Accurately Assessing Goodness-of-Fit of the Covariance Structure in the Linear Mixed Model

Lloyd J. Edwards[1*], Matthew J. Gurka[2], Byron C. Jaeger[1], Keith E. Muller[2]

[1]Department of Biostatistics, University of North Carolina at Chapel Hill,

Chapel Hill, North Carolina 27599-7420

[2]Department of Health Outcomes and Policy, University of Florida, Gainesville, FL, 32610-0177

[*]*email:* Lloyd_Edwards@unc.edu

SUMMARY. The linear mixed model (LMM) is the most widely used statistical tool for the analysis of longitudinal models of Gaussian outcomes. Accurately modeling the variability is critical for accurate tests of hypotheses about fixed effects. In practice, information criteria are used almost exclusively to assess goodness-of-fit when choosing a covariance structure for the LMM. However, the ability of information criteria to select the correct structure has not been definitively shown. Furthermore, awkward interpretation and poor comparability across data sets limits their effective use. We demonstrate that an R^2 statistic developed for fixed effects in LMMs is a viable alternative to information criteria in choosing an appropriate covariance structure. In simulations of conditions often used in practice, the R^2 statistic outperforms information criteria in selecting the correct covariance structure. A key difference is that the approach rarely underfits the covariance, which protects accuracy of tests and confidence intervals. The performance of the R^2 statistic in this setting and its ease of implementation and interpretation combine to make the R^2 statistic an ideal tool for LMM covariance structure selection when interest lies in valid inference about fixed effects.

KEY WORDS: Goodness of fit, Information Criterion, Longitudinal data, Model Selection, R-squared, Signal-to-noise

1. Introduction

In longitudinal data analysis, one must be concerned with modeling both the between-subject effects (fixed effects, the mean structure) and within-subject effects (random effects, the covariance structure). The linear mixed model (LMM) stands as one of the most widely used statistical tools for the analysis of longitudinal data for Gaussian outcomes. The LMM explicitly specifies not only the mean structure, but also the covariance structure. Hence three types of model comparisons can occur. 1) Models with different mean structures and the same covariance structure may be compared. Models with nested mean structures are the most common. 2) Models with the same mean structures and different covariance structures may be compared. Both nested and nonnested covariance structure comparisons are common. 3) Models with different mean structures and different covariance structures may be compared.

Our focus lies entirely on comparing the covariance structures of two models with the same mean structure. We propose and evaluate a new goodness-of-fit criterion for covariance structure selection in the LMM. The motivating example involves comparing distinct covariance structures for the full (saturated) mean structure. In addition, we compare distinct structures across all possible mean structures in the example, and we look at correct and incorrect mean structures in the simulations.

In longitudinal data analysis, modeling the variability can be immensely important for proper analyses because accurate tests of significance require proper choice of the covariance structure. Gurka et al. [1] proved and demonstrated that underfitting the covariance leads to inflated Type I error rates of tests on the fixed effects. Thus a great deal of attention must be given to covariance structure selection, even when primary interest lies on fixed-effects inference. In practice, information criteria such as the AIC [1] and BIC [2] presently dominate the selection of a covariance structure in the LMM. However, information criteria provide only rules of thumb to discriminate models. Consequently we agree with Verbeke and Molenberghs [4, Section 6.4] that information criteria should never be used or interpreted as formal statistical tests of

significance. In addition, the data-dependent scales of the AIC and BIC make comparisons across studies impossible.

Gomez et al. [5] studied the performance of the AIC and BIC in selecting the true covariance structure from a large set. The authors concluded that AIC and BIC are useful tools to help the researcher choose a covariance structure. However, because AIC and BIC do not always point to the correct covariance structure, it is not wise to not depend on them exclusively when choosing a covariance structure. Gomez et al. noted that it is important to be especially careful with small sample sizes because success rates in their simulation studies were very low and Type I error rates were inflated.

Gurka [6] observed that although many covariance structure selection criteria have been suggested, none has been found to be clearly superior. Other resources such as correlograms [7], knowledge about the design, and science should also be brought into play. Gurka et al. [1] concluded that if the data allow it, one should select an unstructured covariance. The authors also concluded that one principle seems clear: controlling type I error for tests of fixed effects demands avoiding an underfitted covariance structure. As illustrated in a US FDA guidance document [8], using an unstructured covariance structure is a strong preference in many food and drug settings, especially in the gold-standard context of randomized clinical trials.

Unfortunately, convergence often becomes an issue with mixed models using unstructured covariance, especially in small samples. Cheng et al. [9] provided practical advice on improving the chances for convergent models. The same authors recommended strategies for building a LMM with a focus on choosing a "good enough" mean structure and assumed the covariance structure remained the same through the process (the most common scenario). However, the authors also noted that the same model building strategies for choosing a "good enough" mean structure also apply to choosing an appropriate covariance structure in the LMM.

When comparing LMMs with the same fixed effects and different covariance structures, we propose using the signal-to-noise ratio (SNR) of the model, the ratio of explained to unexplained

variance, to compare the goodness-of-fit of covariance structures. We propose using the model R^2_β [10], which measures the multivariate association between the repeated outcomes and the fixed effects in the LMM, as a measure of explained variance. By comparing SNR between models, assessing goodness-of-fit of the covariance structures is implemented by comparing R^2_β between models. We prove that the statistic can be used for covariance structure selection. Simulations demonstrate that the statistic outperforms standard information criteria in choosing the correct covariance structure under conditions often used in practice. Perhaps even more importantly, we demonstrate that the statistic does not underfit the covariance structure, thus ensuring accurate inference on the fixed effects of the linear mixed model.

2. The Linear Mixed Model and R^2_β for Fixed Effects

With N independent sampling units (often *persons* in practice), the LMM for person i may be written

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i\,, \tag{1}$$

for $i \in \{1, \ldots, N\}$. Here, $\boldsymbol{y}_i$ is a $p_i \times 1$ vector of observations on person i, $\boldsymbol{X}_i$ is a $p_i \times q$ known, constant design matrix for person i, with full column rank q while $\boldsymbol{\beta}$ is a $q \times 1$ vector of unknown, constant, population parameters. Also $\boldsymbol{Z}_i$ is a $p_i \times m$ known, constant design matrix with rank m for person i corresponding to the $m \times 1$ vector of unknown random effects $\boldsymbol{b}_i$, while $\boldsymbol{e}_i$ is a $p_i \times 1$ vector of unknown random errors. The full rank assumptions simplifies the exposition of the new method but need not meaningfully affect practice. Chapters 11, 12, and 14 in Muller and Stewart [11] contain detailed discussions of estimability with less than full rank design codings, while 15, 16, 18 cover corresponding invariance properties of inference.

Gaussian $\boldsymbol{b}_i$ and $\boldsymbol{e}_i$ are independent with mean $\mathbf{0}$ and

$$\mathcal{V}\left(\begin{bmatrix}\boldsymbol{b}_i \\ \boldsymbol{e}_i\end{bmatrix}\right) = \begin{bmatrix}\boldsymbol{\Sigma}_{bi}(\boldsymbol{\tau}_b) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{ei}(\boldsymbol{\tau}_e)\end{bmatrix}. \tag{2}$$

Here $\boldsymbol{\mathcal{V}}(\,\cdot\,)$ is the covariance operator, while both $\boldsymbol{\Sigma}_{bi}(\boldsymbol{\tau}_b)$ and $\boldsymbol{\Sigma}_{ei}(\boldsymbol{\tau}_e)$ are positive-definite, symmetric covariance matrices. Therefore $\boldsymbol{\mathcal{V}}(\boldsymbol{y}_i)$ may be written $\boldsymbol{\Sigma}_i = \boldsymbol{Z}_i\boldsymbol{\Sigma}_{bi}(\boldsymbol{\tau}_b)\boldsymbol{Z}_i' + \boldsymbol{\Sigma}_{ei}(\boldsymbol{\tau}_e)$.

As a result, $\boldsymbol{y}_i \sim \mathcal{N}_{p_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$. We assume that $\boldsymbol{\Sigma}_i$ can be characterized by a finite set of parameters represented by an $r \times 1$ vector $\boldsymbol{\tau}$ which consists of the unique parameters in $\boldsymbol{\tau}_b$ and $\boldsymbol{\tau}_e$. Throughout $n = \sum_{i=1}^{N} p_i$.

We will also need to refer to a stacked data version of model (1) containing the data for all i:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{e}. \tag{3}$$

Throughout $\boldsymbol{y} = [\,\boldsymbol{y}_1' \cdots \boldsymbol{y}_N'\,]'$ $(n \times 1)$, $\boldsymbol{X} = [\,\boldsymbol{X}_1' \cdots \boldsymbol{X}_N'\,]'$ $(n \times q)$, $\boldsymbol{Z} = \text{diag}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_N)$ $(n \times Nm)$, $\boldsymbol{b} = [\,\boldsymbol{b}_1' \cdots \boldsymbol{b}_N'\,]'$ $(Nm \times 1)$, and $\boldsymbol{e} = [\,\boldsymbol{e}_1' \cdots \boldsymbol{e}_N'\,]'$ $(n \times 1)$. Here $\boldsymbol{b} \sim \mathcal{N}_{Nm}[\boldsymbol{0}, \boldsymbol{\Sigma}_{bi}(\boldsymbol{\tau}_b) \otimes \boldsymbol{I}_N]$ and $\boldsymbol{e} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{\Sigma}_e)$ for $\boldsymbol{\Sigma}_e = \text{diag}[\boldsymbol{\Sigma}_{e1}(\boldsymbol{\tau}_e), \cdots, \boldsymbol{\Sigma}_{eN}(\boldsymbol{\tau}_e)]$, where the operator '$\otimes$' is the Kronceker product between matrices. In turn $\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \boldsymbol{\mathcal{V}}(\boldsymbol{y}) = \text{diag}(\boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_N)$.

In the linear univariate model, R^2 corresponds to comparing two models. The same principle applies to comparing fixed effects in the LMM. The most common situation involves a model including an intercept and a hypothesis excluding the intercept, with $H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0}$ for $\boldsymbol{C} = [\,\boldsymbol{0}_{(q-1)\times 1}\boldsymbol{I}_{q-1}\,]$ of rank $q - 1$. The corresponding F statistic is given by

$$F(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}) = \frac{(\boldsymbol{C}\widehat{\boldsymbol{\beta}})'\big[\boldsymbol{C}\big(\boldsymbol{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\big)^{-1}\boldsymbol{C}'\big]^{-1}(\boldsymbol{C}\widehat{\boldsymbol{\beta}})}{\text{rank}(\boldsymbol{C})}\,. \tag{4}$$

Edwards et al. [10] defined R^2_β for the LMM as

$$R^2_\beta = \frac{(q-1)\nu^{-1}F(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}})}{1 + (q-1)\nu^{-1}F(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}})}\,. \tag{5}$$

The $F(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}})$ statistic used for R^2_β corresponds to a test of the null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{q-1} = 0$. Using the $F(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}})$ statistic allows computing R^2_β using a single model fit for the model of interest, rather than needing to fit a full model and a null model. Edwards et al [10] used the small sample Kenward-Roger F [12] to define R^2_β and hence the statistic is defined for REML estimation. However, R^2_β also is appropriate under maximum likelihood estimation. As defined, $R^2_\beta = 0$ if and only if $F(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}) = 0$. However, for

$0 < F(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}) < \infty$, R^2_β cannot equal 1, but can only near 1 as $F(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}})$ nears ∞. As a result, we have $0 \leq R^2_\beta < 1$.

Edwards et al. [10] proposed R^2_β as a statistic that measures multivariate association between the repeated outcomes and the fixed effects in the LMM. The R^2_β statistic arises as a 1–1 function of an appropriate F statistic for testing all fixed effects (except typically the intercept) in a full model. The statistic compares the full model to a null model with all fixed effects deleted (except typically the intercept) while retaining exactly the same covariance structure. Furthermore, R^2_β leads immediately to a natural definition of a partial R^2 statistic. Since its introduction, R^2_β has been gaining recognition by investigators as useful for their analyses [13-15]. Software is freely available for easily computing R^2_β, both SAS and R programs [16].

3. R^2_β Criterion for Covariance Goodness-of-Fit

We assume that we are comparing two LMMs with the same fixed effects but that have different covariance structures, either nested or non-nested. The different covariance structures may be due to using different random effects which gives rise to different covariance structures or due to assuming different covariance structures using the same random effects. For the case of two LMMs we have

$$\text{Model 1} \quad \boldsymbol{y}_i = \boldsymbol{X}_{1i}\boldsymbol{\beta} + \boldsymbol{Z}_{1i}\boldsymbol{b}_{1i} + \boldsymbol{e}_{1i} \tag{6}$$

$$\text{Model 2} \quad \boldsymbol{y}_i = \boldsymbol{X}_{1i}\boldsymbol{\beta} + \boldsymbol{Z}_{2i}\boldsymbol{b}_{1i} + \boldsymbol{e}_{2i}\,. \tag{7}$$

Under Model 1, we assume $\boldsymbol{y}_i \sim \mathcal{N}_{p_i}(\boldsymbol{X}_{1i}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{1i})$, and under Model 2, we assume $\boldsymbol{y}_i \sim \mathcal{N}_{p_i}(\boldsymbol{X}_{1i}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{2i})$, where $\boldsymbol{\Sigma}_{1i} \neq \boldsymbol{\Sigma}_{2i}$ for $i \in \{1, \ldots, N\}$. In many practical applications, we have $\boldsymbol{Z}_{1i}\boldsymbol{b}_{1i} = \boldsymbol{Z}_{2i}\boldsymbol{b}_{2i}$. The population-average model has $\boldsymbol{Z}_{1i} = \boldsymbol{Z}_{2i} = \boldsymbol{0}$.

For a LMM, we define the model signal-to-noise ratio (SNR) as

$$\text{SNR}_{LMM} = \frac{R^2_\beta(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}})}{1 - R^2_\beta(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}})}\ . \tag{8}$$

SNR_{LMM} is the ratio of the (standardized) explained variance, $R^2_\beta(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}})$, to (standardized) unexplained variance, $1 - R^2_\beta(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}})$, for the LMM. The SNR is a term popular in engineering

settings including amplifier design and acoustics, where it refers to a ratio of the power for signal to the power for noise. The ratio form makes the criterion scale-free and easily compared across situations. Since $0 \leq R^2_\beta < 1$, we have $0 \leq \text{SNR}_{LMM} < \infty$. As a performance measure of the goodness-of-fit criterion for regression, an objective is to maximize the SNR [17,18]. Bérubé and Wu [19] provided an overview of the performance of the SNR for a variety of models. They concluded that the performance of the SNR is model dependent and its validity deteriorates as the true model deviates from the assumed model. The SNR_{LMM} depends on estimated values of all of the parameters in the model as well as the form of the mean-variance relationship.

For the LMM, we consider the case where the signal is fixed, i.e., $\boldsymbol{X\beta}$ is chosen, and we must select a structure for the noise, $\boldsymbol{\Sigma}_s$. The LMM with the best predictive ability, a measure of goodness-of-fit, can be defined as the model that provides the highest signal, $\boldsymbol{X\beta}$, given the noise, $\boldsymbol{\Sigma}_s$, i.e., the LMM with the largest SNR_{LMM}. The worse the SNR_{LMM}, the less predictive ability the model has, based on the estimated mean and covariance structures. Box [17] noted that the SNR could be used as a performance criteria both in the analysis of dispersion as well as location. For determining the best-fitting covariance structure between LMMs with the same mean structures and different covariance structures, we have the following proposition.

Proposition. We consider a set of K LMMs with distinct covariance structures and the same mean structure. Here $k \in \{1, \ldots, K\}$ indexes the candidate covariance structures, $\{\boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_K\}$. Also, $R^2_\beta(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}_k)$ denotes R^2_β for Model k, with $\boldsymbol{\Sigma}_k$ the covariance under the stacked model representation in (3). For the sake of brevity, we drop $\widehat{\boldsymbol{\beta}}$ and use $R^2_\beta(\widehat{\boldsymbol{\Sigma}}_k)$ to denote R^2_β for Model k, giving the set $\{R^2_\beta(\widehat{\boldsymbol{\Sigma}}_1), \cdots, R^2_\beta(\widehat{\boldsymbol{\Sigma}}_K)\}$. Also $R^2_{\beta\text{max}} = \max_k \{R^2_\beta(\widehat{\boldsymbol{\Sigma}}_1), \cdots, R^2_\beta(\widehat{\boldsymbol{\Sigma}}_K)\}$. The covariance structure corresponding to $R^2_{\beta\text{max}}$ has the largest SNR_{LMM} and therefore has the best goodness-of-fit for the covariance structure for the data. *The Appendix contains a proof.*

When considering fixed effects in a LMM, the F statistic always compares two LMMs with different (nested, full and reduced) mean structures and the same covariance structure. The F

test is often used to select mean structures in the LMM. The test statistic F for fixed effects in the LMM is proportional to SNR_{LMM}, i.e., $F = \text{SNR}_{LMM} \cdot \nu/(q-1)$, for ν the denominator degrees of freedom chosen using the small sample Kenward-Roger F [12].

When comparing nested fixed effect structures in the linear univariate model, the F distribution quantiles are $1-1$ functions of quantiles for the beta distribution of R^2. Since quantiles for the F distribution were historically more accessible than for the beta distribution, p-values for the R^2 are computed using the model F statistic in the linear univariate model. The F test assesses goodness-of-fit for location (mean structure) for the assumed Gaussian distribution. Matuszewski [20] demonstrated that the F distribution for the model F statistic in the LMM gives the same results as the beta distribution for R^2_β when comparing nested fixed effects. However, for dispersion (covariance structure), the F statistic cannot be used to assess goodness-of-fit between covariance structures. Instead, R^2_β, which is a $1-1$ transformation of the F statistic (for a fixed sample size), can be used for assessing goodness-of-fit for covariance structures for the LMM based on using SNR_{LMM} as an objective function.

4. Simulation Study

4.1 Description of Simulations

Monte Carlo simulations using SAS Version 9.4 were performed to assess the performance of R^2_β, AIC, and BIC in selecting a valid covariance structure while fitting a linear mixed model using REML estimation. While there has been considerable discussion of the proper forms for the AIC and BIC [6], here the forms used are those by PROC MIXED in SAS Version 9.4.

A total of 10,000 replicates were simulated for three sample sizes, $N \in \{20, 50, 100\}$. The number of repeated measures per independent sampling unit was held constant ($p_i = 5$). In each simulation, data were generated from a population linear mixed model that consisted of the following fixed effects: an intercept, a dummy variable indicating membership in one of two groups (equally allocated), time values with five time points equally spaced across $[0, 1]$, and the interaction between the "group" and "time" predictors. The population mean structure contained

a linear trend over time that differed in intercept and slope between two groups, with $\boldsymbol{\beta} = \delta \cdot [2, 1, 1, 0.5]'$ corresponding to an intercept, contrast in intercepts, a slope, and the contrast in slopes between the groups. In each sample size setting, $\delta \in \{0, 1, 2\}$ values were used. The scaling parameter allows for mean values of R^2_β near 0, 0.5, and 0.75 (no, medium, large proportion of explained variance), which correspond to SNR_{LMM} near 0, 1, and 3, respectively. The population covariance structure included a random intercept ($\sigma^2_{b0} = 1$) and a random slope ($\sigma^2_{b1} = 1$) that were correlated ($\rho = 0.25$), as well as a within-unit error variance term, $\sigma^2_e = 1$.

For all possible nine combinations of N and δ, in each replication AIC, BIC, and R^2_β were calculated for 6 possible covariance structures, with structure 6 being the population structure (Table 1). Mean values of all three selection statistics were calculated, as well as the percentage of times that each statistic "chose" the candidate structures. The preferred model was indicated by the lowest AIC and BIC and by the largest R^2_β for a given replication. Given the desire not to underfit the covariance, the percentage of time that the statistics chose Structure 4 (unstructured) or Structure 6 (population) was also tabulated. Results were tabulated for two settings: a) the true fixed effects were included (intercept, group variable, time variable, and a group $\times$ time interaction; Table 2), and b) the group variable was incorrectly ignored (intercept and time variable; Table 3).

As noted in Section 1, convergence often becomes an issue with mixed models, especially in small samples. For $N = 20$, we had models that did not converge, which was not unexpected. As often done in practice, the models were discarded and we computed goodness-of-fit statistics for the set of models that did converge.

4.2 Simulation Results

Table 2 displays the distribution of covariance candidate structures selected out of 10,000 for each of the N and δ value combinations when the true mean structure was specified, with a focus on how often Structure 6 (and Structure 4 or Structure 6) were chosen. Two observations stand

out: 1) R^2_β is affected by the signal-to-noise ratio, with better performance in selecting the correct covariance structure as the SNR_{LMM} increases, whereas the AIC and BIC are not; and 2) the performance of the AIC and BIC are greatly affected by N, whereas the performance of R^2_β is not. No matter what value of N or δ, the BIC was the inferior covariance structure selection tool. A closer look at the models that were chosen when Structure 6 was not selected indicates an important finding: in a vast majority of instances, R^2_β chose Structure 4 (an unstructured covariance), while the AIC and BIC selected Structure 5 (an uncorrelated random intercept and slope). Thus, the R^2_β statistic rarely underfitted the covariance structure, while the AIC and BIC almost always selects a smaller (inadequate) covariance structure when choosing the incorrect model. The result is important when the focus is on fixed-effects inference and the goal is to not underfit the covariance model [1].

When the fixed-effects structure was underfitted (i.e., $\delta > 0$ and the group variable was omitted), covariance structure selection results were consistent for all three model selection statistics (Table 3). All conclusions made for Table 2 hold true in Table 3.

When the goal is accurate inference on the fixed effects, it is important not underfit the covariance structure [1]. In such settings, our simulations support using the R^2_β statistic for choosing the appropriate covariance structure, particularly with a small sample.

5. Example: Dental Study Data for Choosing Covariance Structure

In Edwards et al. [10], we used a well-known example from Potthoff and Roy (1964) to demonstrate results for R^2_β in assessing the multivariate association between the repeated dental outcomes and the fixed effects used in the LMM. The data come from an orthodontic study with 27 children, 16 boys and 11 girls. For each child, the distance (mm) from the center of the pituitary to the pterygomaxillary fissure was measured at ages 8, 10, 12, and 14 years. The objectives were to determine whether, on the average over time, distances are larger for boys than for girls and whether, on the average over time, the rate of change of the distance is similar for boys and girls.

Edwards et al. [10] fitted LMMs with three different fixed-effect structures (in addition to a fixed intercept) using REML estimation: (I) a model with continuous age effect only; (II) a model with linear Age and Gender effect; (III) a model with linear Age, Gender, and their interaction. We also considered three different covariance structures: (1) random intercepts, $\boldsymbol{\Sigma}_{bi}(\boldsymbol{\tau}_b) = \sigma_b^2$ (scalar), and $\boldsymbol{\Sigma}_{ei}(\boldsymbol{\tau}_e) = \sigma^2 \boldsymbol{I}_{p_i}$; (2) random intercepts and slopes with diagonal covariance, $\boldsymbol{\Sigma}_{bi}(\boldsymbol{\tau}_b)$ (2×2), and $\boldsymbol{\Sigma}_{ei}(\boldsymbol{\tau}_e) = \sigma^2 \boldsymbol{I}_{p_i}$and (3) random intercepts and slopes with unstructured covariance, $\boldsymbol{\Sigma}_{bi}(\boldsymbol{\tau}_b)$ (2×2), and $\boldsymbol{\Sigma}_{ei}(\boldsymbol{\tau}_e) = \sigma^2 \boldsymbol{I}_{p_i}$.

Here we use the proposed covariance goodness-of-fit technique to choose the best fitting covariance structure for the data when the LMMs have the same fixed effects. Table 4 gives the results for R^2_β, AIC, and BIC (smaller is better for AIC and BIC) for comparing the three covariance structures for each of the three fixed-effects structures. We found that R^2_β gave different results than AIC and BIC when comparing the three structures, which are commonly considered in practice. Using R^2_β leads to choosing covariance structure 3 (unstructured) over covariance structures 1 or 2 for each of the 3 fixed-effects structures. In contrast, under fixed-effects structure (I), both AIC and BIC would choose covariance structure 2 (diagonal) and under fixed-effects structures (II) and (III), both AIC and BIC would choose covariance structure 1 (random interecept). Given the small sample size for the study, the simulation results give support that AIC and BIC may underfit the covariance structure while R^2_β may rarely do so. Thus the simulation results support selecting covariance structure 3 over covariance structures 1 and 2 based on R^2_β, rather than AIC or BIC.

6. Conclusions

The linear mixed model is a useful tool for modeling correlated data. With longitudinal data the correlations among observations may follow a complex structure. If so, it is important to properly model the covariance even when primary interest lies in fixed effect inference. It has been shown that underfitting the covariance structure can lead to bias with respect to fixed effect inference. The most commonly used tools in covariance structure selection are information

criteria such as the AIC and BIC. However, the AIC and BIC do not always point to the correct covariance structure. Furthermore, we show here that they are at risk of identifying a model that underfits the true covariance structure, which in turn can lead to biased fixed effect inference. In addition, there are no good guidelines for interpretation and comparison of information criteria values across varying models and data sets. Hence it would not be wise to depend on them when choosing a covariance structure.

The research developed here has possible extensions that we are actively pursuing. We are working to develop distributional theory for the difference between R^2_β in order to conduct statistical tests on the difference in goodness-of-fit between multiple covariance models, thereby introducing a probabilistic method of covariance model selection in the LMM. Also, though focus was on comparing the covariance structures of two models with the same mean structure, the simulation results suggests that with further development, R^2_β may be used to compare models with different mean structures and different covariance structures.

The R^2 statistic is a commonly used tool in linear model selection and interpretation that has broad practical appeal. Edwards et al. [10] described a generalization of the statistic that can be used for fixed effects in LMMs, and continues to gain in popularity. Here we demonstrate that the statistic can also be used in selecting the best covariance structure, and more importantly, outperforms standard information criteria with respect to not underfitting the covariance structure. The performance, accessibility and interpretation of the new approach makes it an excellent tool in LMM covariance structure selection. Using the new statistic will in turn ensure valid inference when comparing means.

Appendix

Proposition. We consider a set of K LMMs with distinct covariance structures and the same mean structure. Here $k \in \{1, \ldots, K\}$ indexes the candidate covariance structures, $\{\boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_K\}$. Also, $R^2_\beta(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}_k)$ denotes R^2_β for Model k, with $\boldsymbol{\Sigma}_k$ the covariance under the stacked model representation in (3). For the sake of brevity, we drop $\widehat{\boldsymbol{\beta}}$ and use $R^2_\beta(\widehat{\boldsymbol{\Sigma}}_k)$ to

denote R^2_β for Model k, giving the set $\{R^2_\beta(\widehat{\boldsymbol{\Sigma}}_1), \cdots, R^2_\beta(\widehat{\boldsymbol{\Sigma}}_K)\}$. Also $R^2_{\beta\max} = \max_k\{R^2_\beta(\widehat{\boldsymbol{\Sigma}}_1), \cdots, R^2_\beta(\widehat{\boldsymbol{\Sigma}}_K)\}$. The covariance structure corresponding to $R^2_{\beta\max}$ has the largest SNR_{LMM} and therefore has the best goodness-of-fit for the covariance structure for the data.

Proof. Without loss of generality, we consider the case of comparing two LMMs with the same fixed effects and different covariance structures (6). If Model 1 has a greater SNR than Model 2, then

$$\frac{R^2_\beta(\widehat{\boldsymbol{\Sigma}}_2)}{1 - R^2_\beta(\widehat{\boldsymbol{\Sigma}}_2)} < \frac{R^2_\beta(\widehat{\boldsymbol{\Sigma}}_1)}{1 - R^2_\beta(\widehat{\boldsymbol{\Sigma}}_1)} .$$

From (8) and assuming $0 \leq R^2_\beta(\widehat{\boldsymbol{\Sigma}}_1) < 1$ and $0 \leq R^2_\beta(\widehat{\boldsymbol{\Sigma}}_2) < 1$, we can deduce the following:

$$\begin{aligned}
&R^2_\beta(\widehat{\boldsymbol{\Sigma}}_2)\big[1 - R^2_\beta(\widehat{\boldsymbol{\Sigma}}_1)\big] < R^2_\beta(\widehat{\boldsymbol{\Sigma}}_1)\big[1 - R^2_\beta(\widehat{\boldsymbol{\Sigma}}_2)\big] \\
&R^2(\widehat{\boldsymbol{\Sigma}}_2) - R^2_\beta(\widehat{\boldsymbol{\Sigma}}_2)R^2_\beta(\widehat{\boldsymbol{\Sigma}}_1) < R^2_\beta(\widehat{\boldsymbol{\Sigma}}_1) - R^2_\beta(\widehat{\boldsymbol{\Sigma}}_1)R^2_\beta(\widehat{\boldsymbol{\Sigma}}_2) \\
&R^2_\beta(\widehat{\boldsymbol{\Sigma}}_2) - R^2_\beta(\widehat{\boldsymbol{\Sigma}}_2)R^2_\beta(\widehat{\boldsymbol{\Sigma}}_1) < R^2_\beta(\widehat{\boldsymbol{\Sigma}}_1) - R^2_\beta(\widehat{\boldsymbol{\Sigma}}_1)R^2_\beta(\widehat{\boldsymbol{\Sigma}}_2) .
\end{aligned}$$

It follows that (8) is true if and only if

$$R^2_\beta(\widehat{\boldsymbol{\Sigma}}_2) < R^2_\beta(\widehat{\boldsymbol{\Sigma}}_1) .$$

If $R^2_\beta(\widehat{\boldsymbol{\Sigma}}_1) = 0$ in (8), then $R^2_\beta(\widehat{\boldsymbol{\Sigma}}_2) = R^2_\beta(\widehat{\boldsymbol{\Sigma}}_1) = 0$ and $\text{SNR}_{LMM} = 0$ for Models 1 and 2. □

Acknowledgements

Edwards' work was supported in part by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant Award Number UL1TR001111. Gurka's work supported in part by NIH/NHLBI R01-HL120960 and NIH/NIGMS U54-GM104942. Muller's work supported in part by NIH/NIDCR R01-DE020832, NIH/NIDCR U54-DE019261, NIH/NCRR/UL1 TR000064, NIH/1R25GM111901-01, PCORI/HPC-1503-27891 and PCORI/National Patient-Centered Clinical Research Network (Shenkman-Hogan, PI).

References

1. Gurka MJ, Edwards LJ, Muller KE (2011). Avoiding bias in mixed model inference for fixed effects. *Statistics in Medicine*, **30**, 2696-707.
2. Akaike H (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, **AC-19**, 716-723.
3. Schwarz G (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.
4. Verbeke G, Molenberghs G (2000). *Linear Mixed Models For Longitudinal Data*. New York: Springer-Verlag.
5. Gómez VE, Schaalje GB, Fellingham GW (2005). Performance of the Kenward-Roger method when the covariance structure is selected using AIC and BIC. *Communications in Statistics-Simulation and Computation*, **34**, 377–392.
6. Gurka MJ (2006). Selecting the best linear mixed model under REML. *The American Statistician*, **60**, 20-26.
7. Littell RC, Pendergast J, Natarajan R (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*, **19**, 1793–1819.
8. Laughren TP (2007). Recommendation for approvable actions for zyprexa pediatric supplements for bipolar disorder (acute mania) and schizophrenia, to File NDA 20-592 (S-040 [bipolar] and S-041. Available at *www.fda.gov/downloads/drugs/developmentapprovalprocess/developmentresources/ucm195881.pdf*.
9. Cheng J, Edwards LJ, Maldonado-Molina MM, Komro KA, Muller KE (2010). Real longitudinal data analysis for real people: Building a good enough mixed model. *Statistics in Medicine*, **29,** 504-520.
10. Edwards LJ, Muller KE, Wolfinger RD, Qaqish BF, Schabenberger O (2008). An R^2 statistic for fixed effects in the linear mixed model. *Statistics in Medicine*, **27**, 6137-57.

11. Muller KE, Stewart PW. *Linear Model Theory: Univariate, Multivariate, and Mixed Models*. John Wiley and Sons, Inc: Hoboken, New Jersey, 2006.

1. Kenward MG, Roger JH (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, **53**, 983-997.

13. O'Reilly D, Navakatikyan MA, Filip M, Greene D, Van Marter LJ (2012). Peak-to-peak amplitude in neonatal brain monitoring of premature infants. *Clinical Neurophysiology*, **123**, 2139–2153.

14. Guglielminotti J, Mentré F, Gaillard J, Ghalayini M, Montravers P, Longrois D (2013). Assessment of pain during labor with pupillometry: a prospective observational study. *Anesthesia and Analgesia*, **116**, 1057–1062.

15. Tzeng YC, MacRae BA, Ainslie PN, Chan GSH (2014). Fundamental relationships between blood pressure and cerebral blood flow in humans. *Journal of Applied Physiology*, **117**, 1037-1048.

16. Ballarini N, Jaeger BC (2015). Software in SAS and R programming languages to calculate model R2 and semi-partial R2 for fixed effects in the linear and generalized linear mixed model. *https://github.com/bcjaeger/R2FixedEffectsGLMM/*, accessed July 26, 2016.

17. Taguchi G (1986). *Introduction to Quality Engineering: Designing Quality Into Products and Processes*. White Plains, NY: Kraus International Publications.

18. Box GEP (1988). Signal-to-noise ratios, performance criteria, and transformations (with discussion). *Technometrics*, **30**, 1-40.

19. Bérubé J, Wu CFJ (2000). Signal-to-Noise ratio and related measures in parameter design optimization: an overview. *Sankhya: The Indian Journal of Statistics, Series* B, **62**, 417-432.2.

20. Matuszewski JM (2012). *Properties of an R^2 Statistic for Fixed Effects in the Linear Mixed Model for Longitudinal Data.* Unpublished doctoral dissertation, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

21. Potthoff RF, Roy SN (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313-326.

Table 1. Simulation study covariance structures

Candidate Model	Covariance Structure
1	Independent observations, equal variance
2	Random intercept only (compound symmetry)
3	No random effects, first-order autoregressive error term
4	Unstructured covariance
5	Random intercept and slope (independent); $\boldsymbol{\Sigma}_{ei}(\boldsymbol{\tau}_e) = \sigma^2 \boldsymbol{I}_{p_i}$
6 (true)	Random intercept and slope (correlated); $\boldsymbol{\Sigma}_{ei}(\boldsymbol{\tau}_e) = \sigma^2 \boldsymbol{I}_{p_i}$

Table 2. Simulation Results: Selection of Candidate Covariance Structures when the True Mean Structure (Group, Time, Group x Time Interaction) was Specified
10,000 Simulated Data Sets Per Sample Size/Fixed Effect Scale Combination

		Fixed Effects Scale Parameter = 0						Fixed Effects Scale Parameter = 1						Fixed Effects Scale Parameter = 2					
		% Structure Chosen			Mean Value			% Structure Chosen			Mean Value			% Structures Chosen			Mean Value		
	Fitted Structure (S)	AIC	BIC	R^2	AIC	BIC	R^2	AIC	BIC	R^2	AIC	BIC	R^2	AIC	BIC	R^2	AIC	BIC	R^2
m = 20[#]	S1: Ind	0	0	0	372	373	0.05	0	0	0	372	373	0.22	0	0	0	372	373	0.50
	S2: CS	22.3	35.0	0	335	337	0.07	22.3	35.0	0.0	335	337	0.32	22.3	34.9	0	335	337	0.63
	S3: AR(1)	8.2	6.4	5.2	335	338	0.09	8.2	6.4	2.4	335	338	0.40	8.2	6.4	1.5	335	338	0.71
	S4: UN	1.3	0	42.6	345	361	0.14	1.3	0	31.3	345	361	0.46	1.3	0.0	26.2	345	361	0.74
	S5: Ran Int + Slope (Ind)	48.5	46.8	3.1	332	335	0.10	48.5	46.8	2.5	332	335	0.40	48.5	46.8	2.2	332	335	0.71
	S6: Ran Int + Slope (Corr)*	19.7	11.8	49.0	333	336	0.12	19.7	11.8	63.8	333	336	0.47	19.7	11.8	70.1	333	336	0.76
	S4+S6*	***21.0***	***11.8***	***91.6***				***21.0***	***11.8***	***95.1***				***21.0***	***11.8***	***96.3***			
m = 50	S1: Ind	0	0	0	947	949	0.02	0	0	0	947	949	0.20	0	0	0	947	949	0.49
	S2: CS	3.6	11.3	0	846	850	0.03	3.6	11.3	0	846	850	0.29	3.6	11.3	0	846	850	0.61
	S3: AR(1)	2.5	1.9	3.2	844	850	0.04	2.4	1.9	0.2	844	850	0.37	2.5	1.9	0.0	844	850	0.69
	S4: UN	1.1	0	43.1	849	880	0.05	1.1	0	24.8	849	880	0.42	1.1	0	18.0	849	880	0.73
	S5: Ran Int + Slope (Ind)	66.3	75.2	0.4	836	842	0.04	66.3	75.2	0.0	836	842	0.37	66.3	75.2	0.3	836	842	0.69
	S6: Ran Int + Slope (Corr)*	26.5	11.6	53.2	837	844	0.05	26.5	11.6	74.7	837	844	0.44	26.5	11.6	81.6	837	844	0.75
	S4+S6*	***27.6***	***11.6***	***96.3***				***27.6***	***11.6***	***99.5***				***27.6***	***11.6***	***99.6***			
m = 100	S1: Ind	0	0	0	1900	1903	0.01	0	0	0	1900	1903	0.19	0	0	0	1900	1903	0.48
	S2: CS	0.2	1.3	0	1694	1700	0.01	0.2	1.3	0	1694	1700	0.28	0.2	1.3	0	1694	1700	0.60
	S3: AR(1)	0.4	0.4	2.9	1690	1698	0.02	0.4	0.3	0	1690	1698	0.36	0.4	0.4	0	1690	1698	0.68
	S4: UN	1.1	0	44.4	1686	1728	0.03	1.0	0	21.8	1686	1728	0.42	1.0	0	14.1	1686	1728	0.73
	S5: Ran Int + Slope (Ind)	61.5	84.1	0.1	1674	1682	0.02	61.5	84.1	0.0	1674	1682	0.36	61.5	84.1	0.0	1674	1682	0.69
	S6: Ran Int + Slope (Corr)*	36.9	14.2	52.6	1674	1684	0.03	36.9	14.2	78.2	1674	1684	0.43	36.9	14.2	85.9	1674	1684	0.75
	S4+S6*	***38.0***	***14.2***	***97.0***				***38.0***	***14.2***	***100***				***38.0***	***14.2***	***100***			

* Structure 6 (Random Intercept and Random Slope, correlation = 0.25) was the true covariance structure

** The percentage of simulated models where the selected covariance was <u>not</u> underfit relative to the true covariance (Structure 6: True Covariance, Structure 4: Unstructured Covariance), [#] Results based on models that converged. See Section 4.1 for further details.

Table 3. Simulation Results: Selection of Candidate Covariance Structures when the Incorrect Mean Structure (no Group Variable) was Specified
10,000 Simulated Data Sets Per Sample Size/Fixed Effect Scale Combination

		Fixed Effects Scale Parameter = 0						Fixed Effects Scale Parameter = 1						Fixed Effects Scale Parameter = 2					
		% Structure Chosen			Mean Value			% Structure Chosen			Mean Value			% Structure Chosen			Mean Value		
	Fitted Structure (S)	AIC	BIC	R^2	AIC	BIC	R^2	AIC	BIC	R^2	AIC	BIC	R^2	AIC	BIC	R^2	AIC	BIC	R^2
m = 20[#]	S1: Ind	0	0	0	378	379	0.01	0	0	0	392	393	0.07	0	0	0	425	426.4	0.17
	S2: CS	21.2	33.2	0	339	341	0.02	18.3	29.2	0	344	346	0.18	12.3	50.5	0	355	326.8	0.45
	S3: AR(1)	7.8	6.3	12.8	339	342	0.04	7.0	5.5	7.4	344	347	0.32	5.0	4.1	8.8	354	357.2	0.64
	S4: UN	1.3	0	40.0	349	365	0.05	1.3	0	27.6	353	369	0.36	1.3	0.1	21.2	362	377.7	0.66
	S5: Ran Int + Slope (Ind)	50.1	48.5	3.6	336	339	0.04	49.1	50.1	3.0	340	343	0.32	42.3	47.9	2.9	350	352.9	0.64
	S6: Ran Int + Slope (Corr)*	19.5	12.1	43.6	336	340	0.04	24.5	15.3	61.9	341	344	0.35	39.1	24.5	67.2	350	353.4	0.67
	S4+S6*	***20.8***	***12.1***	***83.6***				***25.8***	***15.3***	***89.5***				***40.4***	***24.6***	***88.4***			
m = 50	S1: Ind	0	0	0	950	952	0.00	0	0	0	986	988	0.06	0	0	0	1069	1071	0.16
	S2: CS	3.3	10.9	0	848	852	0.01	2.1	8.3	0	860	864	0.18	0.8	3.3	0	888	891	0.45
	S3: AR(1)	2.6	2.0	10.4	847	852	0.02	1.9	1.6	2.3	858	864	0.32	0.7	0.8	3.9	885	890	0.64
	S4: UN	1.1	0	42.1	851	882	0.02	1.1	0	24.5	862	893	0.36	1.2	0	17.2	884	915	0.67
	S5: Ran Int + Slope (Ind)	66.3	75.2	0.6	838	844	0.02	58.3	72.2	0.3	849	855	0.32	31.9	53.6	0.3	874	879	0.65
	S6: Ran Int + Slope (Corr)*	26.7	11.9	46.9	839	846	0.02	36.5	18.0	73.0	849	857	0.37	65.4	42.3	78.6	872	879	0.69
	S4+S6*	***27.8***	***11.9***	***89.0***				***37.6***	***18.0***	***97.5***				***66.6***	***42.3***	***95.8***			
m = 100	S1: Ind	0	0	0	1902	1905	0.00	0	0	0	1973	1976	0.06	0	0	0	2140	2143	0.16
	S2: CS	0.2	1.3	0	1695	1700	0.00	0.1	0.6	0	1719	1725	0.18	0	0	0.1	1774	1779	0.45
	S3: AR(1)	0.4	0.4	10.3	1691	1699	0.01	0.2	0.2	0.4	1715	1722	0.32	0	0	1.0	1767	1775	0.64
	S4: UN	1.0	0	43.4	1687	1729	0.01	1.2	0	22.7	1707	1750	0.36	1.4	0	16.1	1753	1794	0.68
	S5: Ran Int + Slope (Ind)	61.4	84.1	0.1	1674	1682	0.01	45.0	73.7	0	1696	1704	0.32	11.4	34.1	0	1745	1752	0.64
	S6: Ran Int + Slope (Corr)*	37.0	14.3	46.2	1674	1685	0.01	53.5	25.6	76.9	1695	1705	0.37	87.2	65.8	82.9	1740	1750	0.69
	S4+S6*	***38.0***	***14.3***	***89.6***				***54.7***	***25.6***	***99.6***				***88.6***	***65.8***	***99.0***			

* Structure 6 (Random Intercept and Random Slope, correlation = 0.25) was the true covariance structure

** The percentage of simulated models where the selected covariance was not underfit relative to the true covariance (Structure 6: True Covariance, Structure 4: Unstructured Covariance), [#] Results based on models that converged. See Section 4.1 for further details.

Table 4. R^2_β, AIC, BIC Results for Choosing Covariance for Dental Study Data

		GOF Statistic		
Fixed Effects Model[#]	Covariance Model[*]	R^2_β	AIC	BIC
I	1	0.59	451	454
	2	0.61	449	453
	3	0.77	451	456
II	1	0.71	442	444
	2	0.71	443	447
	3	0.73	443	448
III	1	0.67	438	440
	2	0.73	439	443
	3	0.80	441	446

[#]I $\equiv$ Intercept, Age
II $\equiv$ Intercept, Age, Gender
III $\equiv$ Intercept, Age, Gender, Gender x Age

[*]1 $\equiv$ Random intercept only and $\mathbf{\Sigma}_{ei} = \sigma^2 \boldsymbol{I}_{p_i}$ (compound symmetry)
2 $\equiv$ Random intercept and slope with diagonal random effects covariance and $\mathbf{\Sigma}_{ei} = \sigma^2 \boldsymbol{I}_{p_i}$
3 $\equiv$ Random intercept and slope with unstructured random effects covariance and $\mathbf{\Sigma}_{ei} = \sigma^2 \boldsymbol{I}_{p_i}$