Automated Search Algorithm for Optimal Generalized Linear Mixed Models (GLMMs)

Miryeong Koo[1] · **Jinming Zhang**[1]
[1]**University of Illinois at Urbana-Champaign**

Abstract

Only a limited number of predictors can be included in a generalized linear mixed model (GLMM) due to estimation algorithm divergence. This study aims to propose a machine learning based algorithm (e.g., mixed-effects random forest) that can consider all predictors without the convergence issue and automatically searches for optimal GLMMs.

1 Introduction

Educational data typically have a hierarchical structure (Bryk & Raudenbush, 1989; Woltman et al., 2012) due to their sampling scheme. For example, schools in a nation are first selected, and then students in the schools are sampled. As a result, students are nested in schools, schools are nested in nations. Consequently, students from the same school tend to be correlated among themselves (school effect or random effect). A generalized linear mixed model (GLMM) is typically used in such a data set.

GLMMs estimate both fixed and random effects, leading to accommodating school effects. However, GLMMs frequently fail to converge when they need to consider many predictors and their interactions (Bates et al., 2015). To solve the convergence issue, previous research typically considered only a small subset of predictors based on literature review or applying regularization techniques such as Lasso (Tibshirani, 1996). Nevertheless, both approaches have limitations. The former case may exclude some predictors that have a large influence on the outcome, while the latter does not account for random effects.

To address the issues, this study aims to develop an algorithm that can automatically rank all predictors of statistical importance based on the whole dataset without convergence problems and then search for optimal GLMMs according to the ranking. The algorithm applies a machine learning method, called mixed-effects random forest (MERF; Hajjem et al., 2014), to rank all predictors according to their statistical importance in a mixed-effects model. Then, the algorithm searches for a random intercept model with significant predictors sequentially based on the ranking provided by MERF. Next, it searches for significant interaction terms. Lastly, random slopes are explored and possibly added to the models.

Although the proposed algorithm does not directly account for substantive meaning of each predictor, it does provide candidate GLMMs recommended. Thus, the proposed algorithm has the potential to reduce the time and effort otherwise required by researchers to identify optimal GLMMs.

2 Theoretical framework

2.1 Generalized Linear Mixed Models (GLMMs)

The generalized linear model (GLM) (McCullagh & Nelder, 1989) assumes that all observations are independent, while the generalized linear mixed model (GLMM) (Breslow & Clayton, 1993) allows dependency among subjects in the same group. GLMM can handle data hierarchy by including random effects for group dependency (e.g., school effects. The linear mixed model (LMM), also known as hierarchical linear modeling (Raudenbush & Bryk, 2002) or multilevel modeling (Goldstein, 2011), is a special case of the GLMM, where the response variable is continuous.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 352–358

The LMM estimates both fixed effects and random effects and has the form as shown in Equation (1),

$$Y = X\beta + Zb + \varepsilon, \quad (1)$$

where $\varepsilon \sim N(0, \sigma^2 I)$, $b \sim N(0, G)$, and ε and b are independent from each other. Here G is a block-diagonal covariance matrix.

The first two terms of the right-hand side of Equation (1) represent fixed- and random-effects parts, respectively. X is the $n \times (K+1)$ matrix of fixed effects from predictors (here n is the number of observations and K is the number of predictors), β is the $(K+1)$ dimensional fixed effect parameter vector, Z is the design matrix of J groups (schools), b is random effect vector, and ε is the (level-1) residual vector. The vector b for a random intercept model is J×1 vector of random intercept (b_{0j}) for each group, where $b_{0j} \sim N(0, \tau^2)$. For a random intercept and a random slope model, vector b is a J×2 vector of random effects for each group, where $b_j = \begin{pmatrix} b_{0j} \\ b_{1j} \end{pmatrix} \sim N(0, G)$, and Z is an n×$2J$ block-diagonal matrix, for each subject, including 1 for their group's random intercept (b_{0j}) and Z for their group's random slope (b_{1j}). It can be represented as a conditional model, as illustrated in Equation (2),

$$\mu = E(Y|b) = X\beta + Zb. \quad (2)$$

The generalized linear mixed model (GLMM) extends LMM to accommodate non-continuous responses, such as binary or categorical responses and the models are denoted as Equation (3),

$$g(\mu) = X\beta + Zb, \quad (3)$$

where $g()$ is a monotonic increasing and differentiable link function. For example, the logit function widely serves as a link function for binary responses.

The correlation between individuals (students) within the same group (school) is called the intraclass correlation coefficient (ICC) which measures the similarity of within-group individuals (Raudenbush & Bryk, 2002). In a random intercept model, the ICC is calculated by between-group variance (τ^2) divided by total variance, which is the sum of between-group variance (τ^2) and within-group variance (σ^2). Since the ICC is the proportion of total variance due to group differences, higher ICC implies larger group difference.

2.2 Machine learning method: MERF

A machine learning method, RF is a tree-based ensemble method that aggregates a cluster of random decision trees. Unlike standard RF (Breiman, 2001) that considers fixed effects only, MERF (Hajjem et al., 2014) is also capable of taking random effects into account, as shown in Equation (4),

$$Y = f(X) + Zb + \varepsilon, \quad (4)$$

where $f(X)$ is a general and unspecified fixed-effects part. MERF is applied as follows: After calculating the fixed part for the predictors with initial value for $\hat{\beta}_j$, $\hat{\sigma}$, and $\hat{G}$, the algorithm takes bootstrap samples from the training set to build a forest of trees. The predicted fixed part for observation i in group j, $\hat{f}(X_{ij})$, is obtained with the training set of trees in the forest. Next, it computes $\hat{b}_j$ with the updated estimate of the random part of Equation (4) and updates the variance components $\hat{\sigma}$ and $\hat{G}$. The algorithm keeps repeating those steps until convergence. See Hajjem et al. (2014) for detailed explanation.

RF-based methods rank all predictors by their importance in prediction. Specifically, the importance function of LongituRF package (Capitaine, 2020) in R prints two measures of variable importance: the mean decrease of prediction accuracy when a given variable is permutated (permutation-based importance) and the total decrease in node impurity that results from splits over that variable, averaged over all trees (node impurity-based importance) (James et al., 2013). The permutation-based importance criterion is applied to rank predictors to avoid overfitting.

In the proposed algorithm, the ranking of all predictors is utilized to support predictor selection, which serves as a basis of optimal model selection. By sequentially adding a top-ranked predictor to the provisionary model, the algorithm performs predictor selection. Then, based on the predictors selected, the optimal model is finally identified in the last step. Since educational large-scale assessment (LSA) data typically have hierarchical structures in common, MERF is applied for supporting predictor selection from LSA data. The detailed explanation of each step is followed in the next section.

3 Automated search algorithm for optimal GLMM

In this study, we developed an algorithm to automatically search for optimal GLMMs for any large data set with hierarchical structures, especially LSA data. It mimics how experienced researchers identify the best-fitting GLMM. This algorithm utilizes forward selection (Hastie et al., 2017), which begins with a model containing no predictors, and then adds predictors to the model, one at a time, until complex model with the newly added predictor is not significantly different from simpler model. Unlike traditional forward selection that adds the predictor that gives the greatest additional improvement to the fit to the model, we select the predictor based on the ranking of their importance in prediction sorted by MERF. The proposed algorithm contains three main steps, preparation, predictor selection, and model selection, as demonstrated in Figure 1.

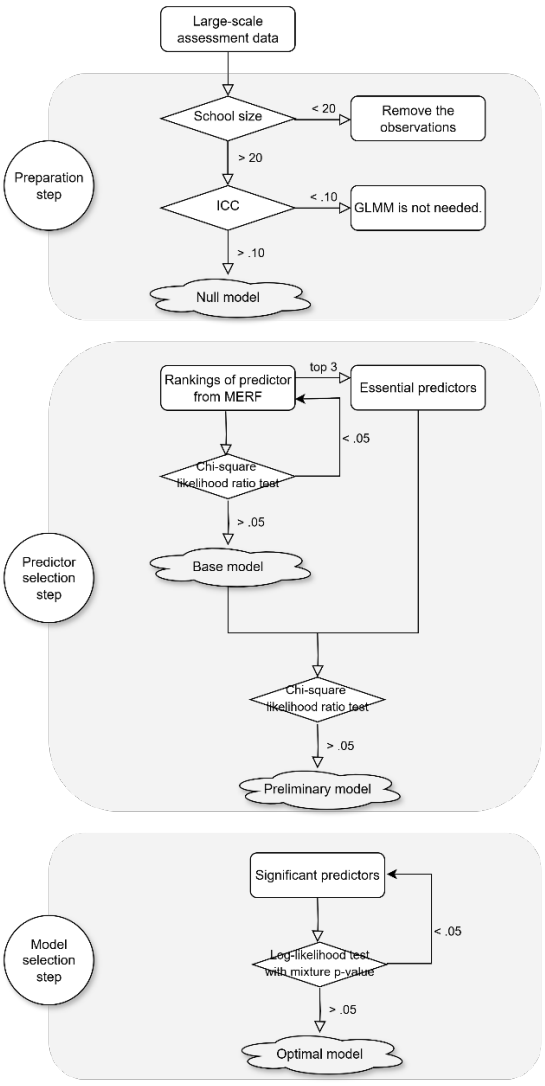

Figure 1: Flowchart of the automated search algorithm.

3.1 Preparation

In the preparation step, data from schools with fewer than 20 students are removed. Next, the size of school effects (i.e., ICC) is measured to decide whether GLMM is needed or not. If the ICC is less than .1, the GLMM is not needed. If the ICC is larger than .1, a random intercept model with no predictors, called the null model, is built. Both the minimum school size and the magnitude of the ICC are tentatively decided, yet they can be set by user.

3.2 Predictor selection

In the predictor selection stage, MERF is applied to rank all predictors of statistical importance, specifically mean decrease of prediction accuracy, from the highest to the lowest. Especially, the three highest ranked predictors are denoted essential predictors, *V1*, *V2*, and *V3*, and further utilized to the selection of possible interaction term(s).

Secondly, the top-ranked predictor is selected among all predictors that are not in the provisional model and added to the model. If the model doesn't converge, the algorithm chooses the next top-ranked predictor instead. Next, a log-likelihood ratio test is performed to determine whether the complex model with the newly added predictor is significantly different from the simpler model. Note that we use forward selection, a greedy algorithm, producing a nested sequence of models. If the complex model with the added predictor is significantly different from the simpler model, the predictor is added to the provisionary model; otherwise, stop predictor selection and fit the GLMM, a random intercept model with all significant predictors. This model is referred to as a base model.

Then, the algorithm searches for significant interaction terms. Three interaction terms of the essential predictors (i.e., *V1*:*V2*, *V1*:*V3*, and *V2*:*V3*) are sequentially added to the provisionary model and tested their significance. If the complex model with the newly added interaction term is significantly different from the simpler model, add the term to the model; otherwise, stop the procedure and identify the current model as the preliminary model.

3.3 Model selection

In the model selection step, the proposed algorithm aims to identify the optimal GLMM. First, it systematically tests each of the significant predictors in the base model as a random slope. To test the number of random effects, e.g., whether a random intercept and random slope model is significantly different from a random intercept model, a likelihood ratio test with mixture *p*-value (Self & Liang, 1987; Stram & Lee, 1994) is conducted.

To be specific, the top-ranked significant predictor is selected and a random slope for the predictor is added to the preliminary model. Then, whether the complex model with the newly added random slope is significantly different from the simpler model (preliminary model) is tested. If it is significant, the random slope is added to the preliminary model. Adding the next ranked significant predictors to the model and testing its significance are repeated until a newly added random slope is not significant; otherwise, stop and identify the current model as the optimal GLMM.

The processes of optimal model selection are summarized as follows.

1. Start with the null model.
2. Rank all predictors based on their importance using machine learning methods (e.g., MERF).
3. Select the top-ranked predictor among all predictors that are not in the provisional model.
4. Add this predictor to the (provisional) model. If the model doesn't converge, choose the next top-ranked predictor, and so on. Conduct a chi-square difference test to determine whether the model with the newly added predictor is significantly different from the simpler model.
5. If there is a significant difference, add the predictor to the current model; otherwise, remove the predictor.
6. Repeat steps 3 to 5 until a newly added predictor is not significant.
7. Add an interaction term of the top-3 ranked predictors, called essential predictors, to the model one-at-a time and test their significance.
8. If there is a significant difference, add the interaction to the model; otherwise, remove the interaction term.
9. Repeat steps 7 to 8 until a newly added interaction term is not significant.
10. Select the top-ranked significant predictor and add a random slope for the predictor to the model. Conduct a likelihood ratio test with mixture *p*-value whether the model with the newly added random slope is significantly different from the simpler model.
11. If there is a significant difference, add the slope to the current model; otherwise, remove the slope.
12. Repeat steps 10 and 11 until a newly added random slope is not significant.
13. Identify the current model as the optimal GLMM.

4. A real data example

We illustrate how the proposed algorithm can be applied to explore optimal GLMMs using real LSA data, the Trends in International Mathematics and Science Study (TIMSS). Specifically, the U.S. eighth grade student data collected in 2019 is utilized to explore optimal GLMM on their achievement scores in mathematics. Note that this section does not intend to compare the proposed algorithm's performance with other existing methods, but to illustrate how the algorithm automatically searches for optimal GLMMs from LSA data with hierarchy step by step. As far as we know, there are no existing methods to perform such a thing so far. Data was already cleaned (e.g., missing data imputation, etc.).

4.1 Preparation

Starting from 232 predictors (independent variables) of 8,698 students, the algorithm examines the schools with less than the minimum number of students and calculates the ICC value to decide whether GLMM is necessary. The minimum number of students and the ICC cut-off values are temporarily set to 20 and 0.1, respectively. We find that there are 44 schools with fewer than 20 students, leading that 600 observations being deleted. The ICC value is 0.44, implying that 44% of total variance in students' achievement scores in mathematics is explained by school differences. Thus, GLMM is needed.

Next, the algorithm fits a random intercept model without any predictors, also denoted as null model (Equation 5),

$$y_{ij} = 515 + b_{0j} + \varepsilon_{ij}. \quad (5)$$

$i = 1, 2, \ldots, I_j, J = 229$, $b_{0j} \sim N(0, 4{,}027)$, $\varepsilon_{ij} \sim N(0, 5{,}239)$, where y_{ij} referring to students' achievement scores in mathematics.

4.2 Predictor selection

In this step, the algorithm aims to search for the best base model, which is a random intercept model including all significant predictors. First, a machine learning method, MERF, ranks all 232 predictors based on their importance. We set hyper-parameters: the number of trees (*ntree*) is set to 2,000, and the number of variables randomly sampled as candidates at each split (*mtry*) is to 73, the floor of the total number of predictors divided by 3, as recommended by Breiman (2001). Note that as *ntree* gets larger, the more stable predictive error can be obtained, at the expense of computational efficiency.

The importance plot of the top-30 ranked predictors obtained by MERF is illustrated in Figure 2. Their names are abbreviated as *V1* to *V30* by their ranks and the original names are presented in Table 1.

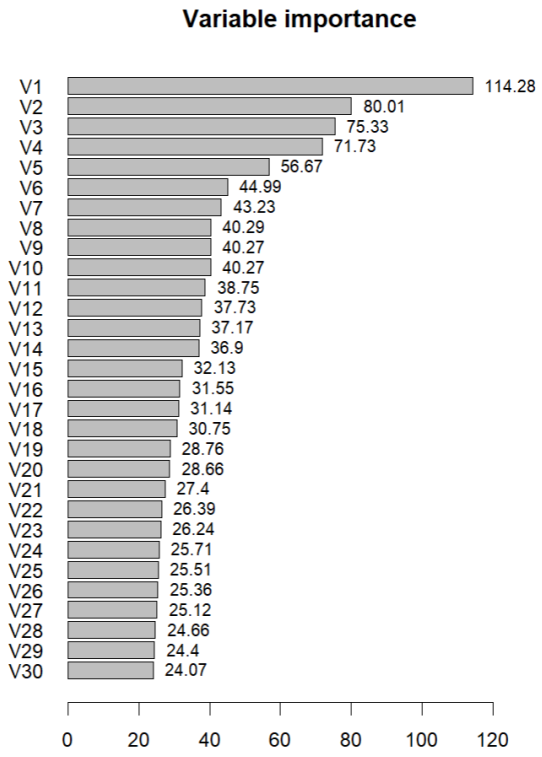

Figure 2: Predictor importance plot.

Then, all the predictors are tested for their significance one-at-a time. The algorithm fits the base model with 18 significant predictors (*V1* to *V18*), including two school-level predictors. A likelihood ratio test is conducted to compare the null model with the base model. The result indicates that the addition of the predictors significantly improved model fit, $\Delta\chi^2(18) = 3366.70$, $p < .001$ (see, Table 2). The model's R_1^2 and R_2^2 are 0.33 and 0.67, respectively. Note that *V9* is no longer significant in the base model.

No.	Variable	No.	Variable
V1	BSBG04	V16	BSBM18F
V2	BSBM19C	V17	BSBM18E
V3	BCBG03A	V18	BSBM15
V4	BSBM19A	V19	BSBS24C
V5	BSBG07	V20	BCBG06B
V6	BSBM19B	V21	BCDGTIHY
V7	BSBM19F	V22	BCBG03B
V8	BSBM19H	V23	BSBM26AA
V9	BCDGSBC	V24	BSBM16I
V10	BSDGEDUP	V25	BSBS27BB
V11	BSBM18C	V26	BSBE03E
V12	BSBM19D	V27	BSBS43BB
V13	BSDAGE	V28	BSBE03F
V14	BSBS24B	V29	BSBM18D
V15	BSBS24G	V30	BSBE01A

Table 1: Predictor names.

Model	*df*	AIC	logLik	$\Delta\chi^2$	*p*
Null	3	93095	-46544		
Base	21	89764	-44861	3367	<.001
Prelim-	22	89759	-44858	7	<.01

Table 2: Model comparison (step 2).

Next, the algorithm sequentially adds the interaction terms of the essential predictors. The result of likelihood test shows that an interaction term, *V1:V2*, significantly improved model fit, $\Delta\chi^2(1) = 6.77, p < .01$. We denote the model with the added interaction term as the preliminary model. The models' R^2 at both levels are also slightly improved.

4.3 Model selection

Based on the preliminary model in the previous step, the algorithm searches for the optimal random slopes. As a result, a random slope for *V1* is added to the preliminary model. The likelihood ratio test with mixture *p*-value indicates that the complex (called intermediate) model with the newly added slope is significantly different from the simpler (preliminary) model, $\Delta\chi^2(2) = 56.69$, *p* (mixture)

Model	*df*	AIC	logLik	$\Delta\chi^2$	p^*
Prelim-	22	89759	-44858	7	
Interim-	24	89707	-44829	57	<.001
Optimal	27	89672	-44809	41	<.001

Table 3: Model comparison (step 3).

< .001 (see, Table 3). Note that the interaction term (*V1*:*V2*) is no longer significant.

Then, a random slope for *V2* is newly added to intermediate model and the likelihood test with mixture *p*-value also demonstrates that the newly added slope significantly improved model fit, $\Delta\chi^2(3) = 41.00$, *p* (mixture) < .001 (see, Table 3). Since the random slope for *V3* is not significant, we stop model searching and the provisionary model is identified as the optimal GLMM.

The optimal GLMM has a random intercept, 18 predictors (two school-level predictors, 16 student-level predictors), an interaction term, and two random slopes for *V1* and *V2*, as shown in Equation (6),

$$y_{ij} = 603 + 9(V1)_{ij} + 6(V2)_{ij} - 17(V3)_j - 12(V4)_{ij} + 7(V5)_{ij} + 5(V6)_{ij} - 3(V7)_{ij} + 5(V8)_{ij} - 4(V9)_j + (V10)_{ij} + 5(V11)_{ij} - 3(V12)_{ij} - 11(V13)_{ij} + 5(V14)_{ij} + 6(V15)_{ij} + 3(V16)_{ij} + 2(V17)_{ij} - 8(V18)_{ij} + (V1)_{ij} \times (V2)_{ij} + b_{1j}(V1)_{ij} + b_{2j}(V2)_{ij} + b_{0j} + \varepsilon_{ij}. \quad (6)$$

$i = 1, 2, \ldots, I_j, J = 229, \varepsilon_{ij} \sim N(0, 3381)$,

$$\begin{pmatrix} b_{0j} \\ b_{1j} \\ b_{2j} \end{pmatrix} = b_j \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, G = \begin{pmatrix} 2080 & - & - \\ -178 & 52 & - \\ -152 & 28 & 48 \end{pmatrix}\right).$$

Note that the interaction term (*V1*:*V2*) is no longer significant in the optimal GLMM. The proposed algorithm is likely to find slightly different optimal GLMM, due to randomness of predictors' importance ranking obtained by MERF. A user can ensure consistent results across runs by setting a random seed.

5. Scientific Importance

It is crucial for researchers to build an adequate optimal model to make valid statistical inferences. Identifying the best-fitting GLMM is more time-consuming and complex than finding the best generalized linear model (GLM), because GLMM also includes random effects. This algorithm automatically evaluates a large number of models during the process of building an optimal GLMM model. One of the major components is a machine learning approach (e.g., MERF), which is applied to sort all predictors based on their importance, allowing for efficient predictor selection.

In addition, all available predictors from LSA data can be utilized in searching for optimal GLMMs without convergence problems using the algorithm developed here. It also provides a systematic and transparent process that can be produced by others, for example, a random intercept model is fitted, interaction terms of essential predictors are searched, and then random slopes are sequentially added to the model. The proposed algorithm has the potential to reduce the time and effort required by researchers and to provide guidelines for exploring the optimal GLMMs. We further update this algorithm by taking more considerations into account to explore best-fitting GLMMs.

References

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv*:1506.04967. https://arxiv.org/pdf/1506.04967

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5-32. https://doi.org/10.1023/A:1010933404324

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88*(421), 9-25. https://doi.org/10.2307/2290687

Bryk, A. S., & Raudenbush, S. W. (1989). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education, 97*(1), 65–108. http://www.jstor.org/stable/1084940

Capitaine, L. (2020). *LongituRF: random forests for longitudinal data*. R package version 0.9.

Goldstein, H. (2011). *Multilevel statistical models*. John Wiley & Sons.

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation, 84*(6), 1313-1328. https://doi.org/10.1080/00949655.2012.741599

Hastie, T., Tibshirani, R., & Tibshirani, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv*. https://doi.org/10.48550/arXiv.1707.08692

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Routledge.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). sage.

Self, S. G., & Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood tests under nonstandard conditions. *Journal of the American Statistical Association, 82*, 605–610. https://doi.org/10.1080/01621459.1987.10478472

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics, 50*, 1171–1177. https://doi.org/10.2307/2533455

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 58*(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology, 8*(1), 52-69. https://modir3-3.ir/article-english/article-m2568.pdf