Convex method for variable selection in high-dimensional linear mixed models

Jozef Jakubík

Institute of Measurement Science
Slovak Academy of Sciences

2nd September 2015

Linear mixed model (LMM)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\,\mathbf{u} + \boldsymbol{\varepsilon},$$

where

- $\mathbf{Y}$ is a $n \times 1$ known vector of observations, $\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$;
- $\boldsymbol{\beta}$ is a $p \times 1$ unknown vector of fixed effects;
- $\mathbf{X}$ is a $n \times p$ known design matrix relating the observations $\mathbf{Y}$ to $\boldsymbol{\beta}$;
- $\mathbf{u}$ is a $q \times 1$ unknown vector of random effects, $\mathrm{E}(\mathbf{u}) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{u}) = \mathbf{D}$;
- $\mathbf{Z}$ is a $n \times q$ known design matrix relating the observations $\mathbf{Y}$ to $\mathbf{u}$;
- $\boldsymbol{\varepsilon}$ is a $n \times 1$ unknown vector of random errors, $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbf{R}$.

Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Institute of Measurement Science, SAS

High-dimensional LMM

$$Y = X \beta + Z u + \varepsilon$$

$$q \ll n \ll p$$

Motivation

LMM allows us to specify the covariance structure of the model, which enables us to capture relationships in data.

For example:

- population structure,
- family relatedness.

This could, for example, be handy in:

- Genome-wide association studies (GWAS)
- Mass spectrometry studies

Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Institute of Measurement Science, SAS

Variable selection

We know that:

- only a small subset of all p variables (in X) influence observations Y. We denote this subset S^0 and $s^0 = |S^0|$;
- all q variables (in Z) influence observations Y, but the effect of some variables can be very small.

We aim for an estimate of the S^0.

Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Institute of Measurement Science, SAS

Example

We investigate which genetics aspects influence the size of soya beans.

DNA with 10^6 variables.

Just a small group of relevant genetics variables.

A few relevant external variables, for example weather, land ...

Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Institute of Measurement Science, SAS

Methods

All of the following methods are primarily $\boldsymbol{\beta}^0$ estimation methods, not selection methods. However, they can be thought of as selection methods if we define selected variables to be those for which $\hat{\boldsymbol{\beta}}_i \neq 0$ for $i = 1, \ldots, p$.

After variable selection:

- Estimation • Henderson's mixed models equation - BLUE for $\boldsymbol{\beta}$ and BLUP for $\mathbf{u}$
- Model selection • Cross-validation, Information criteria, ...

Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Institute of Measurement Science, SAS

Methods

- LASSO [Tibshirani, 1996]

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

- LMMLASSO [Schelldorfer et al., 2011]
- LASSOP [Rohart et al., 2014]

Two new approaches

- Naive transformation to linear regression
- Convex method

Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Institute of Measurement Science, SAS

Existing methods

LMMLASSO

$$(\hat{\beta}, \hat{\mathbf{D}}, \hat{\sigma}^2) = \underset{\beta, \sigma^2>0, \mathbf{D}\succ 0}{\arg\min} \left[\frac{1}{2}\log|\mathbf{\Sigma}| + \frac{1}{2}(\mathbf{Y}-\mathbf{X}\beta)^T\mathbf{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\beta) + \lambda\|\beta\|_1\right]$$

where $\mathbf{\Sigma} = (\mathbf{Z}\mathbf{D}\mathbf{Z}^T + \mathrm{R})$.

LASSOP

$$(\hat{\beta}, \hat{\mathbf{D}}, \hat{\sigma}^2) = \underset{\beta, \sigma^2>0, \mathbf{D}\succ 0}{\arg\min} \left[\frac{1}{2}\log|\mathbf{R}| + \frac{1}{2}(\mathbf{Y}-\mathbf{X}\beta-\mathbf{Z}\mathbf{u})^T\mathbf{R}^{-1}(\mathbf{Y}-\mathbf{X}\beta-\mathbf{Z}\mathbf{u}) \right.$$
$$\left. + \frac{1}{2}\log|\mathbf{D}| + \frac{1}{2}\mathbf{u}^T\mathbf{D}^{-1}\mathbf{u} + \lambda\|\beta\|_1\right]$$

- generally not convex
- similar results
- both implemented in R

Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Institute of Measurement Science, SAS

Naive method

Data transformation that removes random effects:

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

where $\mathbf{Z}^{+}$ is the pseudoinverse matrix.

The transformation allows us to use the LASSO method for linear regression models.

Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Institute of Measurement Science, SAS

Convex method

$$(\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}) = \arg\min_{\boldsymbol{\beta},\mathbf{u}} \left[\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}\|_2^2 - \lambda \|\boldsymbol{\beta}\|_1 - \Lambda \sum_{i=1}^{q^*} \| {}_i\mathbf{u}\|_2^2 \right],$$

where λ and Λ are fixed parameters, q^* is the number of variance components (without error) and ${}_i\mathbf{u}$ is the part of vector $\mathbf{u}$ belonging to the i-th variance component.

- convex
- we use MATLAB with the convex programming modeling system CVX and the solver Mosek

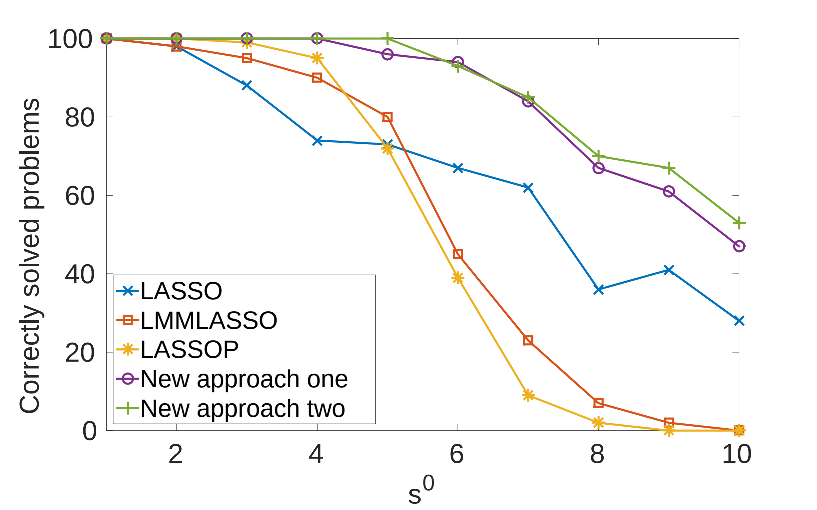
Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Institute of Measurement Science, SAS

Simulation study

- $n = 120$ observations divided into twenty groups of six
- $p = 150$ all available fixed variables
- $s^0 = \{1, \dots, 10\}$ relevant fixed variables
- $q^* = 2$, $q = 40$
- $\mathbf{u}$ consists of two parts, one for each variance component and $\mathbf{u} \sim \mathcal{N}(0, \mathbf{D} = 2 \cdot \mathbf{I})$
- $\varepsilon \sim \mathcal{N}(0, \mathbf{D} = \mathbf{I})$

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 the set S^0.

Result



Jozef Jakubík (jozef.jakubik.jefo@gmail.com)
Institute of Measurement Science, SAS

Summary

- Thanks to convexity, both 'new' methods can solve problems with dimension up to 10^5 variables. On the other hand, neither of the 'old' method can handle with more then 10^3 variables.
- For solving the convex problem, it is possible to use good existing software.
- Both 'new' methods hit exactly the set S^0 more times than the 'old' methods.

References

Bühlmann, P. and Van De Geer, S. (2011).
Statistics for high-dimensional data: methods, theory and applications.
Springer Science & Business Media.

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

Rohart, F., San Cristobal, M. and Laurent, B. (2014).
Selection of fixed effects in high dimensional linear mixed models using a multicycle ECM algorithm.
Computational Statistics & Data Analysis 80, 209--222.

Schelldorfer, J., Bühlmann, P. and van De Geer, S. (2011).
Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ_1-Penalization.
Scandinavian Journal of Statistics 38, 197--214.

Tibshirani, R. (1996).
Regression shrinkage and selection via the lasso.
Journal of the Royal Statistical Society. Series B (Methodological) 58, 267--288.