Structure Estimation, Graphical Modelling and Causal Inference in High Dimensions: Linear Mixed-effects Models

"`lmmlasso`: Estimation for High-dimensional Linear Mixed-Effects Models Using ℓ_1-penalization" [5] is a project of C1 in collaboration with C3, and it builds upon [6] and [2].

1. Introduction

n: number of observations, p: number of variables

	$n > p$	$n << p$
Linear Regression	Ordinary Least Squares	Lasso
Linear Mixed-Effects Models	Maximum Likelihood (ML) Restricted Maximum Likelihood (REML)	`lmmlasso`

Key challenges: high-dimensionality, non-convexity

2. Linear mixed models and ℓ_1-penalized estimation

2.1 High-dimensional Model Set-up

Inhomogeneous data (not independent, but grouped observations)
$i = 1, \ldots, N$ grouping index
$j = 1, \ldots, n_i$ observation index
$N_T = \sum_{i=1}^N n_i \ll p$ total number of observations

For each group i:
- $\boldsymbol{y}_i$: $n_i \times 1$ vector of responses
- $\boldsymbol{X}_i$: $n_i \times p$ fixed-effects design matrix
- $\boldsymbol{Z}_i$: $n_i \times q$ random-effects design matrix
- $\boldsymbol{b}_i$: $q \times 1$ group-specific vector of random regression coefficients

Common for all groups:
- $\boldsymbol{\beta}$: $p \times 1$ vector of fixed regression coefficients

Using the notation from [4], the model can be written as

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{\varepsilon}_i \quad i = 1, \ldots, N \quad , \tag{1}$$

assuming that

i) $\boldsymbol{\varepsilon}_i \sim \mathcal{N}_{n_i}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{n_i})$ and uncorrelated for $i = 1, \ldots, N$,
ii) $\boldsymbol{b}_i \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Psi})$ and uncorrelated for $i = 1, \ldots, N$,
iii) $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_N$ are independent.

$\boldsymbol{\Psi} = \boldsymbol{\Psi}_{\boldsymbol{\theta}}$ is a covariance matrix where $\boldsymbol{\theta}$ is an unconstrained set of parameters (with dimension q^*) such that $\boldsymbol{\Psi}_{\boldsymbol{\theta}}$ is positive definite. From model (1) we deduce that $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ are independent and $\boldsymbol{y}_i \sim \mathcal{N}_{n_i}(\boldsymbol{X}_i \boldsymbol{\beta}, \boldsymbol{V}_i(\boldsymbol{\theta}, \sigma^2))$ with $\boldsymbol{V}_i(\boldsymbol{\theta}, \sigma^2) = \boldsymbol{Z}_i \boldsymbol{\Psi}_{\boldsymbol{\theta}} \boldsymbol{Z}_i^T + \sigma^2 \boldsymbol{I}_{n_i}$.

Denote the stacked vectors $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_N^T)^T$, $\boldsymbol{b} = (\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_N^T)^T$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1^T, \ldots, \boldsymbol{\varepsilon}_N^T)^T$ and the stacked matrices $\boldsymbol{X} = (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_N^T)^T$, $\boldsymbol{Z} = \text{diag}(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N)$ and $\boldsymbol{V} = \text{diag}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_N)$. Then model (1) can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon} \tag{2}$$

and the negative log-likelihood is given by

$$-\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = \frac{1}{2}\Big\{ N_T \log(2\pi) + \log|\boldsymbol{V}| + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \Big\} \tag{3}$$

2.2 ℓ_1-penalized maximum likelihood estimator

Since $N_T \ll p$, consider:

$$Q_\lambda(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) := \frac{1}{2}\log|\boldsymbol{V}| + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{k=2}^p |\beta_k| \quad , \tag{4}$$

β_1: unpenalized intercept, λ: nonnegative regularization parameter
Consequently,

$$\hat{\boldsymbol{\phi}} := (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}^2) = \underset{\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2 > 0, \boldsymbol{\Psi} > 0}{\text{argmin}} \; Q_\lambda(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) \quad , \tag{5}$$

which is a *nonconvex optimization problem!*

2.3 Prediction of the random effects

Predict the random-effects coefficients $\boldsymbol{b}_i$ by the *maximum a posteriori (MAP)* principle, which yields

$$\hat{\boldsymbol{b}}_i = [\boldsymbol{Z}_i^T \boldsymbol{Z}_i + \hat{\sigma}^2 \boldsymbol{\Psi}_{\hat{\boldsymbol{\theta}}}^{-1}]^{-1} \boldsymbol{Z}_i^T (\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}) \quad i = 1, \ldots, N \quad . \tag{6}$$

2.4 Selection of the regularization parameter

Use the Bayesian Information Criterion (BIC) defined by

$$-2\ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}^2) + \log N_T \cdot \hat{df}_\lambda \quad , \tag{7}$$

where $\hat{df}_\lambda := |\{1 \le k \le p; \hat{\beta}_k \neq 0\}| + \dim(\boldsymbol{\theta})$ is the sum of the number of the nonzero fixed regression coefficients and the number of variance components.

3. Coordinate Gradient Descent Algorithm

We calculate the estimator (5) by the coordinate gradient descent algorithm proposed in [7], and used in [3]. The key elements are:

- **Coordinatewise optimization.** Cycle through the coordinates and minimize the objective function $Q_\lambda(.)$ with respect to only one coordinate while keeping the other parameters fixed.
- **Quadratic approximation.** In each step, approximate $Q_\lambda(.)$ by a strictly convex quadratic function.
- **Inexact line search.** Calculate a descent direction and employ an inexact line search to ensure a decrease in the objective function.

Define

$$P(\boldsymbol{\phi}) := \sum_{k=2}^p |\beta_k| \quad , \quad g(\boldsymbol{\phi}) := \frac{1}{2}\log|\boldsymbol{V}| + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{8}$$

Then (5) can be written as

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\text{argmin}} \, Q_\lambda(\boldsymbol{\phi}) := g(\boldsymbol{\phi}) + \lambda P(\boldsymbol{\phi}) \tag{9}$$

Let $\boldsymbol{e}_j$ the jth unit vector.

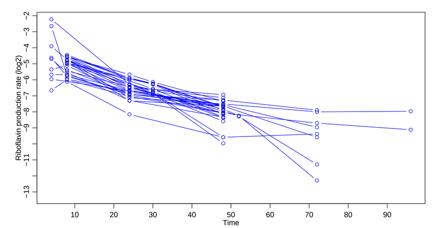
Algorithm. (Coordinate Gradient Descent)

(0) Let $\boldsymbol{\phi}^0 \in \mathbb{R}^{p+q^*+1}$ *be an initial value.*
For $\ell = 0, 1, 2, \ldots$, *let* $\mathcal{S}^\ell$ *be the index cycling through the coordinates* $\{1\}, \{2\}, \ldots, \{p+q^*\}, \{p+q^*+1\}$
(1) Approximate the second derivative $\frac{\partial^2}{\partial(\phi_{\mathcal{S}^\ell})^2} Q_\lambda(\boldsymbol{\phi}^\ell)$ *by* $h^\ell > 0$.
(2) Calculate the descent direction
$d^\ell := \text{argmin}_{d \in \mathbb{R}} \Big\{ g(\boldsymbol{\phi}^\ell) + \frac{\partial}{\partial \phi_{\mathcal{S}^\ell}} g(\boldsymbol{\phi}^\ell) d + 1/2 d^2 h^\ell + \lambda P(\boldsymbol{\phi}^\ell + d\boldsymbol{e}_{\mathcal{S}^\ell}) \Big\}$.
(3) Choose a stepsize $\alpha^\ell > 0$ *and set* $\boldsymbol{\phi}^{\ell+1} = \boldsymbol{\phi}^\ell + \alpha^\ell d^\ell \boldsymbol{e}_{\mathcal{S}^\ell}$ *such there is a decrease in the objective function.*
until convergence.

This algorithm is implemented in the `R` package `lmmlasso`, which is available from the first author's website (`http://stat.ethz.ch/people/schell`) and will be made available on `http://cran.r-project.org`.

4. Application: Riboflavin data

Data description. Data set provided by DSM (Switzerland), see also [1] response: logarithm of the riboflavin production rate of Bacillus subtilis
$p = 4088$ covariates measuring the gene expression levels
$N = 28$ groups with $n_i \in \{2, \ldots, 6\}$ and $N_T = 111$



Model. We fit a random-intercept model

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + b_{i1} + \varepsilon_{ij} \quad i = 1, \ldots, N, \quad j = 1, \ldots, n_i \tag{10}$$

with $b_{i1} \sim \mathcal{N}(0, \tau^2)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.
Result. Compare *lmmlasso*, *cv-Lasso* (standard Lasso using 10-fold cross-validation) and *Lasso* (standard Lasso using BIC).

	lmmLasso	cv-Lasso	Lasso
$\hat{\sigma}^2$	0.32	0.38	0.30
$\hat{\tau}^2$	0.05	-	-
$\lvert S(\hat{\boldsymbol{\beta}})\rvert$	22	18	21

Conclusions. We see that the total variability can be split into 13.2% *between-subject variability* and 86.8% *within-subject variability.*

5. Current Collaborations

We have three ongoing collaborations:

1) ETH Zurich, Prof. Dr. Sara van de Geer (project C3)
2) ZEW Mannheim, Dr. Stephan Dlugosz, Labour Markets, Human Resources and Social Policy (project A5)
3) TU Berlin, Machine Learning Group, Prof. Dr. Klaus-Robert Müller, Berlin Brain Computer Interface (BBCI)

6. Future Work

In the remaining of the first funding period, we are going to generalize the gaussian linear mixed-effects model to non-gaussian response variables, i.e. the logistic and poisson case. We will focus on theoretical as well as computational aspects and set up an `R` package called `glmmlasso`. This is again joint work with C3.

References

[1] Kalisch M. Bühlmann, P. and M.H. Maathuis. Variable selection in high-dimensional linear models: partially faithful distributions and the pc-simple algorithm. *Biometrika*, 97:261–278, 2010.

[2] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* to appear, 2011.

[3] L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society*, 70:53–71, 2008.

[4] José C. Pinheiro and Douglas M. Bates. *Mixed-Effects Models in S and S-Plus.* Springer, New York, 2000.

[5] J. Schelldorfer and P. Bühlmann. Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *arXiv preprint 1002.3784*, 2010.

[6] N. Städler, P. Bühlmann, and S. van de Geer. l_1-penalization for mixture regression models (with discussion). *Test*, Online First, 2010.

[7] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming, Series B*, 117:387–423, 2009.


Jürg Schelldorfer
Peter Bühlmann

Research Group FOR916
Statistical Regularization
Project C1

Statistical Regularization