Estimation for a high-dimensional Mixed-Effects Model using ℓ_1-constraints

Jürg Schelldorfer

joint work with Peter Bühlmann

Seminar für Statistik, ETH Zürich

September 25, 2009

Overview

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

Table of Contents

1. Introduction
2. Linear Mixed-Effects Models and ℓ_1-penalized estimation
3. Theoretical Result: Consistency
4. Numerical Algorithm
5. Simulation Study and Real Data Example

(classical) Multiple Linear Regression

For N independent observations

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i \quad i = 1, ..., N \quad \epsilon_i \sim i.i.d.$$

Assuming $N > p$ and the design matrix has full rank, the LS estimator for β is

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Lasso estimator

For $N << p$ we should not use the LS estimator. We can use the Lasso (Tibshirani, 1996)

$$\hat{\beta}(\lambda) = argmin_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

or equivalently

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

with the following properties:

- The Lasso does variable selection (i.e. some coefficients are set exactly to zero)
- Convex optimiziation problem, which can be solved efficiently

Linear Mixed-Effects Model

Inhomogeneous data:
For $i = 1, ..., N$ independent units, $j = 1, .., n_i$ observations

$$y_{ij} = x_{ij}^T b + z_{ij}^T \beta_i + \epsilon_{ij}$$

or in matrix notation

$$Y_i = X_i b + Z_i \beta_i + \epsilon_i \quad i = 1, .., N$$

Y_i : n_i-dim response vector
X_i : $n_i \times p$ matrix of covariates with fixed effects $b \in \mathbb{R}^p$
Z_i : $n_i \times q$ matrix of covariates with random effects $\beta_i \in \mathbb{R}^q$
ϵ_i : n_i-dim vector of errors.

Linear Mixed-Effects Model

Assumptions:

$$\beta_i \sim N_q(0, G) \quad \epsilon_i \sim N_{n_i}(0, \sigma^2 V_i)$$

and both independent within and across units.

The fixed effects b, the random effects β_i and the covariance parameters (in G and V_i) are estimated by using ML or REML.

Recap

	$N_{Tot} > p$	$N_{Tot} << p$
Linear Regression	Ordinary Least Squares	Lasso
Linear Mixed-Effects Models	Maximum Likelihood (ML) Restricted Maximum Likelihood (REML)	?

High-dimensional Model Set-up

$i = 1, ..., N$ being N independent groups
$j = 1, .., n_i$ observations per group.
$N_{Tot} = \sum_{i=1}^{N} n_i << p$.

$$Y_i = X_i b + Z_i \beta_i + \epsilon_i \qquad i = 1, ..., N$$

and assume

$$\beta_i \sim N_q(0, \tau^2 \mathbf{I}) \quad \epsilon_i \sim N_{n_i}(0, \sigma^2 \mathbf{I})$$

and both being independent within and across groups.

Aim: Estimate $b, \sigma^2, \tau^2, \beta_1, .., \beta_N$

ℓ_1-penalized Maximum Likelihood Estimator

From the likelihood function, estimate the parameters b, σ^2 and τ^2 by minimizing

$$Q_\lambda(b, \sigma^2, \tau^2) := \frac{1}{2} \sum_{i=1}^{N} \left\{ \log(|\Lambda_i|) + (Y_i - X_i b)^T \Lambda_i^{-1} (Y_i - X_i b) \right\} + \lambda \|b\|_1$$
$$:= g(b, \sigma^2, \tau^2) + \lambda \|b\|_1$$

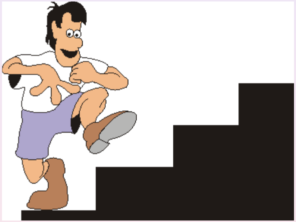
where

$$\Lambda_i = \sigma^2 \mathbf{I} + \tau^2 Z_i Z_i^T \quad i = 1, ..., N$$

$$\hat{b}, \hat{\sigma}^2, \hat{\tau}^2 = argmin_{b,\sigma^2,\tau^2} Q_\lambda(b, \sigma^2, \tau^2)$$

Major Challenge

Make the step



convex $\longrightarrow$ nonconvex

in Computation and Theory!

Estimation of the random effects

Employ the maximum a posteriori (MAP) estimate. Let $Y_i|\beta_i \sim h_1 d\mu$ and $\beta_i \sim h_2 d\mu$, then

$$\hat{\beta}_i = argmax_{\beta_i}\Big\{ \log h_1(Y_i|\beta_i) + logh_2(\beta_i)\Big\}$$
$$= [Z_i^T Z_i + \frac{\sigma^2}{\tau^2}\mathbf{I}_{q\times q}]^{-1} Z_i^T r_i$$

with

$$r_i := (Y_i - X_i b)$$

which corresponds to a Ridge Regression with $\lambda_{Ridge} = \frac{\sigma^2}{\tau^2}$.

Model Selection

- **Choice of the tuning parameter λ**
  Choose a λ-sequence $\lambda_1 < ... < \lambda_K$ and select the optimal λ to be
  $$\lambda^* = argmin_\lambda BIC(\lambda)$$
  or:
  mAIC, cAIC, GIC,... many other suggestions
- **Selection of the random effects**
  We assume that the variables having a random effect are known.
  How to find them, still an open problem...
  We assume that $q << p$.

Notation

Let $i = 1, .., N$ as before and set $n = n_i$ fixed.
$Y \in \mathcal{Y} \subset \mathbb{R}^n$, $X \in \mathcal{X}^n \subset \mathbb{R}^{n \times p}$
Define the parameter

$$\theta^T := (b^T, \eta^T) = (b^T, 2\log\sigma, 2\log\tau)$$

and the parameter space

$$\Theta = \{(b^T, \eta^T); \sup_{x \in \mathcal{X}} |x^T b| \le K, \|\eta\|_\infty \le K\} \subset \mathbb{R}^{p+2} \quad \text{for some } K > 0$$

Let $\{f_\theta, \theta \in \Theta\}$ be the density with respect to the new parametrization.

Notation

Define the so-called excess risk

$$\mathcal{E}(\theta|\theta_0) := \int \log\left[\frac{f_{\theta_0}}{f_\theta}\right] f_{\theta_0} d\mu$$

and for fixed $X_1, ..., X_N, Z_1, ..., Z_N$ we define the average excess risk

$$\overline{\mathcal{E}}(\theta|\theta_0) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{E}\Big(\theta(X_i, Z_i)|\theta_0(X_i, Z_i)\Big)$$

Rewrite our penalized estimator:

$$\hat{\theta}_\lambda := argmin_{\theta\in\Theta}\Big\{-\sum_{i=1}^{N} logf_\theta(Y_i, X_i, Z_i) + \lambda\|b\|_1\Big\}$$

Consistency

Theorem

Under some regularity conditions and assuming that

$$\|b\|_1 = o\left(\sqrt{\frac{N}{log^5 N}}\right) \quad \lambda = C\sqrt{\frac{log^5 N}{N}} \quad \textit{for some } C > 0$$

Then for $\hat{\theta}_\lambda$ *holds*

$$\overline{\mathcal{E}}(\hat{\theta}_\lambda|\theta_0) \xrightarrow{P} 0 \quad (N \longrightarrow \infty)$$

An oracle inequality can be established as well.

Algorithm

Set $P(\theta) := \|b\|_1$. Solving

$$argmin_\theta Q_\lambda(\theta) = argmin_\theta \Big\{ g(\theta) + \lambda P(\theta) \Big\}$$

is challenging.

Coordinate Gradient Descent from Tseng and Yun (2007).
Key elements:

- **Coordinatewise optimization**
- **Quadratic approximation** of the objective function
- **Inexact line search** using the Armijo rule

Algorithm

Coordinate Gradient Descent Algorithm

0. *Let* $\theta^0 \in \mathbb{R}^{p+2}$ *be an initial value.*

For $k = 1, 2, ...$ *let* $\mathcal{J}^k$ *be the set cycling through the coordinates* $\{1, ..., p, p+1, p+2\}$

1. *Choose an appropriate hessian* $H^k > 0$
2. $d^k := argmin_d \left\{ g(\theta^k) + \nabla g(\theta^k) d + 1/2 d^2 H^k + \lambda P(\theta^k + d e_{\mathcal{J}_k}) \right\}$
3. *Choose a stepsize* $\alpha^k > 0$ *by the Armijo rule and set*

$$\theta^{k+1} = \theta^k + \alpha^k d^k e_{\mathcal{J}_k}$$

Algorithm

The Armijo rule is defined as follows:

Armijo Rule

Choose $\alpha^k_{init} > 0$ *and let* α^k *be the largest element of* $\{\alpha^k_{init}\beta^j\}_{j=0,1,2,..}$ *satisfying*

$$Q_\lambda(\theta^k + \alpha^k d^k) \leq Q_\lambda(\theta^k) + \alpha^k \sigma \triangle^k$$

where

$$\triangle^k := \nabla g(\theta^k) d^k + \gamma (d^k)^2 H^k + \lambda P(\theta^k + d^k e_{J_k}) - \lambda P(\theta^k)$$

Convergence Results

Numerical Convergence

If $(\theta^k)_{k \geq 0}$ is chosen according to the Coordinate Gradient Descent Algorithm, then every cluster point of $\{\theta^k\}_{k \geq 0}$ is a stationary point of $Q_\lambda(\theta)$.

Remarks:

- Due to the nonconvex form of $Q_\lambda(\theta)$, the convergence can be slow
- The result depends on the starting value

Small Simulation Study

Random-intercept model

$$y_{ij} = (b_0 + \beta_i) + x_{ij}^T b + \epsilon_{ij} \quad i = 1, .., N, \quad j = 1, ..., n_i$$

with $N = 30$, $n_i = n = 6$, $p = 500 (p > N_{Tot})$ and $\sigma = \tau = 1$ and $b = (1, 2, 3, 1, 0, ...0)$.

We simulated the covariates $k = 1, .., p$ by $x^{(k)} \sim N_n(0, \Sigma)$ with $\Sigma_{ij} = \rho^{|i-j|}$ and $\rho = 0.2$, so the signal-to-noise ratio is 18.8.

In each run we have chosen the optimal model using BIC over a grid $\lambda_{min} < ... < \lambda_{max}$.

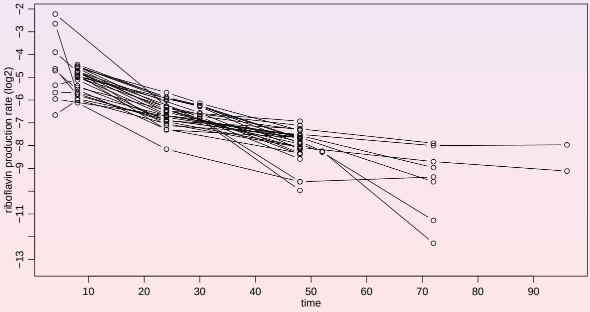
Small Simulation Study

Results:

quantity	true value	median	mean	sd
$\|\mathcal{A}\|$:	4	5	25.13	(97.44)
TP :	4	4	4	(0)
σ:	1	1.05	1.01	(0.22)
τ:	1	0.94	0.94	(0.15)
b_0:	1	0.98	0.98	(0.2)
b_1:	2	1.78	1.79	(0.1)
b_2:	3	2.86	2.84	(0.1)
b_3:	1	0.79	0.79	(0.1)

Riboflavin Production in Bacillus Subtilis

A data set provided by DSM (Switzerland).
The response variable $Y \in \mathbb{R}$: Riboflavin production rate
covariates $X \in \mathbb{R}^p$: expressions from genes
$N = 28$, $N_{Tot} = 111$, $n_i \in \{2, ..., 6\}$, $p = 4088$



Analysis of the Riboflavin Data Set

We fit a random-intercept model

$$y_{ij} = (b_0 + \beta_i) + x_{ij}^T b + \epsilon_{ij} \quad i = 1, .., N, \quad j = 1, ..., n_i$$

We get:
$\sigma = 0.69 \quad \tau = 0.23$
$|\mathcal{A}| = 9$
$\beta_1, ..., \beta_N \in [-0.33, 0.2]$
Comment:

- A very sparse model is chosen ($|\mathcal{A}_{Lasso}| = 18$, $|\mathcal{A}_{adaptiveLasso}| = 12$)
- 6 out of 9 variables are also selected by the Lasso (the remaining three have a small absolute value)

Analysis of the Riboflavin Data Set

Compare the predictive performance to the Lasso.

Choose a subset for which $n_i = 4$, i.e. each unit has measurements at four time points.

Carry out a **leave-one-time-point-out Cross-Validation**.

Use AIC for choosing the optimal λ-parameter.

The prediction error was reduced by about 12%.

Conclusions

- ”convex $\longrightarrow$ nonconvex Lasso”
- We suggested an algorithm using ℓ_1-constraints in order to estimate the parameters of a simple Mixed-Effects Model.
- Under regularity conditions, the ℓ_1-penalized estimator is consistent
- The numerical algorithm converges to a stationary point.

Thanks

- to my supervisor Peter Bühlmann
- the audience

References

- P. Tseng and S. Yun ; A Coordinte Gradient Descent Method for Nonsmooth Separable Minimization ; Mathematical Programming ; (2007)
- Regression Shrinkage and Selection via the Lasso ; R. Tibshirani ; J. R. Stat. Soc. ; (1996)
- ℓ_1-Penalization for Mixture Regression Models ; N. Städler, P. Bühlmann, S. van de Geer ; to appear ; (2009)