchical linear modeling. *Multivariate Behavioral Research*. 1995; 30:1–20.
25. van der Leeden R, Vrijburg K, Leeuwde J. A review of two different approaches for the analysis of growth data using longitudinal mixed linear models. *Computational Statistics & Data Analysis*. 1996; 21:583–605.
26. Morrell CH, Pearson JD, Brant LJ. Linear transformations of linear mixed-effects models. *American Statistician*. 1997; 51:338–343.
27. Zhang D, Davidian M. Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*. 2001; 57:795–802. [PubMed: 11550930]
28. Marquardt DW, Snee RD. Ridge regression in practice. *American Statistician*. 1975; 29:3–20.
29. Bradley RA, Srivastava SS. Correlation in polynomial regression. *American Statistician*. 1979; 33:11–14.
30. Draper, NR.; Smith, H. *Applied Regression Analysis*. 3rd. Wiley-Interscience; 1998.
31. Jacqmin-Gadda H, Sibillot S, Proust C, Molina JM, Thiebaut R. Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics & Data Analysis*. 2007; 51:5142–5154.
32. Goldstein, H. *Multilevel Statistical Models*. 3rd. Wiley; 2002.

33. Hwang YT, Wei PF. A novel method for testing normality in a mixed model of a nested classification. *Computational Statistics & Data Analysis*. 2006; 51:1163–1183.
34. Belsley DA, Oldford RW. The general problem of ill conditioning and its role in statistical analysis. *Computational Statistics & Data Analysis*. 1986; 4:103–120.
35. Sengupta D, Bhimasankaram P. On the roles of observations in collinearity in the linear model. *Journal of the American Statistical Association*. 1997; 92:1024–1032.
36. Trefethen, LN.; Bau, D. *Numerical Linear Algebra*. SIAM: Society for Industrial and Applied Mathematics; 1997.
37. Yuan KH, Chan W. Structural equation modeling with near singular covariance matrices. *Computational Statistics & Data Analysis*. 2008; 52:4842–4858.
38. Chen WW, Hurvich CM, Lu Y. On the correlation matrix of the discrete Fourier transform and the fast solution of large Toeplitz systems for long-memory time series. *Journal of the American Statistical Association*. 2006; 101:812–822.
39. Pinheiro JC, Bates DM, DebRoy S, Sarkar D. R Development Core Team. Chapter title. nlme: Linear and Nonlinear Mixed Effects Models R 2.9.2. 2009
40. Davidian, M.; Giltinan, DM. *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall; London: 1995.
41. Pinheiro JC, Bates DM. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*. 1996; 6:289–296.
42. Abellana R, Carrasco J, Jover L, Ascaso C. Improving the convergence rate in conditional autoregressive models. *Computational Statistics & Data Analysis*. 2006; 50:1153–1163.
43. Verbeke G, Lesaffre E. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*. 1997; 23:541–556.
44. Verbeke G, Lesaffre E, Brant LJ. The detection of residual serial correlation in linear mixed models. *Statistics in Medicine*. 1998; 17:1391–1402. [PubMed: 9682327]
45. Lange N, Laird NM. The effect of covariance structure on variance-estimation in balanced growth-curve models with random parameters. *Journal of the American Statistical Association*. 1989; 84:241–247.
46. Schielzeth H, Forstmeier W. Conclusions beyond support: overconfident estimates in mixed models. *Behavioral Ecology*. 2009; 20:416–420. [PubMed: 19461866]

Fig. 1. Random-effects correlation in the transformed space as a function of location shift (δ) from a setting with an initial correlation (ρ_0) of zero, for different variance component ratios

Fig. 2. Random-effects correlation level in the transformed space, showing reduction within the neighborhood of the optimal shift (d)

Fig. 3.

Fig. 3A The non-convergence rate before adaptive fitting (AF) as a function of population correlation level, with each of the nine panels corresponding to one of the nine variance component combinations, controlling the total number of observations for three sample size combinations to 500

Fig. 3B The non-convergence rate before adaptive fitting (AF) as a function of population correlation level, with each of the nine panels corresponding to one of the nine sample size combinations, controlling the product of two variance components for three variance component combinations to 0.25

Table 1
Summary of convergence improvements after adaptive fitting (AF)

Estimation Space	# of AF	# (%) of scenarios ^{**} with failed runs	Total # of failed runs ^{***}	non-convergence rate per scenario mean (range)
Untransformed	0	324 (80.0%)	69,787	17.23% (0.0%, 60.3%)
Transformed	1	52 (12.8%)	165	0.041% (0.0%, 1.3%)
Transformed	2	2 (0.5%)	2	0.00049% (0.0%, 0.1%)

* Total 405 scenarios were simulated and each scenario with 1000 replication runs

** Scenarios with at least one non-converged run

*** Out of all 405×1000 runs

Table 2
Properties of estimated random-effects covariance matrix in the original and optimal transformed spaces

Properties	Estimation space	
	Original	Optimal transformed
Non-positive definite (PD)		
I. Overall non-PD **		
# of scenarios involved ***	321	1
Total # runs involved ****	48,278	1
Rate per scenario		
range	(0.0%, 34.8%)	(0.0%, 0.1%)
mean	11.92%	0.00025%
II. “Nominally converged” but non-PD		
# of scenarios involved ***	303	0
Total # runs involved ****	13,614	0
Rate per scenario		
range	(0.0%, 19.8%)	0.0%
mean	3.36%	0.00%
Random-effects correlation		
III. Observed $ \rho \geq 0.99$		
# of scenarios involved ***	334	13
Total # runs involved ****	106,520	17
Rate per scenario		
range	(0.0%, 66.4%)	(0.0%, 0.2%)
mean	26.36%	0.0042%
IV. Observed $ \rho < 0.10$		
# of scenarios involved ***	240	405
Total # runs involved ****	38,840	346,522
Rate per scenario		
range	(0.0%, 64.2%)	(45.0%, 100.0%)
mean	9.59%	85.56%

* Observed values at last iteration

** A non-PD run can be either a non-converged run or a “nominally converged” run

*** Scenarios with at least one run associated with that specified event

**** Out of all 405×1000 runs

Table 3
LMM fitting results of IGF data in the original, centering and optimal transformed spaces

Fitting results	Estimation space		
	Original	Centering	Optimal transformed
I. Diagnosis information			
Convergence	No	No	Yes
Positive definite of G [*]	No	No	Yes
Random-effects correlation	-1.000	1.000	-0.187
Minimal eigenvalue of G	-1.590E-11	-2.838E-12	1.961E-09
Condition number of G	4.309E+08	3.487E+09	3.335E+04
Condition number ratio relative to the transformed space	1.290E+04	1.046E+05	1
Log-likelihood	-297.1831	-297.1832	-297.1831
AIC	606.3663	606.3664	606.3662
BIC	627.1238	627.1239	627.1237
# of iterations	75	64	22
II. Parameter estimates at last iteration Beta1(standard error)	-0.002535 (0.005045)	-0.002535 (0.005044)	-0.002534 (0.005043)
Var(b1)	6.5495E-05	6.5493E-05	6.5401E-05
Var(b0)	6.7855E-03	9.8314E-03	2.0322E-09
Cov(b0,b1)	-6.6575E-04	8.0243E-04	-6.8174E-08
Var(noise)	0.6734	0.6734	0.6734

* Estimated random-effects covariance matrix