Linear mixed models for high-dimensional data: extending the functionalities of the LMER package

Matteo Amestoy, Mark van de Wiel, Wessel van Wieringen

13 May 2021

Introduction

- Linear mixed model

$$Y = X\boldsymbol{\beta} + Z\boldsymbol{\gamma} + \epsilon, \quad \boldsymbol{\gamma} \sim \mathcal{N}(0, \Sigma), \; \epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$

$$Y \sim \mathcal{N}(X\boldsymbol{\beta}, Z\Sigma Z^\top + \sigma^2 I_n)$$

Introduction

- Linear mixed model

$$Y = X\beta + Z\gamma + \epsilon, \;\; \gamma \sim \mathcal{N}(0, \Sigma), \;\; \epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$

$Y \sim \mathcal{N}(X\beta, Z\Sigma Z^\top + \sigma^2 I_n)$

- Shrinkage
  - Solve identifiability issues (high dimensionality)
  - Stabilise estimator - trade bias for variance

Introduction

- Linear mixed model

$$Y = X\beta + Z\gamma + \epsilon, \;\; \gamma \sim \mathcal{N}(0, \Sigma), \;\; \epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$

  $Y \sim \mathcal{N}(X\beta, Z\Sigma Z^\top + \sigma^2 I_n)$

- Shrinkage
  - Solve identifiability issues (high dimensionality)
  - Stabilise estimator - trade bias for variance
- Regularising LMM
  - Fixed effects - high dimensionality and colinearity
  - Random effects - not well defined

Introduction - Regularisation of the random effects

High dimensionality

Data: 10 individuals observed at 10 time-points

$$Y \sim 1 + t + t^2 + t^3 + (1 + t + t^2 + t^3 | ind)$$

Introduction - Regularisation of the random effects

High dimensionality

Data: 10 individuals observed at 10 time-points

$$Y \sim 1 + t + t^2 + t^3 + (1 + t + t^2 + t^3 | ind)$$

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: Y ~ 1 + t + t2 + t3 + (1 + t + t2 + t3 | ind)
   Data: data
      AIC       BIC    logLik  deviance  df.resid
 311.8980  350.9756 -140.9490  281.8980        85
Random effects:
 Groups   Name        Std.Dev. Corr
 ind      (Intercept) 0.2999
          t           1.2829   -1.00
          t2          0.8992   -0.93  0.95
          t3          1.8525    0.91 -0.94 -1.00
 Residual             0.9273
Number of obs: 100, groups:  ind, 10
Fixed Effects:
(Intercept)            t           t2           t3
    -0.1696      -0.6580       0.8416       0.2623
```

Introduction - Regularisation of the random effects

High dimensionality

Data: 10 individuals observed at 10 time-points

$$Y \sim 1 + t + t^2 + t^3 + (1 + t + t^2 + t^3 | ind)$$

What if we have 25 extra individuals observed only once?

Introduction - Regularisation of the random effects

High dimensionality

Data: 10 individuals observed at 10 time-points

$$Y \sim 1 + t + t^2 + t^3 + (1 + t + t^2 + t^3 | ind)$$

What if we have 25 extra individuals observed only once?

```
Error: number of observations (=125) <= number of random
 effects (=140) for term (1 + t + t2 + t3 | ind); the ra
ndom-effects parameters and the residual variance (or sc
ale parameter) are probably unidentifiable
```

Introduction - Regularisation of the random effects

$$Y \sim X + (\tilde{Z}|ind) \iff Y = X\boldsymbol{\beta} + Z\boldsymbol{\gamma} + \epsilon$$

$$ind = \begin{bmatrix} 1 \\ 2 \\ 1 \\ 3 \\ 3 \end{bmatrix}, \tilde{Z} = \begin{bmatrix} \tilde{Z}_1 \\ \tilde{Z}_2 \\ \tilde{Z}_3 \\ \tilde{Z}_4 \\ \tilde{Z}_5 \end{bmatrix}, \tilde{\gamma}_i \sim \mathcal{N}(0, \tilde{\Sigma})$$

Introduction - Regularisation of the random effects

$$Y \sim X + (\tilde{Z}|ind) \iff Y = X\boldsymbol{\beta} + Z\boldsymbol{\gamma} + \epsilon$$

$$ind = \begin{bmatrix} 1 \\ 2 \\ 1 \\ 3 \\ 3 \end{bmatrix}, Z = \begin{bmatrix} \tilde{Z}_1 & . & . \\ . & \tilde{Z}_2 & . \\ \tilde{Z}_3 & . & . \\ . & . & \tilde{Z}_4 \\ . & . & \tilde{Z}_5 \end{bmatrix}, \gamma = \begin{bmatrix} \tilde{\gamma}_1 \\ \tilde{\gamma}_2 \\ \tilde{\gamma}_3 \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \tilde{\Sigma} & . & . \\ . & \tilde{\Sigma} & . \\ . & . & \tilde{\Sigma} \end{bmatrix}\right)$$

Such that $(Z\gamma)_i = \tilde{Z}_i \tilde{\gamma}_{ind(i)}$

Introduction - Regularisation of the random effects

$$Y \sim X + (\tilde{Z}|ind) \iff Y = X\boldsymbol{\beta} + Z\boldsymbol{\gamma} + \epsilon$$

$$ind = \begin{bmatrix} 1 \\ 2 \\ 1 \\ 3 \\ 3 \end{bmatrix}, Z = \begin{bmatrix} \tilde{Z}_1 & . & . \\ . & \tilde{Z}_2 & . \\ \tilde{Z}_3 & . & . \\ . & . & \tilde{Z}_4 \\ . & . & \tilde{Z}_5 \end{bmatrix}, \gamma = \begin{bmatrix} \tilde{\gamma}_1 \\ \tilde{\gamma}_2 \\ \tilde{\gamma}_3 \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \tilde{\Sigma} & . & . \\ . & \tilde{\Sigma} & . \\ . & . & \tilde{\Sigma} \end{bmatrix}\right)$$

Such that $(Z\gamma)_i = \tilde{Z}_i \tilde{\gamma}_{ind(i)}$

Z has $(10 + 25) \times 4 = 140$ columns and $10 \times 10 + 25 = 125$ lines.

Introduction - Regularisation of the random effects

$$Y \sim X + (\tilde{Z}|ind) \iff Y = X\boldsymbol{\beta} + Z\boldsymbol{\gamma} + \epsilon$$

$$ind = \begin{bmatrix} 1 \\ 2 \\ 1 \\ 3 \\ 3 \end{bmatrix}, Z = \begin{bmatrix} \tilde{Z}_1 & . & . \\ . & \tilde{Z}_2 & . \\ \tilde{Z}_3 & . & . \\ . & . & \tilde{Z}_4 \\ . & . & \tilde{Z}_5 \end{bmatrix}, \gamma = \begin{bmatrix} \tilde{\gamma}_1 \\ \tilde{\gamma}_2 \\ \tilde{\gamma}_3 \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \tilde{\Sigma} & . & . \\ . & \tilde{\Sigma} & . \\ . & . & \tilde{\Sigma} \end{bmatrix}\right)$$

Such that $(Z\gamma)_i = \tilde{Z}_i \tilde{\gamma}_{ind(i)}$

Z has $(10 + 25) \times 4 = 140$ columns and $10 \times 10 + 25 = 125$ lines.

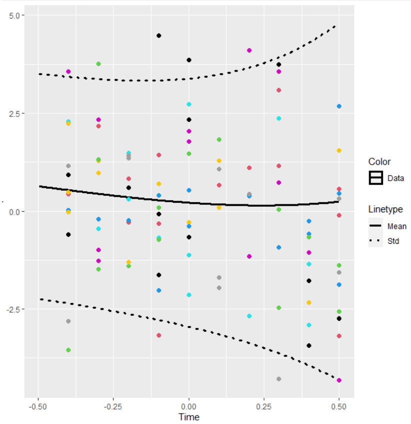
Identifiability is less restrictive and depends on

- the number of individuals
- the number of repeats

Introduction - Regularisation of the random effects

Overfitting

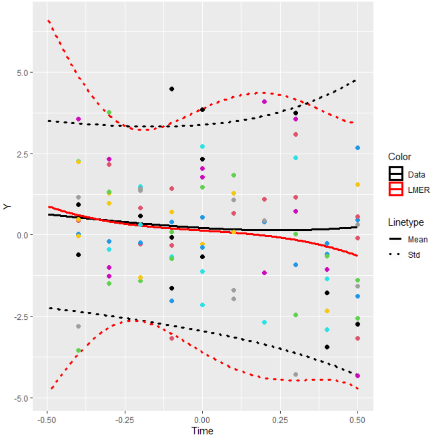
- $Y \sim 1 + t + t^2 + t^3 + (1 + t + t^2 + t^3|ind)$
- 50 individuals
- 2 repeated measurements



Introduction - Regularisation of the random effects

Overfitting

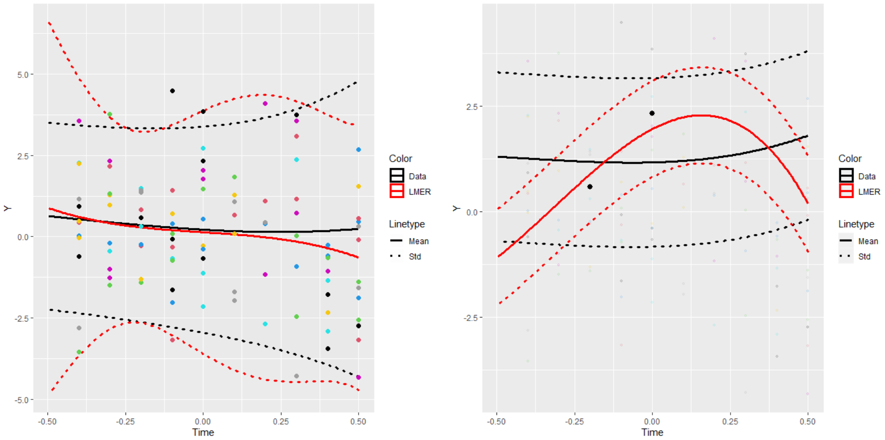
- $Y \sim 1 + t + t^2 + t^3 + (1 + t + t^2 + t^3 | ind)$
- 50 individuals
- 2 repeated measurements



Introduction - Regularisation of the random effects

Overfitting

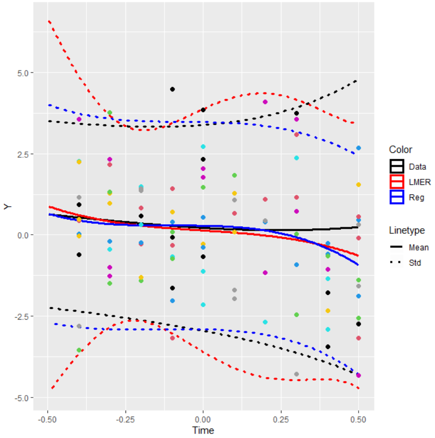
- $Y \sim 1 + t + t^2 + t^3 + (1 + t + t^2 + t^3 | ind)$
- 50 individuals
- 2 repeated measurements

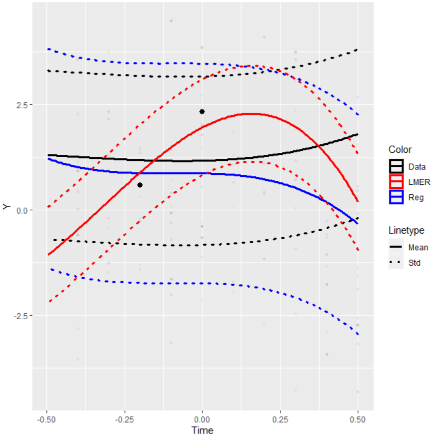


Introduction - Regularisation of the random effects

Overfitting

- $Y \sim 1 + t + t^2 + t^3 + (1 + t + t^2 + t^3 | ind)$
- 50 individuals
- 2 repeated measurements





Bayesian inference - Empirical Bayes

- Bayesian inference - priors on $(\beta, \Sigma) = \theta \sim p(.|\Theta)$

Bayesian inference - Empirical Bayes

- Bayesian inference - priors on $(\boldsymbol{\beta}, \Sigma) = \theta \sim p(.|\Theta)$
  - Existing solvers
    - STAN - draws from the posterior
    - INLA - approximate the posterior

Bayesian inference - Empirical Bayes

- Bayesian inference - priors on $(\beta, \Sigma) = \theta \sim p(.|\Theta)$
  - Existing solvers
    - STAN - draws from the posterior
    - INLA - approximate the posterior
  - How to choose the prior?
    - Shape - problem specific
    - Values of the hyperparameters

Bayesian inference - Empirical Bayes

- Bayesian inference - priors on $(\beta, \Sigma) = \theta \sim p(.|\Theta)$
  - Existing solvers
    - STAN - draws from the posterior
    - INLA - approximate the posterior
  - How to choose the prior?
    - Shape - problem specific
    - Values of the hyperparameters

- Choice of the hyperparameters Θ with empirical Bayes

We maximise the marginal likelihood $p(Y|\Theta)$

$$\Theta^* = \arg\max \int p(Y|\theta)p(\theta|\Theta)d\theta$$

Bayesian inference - Empirical Bayes

- Bayesian inference - priors on $(\beta, \Sigma) = \theta \sim p(.|\Theta)$
  - Existing solvers
    - STAN - draws from the posterior
    - INLA - approximate the posterior
  - How to choose the prior?
    - Shape - problem specific
    - Values of the hyperparameters
- Choice of the hyperparameters Θ with empirical Bayes

We maximise the marginal likelihood $p(Y|\Theta)$

$$\Theta^* = \arg\max \int p(Y|\theta)p(\theta|\Theta)d\theta$$

Sampling from the posterior is too slow.

Marginal Likelihood maximization

- Empirical Bayes maximises the marginal likelihood

$$\Theta^* = \arg\max \int p(Y|\theta)p(\theta|\Theta)d\theta$$
$$= \arg\max \int \exp(ll(\theta; Y, \Theta))d\theta$$

Marginal Likelihood maximization

- Empirical Bayes maximises the marginal likelihood

$$
\begin{aligned}
\Theta^* &= \arg\max \int p(Y|\theta)p(\theta|\Theta)d\theta \\
&= \arg\max \int \exp(ll(\theta; Y, \Theta))d\theta
\end{aligned}
$$

- Laplace approximation to estimate the integral

$$
\int \exp(ll(\theta; Y, \Theta))dx \simeq (2\pi)^{d/2} \frac{\exp(ll(\theta^*))}{|-H(ll)(\theta^*)|^{1/2}}
$$

where $\theta^*(\Theta)$ is the MAP and $H(ll)$ is the Hessian matrix of ll.

Marginal Likelihood maximization

- Empirical Bayes maximises the marginal likelihood

$$\Theta^* = \arg\max \int p(Y|\theta)p(\theta|\Theta)d\theta$$
$$= \arg\max \int \exp(ll(\theta; Y, \Theta))d\theta$$

- Laplace approximation to estimate the integral

$$\int \exp(ll(\theta; Y, \Theta))dx \simeq (2\pi)^{d/2} \frac{\exp(ll(\theta^*))}{|-H(ll)(\theta^*)|^{1/2}}$$

where $\theta^*(\Theta)$ is the MAP and $H(ll)$ is the Hessian matrix of ll.

We need a fast estimation of the MAP θ^*

Choice of priors and EM MAP estimation

Conjugate priors

- $\beta \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda} I_p)$
- $\Sigma \sim \mathcal{IW}(\eta, \Phi)$

Choice of priors and EM MAP estimation

Conjugate priors

- $\beta \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda} I_p)$
- $\Sigma \sim \mathcal{IW}(\eta, \Phi)$

EM update leads to a intuitive parameterisation.
Update rule without prior (likelihood maximisation):

$$\Sigma_{k+1} = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{\boldsymbol{\gamma}|\theta_k, Y} \left[\boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^\top \right]$$

Choice of priors and EM MAP estimation

Conjugate priors

- $\beta \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda} I_p)$
- $\Sigma \sim \mathcal{IW}(\eta, \Phi) \longrightarrow \Sigma \sim \mathcal{IW}(b, A)$

EM update leads to a intuitive parameterisation.
Update rule with (Maximum a posteriori):

$$\Sigma_{k+1} = bA + (1-b)\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\gamma|\theta_k, Y}\left[\gamma_i \gamma_i^\top\right]$$

with $A = \frac{\Phi}{\eta+q+1}$, $b = \frac{\eta+q+1}{m+\eta+q+1}$

Choice of priors and EM MAP estimation

Conjugate priors

- $\beta \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda} I_p)$
- $\Sigma \sim \mathcal{IW}(b, A)$

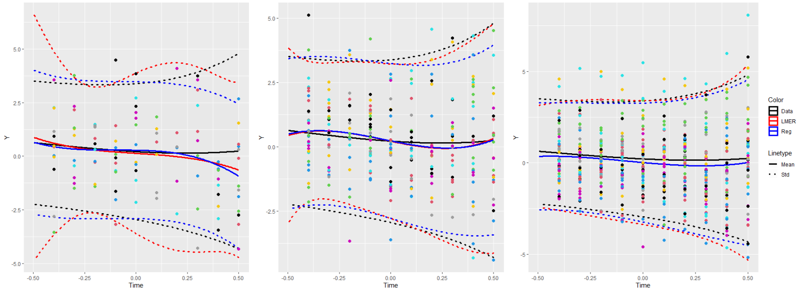
Given $\Theta = \{\lambda, b, A\}$ we can compute the MAP $\theta^*(\Theta) = \{\beta, \sigma, \Sigma\}$ and solve:

$$\Theta^* = \arg\max_{\Theta} (2\pi)^{d/2} \frac{\exp(ll(\theta^*))}{|-H(ll)(\theta^*)|^{1/2}}$$

Application

- $Y \sim 1 + t + t^2 + t^3 + (1 + t + t^2 + t^3 | ind)$
- 50 individuals

2 meas., $b = 0.99$ 5 meas., $b = 0.32$ 10 meas., $b = 0.20$



Results - RE shrinkage influence of repeats

Set up:

- 40 individuals
- **FE**: $\beta \sim \mathcal{N}(0, I_2)$, $X_i \sim \mathcal{N}(0, I_n)$
- **RE**: $\Sigma \sim \mathcal{IW}(\nu, \Phi)$ such that $\boldsymbol{E}(\Sigma) = I_4$

Results - RE shrinkage influence of repeats

Set up:

- 40 individuals
- **FE**: $\beta \sim \mathcal{N}(0, I_2)$, $X_i \sim \mathcal{N}(0, I_n)$
- **RE**: $\Sigma \sim \mathcal{IW}(\nu, \Phi)$ such that $\boldsymbol{E}(\Sigma) = I_4$

Median of 30 experiments:

Nb. repeats	RMSE β ratio	KL ratio	hp b
2	0.97	2.26	0.38
3	1.07	1.49	0.34
5	1.02	1.20	0.29
8	1.01	1.05	0.29

Results - Interaction FE/RE with high dimensionality

Set up:

- 280 observations - 40 individuals - 7 repeats
- **FE**: $\beta \sim \mathcal{N}(0, I_q)$, $X_i \sim \mathcal{N}(0, I_n)$, $q = \{2, 500\}$
- **RE**: $\Sigma \sim \mathcal{IW}(\nu, \Phi)$ such that $\boldsymbol{E}(\Sigma) = I_2$

Results - Interaction FE/RE with high dimensionality

Set up:

- 280 observations - 40 individuals - 7 repeats
- **FE:** $\beta \sim \mathcal{N}(0, I_q)$, $X_i \sim \mathcal{N}(0, I_n)$, $q = \{2, 500\}$
- **RE:** $\Sigma \sim \mathcal{IW}(\nu, \Phi)$ such that $\boldsymbol{E}(\Sigma) = I_2$

Median of 10 experiments:

q	b	λ
2	0.17	0.09
500	0.13	1.57

Conclusion

- We propose a LMM regularisation framework
  - Data driven hyperparameter learning
  - Combined regularisation of FE and RE

Conclusion

- We propose a LMM regularisation framework
  - Data driven hyperparameter learning
  - Combined regularisation of FE and RE
- Allows to model complex data
  - High dimensional fixed effects
  - Complex correlation structures
    - High number of covariates / multiple random effects
    - Unevenly distributed observations

Conclusion

- We propose a LMM regularisation framework
  - Data driven hyperparameter learning
  - Combined regularisation of FE and RE
- Allows to model complex data
  - High dimensional fixed effects
  - Complex correlation structures
    - High number of covariates / multiple random effects
    - Unevenly distributed observations
- Can be extended to multivariate outcomes

Conclusion

Thank you!

Contact: m.amestoy@amsterdamumc.nl