Comparison of Methods for Variable Selection in High-Dimensional Linear Mixed Models

J. Jakubík

Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia,
Email: jozef.jakubik.jefo@gmail.com

Abstract. *Currently is the analysis of high-dimensional data a popular field of research, thanks to many applications e.g. in genetics. At the same time, the type of problems that tend to arise in genetics, can often be modeled using LMMs in conjunction with high-dimensional data. In this paper we introduce two new methods and briefly compare them to existing methods, which can be used for variable selection in high-dimensional linear mixed models. As we will show in a small simulation study, both methods perform well compared to existing methods.*

Keywords: Linear mixed model, Variable selection, High-dimensional data

1. Introduction

Linear mixed model (LMM) allow us to specify the covariance structure of the model, which enables us to capture relationships in data, for example population structure, family relatedness etc. Therefore, LMMs are often preferred to linear regression models. Consider a LMM of the form

$$Y = X\beta + Zu + \varepsilon,$$

where

Y is $n \times 1$ vector of observations,

X is $n \times p$ matrix of regressors,

β is $p \times 1$ vector of unknown fixed effects,

Z is $n \times q$ matrix of predictors,

u is $q \times 1$ vector of random effects with the distribution $\mathcal{N}(0, \sigma_D^2 I)$,

ε is $n \times 1$ error vector with the distribution $\mathcal{N}(0, \sigma^2 I)$ and independent from u.

In genome-wide association studies in genetics, one studies the dependence of phenotype on the genotype. Genetic information can consist of 10^6 variables, but only information about the genotype of a small group of subjects is available. Variable selection in high-dimensional data refers to the selection of a small group of variables (denote it S^0, and $s^0 = |S^0|$ the number of relevant variables) which influence observations. In our case LMM we assume, that matrix X is high-dimensional and we select only variables from matrix X.

More information about the model can be found in Section 3.

2. Methods

In this paper we compare four methods for variable selection in high-dimensional LMMs.
All of the mentioned methods are primarily β estimation methods, not selection methods. However they can be thought of as selection methods if we define selected variables to be those for which $\beta_i \neq 0$ for $i = 1, \dots, p$.

LASSO

Least absolute shrinkage and selection operator [1, 2] is an established method for selecting variables in linear regression models. LASSO corresponds to the ℓ_1-penalized ordinary least squares estimate:

$$\hat{\beta} = \arg\min_{\beta} \left[\|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right],$$

where λ is a fixed parameter.
In this study we use the LASSO as the reference, as it ignores LMM data structure.

LMM-LASSO

In [3], authors propose a data transformation, which eliminates correlation between observations. We first estimate σ_D^2, σ^2 by Maximum Likelihood under the null model, ignoring the effect of variables in matrix X. Let $K = 1/q \cdot ZZ^\mathsf{T}$. Having fixed $\hat{\gamma} = \hat{\sigma_D^2}/\hat{\sigma^2}$, we use the spectral decomposition of $K = U\Lambda U^\mathsf{T}$ to rotate our data, so that the covariance matrix becomes isotropic:

$$\tilde{X} = (\hat{\gamma}\Lambda + I)^{-\frac{1}{2}} U^\mathsf{T} X$$
$$\tilde{Y} = (\hat{\gamma}\Lambda + I)^{-\frac{1}{2}} U^\mathsf{T} Y.$$

After transforming the data we use the LASSO method

$$\hat{\beta} = \arg\min_{\beta} \left[\frac{1}{\hat{\sigma^2}} \|\tilde{Y} - \tilde{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right].$$

New Approach One

The first approach consists in a transformation that removes group effects from data. The principle of this transformation is widely used in data analysis, for example in restricted/residual maximum likelihood (REML). In our case we transform the data as follows

$$\tilde{X} = (I - ZZ^+)X,$$
$$\tilde{Y} = (I - ZZ^+)Y,$$

where Z^+ is the pseudoinverse matrix. The transformation eliminates random segments of the problem (associated with matrix Z) and which allows us to use LASSO method for linear regression model.

New Approach Two

Recently, a publication [4, 5] in the field of variable selection for high-dimensional LMM data, presents methods based on non-convex optimization problem with one penalty parameter. For problems of dimension higher than 10^4 are methods based on non-convex optimization problem almost unusable, because their computational complexity is beyond the capabilities of current computers. One of the possible solutions to this problem is the simplification of the optimized function to convex function. Therefore, we have proposed a method based on the solution to the following convex problem

$$(\hat{\beta}, \hat{u}) = \arg\min_{\beta, u} \left[\|Y - X\beta - Zu\|_2^2 - \lambda \|\beta\|_1 - \Lambda \|u\|_2^2 \right],$$

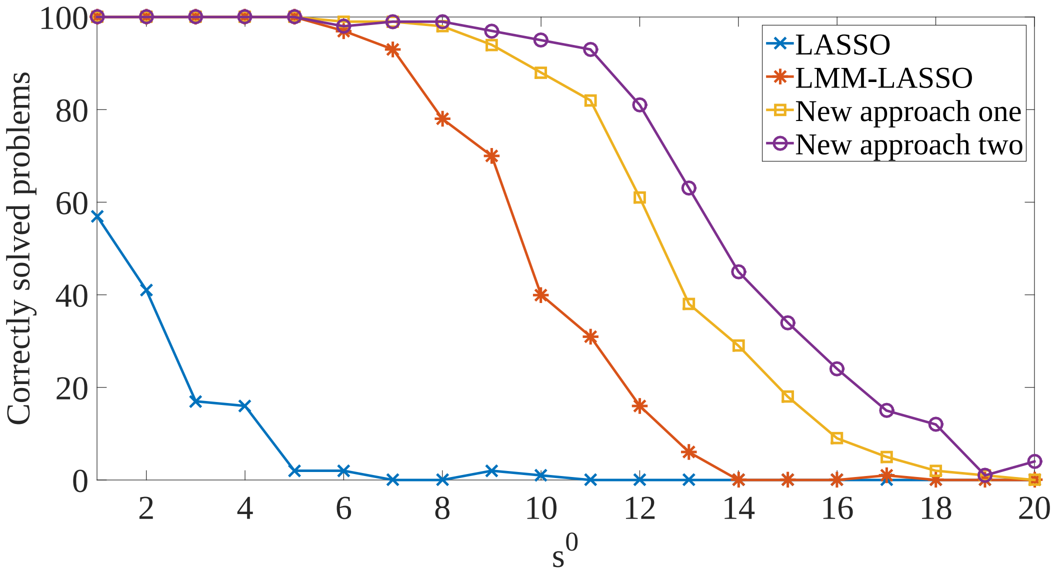

Fig. 1: Comparison of the number of correctly solved problems for different s^0 with four different methods.

where λ and Λ are fixed parameters.
Basically we are exchanging computational complexity for the need to inspect a two-dimensional parameter space.

3. Simulation Study

Data in our simulation study are divided into twenty groups of ten observations. Together we have $n = 200$ observations. For each observation we observe $p = 5000$ variables, but only $s^0 = \{1,\ldots,20\}$ variables influence observations. Relevant variables are randomly selected from all variables and effect of relevant variables is one. The effect of other variables is zero. Matrix Z captures group structure of the data. $Z_{i,j} = 1$ if the i-th observation belongs to the j-th group, 0 otherwise. Random effects u are randomly selected from $\mathcal{N}(0,I)$. Errors are from $\mathcal{N}(0,0.2\cdot I)$.
For all mentioned methods we get different sets of selected variables for different parameters λ or Λ. We generate a hundred problems as described in previous paragraph. As a correctly solved problem we consider only a problem for which the method gives for at least one parameter or parameter combination as the selected variable set exactly set S^0. Figure 1 shows the number of correctly solved problems for all four methods for different numbers of relevant variables (from 1 to 20). The methods from [4, 5] were not compared because they were not able to solve problems of dimension $p = 5000$.

4. Conclusion

In Figure 1 we can see that the LASSO method is not suitable if the problem has the structure of a LMM.
For small numbers of relevant variables the remaining methods are almost infallible. With the increasing number of relevant variables, the accuracy of methods decreases to almost zero. This is understandable, because with more relevant variables the correlation of each relevant variable with vector of observations decreases. Therefore, it is more difficult to identify correctly the exact set of variables.

This simple study hints at the potential of the newly proposed methods to significantly outperform both the LASSO and the LMM-Lasso.
However, this is only a preliminary study and one of the first addressing the question. A more extensive analysis can be expected in the future.

Acknowledgements

The work was supported by the Scientific Grant Agency VEGA of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, by the projects VEGA 2/0047/15 and VEGA 2/0043/13.

References

[1] Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288, 1996.

[2] Bühlmann, P, Van De Geer S. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media. 2011.

[3] Lippert C. Linear mixed models for genome-wide association studies. `https://publikationen.uni-tuebingen.de/xmlui/handle/10900/50003`, 2013.

[4] Schelldorfer J, Bühlmann P, van De Geer S. Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scandinavian Journal of Statistics* 38(2): 197–214, 2011.

[5] Rohart F, San Cristobal M, Laurent B. Selection of fixed effects in high dimensional linear mixed models using a multicycle ECM algorithm. *Computational Statistics & Data Analysis* 80, 209–222, 2014.