Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Revolutionizing High-Dimensional Regularization: EMLMLasso Algorithm for Linear Mixed-Effects Models

Daniela C. R. Oliveira,[1]
DEMAT/UFSJ, São João del Rei, MG
Fernanda L. Schumacher[2]
Division of Biostatistics, College of Public Health, The Ohio State University, USA
Marcos S. Oliveira[3]
DEMAT/UFSJ, São João del Rei, MG
Daiane A. Zuanetti[4]
DES/UFSCar, São Carlos, SP
Victor H. Lachos[5]
Department of Statistics, College of Liberal Arts and Sciences, University of Connecticut, USA

Abstract. The expectation–maximization (EM) algorithm, often used for maximum likelihood estimation, has not seen much application in addressing high-dimensional regularization challenges within linear mixed-effects models. This study introduces the `EMLMLasso` algorithm, which merges the EM algorithm with the widely used and efficient R package `glmnet`, enabling Lasso variable selection for fixed effects in such models. We thoroughly evaluate its performance, comparing it to two existing algorithms from R packages `glmmLasso` and `splmm`. Our findings, based on simulations and real-world cases, demonstrate the robustness and effectiveness of our approach, even when the number of predictors (p) exceeds the number of observations (n). Notably, across most scenarios, the `EMLMLasso` algorithm consistently outperforms both `glmmLasso` and `splmm`. Moreover, our method is versatile and straightforward to implement, with the potential for extensions to include ridge and elastic net penalties in linear mixed-effects models.

Keywords. EM algorithm, High-dimensional data, Mixed-effects models, R package `glmnet`, Regularized variable selection methods

1 Introduction

The linear mixed-effects models (LMM) represent a significant statistical tool for analyzing relationships between responses and covariates in clustered or longitudinal data scenarios [7]. This approach finds widespread use across various domains such as genetics, health, finance, ecology, and image processing, underscoring the need for selecting the most appropriate LMM for such data. However, challenges arise when dealing with high-dimensional variable selection, where the number of predictors (p) exceeds the number of observations (n) [4]. Despite advancements in computational and statistical methods, selecting fixed effects in LMM remains a daunting task, particularly under high dimensionality.

[1] daniela@ufsj.edu.br
[2] schumacher.313@osu.edu
[3] mso@ufsj.edu.br
[4] dzuanetti@ufscar.br
[5] hlachos@uconn.edu

DOI: 10.5540/03.2025.011.01.0472

Among the statistical methods proposed for variable selection, regularization-based approaches stand out for their ability to simultaneously identify crucial variables. In the realm of fixed effects selection within LMM, notable works by [10] and [6] employ L1-penalization to maximize the penalized log-likelihood (PML) function through computational techniques. Moreover, there exist various strategies for concurrently selecting fixed and random effects in LMM, like `glmmLasso` [6] and `splmm` [13], which can offer comparative insights.

This work presents an approach for fixed effects selection merging the EM algorithm with PML estimation using the Lasso penalty, with flexibility to extend the methodology to other types of penalties. Leveraging the `glmnet` [5] package, which efficiently fits generalized linear models via PML, our `EMLMLasso` algorithm determines the regularization path for the Lasso penalty. We use the Bayesian Information Criterion (BIC) to identify the optimal tuning parameter. The resulting model, comprising only non-zero fixed effects variables, can be easily fitted using established R packages such as `lme4` [2] or `skewlmm` [11].

2 Methods

2.1 The Linear Mixed-Effects Model

The normal linear mixed model (LMM) is specified as follows [7]:

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i, \tag{1}$$

where $Y_i = (Y_{i1}, \ldots, Y_{in_i})^\top$ is a $n_i \times 1$ vector of observed continuous responses for subject i, with $i = 1, 2, ..., n$, X_i is the $n_i \times p$ design matrix corresponding to the fixed effects, β of dimension $p \times 1$; Z_i is the $n_i \times q$ design matrix corresponding to the $q \times 1$ vector of random effects b_i; $b_i \overset{iid}{\sim} N_q(0, D)$ is independent of $\epsilon_i \overset{indep.}{\sim} N_{n_i}(0, \sigma^2 I_{n_i})$, the dispersion matrix $D = D(\alpha)$ depends on unknown and reduced parameters α; and ϵ_i of dimension $(n_i \times 1)$ is the vector of random errors.

To account for high-dimension problems, we allow the general framework where the number p of fixed-effects regression coefficients can be larger than the total number of observations, that is, $p > n$. To perform PML estimation in the general LMM specification from (1), we now present a proposal based on the EM algorithm.

2.2 EMLMLasso Algorithm

Define a grid of possible values for λ. Traverse all values of lambda within this grid. The algorithm initializes with the first fixed value for λ as follows:

```
Initialization
$\beta^{(0)} = \arg\, min_\beta[(Y - X\beta)^\top(Y - X\beta) + \lambda||\beta||_1]$
$\sigma^{2(0)} = \frac{1}{n}(Y - X\beta^{(0)})^\top(Y - X\beta^{(0)})$,
$D^{(0)} = I_q$,
$b^{(0)} = 0$.
While the stopping criterion is not satisfied do
E-Step:
$\tilde{y}_i^{(k)} = y_i - Z_i b_i^{(k)}$,  $\Lambda_i^{(k)} = (D^{-1(k)} + Z_i^\top Z_i/\sigma^{2(k)})^{-1}$,
$b_i{}^{(k)} = \frac{1}{\sigma^{2(k)}}\Lambda_i^{(k)} Z_i^\top(\tilde{y}_i^{(k)} - X_i\beta^{(k)})$,
$\lambda_1^{(k)} = 2\lambda\sigma^{2(k)}$.
```

```
M-Step:
    $\beta^{(k+1)} = \arg\ min_\beta[(\tilde{y}^{(k)} - X\beta)^\top(\tilde{y}^{(k)} - X\beta) + \lambda_1^{(k)}\Psi_p(\beta)]$,
    $\sigma^{2(k+1)} = \frac{1}{N}[(\tilde{y}^{(k)} - X\beta^{(k)})^\top(\tilde{y}^{(k)} - X\beta^{(k)}) + \sum_{i=1}^n tr(Z_i\Lambda_i^{(k)}Z_i^\top)]$,
    $D^{(k+1)} = \frac{1}{n}\sum_{i=1}^n(b_i^{(k)}b_i^{\top(k)} + \Lambda_i^{(k)})$.
end
```

The stopping criterion is defined using the Bayesian Information Criterion (BIC), determining the optimal value of λ [12] and, finally, $\theta = (\beta, \sigma^2, D)$.

The EM algorithm's "E-Step" (Expectation Step) calculates the expectation of the complete-data log-likelihood with respect to the distribution of the latent variables, given the observations and the current parameter estimates, and the "M-Step" (Maximization Step) maximizes this expectation calculated in the E-Step to update the parameter estimates. EMLMLasso can be summarized as follows: first, a grid with possible λ values is defined (regularization parameter or parameter that controls the penalty of coefficient estimates), then for each λ value in the grid, iterations are made using the EM algorithm, in which the optimization of the β coefficients is done with the aid of the R package glmnet and, finally, an optimal value of λ is selected using BIC, and the estimate of the desired θ, for the selection variables. The main innovation of the EMLMLasso algorithm compared to existing algorithms is the incorporation of the R package glmnet within the EM algorithm for the optimization of the β coefficients, a simple method of adaptive regularization that improved accuracy in situations of high dimensionality and multicollinearity.

3 Simulation Studies

This section evaluates the performance of the EMLMLasso algorithm in three scenarios, comparing the results with the glmmLasso and splmm packages.

Due to page limitations for this article, we highlight the main results found in the evaluated scenarios. More quantitative details about the scenarios can be seen in the article by [9].

In Scenario 1, with fixed effects ($p = 9$) and random effects ($q = 2$), both EMLMLasso and splmm accurately identify significant variables (β_1 and β_2). However, glmmLasso tends to misclassify non-significant coefficients. Root Mean Square Error (RMSE) values indicate that EMLMLasso and splmm have similar performance, with glmmLasso slightly better. Results improve with larger samples. Scenario 2 examines the impact of the presence of categorical covariates. Here, all algorithms successfully identify significant variables. EMLMLasso demonstrates superior accuracy in parameter estimation compared to glmmLasso and splmm. In Scenario 3, high-dimensional predictors ($p = 50$) are analyzed. EMLMLasso consistently excels at accurately identifying significant variables, with sensitivity values of 1 across all sample sizes. It outperforms other methods even with estimates of $\beta = 0$. Sensitivity is generally higher but specificity is lower for all algorithms. glmmLasso shows an offset between these measurements.

Furthermore, the use of 10-fold cross-validation reveals the superior performance of EMLMLasso in the selection of fixed effects, with lower median, range and interquartile range of RMSE compared to glmmLasso and splmm.

Overall, the EMLMLasso algorithm presents robustness and efficiency in variable selection, especially in scenarios involving high-dimensional predictors and categorical variables.

4 Application in Genetics

Gene expression experiments study how genes are turned on and off and how this controls what substances are made in a cell. This dataset concerns the response of riboflavin (vitamin B2) production of bacillus subtilis (b. subtilis), a single celled organism (bacterium) found in the human digestive tract. The final goal of researchers is to increase the riboflavin production rate of b. subtilis by editing relevant genes. To facilitate this goal, we used the riboflavinV100 dataset, which contains the genes that most strongly influence the rate of riboflavin production [10]. The data is provided by DSM (Switzerland) and made publicly available in the supplemental materials of [4]. This dataset was previously analyzed by [10], [4], [3], [1], among others. We also use glmmLasso to select relevant covariates for this dataset and compare the results with the ones obtained via EMLMLasso.

Given the longitudinal character of the dataset, we consider the following linear mixed-effects model:

$$y_{ij} = \sum_{k=1}^{100} \beta_k x_{ijk} + \beta_{101} t_{ij} + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}, \tag{2}$$

where the response variable is the log of the rate of riboflavin produced, and there are 100 covariates representing the log of the expression level of 100 genes and the covariate time. This dataset consists of $n = 28$ different strains (species subtypes) of b. subtilis measured between two and four times over the course of 96 hours ($n_i \in \{2, 3, 4\}$), totalizing 71 observations. We standardize the response and all covariates to have mean zero and variance one.

In this application, the number of correlated covariates is larger. For this reason, we decided to evaluate the results of the algorithms with two methods: 1) (Method 1) complete matrix X, and 2) (Method 2) reduced matrix X^*. The matrix X, of dimension 71 × 101, is obtained from the riboflavinV100 dataset. The reduced matrix X^*, of dimension 71 × 70, is obtained using the package findLinearCombos in R, which removes columns that have linear combinations among them in a matrix X.

Table 1: RiboflavinV100 dataset. Gene selections.

Methods	Variable list
$^{(a)}$glmmLasso	TIME XHLA_at XHLB_at
$^{(b)}$findLinearCombos + glmmLasso	TIME
$^{(c)}$EMLMLasso	YHZA_at YHFH_r_at NADC_at YPUF_at ACOA_at YPUD_at YCGN_at YXLE_at YTGD_at PURC_at XLYA_at YCGO_at GSIB_at YTCF_at GAP_at YRDD_i_at CARA_at YCIB_at YOSJ_at ALD_at TRXA_at PCKA_at
$^{(d)}$findLinearCombos + EMLMLasso	TIME YHZA_at YRZI_r_at DEGQ_r_at YXLE_at ARGF_at YTGD_at GUAB_at AHPC_at XLYA_at YCGO_at YTCF_at GAP_at
$^{(e)}$splmm	TIME YHZA_at YHFH_r_at YRPE_at YCGN_at YXLE_at ARGF_at XLYA_at YTCF_at YTGA_at YTGB_at PCKA_at YCKE_at
$^{(f)}$findLinearCombos + splmm	TIME YHZA_at YHFH_r_at YRPE_at YCGN_at YXLE_at ARGF_at GUAB_at XLYA_at YTCF_at YTGA_at

After running additional tests, for the estimation of λ in the glmmLasso, we considered a sequence from 0 to 500 by 1, for the EMLMLasso, we used a sequence from 0.001 to 0.5 with length out equal to 500, and a sequence from 1.541 to 1.701 by 0.001 for the splmm. When we work with the complete matrix X (Method 1), the optimal values for λ are 43, 0.22 and 1.557 for glmmLasso, EMLMLasso and splmm, respectively. For the reduced matrix X^* (Method 2), we kept a sequence from 1.4 to 1.7, by 0.01 for the splmm. In this case, the optimal values for λ are 43, 0.404 and 1.41 for glmmLasso, EMLMLasso and splmm, respectively. Table 1 shows that with Method 1, EMLMLasso selected 22 genes, the glmmLasso selected 3 covariates, and the splmm selected 13 covariates. However, when we use Method 2, the EMLMLasso selected 12 genes and the covariate TIME, the glmmLasso selected only 1 covariate, and the splmm selected 11 covariates. The selected variables have these labels because they are the names given to the genes in the dataset. These labels were made available by [4].

We use the riboflavinV100 dataset, the **R** package lme4, and fitted a LMM with the predictors obtained with Method 1 and another with the predictors from Method 2, for each algorithm. We used the R packages joineR, lme4, splines, and caret to perform the 4-fold cross-validation and compare the predictive power of each method, calculating the mean squared error (MSE) of y as

$$\mathrm{MSE}_y = (y - \hat{y})^\top (y - \hat{y}). \tag{3}$$

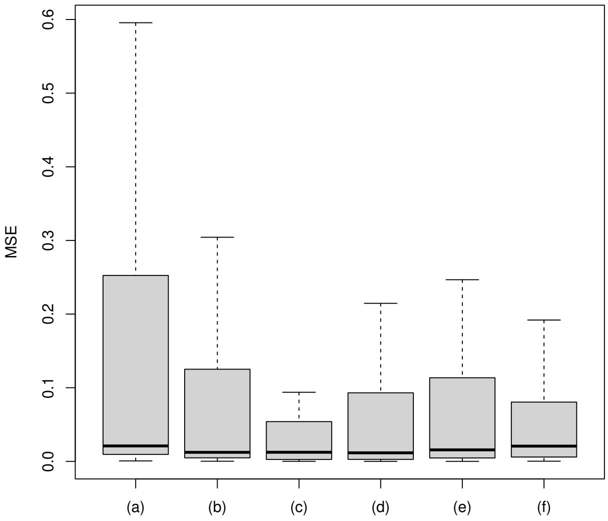

Figure 1: RiboflavinV100 dataset. 4-fold cross-validation to evaluate performance using the mean squared error defined in (3) of the glmmLasso, EMLMLasso and splmm algorithms. Scenarios (a), (b), (c), (d), (e), and (f) as defined in Table 1. Source: by the authors.

When comparing the MSE results (see Figure 1) between matrix X (without applying the findLinearCombos function) and matrix X^* (with the application of the findLinearCombos function), the splmm algorithm performed better with matrix X in terms of a lower mean MSE value. Conversely, the glmmLasso and EMLMLasso algorithms showed improved performance with matrix X^*. Regarding dispersion, i.e., lower interquartile range of MSE, EMLMLasso showed better results with matrix X, while the splmm and glmmLasso algorithms performed better with matrix X^*. The results indicated that the EMLMLasso algorithm was more sensitive to the removal of linear combinations, which impacted the MSE dispersion. In summary, when using the EMLMLasso algorithm with matrices X and X^*, the median MSE values were close to zero (0.0124 and 0.0116,

DOI: 10.5540/03.2025.011.01.0472

respectively), with lower dispersion observed for matrix X. Therefore, we conclude that a viable strategy for achieving better predictive power is to use the `EMLMLasso` algorithm without applying the `findLinearCombos` function (Method (c), which selected 22 covariates).

5 Discussion

In this work, we propose a novel algorithm for variable selection in linear mixed models based on the EM algorithm and the Lasso penalty, where the Lasso estimation step depends on R package `glmnet`. We call the proposed algorithm `EMLMLasso`. Even though other complex solutions have been proposed to deal with variable selection problems in linear mixed models under low or high-dimensional settings, to the best of our knowledge, it is the first attempt to propose a straightforward implementation relying on existing packages. We focus on the Lasso penalty, but it certainly can be implemented for other kinds of penalties, such as ridge and elastic net. We provide a publicly available R code to compute the methods introduced in this work, which is available for download from GitHub at `https://github.com/fernandalschumacher/EMLMLasso`.

For comparison purposes, we chose to use the publicly available R package `glmmLasso` [6], which is a well-known package for variable selection in generalized mixed-effects models, and `splmm` [13], that fits linear mixed-effects models for high-dimensional data ($p >> n$) with penalty for both the fixed effects and random effects for variable selection. Under three scenarios, we investigate the performance of the proposed algorithm to select significant fixed effects through a set of simulations. In the first scenario, we simulated covariates from the normal distribution and evaluated the capability of the `EMLMLasso`, `glmmLasso`, and `splmm` algorithms to select the fixed effects. In a second scenario, we evaluated the ability of the algorithms to select fixed effects in the presence of categorical covariates. In a third scenario, we consider a large vector of fixed effects and evaluate the sensitivity and specificity of the algorithms. Finally, we use 10-fold cross-validation to evaluate the performance of algorithms under a high-dimensional setting ($p > n$). The results of the simulations demonstrated good properties of the proposed variable selection procedure. The `EMLMLasso` algorithm outperformed `glmmLasso`, and `splmm` in the majority of the scenarios considered, especially when evaluating the specificity.

We also analyzed one dataset applied in Genetics. These are gene expression data ($p > n$), where we are interested in relevant genes responsible for increasing the production of the riboflavin (vitamin B2) of *bacillus subtilis*, a bacterium found in the human digestive tract. In this study, as the covariates are correlated and $p > n$, we evaluated the three algorithms by adopting two configurations: 1) considering the original data and 2) using a function from R to eliminate linear correlations. The `EMLMLasso` made the selection of genes under the two considered strategies and presented the best predictive power.

The algorithm developed here does not consider censoring and/or missing responses, a typical problem in longitudinal studies. [8] have proposed a likelihood-based treatment based on the EM algorithm for parameter estimation in linear and nonlinear mixed-effects models with censored data (LMEC/NLMEC). Therefore, it would be a worthwhile task to investigate the applicability of variable selection in the context of LMEC/NLMEC models. The selection of the random effects is also a topic of our future research.

Acknowledgements

Daniela Carine Ramires de Oliveira's research was partially supported by research project n$^{\underline{o}}$. 401418/2022-7 from the National Council for Scientific and Technological Development (CNPq) – Brazil. Victor H. Lachos acknowledges partial financial support from the Office of the Vice

President for Research (REP) and UConn - CLAS's Summer Research Funding Initiative 2023. All authors of this work are supported by research project no. APQ-04657-23 from the Minas Gerais Research Support Foundation (Fapemig) – Brazil.

References

[1] A. Alabiso and J. Shang. "High-dimensional linear mixed model selection by partial correlation". In: **Communications in Statistics - Theory and Methods** 52.18 (2023), pp. 6355–6380. DOI: 10.1080/03610926.2022.2028838.

[2] D. Bates, M. Mächler, B. Bolker, and S. Walker. "Fitting Linear Mixed-Effects Models Using lme4". In: **Journal of Statistical Software** 67.1 (2015), pp. 1–48. DOI: 10.18637/jss.v067.i01.

[3] J. Bradic, G. Claeskens, and T. Gueuning. "Fixed Effects Testing in High-Dimensional Linear Mixed Models". In: **Journal of the American Statistical Association** 115.532 (2020), pp. 1835–1850. DOI: 10.1080/01621459.2019.1660172.

[4] P. Bühlmann, M. Kalisch, and L. Meier. "High-Dimensional Statistics with a View Toward Applications in Biology". In: **Annual Review of Statistics and Its Application** 1.1 (2014), pp. 255–278. DOI: 10.1146/annurev-statistics-022513-115545.

[5] J. Friedman, T. Hastie, and R. Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: **Journal of Statistical Software** 33.1 (2010), pp. 1–22. DOI: 10.18637/jss.v033.i01.

[6] A. Groll. **glmmLasso: Variable selection for generalized linear mixed models by L1-penalized estimation**. Online. https://cran.r-project.org/package=glmmLasso, Accessed May 05, 2023.

[7] N. M. Laird and J. H. Ware. "Random-Effects Models for Longitudinal Data". In: **Biometrics** 38.4 (1982), pp. 963–974. DOI: 10.2307/2529876.

[8] L. A. Matos, V. H. Lachos, N. Balakrishnan, and F. V. Labra. "Influence diagnostics in linear and nonlinear mixed-effects models with censored data". In: **Computational Statistics & Data Analysis** 57.1 (2013), pp. 450–464. DOI: 10.1016/j.csda.2012.06.021.

[9] D. C. R. Oliveira, F. L. Schumacher, and V. H. Lachos. "EMLMLasso: The use of the EM algorithm for regularization problems in high-dimensional linear mixed-effects models". In: (2023). URL: https://arxiv.org/abs/2308.01518.

[10] J. Schelldorfer, P. Bühlmann, and S. V. De Geer. "Estimation for high−dimensional linear mixed−effects models using l_1−penalization". In: **Scandinavian Journal of Statistics** 38.2 (2011), pp. 197–214. DOI: 10.1111/j.1467-9469.2011.00740.x.

[11] F. L. Schumacher, V. H. Lachos, and L. A. Matos. "Scale mixture of skew-normal linear mixed models with within-subject serial dependence". In: **Statistics in Medicine** 40.7 (2021), pp. 1790–1810. DOI: 10.1002/sim.8870.

[12] H. Wang, R. Li, and C. L. Tsai. "Tuning parameter selectors for the smoothly clipped absolute deviation method". In: **Biometrika** 94.3 (2007), pp. 553–568. DOI: 10.1093/biomet/asm053.

[13] L. Yang and T. T. Wu. "Model-based clustering of high-dimensional longitudinal data via regularization". In: **Biometrics** 79.2 (2022), pp. 761–774. DOI: 10.1111/biom.13672.

DOI: 10.5540/03.2025.011.01.0472