High-Dimensional Gaussian and Generalized Linear Mixed Models

Jürg Schelldorfer

Seminar für Statistik

December 6, 2011

Riboflavin Production in Bacillus Subtilis

A data set provided by DSM Nutritional Products

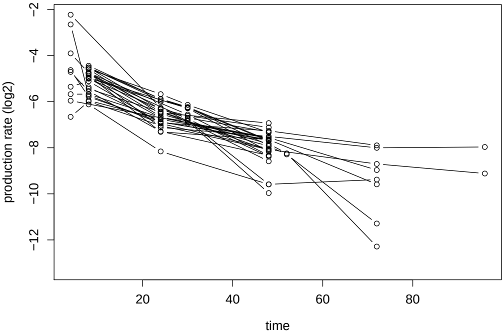
Goal:
improve riboflavin production rate by genetic engineering

Data:
response variable $Y \in \mathbb{R}$: riboflavin (log-)production rate
covariates $X \in \mathbb{R}^p$: expressions from genes
$n = 111$ observations and $p = 4088$ variables

$\hookrightarrow$ "simple" high-dimensional regression problem, but...

Riboflavin Production in Bacillus Subtilis

...we know more about the data...

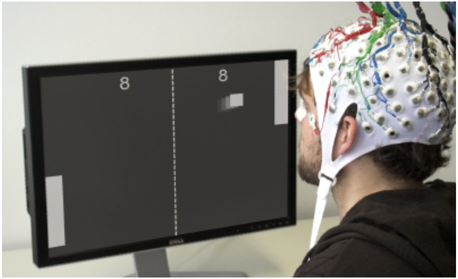


$n = 111$ observations and $p = 4088$ variables
$N = 28$ groups with $\{2, \ldots, 6\}$ observations per group

$\hookrightarrow$ high-dimensional longitudinal data

Brain Computer Interface (BCI)

- BCI is muscle-independent communication



- measure EEG signals at different locations on the scalp

Data:

- 83 subjects
- 150 trials per subject
- response $Y \in \{\text{left}, \text{right}\}$

this results in

$$\boldsymbol{y} \in \mathbb{R}^{12'450}, \boldsymbol{X} \in \mathbb{R}^{12'450 \times 1494}$$

Goal:
variable selection

$\hookrightarrow$ grouped data with many covariates

Administrative Data

A data set about employment from the Centre for European Economic Research (ZEW Mannheim)

Data:
binary response variable $Y \in \{\text{employed}, \text{unemployed}\}$

covariates X: income, sex, age group, employment duration,....

quarterly results of (Y, X) of many workers over several years

$n \approx 120'000$ and $p \approx 30'000$

Goal:
variable selection

$\hookrightarrow$ longitudinal data with many covariates

General Framework

Data:

$\hookrightarrow$ truly high-dimensional grouped data

$\hookrightarrow$ grouped data with many covariates

Goals:

1. variable selection
2. parameter estimation

Table of Contents

1. Motivating Examples
2. Context
3. High-Dimensional (Gaussian) Mixed-Effects Models
4. High-Dimensional Generalized Linear Mixed Models
5. Summary

Context

Parameter estimation:

	$n > p$	$n \ll p$
Linear Models	Least Squares [`lm`]	Lasso [`lars`]
Linear Mixed-Effects Models	ML or REML [`lmer`]	?

n: number of observations
p: number of variables

(Classical) Linear Model

For n independent observations

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i \quad i = 1, \ldots, n$$

with ε_i inpedendent and $\mathbb{E}[\varepsilon_i] = 0$

if $rank(\boldsymbol{X}) = p$, the Least Squares estimator is

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \|_2^2 = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

The Lasso Estimator

For $n \ll p$ we should not use the LS estimator. Use the Lasso (Tibshirani, 1996)

$$\hat{\beta}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\} \quad \lambda \geq 0$$

or equivalently

$$\hat{\beta}(s) = \arg\min_{\boldsymbol{\beta}, \|\boldsymbol{\beta}\|_1 \leq s} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$

with the following properties:

- some coefficients $\hat{\beta}_j(\lambda)$ are exactly zero
- convex optimization problem

Linear Mixed-Effects Model

Model equation

Inhomogeneous data:
for $i = 1, \ldots, N$ independent units/groups, $j = 1, \ldots, n_i$ observations

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \beta + \boldsymbol{Z}_i \boldsymbol{b}_i + \varepsilon_i \quad i = 1, \ldots, N$$

$\boldsymbol{Y}_i$: n_i-dim response vector

$\boldsymbol{X}_i$: $n_i \times p$ fixed-effects design matrix
$\boldsymbol{Z}_i$: $n_i \times q$ random-effects design matrix

β : fixed-effects coefficient
$\boldsymbol{b}_i$: random-effects coefficients

ε_i : n_i-dim error vector

- $\varepsilon_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2 \boldsymbol{I}_{n_i})$ and uncorrelated
- $\boldsymbol{b}_i \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Psi}_{\boldsymbol{\theta}})$ and uncorrelated
- $\varepsilon_1, \ldots, \varepsilon_N, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_N$ independent

$\boldsymbol{\Psi}_{\boldsymbol{\theta}} > 0$ is parametrized by $\boldsymbol{\theta} \in \mathbb{R}^d$

parameter estimation:

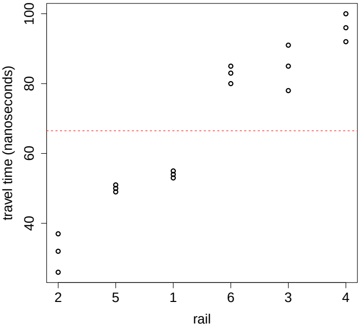
$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}^2)_{ML} = \underset{\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2 > 0, \boldsymbol{\Psi} > 0}{\arg\min} \; -\ell_{ML}(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)$$

Linear Mixed-Effects Model

Example from Pinheiro and Bates (2000)

Travel time of a specific type of wave through the length of railway rails

6 rails, 3 measurements per rail



model:

$$y_{ij} = \beta + b_i + \varepsilon_{ij} \quad i = 1, \ldots, 6, \quad j = 1, \ldots, 3$$

with $b_i \sim \mathcal{N}(0, \theta^2)$ independent of $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$

Recap

Parameter estimation:

	$n > p$	$n \ll p$
Linear Models	Least Squares [lm] ✔	Lasso [lars] ✔
Linear Mixed-Effects Models	ML or REML [lmer] ✔	!

High-Dimensional (Gaussian) Mixed-Effects Models

High-Dimensional Model Set-up

Additionally to a linear mixed-effects model, assume

- the true $\boldsymbol{\beta}_0$ is sparse
- $d = dim(\boldsymbol{\theta})$ small

Aim: Estimate $\boldsymbol{\beta}, \boldsymbol{\theta}$, σ^2 and predict $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N$

The LMMLasso Estimator

Objective function:

$$Q_\lambda(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) := \underbrace{\frac{1}{2} \sum_{i=1}^{N} \left\{ \log(|\boldsymbol{V}_i|) + (\boldsymbol{Y}_i - \boldsymbol{X}_i \beta)^T \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i \beta) \right\}}_{\text{non-convex loss function } \boldsymbol{\rho}} + \underbrace{\lambda \|\boldsymbol{\beta}\|_1}_{\text{convex penalty}}$$

$$:= \boldsymbol{\rho}(\cdot) + \lambda pen(\cdot)$$

where

$$\boldsymbol{V}_i = \boldsymbol{Z}_i \Psi_{\boldsymbol{\theta}} \boldsymbol{Z}_i^T + \sigma^2 \boldsymbol{I}_{n_i} \quad i = 1, \ldots, N$$

LMMLasso estimate:

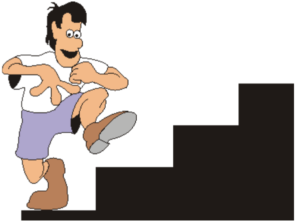
$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}^2) = \underset{\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2}{\arg\min}\, Q_\lambda(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)$$

Questions to Address

1. How to compute the LMMLasso estimator?
2. Numerical convergence of the algorithm?
3. Theoretical properties of the LMMLasso estimator?
4. Prediction of the random effects?
5. How to perform model selection?

Major Challenge

Make the step



convex $\longrightarrow$ non-convex

in Computation and Theory!

How to calculate

$$\hat{\phi} := (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}^2) := \underset{\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2}{\arg\min}\, Q_\lambda(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)?$$

use a coordinate gradient descent algorithm!

i.e. optimize Q_λ w.r.t. one coordinate keeping all other coordinates fixed

key elements (Tseng and Yun, 2009):

- **Gauss-Seidel coordinatewise optimization**
- **Quadratic approximation** of the loss function
- **Inexact line search** using the Armijo rule

Computational Algorithm

Gauss-Seidel coordinatewise optimization

$$
\begin{aligned}
\boldsymbol{\phi}^T = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T, \sigma^2) &= ({\color{red}\beta_1}, \ldots, \beta_p, \theta_1, \ldots, \theta_d, \sigma^2) \in \mathbb{R}^{p+d+1} \\
&\ldots \\
&= (\beta_1, \ldots, {\color{red}\beta_p}, \theta_1, \ldots, \theta_d, \sigma^2) \\
&= (\beta_1, \ldots, \beta_p, {\color{red}\theta_1}, \ldots, \theta_d, \sigma^2) \\
&\ldots \\
&= (\beta_1, \ldots, \beta_p, \theta_1, \ldots, {\color{red}\theta_d}, \sigma^2) \\
&= (\beta_1, \ldots, \beta_p, \theta_1, \ldots, \theta_d, {\color{red}\sigma^2}) \\
&= ({\color{red}\beta_1}, \ldots, \beta_p, \theta_1, \ldots, \theta_d, \sigma^2) \\
&\ldots \\
&= (\beta_1, \ldots, \beta_p, \theta_1, \ldots, \theta_d, {\color{red}\sigma^2}) \\
&= ({\color{red}\beta_1}, \ldots, \beta_p, \theta_1, \ldots, \theta_d, \sigma^2) \\
&\ldots
\end{aligned}
$$

Computational Algorithm

loss function $\rho(\phi)$, penalty $pen(\phi)$, e_j unit vector

LMMLasso algorithm

0. $\phi^0 \in \mathbb{R}^{p+d+1}$ *an initial value*

For $\ell = 0, 1, 2, \ldots$, *let* $\mathcal{S}^\ell$ *cycling through* $\{1\}, \{2\}, \ldots, \{p+d+1\}$

1. *Quadratic approximation*
   a) *Calculate the derivative* $\nabla \rho$
   b) *Choose an appropriate hessian* $h^\ell > 0$
2. *Inexact line search*
   a) *Calculate the descent direction*
   b) *Choose a stepsize* $\alpha^\ell > 0$ *by the Armijo rule and set*

$$\phi^{\ell+1} = \phi^\ell + \alpha^\ell d^\ell e_{\mathcal{S}^\ell}$$

until convergence

Computational Algorithm

The Armijo rule is defined as follows:

Armijo rule

Choose the stepsize $\alpha^\ell > 0$ *in a way such that*

$$Q_\lambda(\phi^\ell + \alpha^\ell d^\ell e_{\mathcal{S}^\ell}) \leq Q_\lambda(\phi^\ell) + \alpha^\ell \xi^\ell$$

for ξ^ℓ *depending on* $\rho, pen, d^\ell, h^\ell, \phi^\ell$ *and some constants.*

Numerical Convergence

Theorem

If $(\phi^\ell)_{\ell \geq 0}$ is chosen according to the LMMLasso algorithm, then every cluster point of $(\phi^\ell)_{\ell \geq 0}$ is a stationary point of $Q_\lambda(\phi)$.

remarks:

- convergence can be slow
- result depends on the starting value

Theoretical Results

Notation

Set $n_C = n_i$ fixed for $i = 1, \ldots, N$

$\boldsymbol{Y}_i \in \mathcal{Y} \subset \mathbb{R}^{n_C}$, $\boldsymbol{X}_i \in \mathcal{X}^{n_C} \subset \mathbb{R}^{n_C \times p}$

define the parameter

$$\phi^T := (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T) = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T, 2\log\sigma) = (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T) \in \mathbb{R}^{p+d+1}$$

and the parameter space for $K > 0$

$$\mathbf{\Phi} = \{(\boldsymbol{\beta}^T, \boldsymbol{\eta}^T); \sup_{x \in \mathcal{X}} |x^T \boldsymbol{\beta}| \leq K, \|\boldsymbol{\eta}\|_\infty \leq K, \mathbf{\Psi}_{\boldsymbol{\eta}} > 0\} \subset \mathbb{R}^{p+d+1}$$

LMMLasso estimator

$$\hat{\phi} := \arg\min_{\phi \in \mathbf{\Phi}} \left\{ \boldsymbol{\rho}(\boldsymbol{\beta}, \boldsymbol{\eta}) + \lambda \|\boldsymbol{\beta}\|_1 \right\}$$

$\{f_\phi, \phi \in \Phi\}$ Gaussian density, ϕ_0 true parameter vector

excess risk

$$\mathcal{E}_{\boldsymbol{X},\boldsymbol{Z}}(\phi|\phi_0) := \int \log\left(\frac{f_{\phi_0,\boldsymbol{X},\boldsymbol{Z}}}{f_{\phi,\boldsymbol{X},\boldsymbol{Z}}}\right) f_{\phi_0,\boldsymbol{X},\boldsymbol{Z}}\, d\mu$$

for fixed $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$

average excess risk

$$\overline{\mathcal{E}}(\phi|\phi_0) := \frac{1}{N}\sum_{i=1}^{N} \mathcal{E}_{\boldsymbol{X}_i,\boldsymbol{Z}_i}(\phi|\phi_0)$$

Consistency

Statement

Theorem

Under some regularity conditions on $\mathbf{Z}_i$ *and assuming that*

$$\|\beta_{0,N}\|_1 = o\left(\sqrt{\frac{N}{\log^4 N \log(p \vee N)}}\right) \quad \lambda_N = C\sqrt{\frac{\log^4 N \log(p \vee N)}{N}}$$

for some $C > 0$*, any global minimizer* $\hat{\phi}$ *satisfies*

$$\overline{\mathcal{E}}(\hat{\phi}|\phi_0) = o_P(1) \quad (N \longrightarrow \infty)$$

for n_C *fixed.*

- linear models: $\|\beta_0\|_1 = o\left(\sqrt{\frac{n}{\log p}}\right)$
- finite mixture models: $\|\beta_0'\|_1 = o\left(\sqrt{\frac{n}{\log^3 n \log(p \vee n)}}\right)$
- key argument: non-central $\chi^2_\nu(\delta)$-distribution

an oracle inequality can be established as well

Prediction of the random effects

maximum a posteriori (MAP) estimate

$$\tilde{\boldsymbol{b}}_i = \arg\max_{\boldsymbol{b}_i} f(\boldsymbol{b}_i | \boldsymbol{y}_i, \beta, \boldsymbol{\theta}, \sigma^2)$$
$$= [\boldsymbol{Z}_i^T \boldsymbol{Z}_i + \sigma^2 \Psi_{\boldsymbol{\theta}}^{-1}]^{-1} \boldsymbol{Z}_i^T (\boldsymbol{Y}_i - \boldsymbol{X}_i \beta)$$

hence

$$\hat{\boldsymbol{b}}_i = [\boldsymbol{Z}_i^T \boldsymbol{Z}_i + \hat{\sigma}^2 \Psi_{\hat{\boldsymbol{\theta}}}^{-1}]^{-1} \boldsymbol{Z}_i^T (\boldsymbol{Y}_i - \boldsymbol{X}_i \hat{\beta})$$

Model Selection

- **Choice of the tuning parameter λ**
  use a grid of λ-values and select the optimal λ to be

$$\lambda^* = \arg\min_{\lambda_k} BIC(\lambda_k)$$

  or:
  mAIC, cAIC, mBIC, GIC,...

- **Selection of the random effects structure**
  assume the random effects structure is known
  assume $q < n$ small

Riboflavin Production in Bacillus Subtilis

Gaussian linear mixed model:

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + b_i^{k_1} z_{ij}^{k_1} + b_i^{k_2} z_{ij}^{k_2} + \varepsilon_{ij} \quad i = 1, \ldots, N, \quad j = 1, \ldots, n_i$$

with $b_i^{k_1} \sim \mathcal{N}(0, \theta_{k_1}^2)$, $b_i^{k_2} \sim \mathcal{N}(0, \theta_{k_2}^2)$, $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ mutually independent

conclusions:

- variability between groups ($\hat{\sigma}^2 = 0.15$, $\hat{\theta}_{k_1}^2 = 0.03$, $\hat{\theta}_{k_2}^2 = 0.06$)
- one dominating gene

Summary LMMLasso

- Coordinate gradient descent algorithm
- Numerical convergence to a stationary point
- Consistency of the LMMLasso estimator

High-Dimensional Generalized Linear Mixed Models

Generalized Linear Model (GLM)

For n realisations of (Y_i, X_i)

- (y_i, x_i^T) independent for $i = 1, \ldots, n$
- y_i has density from exponential family

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

- $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$ with $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$

estimate $\boldsymbol{\beta}$ by

$$\hat{\boldsymbol{\beta}}_{MLE} = \arg\min_{\boldsymbol{\beta}} -\ell(\boldsymbol{\beta})$$

$g = 1, \dots, N$ independent units/groups
$j = 1, \dots, n_g$ observations for unit/group 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
ϕ dispersion parameter

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

Generalized Linear Mixed Model (GLMM)

Model assumptions

- $y_i|\boldsymbol{b}$ are independent for $i = 1, \ldots, n$
- $y_i|\boldsymbol{b}$ has density from exponential family

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

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

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

High-Dimensional GLMM Set-up

Additionally to a GLMM, assume

- the true $\boldsymbol{\beta}_0$ is sparse
- $d = dim(\boldsymbol{\theta})$ small

Aim: Estimate $\boldsymbol{\beta}, \boldsymbol{\theta}$, ϕ and predict $\boldsymbol{b}$

☛ sparse fixed-effects estimates

The GLMMLasso Estimator

Key Idea 1: Lasso-type penalty

objective function

$$Q_\lambda(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi) := -2 \log L(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi) + \lambda \|\boldsymbol{\beta}\|_1 \qquad \lambda \geq 0$$

estimate $(\boldsymbol{\beta}, \theta, \phi)$ by

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

☛ in general, $L(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi)$ cannot be computed explicitly

The GLMMLasso Estimator

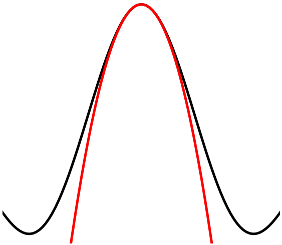
Key Idea 2: Laplace approximation

Calculate

$$I = \int_{\mathbb{R}^q} e^{S(\boldsymbol{b})} d\boldsymbol{b}$$

Idea: approximate $S(\boldsymbol{b})$ by a quadratic function at the mode

$$\tilde{\boldsymbol{b}} := \arg\max_{\boldsymbol{b}} S(\boldsymbol{b})$$



The GLMMLasso estimator

Then

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

with $\tilde{\boldsymbol{b}} = \arg\max_{\boldsymbol{b}} S(\boldsymbol{b})$

hence

$$Q_\lambda(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi) \rightsquigarrow Q_\lambda^{LA}(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi)$$

GLMMLasso estimator:

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

☛ high-dimensional, non-convex optimization problem

The GLMMLasso Algorithm

Key Idea 3: coordinatewise optimization with inexact line search

GLMMLasso algorithm

0. $\psi^0 \in \mathbb{R}^{p+d+1}$ *an initial value*

Repeat *for* $s = 0, 1, 2, \dots$

1. *Fixed-effects parameter optimization*
   For $k = 1, \dots, p$
   a) *Laplace approximation*
   b) *Quadratic approximation and inexact line search*
2. *Covariance parameter optimization*
3. *Dispersion parameter optimization*

until *convergence*

Approximate Algorithm

- ☛ Step 1. b) Quadratic approximation and inexact line search is computationally expensive, since for the mode

$$\tilde{\boldsymbol{b}} = \tilde{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi)$$

Key Idea 4: Approximate algorithm

regard $\tilde{\boldsymbol{b}}$ as fixed for quadratic approximation
$\rightsquigarrow$ remarkably reduction in speed ($\approx$ 50%)
$\rightsquigarrow$ slightly biased parameter estimates

- Bias from the Lasso as well as the approximate algorithm

Stage 1: Variable Screening by GLMMLasso
imposed by the Lasso (variable selection too restrictive)

Stage 2: Parameter Estimation
Key Idea 5: Refitting by ML
with the selected (non-zero) variables to get accurate parameter estimates

$\hat{S}$ the set of selected variables, $\boldsymbol{\beta}_{init}$ fixed effect from Stage 1

1. The GLMMLasso^{LA}-MLE hybrid estimator:

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

   $$(\hat{\boldsymbol{\beta}}^{LA}, \hat{\boldsymbol{\theta}}^{LA}, \hat{\phi}^{LA})_{hybrid} := \underset{\boldsymbol{\beta}_{\hat{S}_{init}}, \boldsymbol{\theta}, \phi}{\arg\min} -2 \log L(\boldsymbol{\beta}_{\hat{S}_{init}}, \boldsymbol{\theta}, \phi)$$

2. The thresholded GLMMLasso^{LA} estimator:

   $\hat{S} = \hat{S}_{thres} := \{k : |\hat{\beta}_{init,k}| > \lambda_{thres}\}$

   $$(\hat{\boldsymbol{\beta}}^{LA}, \hat{\boldsymbol{\theta}}^{LA}, \hat{\phi}^{LA})_{thres} := \underset{\boldsymbol{\beta}_{\hat{S}_{thres}}, \boldsymbol{\theta}, \phi}{\arg\min} -2 \log L(\boldsymbol{\beta}_{\hat{S}_{thres}}, \boldsymbol{\theta}, \phi)$$

Summary GLMMLasso

- Approximate the likelihood using the Laplace approximation
- Coordinatewise optimization
- Approximate algorithm to speed up
- Refitting by maximum likelihood methods

Take-Home Message

Part I: some words

- Truly high-dimensional generalized linear mixed models
- Non-convex loss function and convex penalty
- Fast computational algorithms
- Theoretical results for the Gaussian case

Take-Home Message

Part II: “picture”

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

Acknowledgments

It has been a great honor to work with

Peter Bühlmann

Sara van de Geer

... and I thank all members of the Seminar für Statistik!

Questions?

References

- R. Tibshirani ; Regression Shrinkage and Selection via the Lasso ; J. R. Stat. Soc. (1996)
- P. Tseng and S. Yun ; A Coordinte Gradient Descent Method for Nonsmooth Separable Minimization ; Mathematical Programming (2009)
- ℓ_1-Penalization for Mixture Regression Models (with discussion); N. Städler, P. Bühlmann, S. van de Geer ; Test ; (2010)
- D. Bates ; Computational Methods for Mixed Models ; Vignette for lme4 (2011)

Extensions / Future Work

We see the following points as interesting for further research:

- group Lasso penalty
- elastic net penalty
- nonlinear mixed models
- Selection of random effects
- Model Selection Criteria

(Assumption 2)

(a) Let $\left(\omega_j^{(i)}\right)_{j=1}^{n}$ be the eigenvalues of $\boldsymbol{Z}_i \Psi \boldsymbol{Z}_i^T$ for $i = 1, \ldots, N$. At least two eigenvalues are different, i.e. for all i $\exists j_1 \neq j_2 \in \{1, \ldots, n\}$ such that $\omega_{j_1}^{(i)} \neq \omega_{j_2}^{(i)}$.

(b) For $i = 1, \ldots, N$, the matrices $\boldsymbol{\Omega}_i$ defined by

$$(\boldsymbol{\Omega}_i)_{r,s} = \operatorname{tr}\left(\boldsymbol{V}_i^{-1} \frac{\partial \boldsymbol{V}_i}{\partial \phi_{p+r}} \boldsymbol{V}_i^{-1} \frac{\partial \boldsymbol{V}_i}{\partial \phi_{p+s}} \right) \quad r, s = 1, \ldots, q^* + 1$$

are strictly positive definite.

(Restricted Eigenvalue Condition)

There exists a constant $\kappa \geq 1$, such that for all $\boldsymbol{\beta} \in \mathbb{R}^p$ satisfying $\|\boldsymbol{\beta}_{S_0^c}\|_1 \leq 6\|\boldsymbol{\beta}_{S_0}\|_1$ it holds that $\|\boldsymbol{\beta}_{S_0}\|_2^2 \leq \kappa^2 \boldsymbol{\beta}^T \boldsymbol{\Sigma}_{N,n} \boldsymbol{\beta}$.

Theorem

Consider the weighted ℓ_1-penalized estimator. Suppose that for some $\delta > 0$,

$$w_k \begin{cases} \leq 1/\delta & k \in S_0, \\ \geq 1/\delta & k \notin S_0. \end{cases}$$

Under Assumptions 1, 2 and 3, for $\lambda \geq 2T\delta\lambda_0$ and a constant c_0, we have on the set $\mathcal{J}$ defined in (A.6),

$$\bar{\mathcal{E}}(\hat{\phi}_{weight}|\phi_0) + 2(\lambda/\delta - T\lambda_0)\|\hat{\beta}_{weight} - \beta_0\|_1 \leq 9(\lambda/\delta + T\lambda_0)^2 c_0^2 \kappa^2 s_0,$$

Likelihood of GLMMs

For $\xi_i(\mu_i) = \xi_i(\boldsymbol{\beta}, \boldsymbol{\theta})$, the likelihood function of a GLMM is

$$
\begin{aligned}
L(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi) &= \int_{\mathbb{R}^q} \prod_{i=1}^{n} \left[\exp\left\{ \frac{y_i \xi_i(\boldsymbol{\beta}, \boldsymbol{\theta}) - b(\xi_i(\boldsymbol{\beta}, \boldsymbol{\theta}))}{\phi} + c(y_i, \phi) \right\} \right] \\
&\quad \times \frac{1}{(2\pi)^{q/2}} \exp\Big\{ -\frac{1}{2} \|\boldsymbol{u}\|_2^2 \Big\} d\boldsymbol{u} \\
&= \frac{1}{(2\pi)^{q/2}} \int_{\mathbb{R}^q} \exp\left\{ \sum_{i=1}^{n} \Big(\frac{y_i \xi_i(\boldsymbol{\beta}, \boldsymbol{\theta}) - b(\xi_i(\boldsymbol{\beta}, \boldsymbol{\theta}))}{\phi} + c(y_i, \phi) \Big) \right. \\
&\quad \left. - \frac{1}{2} \|\boldsymbol{u}\|_2^2 \right\} d\boldsymbol{u}.
\end{aligned}
$$