# Variable Screening and Parameter Estimation for High-Dimensional Generalized Linear Mixed Models Using $\ell_1$-Penalization

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Jürg Schelldorfer and Peter Bühlmann

## 1. Overview

| | $n > p$ | $n \ll p$ |
|---|---|---|
| Generalized Linear Models (GLMs) | MLE [glm] | Lasso [glmnet] |
| Generalized Linear Mixed Models (GLMMs) | MLE [glmer] | GLMMLasso [glmmlasso] |

$n$: number of observations
$p$: number of variables

## 2. Generalized Linear Mixed Models and $\ell_1$-Penalized Estimation

### Classical Model Set-up

*Notation:*
$g = 1, ..., N$ independent groups/clusters/subjects
$j = 1, ..., n_g$ observations for group/cluster/subject $g$
$n = \sum_{g=1}^{N} n_g$ total number of observations

$\boldsymbol{y}$ : $n$-dim response variable
$\boldsymbol{b}$ : $q$-dim (correlated) random effects

$\boldsymbol{\beta} \in \mathbb{R}^p$ fixed-effects parameters
$\boldsymbol{\theta} \in \mathbb{R}^d$ covariance parameters
$\phi$ dispersion parameter

$\boldsymbol{X}$ : $n \times p$ model matrix for $\boldsymbol{\beta}$
$\boldsymbol{Z}$ : $n \times q$ model matrix for $\boldsymbol{b}$
$\boldsymbol{\Sigma_\theta}$ : $q \times q$ covariance matrix, determined by $\boldsymbol{\theta}$

*Model Assumptions (MA 1):*

- $y_i|\boldsymbol{b}$ are independent for $i = 1, \dots, n$
- $y_i|\boldsymbol{b}$ has a density of the form

$$\exp\left\{\phi^{-1}\Big(y_i\xi_i - b(\xi_i)\Big) + c(y_i, \phi)\right\} \text{ with } \boldsymbol{\mu}_i = \mathbb{E}[y_i|\boldsymbol{b}]$$

- $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$ with $\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zb}$
- $\boldsymbol{b} \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Sigma_\theta})$ with $\boldsymbol{\Sigma_\theta} \geq 0$ for $\boldsymbol{\theta} \in \mathbb{R}^d$

*Spherical Coordinates (Bates, 2011):*
Write $\boldsymbol{\Sigma_\theta} = \boldsymbol{\Lambda_\theta}\boldsymbol{\Lambda_\theta}^T$ and define $\boldsymbol{u}$ by $\boldsymbol{b} := \boldsymbol{\Lambda_\theta u}$ where $\boldsymbol{u} \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{1}_q)$.

*Parameter Estimation (Bates, 2011):*

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\phi})_{MLE} = \arg\min_{\boldsymbol{\beta},\boldsymbol{\theta},\phi} -\log L(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi)$$

where $L(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi)$ is the likelihood function.

### High-dimensional Model Set-up

*Model Assumptions (MA 2):*

- $n = \sum_{i=1}^{N} n_g \ll p$
- the true $\boldsymbol{\beta}_0$ is sparse
- $d$ small, say $d \leq 10$

$\hookrightarrow$ *Goal*: Assuming (MA 1) and (MA 2), estimate $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, $\phi$ and predict $\boldsymbol{b}$.

### The GLMMLasso Estimator

To cope with high-dimensionality and to enforce sparsity, we use a **Lasso-type penalty** (Tibshirani, 1996). Hence

$$Q_\lambda(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi) := -2\log L(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi) + \lambda\|\boldsymbol{\beta}\|_1, \qquad (1)$$

for $\lambda \leq 0$ and

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\phi}) := \arg\min_{\boldsymbol{\beta},\boldsymbol{\theta},\phi} Q_\lambda(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi).$$

In general, $L(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi)$ has no analytical form, so we use the **Laplace approximation** to approximate the integrand of $L(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi)$ by a quadratic function, i.e.

$$I = \int_{\mathbb{R}^q} e^{-S(\boldsymbol{u})} d\boldsymbol{u} \approx \tilde{I}^{LA} = (2\pi)^{q/2}|S''(\tilde{\boldsymbol{u}})|^{-1/2}e^{-S(\tilde{\boldsymbol{u}})},$$

where $\tilde{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}} S(\boldsymbol{u})$ is the mode of $-S(\boldsymbol{u})$.

The Laplace approximation of $Q_\lambda(.)$ in (1) is

$$\tilde{Q}_\lambda^{LA}(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi) = -2\sum_{i=1}^{n}\left\{\frac{y_i\xi_i(\boldsymbol{\beta}, \boldsymbol{\theta}) - b(\xi_i(\boldsymbol{\beta}, \boldsymbol{\theta}))}{\phi} + c(y_i, \phi)\right\} + \log|(\boldsymbol{Z\Lambda_\theta})^T\boldsymbol{W}_{\boldsymbol{\beta},\boldsymbol{\theta},\phi}(\boldsymbol{Z\Lambda_\theta}) + \boldsymbol{1}_q| + \|\tilde{\boldsymbol{u}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi)\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1,$$

where $\boldsymbol{W}_{\boldsymbol{\beta},\boldsymbol{\theta},\phi} = \text{diag}^{-1}\Big(\phi v(\mu_i(\boldsymbol{\beta}, \boldsymbol{\theta}))g'(\mu_i(\boldsymbol{\beta}, \boldsymbol{\theta}))^2\Big)_{i=1}^{n}$ and $v(.)$ is the variance function.

The GLMMLasso estimator is defined by

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\phi}) := \arg\min_{\boldsymbol{\beta},\boldsymbol{\theta},\phi} \tilde{Q}_\lambda^{LA}(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi) \qquad (2)$$

This is a high-dimensional, non-convex optimization problem!

## 3. Computational Algorithm

To calculate the GLMMLasso estimator (2), we use **coordinatewise optimization with inexact line search**, i.e. optimizing $\tilde{Q}_\lambda^{LA}(.)$ with respect to one coordinate keeping all other coordinates fixed (Tseng and Yun, 2009). An overview of the algorithm may be described as follows:

*GLMMLasso algorithm*

(0) Choose a starting value $(\boldsymbol{\beta}^{(0)}, \boldsymbol{\theta}^{(0)}, \phi^{(0)})$.
**Repeat** for $s = 1, 2, \dots$
(1) (Fixed-effects parameter optimization)
For $k = 1, \dots, p$
a) *(Laplace approximation)*
Calculate the Laplace approximation $\tilde{Q}_\lambda^{LA}(.)$ based on the current parameter estimates.
b) *(Quadratic approximation and inexact line search)*
i) Approximate the second derivative by $h_k^{(s)} > 0$.
ii) Calculate the descent direction $d_k^{(s)} \in \mathbb{R}$
iii) Choose a step size $\alpha_k^{(s)} > 0$ such that there is a decrease in the objective function.
(2) (Covariance parameter optimization)
For $l = 1, \dots, d$

$$\theta_l^{(s)} = \arg\min_{\theta_l} \tilde{Q}_\lambda^{LA}(.)$$

(3) (Dispersion parameter optimization)

$$\phi^{(s)} = \arg\min_{\phi} \tilde{Q}_\lambda^{LA}(.)$$

**until** convergence.

The above algorithm solves (2) exactly. In order to speed up the algorithm, we suggest an **approximate algorithm** comprising the following two parts:

- Regard $\tilde{\boldsymbol{u}}$ as fixed for the quadratic approximation with respect to the derivatives of $\beta_k$ in (1) b).
- Active-set algorithm: cycle through the non-zero coefficients $\beta_k^{(s)}$, and through all $p$ coefficients only every $D$th iteration.

## 4. Two-stage GLMMLasso estimator(s)

Apart from good variable selection properties accomplished by the Lasso, we advocate a two-stage procedure to get accurate parameter estimates. Hence the first stage aims at estimating a candidate set of variables (*variable screening*). The goal of the second step is unbiased parameter estimation (*parameter estimation*). Therefore we propse a **refitting by ML methods**. This two-stage procedure can be summarized as follows:

*Two-stage GLMMLasso*
Stage 1: Compute the GLMMLasso estimator (2).
Stage 2: Perform a ML method as in (3) or (4).

Let $(\hat{\boldsymbol{\beta}}_{init}, \hat{\boldsymbol{\theta}}_{init}, \hat{\phi}_{init})$ denote the estimate from (2).

### The GLMMLasso-MLE hybrid estimator

Define $\hat{S}_{hybrid} := \{k : |\hat{\beta}_{k,init} \neq 0\}$. Then

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\phi})_{hybrid} := \arg\min_{\boldsymbol{\beta}_{\hat{S}_{hybrid}},\boldsymbol{\theta},\phi} -2\log L(\boldsymbol{\beta}_{\hat{S}_{hybrid}}, \boldsymbol{\theta}, \phi) \qquad (3)$$

where for $S \subseteq \{1, \dots, p\}$, $(\boldsymbol{\beta}_S)_k = \beta_k$ if $k \in S$ and $(\boldsymbol{\beta}_S)_k = 0$ if $k \notin S$.

### The thresholded GLMMLasso estimator

The thresholded Lasso with refitting was examined in Geer et al (2010). Define $\hat{S}_{thres} := \{k : |\hat{\beta}_{k,init}| > \lambda_{thres}\}$. The thresholded GLMMLasso estimator is defined by

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\phi})_{thres} := \arg\min_{\boldsymbol{\beta}_{\hat{S}_{thres}},\boldsymbol{\theta},\phi} -2\log L(\boldsymbol{\beta}_{\hat{S}_{thres}}, \boldsymbol{\theta}, \phi) \qquad (4)$$

### Selection of the regularization parameters

For the choice of the regularization parameters $\lambda$ and $\lambda_{thres}$, we propose the BIC and the AIC

$$c_{n,\lambda} = -2\log L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\phi}) + a(n) \cdot \hat{df}_\lambda$$

where $a(n) = \log(n)$ for the BIC, $a(n) = 2$ for the AIC and $\hat{df}_\lambda = |\{1 \leq k \leq p : \hat{\beta}_k \neq 0\}| + \dim(\hat{\boldsymbol{\theta}})$.

## 6. Illustration

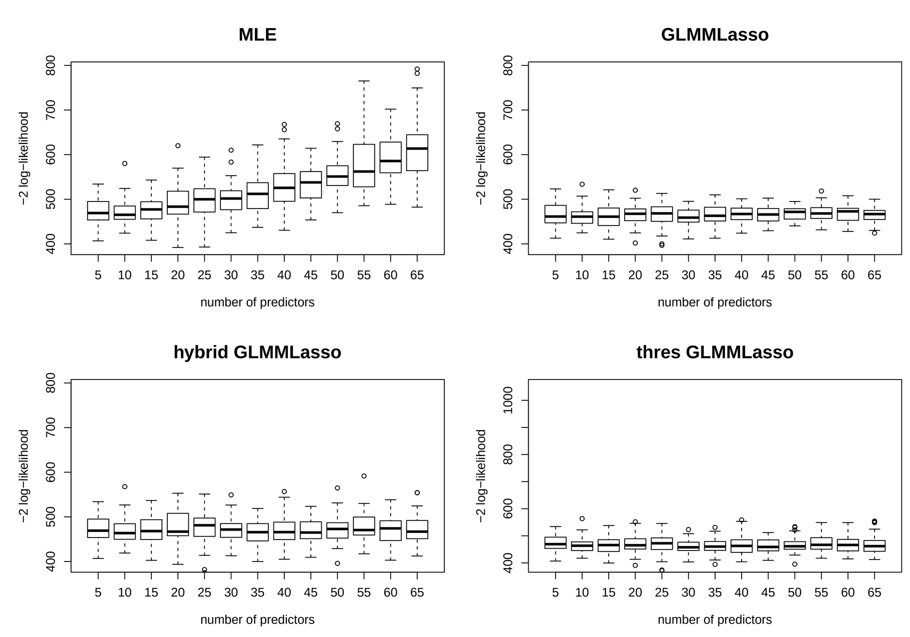


Figure 1: Minus twice out-of-sample log-likelihood for a growing number of covariates. The MLE performs badly whereas the GLMMLasso estimators remain stable. We use a random-intercept logistic mixed model with $n = 400$, $N = 40$, $n_g = 10$, $\theta^2 = 1$, $\boldsymbol{\beta}_0 = (0, 1, -1, 1, -1)$.

## 6. R package

An implementation of the algorithm will be available online in the R package glmmlasso. The Gaussian case is implemented in the standalone R package lmmlasso (Schelldorfer et al, 2011), which is available from CRAN.

## References

Bates D. M. (2011) Computational Methods for Mixed Models. available at http://cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf

Tibshirani R. (1996) Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society, Series B 58:267–288

Tseng P. and Yun S. (2009) A Coordinte Gradient Descent Method for Nonsmooth Separable Minimization. Mathematical Programming 117:387–423

Schelldorfer J., Bühlmann P., van de Geer S. (2011) Estimation for High-Dimensional Linear Mixed-Effects Models Using $\ell_1$-penalization. The Scandinavian Journal of Statistics 38:197–214

van de Geer S, Bühlmann P, Zhou S. (2010) The Adaptive and the Thresholded Lasso for Potentially Misspecified Models. arXiv:1001.5176v3