# A Relaxation Approach to Feature Selection for Mixed-Effects Models

Aleksei Sholokhov, James V. Burke, Peng Zheng, and Aleksandr Aravkin

Monday $4^{\text{th}}$ July, 2022

# Linear Mixed-Effect Models

Linear Mixed-Effect (LME) models are often used for analyzing combined data across a range of groups.

They use use covariates to separate the population variability (fixed-effects) from the group variability (random effects).

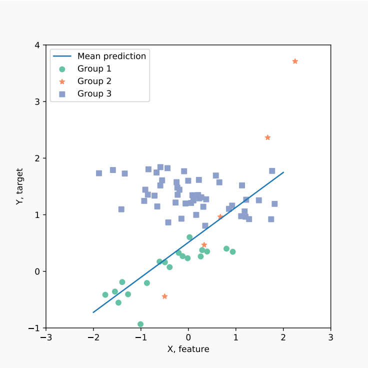
LMEs borrow strength across groups to estimate key statistics in cases when data within units are sparse or highly variable.

# Linear Mixed-Effect Models

Dataset: $m$ groups $(X_i, Z_i, y_i)$, $\quad i = 1, \ldots m$, each has $n_i$ observations

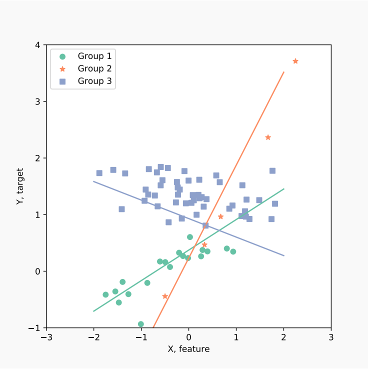
- $X_i \in \mathbb{R}^{n_i \times p}$ – group $i$ design matrix for fixed features
- $Z_i \in \mathbb{R}^{n_i \times q}$ – group $i$ design matrix for random effects
- $y_i \in \mathbb{R}^{n_i}$ – group $i$ observations

Standard Linear Regression:




$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Lambda)$$

Linear Mixed-Effect Model:




$$y_i = X_i\beta + Z_i u_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \Lambda_i)$$

$$u_i \sim \mathcal{N}(0, \Gamma 4)$$

# Notation

$$
\begin{aligned}
y_i &= X_i\beta + Z_i u_i + \varepsilon_i \quad i = 1 \dots m \\
\varepsilon_i &\sim \mathcal{N}(0, \Lambda_i) \\
u_i &\sim \mathcal{N}(0, \Gamma)
\end{aligned} \tag{1}
$$

- $p$ – number of fixed features, $q$ – number of random effects.
- $\beta \in \mathbb{R}^p$ – fixed effects, or mean effects
- $u_i \in \mathbb{R}^q$ – random effects
- $\Gamma \in \mathbb{R}^{q \times q}$ – covariance matrix of random effects, often $\Gamma = \text{Diag}(\gamma)$
- $\varepsilon_i \in \mathbb{R}^{n_i}$ – observation noise
- $\Lambda_i \in R^{n_i \times n_i}$ – covariance matrix for noise

Unknowns: $\beta$, $u_i$, $\gamma$, sometimes $\Lambda_i$.

# Likelihood for Mixed Models

Negative log-likelihood:

$$\begin{aligned}\mathcal{L}(\beta,\gamma) = \sum_{i=1}^{m} \frac{1}{2}(y_i - X_i\beta)^T (Z_i \Gamma Z_i^T + \Lambda_i)^{-1}(y_i - X_i\beta) + \\ + \frac{1}{2}\log\det\left(Z_i\Gamma Z_i^T + \Lambda_i\right), \quad \Gamma = \mathrm{Diag}\left(\gamma\right)\end{aligned} \tag{2}$$

Maximum likelihood estimates for $\beta$ and $\gamma$ solve the problem:

$$\mathcal{LME} \quad \min_{\beta\in\mathbb{R}^p,\, \gamma\in\mathbb{R}_+^q} \mathcal{L}(\beta,\gamma) \tag{3}$$

To select covariates we add a sparsity-promoting regularizer $R(\beta,\gamma)$

$$\mathcal{FS}-\mathcal{LME} \quad \min_{\beta\in\mathbb{R}^p,\, \gamma\in\mathbb{R}_+^q} \mathcal{L}(\beta,\gamma) + R(\beta,\gamma) \tag{4}$$

- $\mathcal{L}(\beta,\gamma)$ is smooth on its domain, quadratic w.r.t. $\beta$ and $\bar{\eta}$-weakly-convex w.r.t. $\gamma$.
- $R(\beta,\gamma)$ is closed, proper, convex, with easily computed *prox operator*

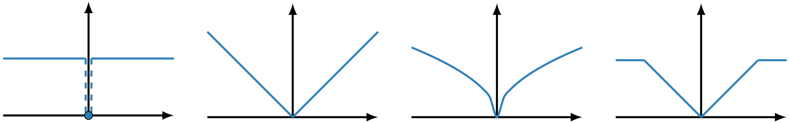
# Regularization

- $R(\beta, \gamma)$ is closed, proper, convex, with easily computed *prox operator*

$$\text{prox}_{\alpha R + \delta_{\mathcal{C}}}(\tilde{\beta}, \tilde{\gamma}) := \underset{(\beta,\gamma)\in\mathcal{C}}{\text{argmin}}\, R(\beta, \gamma) + \frac{1}{2\alpha}\|(\beta, \gamma) - (\tilde{\beta}, \tilde{\gamma})\|_2^2, \quad \text{where } \mathcal{C} := \mathbb{R}^p \times R_+^q \tag{5}$$

Examples:

- $R(x) = \lambda \sum_{j=1}^{p} w_j \|x_j\|_1$ – LASSO and Adaptive LASSO penalties [1, 5]
- $R(x) = \lambda \|x\|_0$ – $\ell_0$ penalty [6, 4]
- $R(x)$ – SCAD penalty ([2, 3])



(a) $\ell_0$ (b) $\ell_1$ (c) $\ell_p,\ p = 1/2$ (d) SCAD

Figure: Three simple graphs

# SR3-Relaxation for Mixed-Effect Models

Original problem $\mathcal{FS}-\mathcal{LME}$:

$$\min_{\beta\in\mathbb{R}^p,\,\gamma\in\mathbb{R}^q_+} \mathcal{L}(\beta,\gamma)+R(\beta,\gamma) \tag{6}$$

Relaxed problem $\mathcal{MSR}3$:

$$\min_{\beta,\tilde\beta\in\mathbb{R}^p,\,\gamma,\tilde\gamma\in\mathbb{R}^q_+} \mathcal{L}(\beta,\gamma)+\phi_\mu(\gamma)+\kappa_\eta(\beta-\tilde\beta,\gamma-\tilde\gamma)+R(\tilde\beta,\tilde\gamma) \tag{7}$$

where the *relaxation* $\kappa_\eta$ decouples the likelihood and the regularizer

$$\kappa_\eta(\beta-\tilde\beta,\gamma-\tilde\gamma):=\frac{\eta}{2}\|\beta-\tilde\beta\|_2^2+\frac{\eta+\bar\eta}{2}\|\gamma-\tilde\gamma\|_2^2 \tag{8}$$

and the *projection function* $\phi_\mu$ replaces $\gamma\geq 0$ with a log-barrier

$$\phi_\mu(\gamma):=\begin{cases}-\mu\sum_{i=1}^q \ln(\gamma_i/\mu), & \mu>0\\ \delta_{\mathbb{R}^q_+}(\gamma), & \mu=0\\ +\infty, & \mu<0\end{cases} \tag{9}$$

# Value Function Reformulation

$\mathcal{MSR}3$-relaxation replaces the original likelihood $\mathcal{L}$ with a *value function* $u_{\eta,\mu}$:

$$\begin{aligned} u_{\eta,\mu}(\tilde{\beta},\tilde{\gamma}) &:= \min_{(\beta,\gamma)} \mathcal{L}_{\eta,\mu}((\beta,\gamma),(\tilde{\beta},\tilde{\gamma})) \\ &:= \min_{(\beta,\gamma)} \mathcal{L}(\beta,\gamma) + \phi_\mu(\gamma) + \kappa_\eta(\beta-\tilde{\beta},\gamma-\tilde{\gamma}) \end{aligned} \tag{10}$$

so $\mathcal{MSR}3$-formulation (7) becomes

$$\min_{(\tilde{\beta},\tilde{\gamma})\in\mathcal{C}} u_{\eta,\mu}(\tilde{\beta},\tilde{\gamma}) + R(\tilde{\beta},\tilde{\gamma}) \tag{11}$$

When $\bar{\eta}$ is larger than the weak-convexity constant

- $u_{\eta,\mu}$ is well-defined and continuously differentiable.
- Solutions $(\tilde{\beta}^*,\tilde{\gamma}^*)$ for $\mathcal{MSR}3$ converge to solutions $(\beta^*,\gamma^*)$ of $\mathcal{FS}-\mathcal{LME}$ when $\mu\to 0$ and $\eta\to\infty$.

**Key observation**: in practice, we don't need accurate solutions for (10): a few Newton iterations keep the solution close to the central path.

# Value Function Reformulation

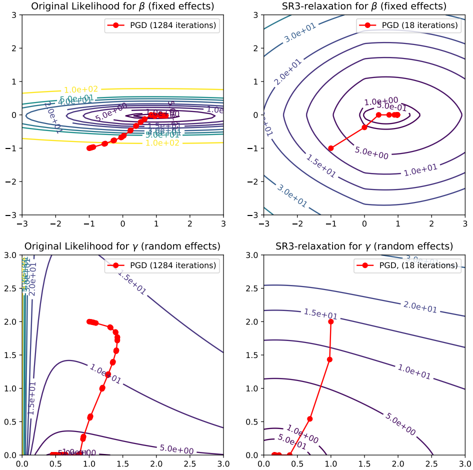



Figure: Comparison of the level-sets for the original likelihood (left) and $\mathcal{MSR}3$-likelihood (right), for fixed (top) and random (bottom) effects.

# Designing an Algorithm

Gradient of a Lagrangian:

$$G_{\nu,\eta}((\beta,\gamma,v),(\tilde{\beta},\tilde{\gamma})) := \begin{bmatrix} \nabla_{\beta}\,\mathcal{L}(\beta,\gamma) + \eta(\beta-\tilde{\beta}) \\ \nabla_{\gamma}\,\mathcal{L}(\beta,\gamma) + (\bar{\eta}+\eta)(\gamma-\tilde{\gamma}) - v \\ v \bigodot \gamma - \mu\mathbf{1} \end{bmatrix} \tag{12}$$

**Lemma:** For every $(\mu,\eta) \in \mathbb{R}_{+} \times \mathbb{R}_{++}$,

$$\begin{aligned} (\hat{\beta},\hat{\gamma}) &= \underset{(\beta,\gamma)}{\operatorname{argmin}}\, \mathcal{L}_{\eta,\mu}((\beta,\gamma),(\tilde{\beta},\tilde{\gamma})) \\ &\Longleftrightarrow \\ \exists \hat{v} &\in \mathbb{R}^{q}_{+} \text{ s.t. } G_{\nu,\eta}((\beta,\gamma,\hat{v}),(\tilde{\beta},\tilde{\gamma})) = 0 \end{aligned} \tag{13}$$

If $\mu > 0$, then $\hat{v} = -\nabla\phi_{\mu}(\hat{\gamma})$, and if $\mu = 0$, then $\hat{v}$ is the unique KKT multiplier associated with the constraint $0 \leq \gamma$.

# $\mathcal{MSR}3$-fast Algorithm

progress $\leftarrow$ **True**;   iter = 0;

$\beta^+, \tilde{\beta}^+ \leftarrow \beta_0$;   $\gamma^+, \tilde{\gamma}^+ \leftarrow \gamma_0$;   $v^+ \leftarrow 1 \in \mathbb{R}^q$;   $\mu \leftarrow \frac{{v^+}^T \gamma^+}{10q}$

**while** $\underline{\texttt{iter} < \texttt{max\_iter and } \|G_\mu(\beta^+, \gamma^+, v^+)\| > \texttt{tol and progress}}$ **do**

  $\beta \leftarrow \beta^+$;   $\gamma \leftarrow \gamma^+$;   $\tilde{\beta} \leftarrow \tilde{\beta}^+$;   $\tilde{\gamma} \leftarrow \tilde{\gamma}^+$

  $[dv, d\beta, d\gamma] \leftarrow \nabla G_\mu((\beta, \gamma, v), (\tilde{\beta}, \tilde{\gamma}))^{-1} G_\mu((\beta, \gamma, v), (\tilde{\beta}, \tilde{\gamma}))$

  $\alpha \leftarrow 0.99 \times \min\left(1, -\frac{\gamma_i}{d\gamma_i}, \forall i: \ d\gamma_i < 0\right)$

  $\beta^+ \leftarrow \beta + \alpha d\beta$;   $\gamma^+ = \gamma + \alpha d\gamma$;   $v^+ \leftarrow v + \alpha dv$

  **if** $\underline{\|\gamma^+ \odot v^+ - q^{-1}{\gamma^+}^T v^+ 1\| > 0.5 q^{-1} {v^+}^T \gamma^+}$ **then** continue;

  **else**

    $\tilde{\beta}^+ = \text{prox}_{\alpha R}(\beta^+)$;  $\tilde{\gamma}^+ = \text{prox}_{\alpha R + \delta_{\mathbb{R}_+}}(\gamma^+)$;  $\mu = \frac{1}{10} \frac{{v^+}^T \gamma^+}{q}$

  **end**

  progress $= (\|\beta^+ - \beta\| \geq \text{tol or } \|\gamma^+ - \gamma\| \geq \text{tol or } \|\tilde{\beta}^+ - \tilde{\beta}\| \geq \text{tol or } \|\tilde{\gamma}^+ - \tilde{\gamma}\| \geq \text{tol})$

  iter += 1

**end**

**return** $\underline{\tilde{\beta}^+, \tilde{\gamma}^+}$

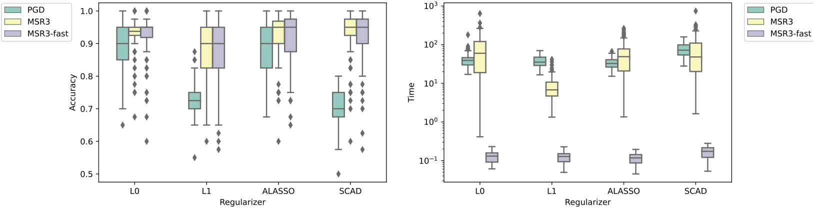
# Application to Synthetic Problems

**The Experiment**

- The number of fixed effects $p$ and random effects $q$ is 20.
- $\beta = \gamma = \frac{1}{2}[1, 2, 3, \ldots, 10, 0 \ldots, 0]$
- 9 groups with sizes [10, 15, 4, 8, 3, 5, 18, 9, 6]
- $X_i \sim \mathcal{N}(0, I)^p$, $Z_i = X_i$, $\varepsilon_i \sim \mathcal{N}(0, 0.3^2 I)$
- Each experiment is repeated 100 times.
- Grid-search for $\eta \in [10^{-4}, 10^2]$, golden search for $\lambda \in [0, 10^5]$
- Final model is chosen to maximize BIC

| Regularizer | Model Metric | PGD | MSR3 | MSR3-fast |
|---|---|---|---|---|
| L0 | Accuracy | 0.89 | **0.92** | **0.92** |
| | Time | 41.68 | 88.54 | **0.13** |
| L1 | Accuracy | 0.73 | **0.88** | **0.88** |
| | Time | 38.39 | 9.13 | **0.13** |
| ALASSO | Accuracy | 0.88 | **0.92** | 0.91 |
| | Time | 34.55 | 65.19 | **0.12** |
| SCAD | Accuracy | 0.71 | **0.93** | 0.92 |
| | Time | 77.62 | 84.67 | **0.17** |

# Application to Synthetic Problems




Benefits:

- $\mathcal{MSR}3$-relaxation has similar (and sometimes better!) feature selection performance than the original likelihood.
- $\mathcal{MSR}3$-fast optimization accelerates the compute time by $\sim 10^2$.

Setbacks:

- $\eta$ is a new hyperparameter to tune.

# $\ell_0$-based Covariate Selection for Bullying Study from GBD

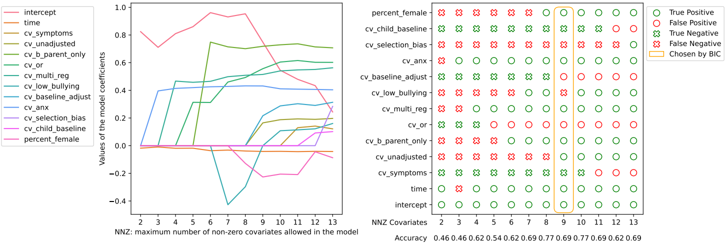



Figure: Fixed and random covariate selection for Bullying dataset from [? ]. The model selected 9 covariates, 7 of which were historically significant, and did not select 4 covariates, 1 of which was historically significant.

# Thank You!

The code is available on GitHub:
https://github.com/aksholokhov/pysr3

- All estimators are fully compatible to `sklearn` library.
- Implements SR3 for linear, generalized-linear, and linear mixed-effect models.
- Has tutorials, tests, and documentation.

**References**:


[1] Howard D. Bondell, Arun Krishna, and Sujit K. Ghosh. Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models. Biometrics, 66(4):1069–1077, dec 2010.

[2] Jianqing Fan and Runze Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. Journal of the American Statistical Association, 96(456):1348–1360, dec 2001.

[3] Yingying Fan and Runze Li. Variable selection in linear mixed effects models. The Annals of Statistics, 40(4):2043–2068, aug 2012.

[4] Richard H. Jones. Bayesian information criterion for longitudinal and clustered data. Statistics in Medicine, 30(25):3050–3056, nov 2011.

[5] Bingqing Lin, Zhen Pang, and Jiming Jiang. Fixed and random effects selection by REML and pathwise coordinate optimization. Journal of Computational and Graphical Statistics, 22(2):341–355, 2013.

[6] Florin Vaida and Suzette Blanchard. Conditional Akaike information for mixed-effects models. Biometrika, 92(2):351–370, jun 2005.