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

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

Thursday 8^{th} June, 2023

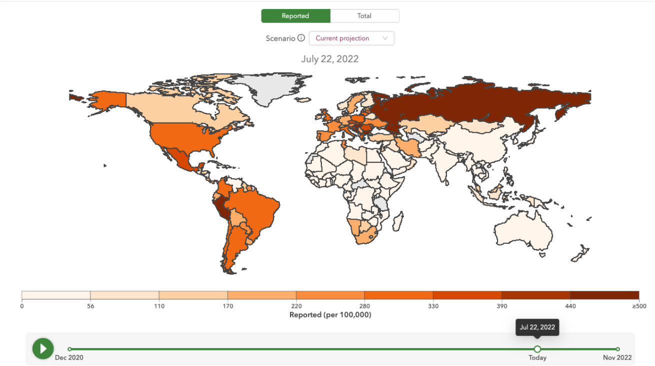
APPLIED MATHEMATICS
UNIVERSITY of WASHINGTON

IHME

Feature Selection for Mixed-Effect Models

Mixed-effect models

- Used for analyzing **combined data** across a range of **groups**.
- Use covariates to separate the **population variability** from the **group variability**.
- **Borrow strength** across groups to estimate key statistics.



[1]

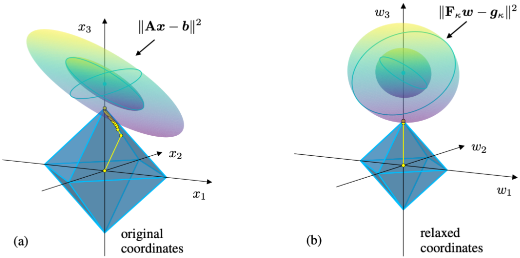
[1] Picture is taken from covid19.healthdata.org

Feature Selection for Mixed-Effect Models

Practitioners:

- Often seek **sparse models** that select **most informative** covariates.
- Want to be **flexible but efficient** in using various sparsity-promoting terms.
- Want a library to be **universal and compatible** with e.g. `sklearn`.

Sparse Relaxed Regularized Regression ($\mathcal{SR}3$)[2] showed great results for t linear models:



***Goal**: create a feature selection library that uses a relaxation approach for feature-selection in mixed-effect models.*

[2] Zheng and Aravkin, "Relax-and-split method for nonconvex inverse problems".

Linear Mixed-Effect (LME) 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 features
- $y_i \in \mathbb{R}^{n_i}$ – group i observations

Model:

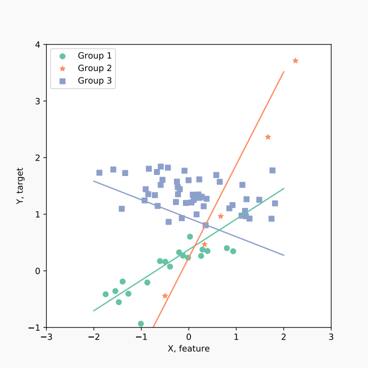
$$y_i = X_i\beta + Z_i u_i + \varepsilon_i$$
$$\varepsilon_i \sim \mathcal{N}(0, \Lambda_i)$$
$$u_i \sim \mathcal{N}(0, \Gamma)$$

Equivalently:

$$y_i = X_i\beta + \omega_i$$
$$\omega_i \sim \mathcal{N}(0, Z_i \Gamma Z_i^T + \Lambda_i)$$

Simplifying assumption:

$$\Gamma = \mathrm{diag}\left(()\, \gamma\right)$$



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 $\Gamma = \mathrm{diag}\left((\gamma)\right)$
- $\varepsilon_i \in \mathbb{R}^{n_i}$ – observation noise
- $\Lambda_i \in R^{n_i \times n_i}$ – covariance matrix for noise

Unknowns: β, u_i, γ, sometimes Λ_i.

Likelihood for Mixed Models

Optimization problem:

$$\mathcal{FS} - \mathcal{LME} \qquad \min_{\beta \in \mathbb{R}^p,\, \gamma \in \mathbb{R}^q_+} \mathcal{L}(\beta, \gamma) + R(\beta, \gamma) \tag{2}$$

Where $\mathcal{L}$:

$$\begin{aligned}\mathcal{L}(\beta, \gamma) = \sum_{i=1}^{m} \frac{1}{2}(y_i - X_i\beta)^T (Z_i \Gamma Z_i^T + \Lambda_i)^{-1} (y_i - X_i\beta) + \\ + \frac{1}{2} \log\det\left(Z_i \Gamma Z_i^T + \Lambda_i\right), \quad \Gamma = \operatorname{diag}\left((\gamma)\right)\end{aligned} \tag{3}$$

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

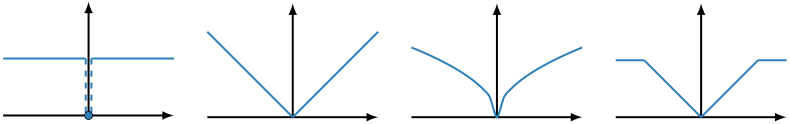
Regularization

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

$$\operatorname{prox}_{\alpha R + \delta_{\mathcal{C}}}(\tilde{\beta}, \tilde{\gamma}) := \underset{(\beta,\gamma)\in\mathcal{C}}{\operatorname{argmin}}\, R(\beta, \gamma) + \frac{1}{2\alpha}\|(\beta, \gamma) - (\tilde{\beta}, \tilde{\gamma})\|_2^2, \quad \text{where } \mathcal{C} := \mathbb{R}^p \times R_+^q \tag{4}$$

Examples:

- $R(x) = \lambda \sum_{j=1}^{p} w_j \|x_j\|_1$ – LASSO and Adaptive LASSO penalties[3]
- $R(x) = \lambda \|x\|_0$ – ℓ_0 penalty[4]
- $R(x)$ – SCAD penalty ([5])



(a) ℓ_0 (b) ℓ_1 (c) $\ell_p,\ p = 1/2$ (d) SCAD

Figure: Four commonly-used regularizers which promote sparsity

[3] Bondell, Krishna, and Ghosh, "Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models"; Lin, Pang, and Jiang, "Fixed and random effects selection by REML and pathwise coordinate optimization".

[4] Vaida and Blanchard, "Conditional Akaike information for mixed-effects models"; Jones, "Bayesian information criterion for longitudinal and clustered data".

[5] J. Fan and Li, "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties"; Y. Fan and Li, "Variable selection in linear mixed effects models".

SR3-Relaxation for Mixed-Effect Models ($\mathcal{MSR}3$)

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

$$\min_{\beta \in \mathbb{R}^p,\, \gamma \in \mathbb{R}^q_+} \mathcal{L}(\beta, \gamma) + R(\beta, \gamma) \tag{5}$$

Relaxed problem $\mathcal{MSR}3$:

$$\min_{\beta, \tilde{\beta} \in \mathbb{R}^p,\, \gamma, \tilde{\gamma} \in \mathbb{R}^q_+} \mathcal{L}(\beta, \gamma) + \phi_\mu(\gamma) + \kappa_\eta(\beta - \tilde{\beta}, \gamma - \tilde{\gamma}) + R(\tilde{\beta}, \tilde{\gamma}) \tag{6}$$

where the *relaxation* κ_η decouples the likelihood and the regularizer

$$\kappa_\eta(\beta - \tilde{\beta}, \gamma - \tilde{\gamma}) := \frac{\eta}{2}\|\beta - \tilde{\beta}\|_2^2 + \frac{\eta}{2}\|\gamma - \tilde{\gamma}\|_2^2, \quad \eta > \bar{\eta} \tag{7}$$

and the *perspective mapping* ϕ_μ replaces $\gamma \geq 0$ with a log-barrier

$$\phi_\mu(\gamma) := \begin{cases} -\mu \sum_{i=1}^q \ln(\gamma_i/\mu), & \mu > 0 \\ \delta_{\mathbb{R}^q_+}(\gamma), & \mu = 0 \\ +\infty, & \mu < 0 \end{cases} \tag{8}$$

Value Function Reformulation

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

$$v_{\eta,\mu}(\tilde{\beta},\tilde{\gamma}) := \min_{(\beta,\gamma)} \mathcal{L}_{\eta,\mu}((\beta,\gamma),(\tilde{\beta},\tilde{\gamma})) \\ := \min_{(\beta,\gamma)} \mathcal{L}(\beta,\gamma) + \phi_\mu(\gamma) + \kappa_\eta(\beta-\tilde{\beta},\gamma-\tilde{\gamma}) \tag{9}$$

so $\mathcal{MSR}3$-formulation (6) becomes

$$\min_{\beta\in\mathbb{R}^p,\ \gamma\in\mathbb{R}^q_+} v_{\eta,\mu}(\tilde{\beta},\tilde{\gamma}) + R(\tilde{\beta},\tilde{\gamma}) \tag{10}$$

When η is larger than the weak-convexity constant

- $v_{\eta,\mu}$ is well-defined and continuously differentiable.
- As $\mu \to 0$ and $\eta \to \infty$, cluster points of solutions to $\mathcal{MSR}3$ are first-order stationary points for $\mathcal{FS}-\mathcal{LME}$
- $v_{\eta,\mu}$ don't need to be evaluated precisely.

Value Function Reformulation

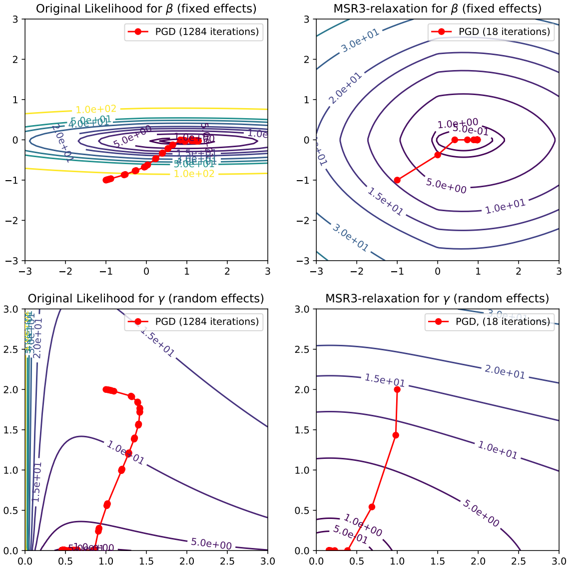


Figure: Comparison of the level-sets for the original likelihood (left) and $\mathcal{MSR}3$-likelihood (right), for fixed (top) and random (bottom) effects.

Designing an Algorithm

$G_{\nu,\eta}$ encodes both gradient of a Lagrangian (lines 1-2) and the complementarity condition (line 3):

$$G_{\nu,\eta}((\beta,\gamma,v),(\tilde{\beta},\tilde{\gamma})) := \begin{bmatrix} \nabla_\beta \mathcal{L}(\beta,\gamma) + \eta(\beta-\tilde{\beta}) \\ \nabla_\gamma \mathcal{L}(\beta,\gamma) + \eta(\gamma-\tilde{\gamma}) - v \\ v \bigodot \gamma - \mu\mathbf{1} \end{bmatrix} \tag{11}$$

We apply Newton method to G while geometrically decreasing μ.
Lemma: For every $(\mu,\eta) \in \mathbb{R}_+ \times \mathbb{R}_{++}$,

$$\begin{aligned} (\hat{\beta},\hat{\gamma}) &= \operatorname*{argmin}_{(\beta,\gamma)} \mathcal{L}_{\eta,\mu}((\beta,\gamma),(\tilde{\beta},\tilde{\gamma})) \\ &\iff \\ \exists \hat{v} &\in \mathbb{R}_+^q \text{ s.t. } G_{\nu,\eta}((\beta,\gamma,\hat{v}),(\tilde{\beta},\tilde{\gamma})) = 0 \end{aligned} \tag{12}$$

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**

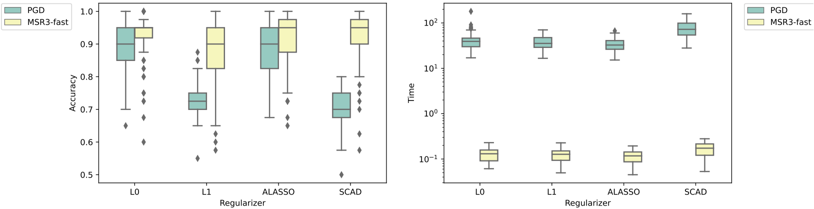
> $\beta \leftarrow \beta^+$; $\gamma \leftarrow \gamma^+$; $\tilde\beta \leftarrow \tilde\beta^+$; $\tilde\gamma \leftarrow \tilde\gamma^+$
>
> $[dv, d\beta, d\gamma] \leftarrow \nabla G_\mu((\beta, \gamma, v), (\tilde\beta, \tilde\gamma))^{-1} G_\mu((\beta, \gamma, v), (\tilde\beta, \tilde\gamma))$
> $\alpha \leftarrow 0.99 \times \min\left(1, -\frac{\gamma_i}{d\gamma_i}, \forall i: \ d\gamma_i < 0\right)$
>
> $\beta^+ \leftarrow \beta + \alpha d\beta$; $\gamma^+ = \gamma + \alpha d\gamma$; $v^+ \leftarrow v + \alpha dv$
>
> **if** $\underline{\|\gamma^+ \odot v^+ - q^{-1} {\gamma^+}^T v^+ \mathbf{1}\| > 0.5 q^{-1} {v^+}^T \gamma^+}$ **then** continue;
>
> **else**
>
> > $\tilde\beta^+ = \mathrm{prox}_{\alpha R}(\beta^+)$; $\tilde\gamma^+ = \mathrm{prox}_{\alpha R + \delta_{\mathbb{R}_+}}(\gamma^+)$; $\mu = \frac{1}{10} \frac{{v^+}^T \gamma^+}{q}$
>
> **end**
>
> `progress` $= (\|\beta^+ - \beta\| \geq$ tol or $\|\gamma^+ - \gamma\| \geq$ tol or $\|\tilde\beta^+ - \tilde\beta\| \geq$ tol or $\|\tilde\gamma^+ - \tilde\gamma\| \geq$ tol)
>
> `iter += 1`

end

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

Application to Synthetic Problems

- The number of fixed effects p and random effects q is 20.
- $\beta = \gamma = \frac{1}{2}[1, 2, 3, \dots, 10, 0 \dots, 0]$
- 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



+ $\mathcal{MSR}3$-relaxation improves feature selection performance of the original likelihood.
+ $\mathcal{MSR}3$-fast optimization accelerates the compute time by $\sim 10^2$.
− Initialization of η is problem-specific

Comparison to Other Libraries

Algorithm	MSR3-Fast (ℓ_1)	glmmLasso[6][7]	lmmLasso[8][9]	PGD (ℓ_1)
Accuracy, %	**88**	48	66	73
FE Accuracy, %	**86**	52	47	56
RE Accuracy, %	**91**	45	84	**91**
Time, sec	**0.19**	1.37	11.51	38.39
Iterations, num	34	50	-	7693

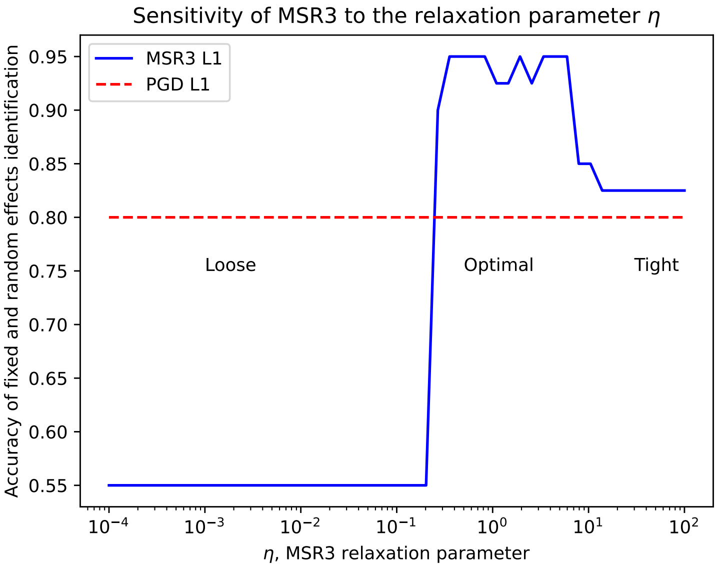
[6] https://rdrr.io/cran/glmmLasso/man/glmmLasso.html

[7] Groll and Tutz, "Variable selection for generalized linear mixed models by L 1-penalized estimation".

[8] https://rdrr.io/cran/lmmlasso/

[9] Schelldorfer, Bühlmann, and DE GEER, "Estimation for high-dimensional linear mixed-effects models using l1-penalization".

Choice of η



ℓ_0-based Covariate Selection for Bullying Study from GBD

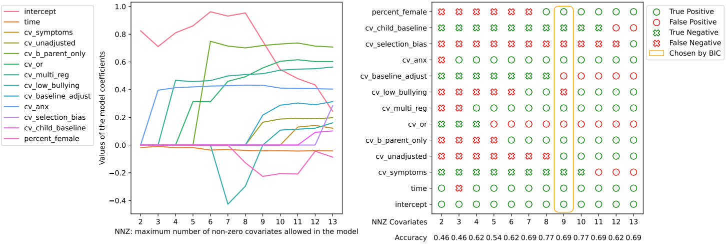


Figure: Fixed and random covariate selection for Bullying dataset from [10]. The model selected 9 covariates, 7 of which were historically significant, and did not select 4 covariates, 1 of which was historically significant.

[10] Institute for Health Metrics and Evaluation (IHME). Bullying Victimization Relative Risk Bundle GBD 2020. Seattle, United States of America (USA), 2021.

Software

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 I

Bondell, Howard D., Arun Krishna, and Sujit K. Ghosh. "Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models". In: Biometrics 66.4 (Dec. 2010), pp. 1069–1077. ISSN: 0006341X. DOI: 10.1111/j.1541-0420.2010.01391.x. arXiv: NIHMS150003. URL: http://doi.wiley.com/10.1111/j.1541-0420.2010.01391.x.

Fan, Jianqing and Runze Li. "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties". In: Journal of the American Statistical Association 96.456 (Dec. 2001), pp. 1348–1360. ISSN: 0162-1459. DOI: 10.1198/016214501753382273. URL: http://www.tandfonline.com/doi/abs/10.1198/016214501753382273.

Fan, Yingying and Runze Li. "Variable selection in linear mixed effects models". In: The Annals of Statistics 40.4 (Aug. 2012), pp. 2043–2068. ISSN: 0090-5364. DOI: 10.1214/12-AOS1028. URL: http://projecteuclid.org/euclid.aos/1351602536.

Groll, Andreas and Gerhard Tutz. "Variable selection for generalized linear mixed models by L 1-penalized estimation". In: Statistics and Computing 24.2 (2014), pp. 137–154.

Jones, Richard H. "Bayesian information criterion for longitudinal and clustered data". In: Statistics in Medicine 30.25 (Nov. 2011), pp. 3050–3056. ISSN: 02776715. DOI: 10.1002/sim.4323. URL: http://doi.wiley.com/10.1002/sim.4323.

Lin, Bingqing, Zhen Pang, and Jiming Jiang. "Fixed and random effects selection by REML and pathwise coordinate optimization". In: Journal of Computational and Graphical Statistics 22.2 (2013), pp. 341–355. ISSN: 10618600. DOI: 10.1080/10618600.2012.681219.

Schelldorfer, Jürg, Peter Bühlmann, and SARA VAN DE GEER. "Estimation for high-dimensional linear mixed-effects models using l1-penalization". In: Scandinavian Journal of Statistics 38.2 (2011), pp. 197–214.

References II

Vaida, Florin and Suzette Blanchard. "Conditional Akaike information for mixed-effects models". In: Biometrika 92.2 (June 2005), pp. 351–370. ISSN: 1464-3510. DOI: 10.1093/biomet/92.2.351. URL: http://academic.oup.com/biomet/article/92/2/351/233128/Conditional-Akaike-information-for-mixedeffects.

Zheng, Peng and Aleksandr Aravkin. "Relax-and-split method for nonconvex inverse problems". In: Inverse Problems 36.9 (2020). ISSN: 13616420. DOI: 10.1088/1361-6420/aba417.