Feature Selection for Mixed-Effects Models

Aleksei Sholokhov

Monday 4th July, 2022

Plan

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).

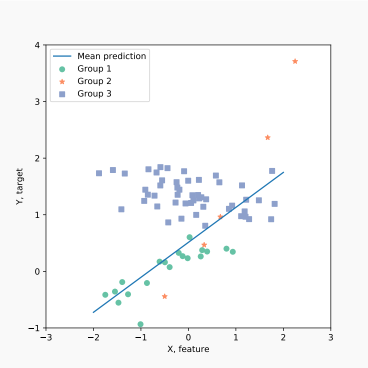
LMEs borrow strength across groups to estimate key statistics in cases when data within units are sparse or highly variable.

Linear Mixed-Effect Models

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

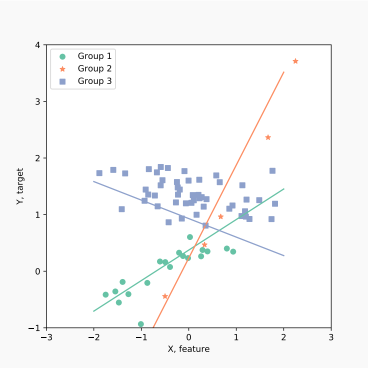
- $X_i \in \mathbb{R}^{n_i \times p}$ – group i design matrix for fixed features
- $Z_i \in \mathbb{R}^{n_i \times q}$ – group i design matrix for random effects
- $y_i \in \mathbb{R}^{n_i}$ – group i observations

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

Unknowns: β, u_i, γ, sometimes Λ_i.

Likelihood for Mixed Models

Negative log-likelihood:

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

Maximum likelihood estimates for β and γ solve the problem:

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

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

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

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

Regularization

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

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

Examples:

- $R(x) = \lambda \sum_{j=1}^{p} w_j \|x_j\|_1$ – LASSO and Adaptive LASSO penalties [BKG10, LPJ13]
- $R(x) = \lambda \|x\|_0$ – ℓ_0 penalty [VB05, Jon11]
- $R(x)$ – SCAD penalty ([FL01, FL12])

TODO: add picture

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 $\mathcal{MSR}3$:

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

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

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

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

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

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 (β^*,γ^*) 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 is enough.

Value Function Reformulation

TODO: insert image here

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 $(\mu,\eta) \in \mathbb{R}_+ \times \mathbb{R}_{++}$,

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

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 $0 \leq \gamma$.

$\mathcal{MSR}3$ 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 iter $<$ max_iter and $\|G_\mu(\beta^+, \gamma^+, v^+)\| >$ tol and progress **do**

- $\beta \leftarrow \beta^+$; $\gamma \leftarrow \gamma^+$; $\tilde{\beta} \leftarrow \tilde{\beta}^+$; $\tilde{\gamma} \leftarrow \tilde{\gamma}^+$
- $[dv, d\beta, d\gamma] \leftarrow \nabla G_\mu((\beta, \gamma, v), (\tilde{\beta}, \tilde{\gamma}))^{-1} G_\mu((\beta, \gamma, v), (\tilde{\beta}, \tilde{\gamma}))$ // Newton Iteration
- $\alpha \leftarrow 0.99 \times \min\left(1, -\frac{\gamma_i}{d\gamma_i}, \forall i : \ d\gamma_i < 0\right)$
- $\beta^+ \leftarrow \beta + \alpha d\beta$; $\gamma^+ = \gamma + \alpha d\gamma$; $v^+ \leftarrow v + \alpha dv$
- **if** $\|\gamma^+ \odot v^+ - q^{-1}{\gamma^+}^T v^+ 1\| > 0.5 q^{-1} {v^+}^T \gamma^+$ **then** continue;
- **else**
  - $\tilde{\beta}^+ = \mathrm{prox}_{\alpha R}(\beta^+)$; $\tilde{\gamma}^+ = \mathrm{prox}_{\alpha R + \delta_{\mathbb{R}_+}}(\gamma^+)$; $\mu = \frac{1}{10} \frac{{v^+}^T \gamma^+}{q}$ // Near central path
- **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 $\tilde{\beta}^+, \tilde{\gamma}^+$

Performance in Comparison to Other Algorithms

- **Scenario 1**: $n = 30$, $n_i = 5$, $p = 9$, $q = 4$, with true parameters $\beta = (1, 1, 0, \ldots, 0)$ and the covariance matrix Γ being:

$$\Gamma = \begin{bmatrix} 9 & 4.8 & 0.6 & 0 \\ 4.8 & 4 & 1 & 0 \\ 0.6 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{14}$$

- **Scenario 2**: everything as in Scenario 1, but $n = 60$ and $n_i = 10$.

Competitors:

- **ALASSO:** 2 stage: A-LASSO+Newton and A-LASSO+PCO
- **M-ALASSO:** Adaptive LASSO + EM Algorithm
- **SCAD-P:** SCAD + Proxy Matrix for Γ
- **rPQL:** Quasi-Likelihood + Adaptive LASSO (for GLMMs)

Performance in Comparison to Other Algorithms

Setup	Algoritm	% C	% CF	% CR	MSE	TIME
$n = 30, \ n_i = 5$	$\mathcal{MSR}3$	58	72	78	0.66	0.015
	rPQL	88	98	88	0.88	26-59
	M-ALASSO	71	73	79	-	-
	ALASSO	79	81	96	-	-
	SCAD-P	-	90	86	-	-
$n = 60, \ n_i = 10$	$\mathcal{MSR}3$	98	100	98	0.69	0.018
	rPQL	98	99	98	0.97	26-59
	M-ALASSO	83	83	89	-	-
	ALASSO	95	96	99	-	-
	SCAD-P	100	100	100	-	-

Table: Comparison of feature selection algorithms. % CF – percent of models where true fixed effects were identified correctly, % CR – percent of models where true random effects were identified correctly, % C – both fixed and random effects were identified correctly.

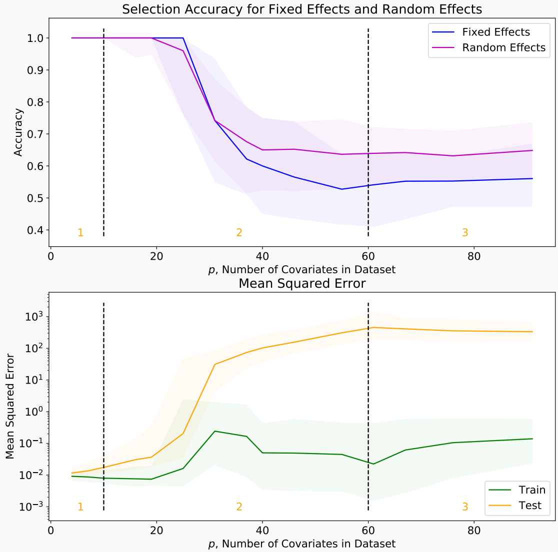
Scalability Experiment

- $n = 60,\ n_i = 10$
- $p = q \in [4, 7, 10, \dots, 90]$, 200 experiments for each.
- $X_i = Z_i$, columns are drawn form $\mathcal{N}(0, \Psi)$ where

$$\Psi = \begin{bmatrix} 9 & 4.8 & 0.6 \\ 4.8 & 4 & 1 \\ 0.6 & 1 & 1 \end{bmatrix}$$

- 50% random coordinate in β are active
- 70% of those are also active in γ

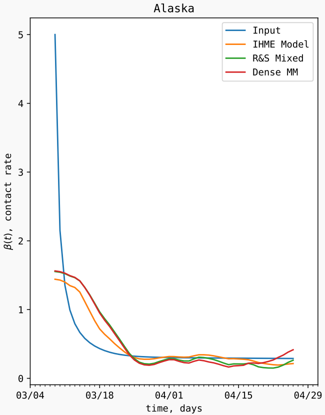
Scalability Experiment



Contact Rate Modeling for COVID-19 Forecasting

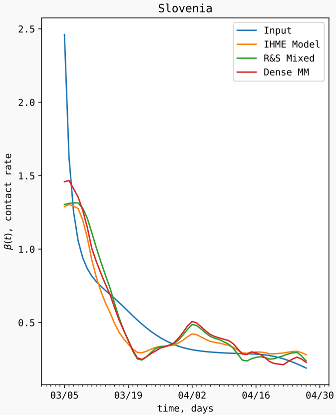
- $n = 60$ groups (countries and US states), $n_i \approx 50$
- Y_i – contact rate for COVID SEIIR Model
- $p = q = 5$ covariates related to temperature, mobility, population, testing; plus intercept.

Contact Rate Modeling for COVID-19 Forecasting



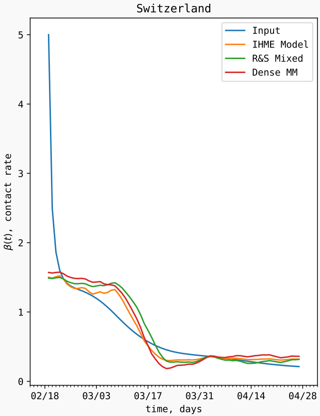

```
RMSE:
  IHME      : 3.85e+00
  Dense MM  : 3.98e+00   +3%
  R&S Mixed : 4.00e+00   +4%

Full MM Coefficients:
name                        local      mean        RE       Var
  intercept              1.23e+01  1.34e+01 -1.08e+00  5.66e-01
  temperature           -6.75e+02 -6.75e+02  0.00e+00  2.19e-20
  mobility_lift          6.44e+01  6.13e+01  3.09e+00  1.83e+03
  proportion_over_1k    -7.14e+00 -7.14e+00  0.00e+00  9.61e-21
  testing_reference      6.11e+00 -2.21e+00  8.32e+00  5.09e+02

R&S Mixed Coefficients:
name                        local      mean        RE       Var
  intercept              1.22e+01  1.43e+01 -2.05e+00  7.95e+01
  temperature           -6.75e+02 -6.75e+02  0.00e+00  0.00e+00
  mobility_lift          5.78e+01  6.00e+01 -2.20e+00  1.83e+03
  proportion_over_1k     0.00e+00  0.00e+00  0.00e+00  0.00e+00
  testing_reference      0.00e+00  0.00e+00  0.00e+00  0.00e+00

Legend:
  Both Fixed and Random
  Fixed Only
  Excluded
```

Contact Rate Modeling for COVID-19 Forecasting




```
RMSE:
  IHME      : 1.46e+00
  Dense MM  : 1.40e+00   -4%
  R&S Mixed : 1.56e+00   +7%

Full MM Coefficients:
name                       local      mean        RE       Var
  intercept             1.31e+01  1.34e+01 -3.62e-01  5.66e-01
  temperature          -6.75e+02 -6.75e+02  0.00e+00  2.19e-20
  mobility_lift         3.28e+01  6.13e+01 -2.85e+01  1.83e+03
  proportion_over_1k   -7.14e+00 -7.14e+00  0.00e+00  9.61e-21
  testing_reference    -3.16e+01 -2.21e+00 -2.93e+01  5.09e+02

R&S Mixed Coefficients:
name                       local      mean        RE       Var
  intercept             1.27e+01  1.43e+01 -1.55e+00  7.95e+01
  temperature          -6.75e+02 -6.75e+02  0.00e+00  0.00e+00
  mobility_lift         4.27e+01  6.00e+01 -1.73e+01  1.83e+03
  proportion_over_1k    0.00e+00  0.00e+00  0.00e+00  0.00e+00
  testing_reference     0.00e+00  0.00e+00  0.00e+00  0.00e+00

Legend:
  Both Fixed and Random
  Fixed Only
  Excluded
```

Contact Rate Modeling for COVID-19 Forecasting



RMSE:
- IHME : 3.79e+00
- Dense MM : 3.88e+00 +2%
- R&S Mixed : 3.98e+00 +5%

Full MM Coefficients:

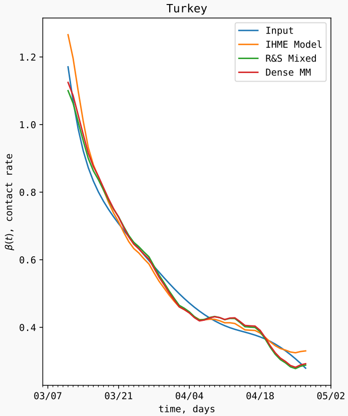
name	local	mean	RE	Var
intercept	1.29e+01	1.34e+01	-5.34e-01	5.66e-01
temperature	-6.75e+02	-6.75e+02	0.00e+00	2.19e-20
mobility_lift	4.82e+01	6.13e+01	-1.31e+01	1.83e+03
proportion_over_1k	-7.14e+00	-7.14e+00	0.00e+00	9.61e-21
testing_reference	-2.31e+01	-2.21e+00	-2.09e+01	5.09e+02

R&S Mixed Coefficients:

name	local	mean	RE	Var
intercept	1.27e+01	1.43e+01	-1.61e+00	7.95e+01
temperature	-6.75e+02	-6.75e+02	0.00e+00	0.00e+00
mobility_lift	6.65e+01	6.00e+01	6.49e+00	1.83e+03
proportion_over_1k	0.00e+00	0.00e+00	0.00e+00	0.00e+00
testing_reference	0.00e+00	0.00e+00	0.00e+00	0.00e+00

Legend:
- Both Fixed and Random
- Fixed Only
- Excluded

Contact Rate Modeling for COVID-19 Forecasting

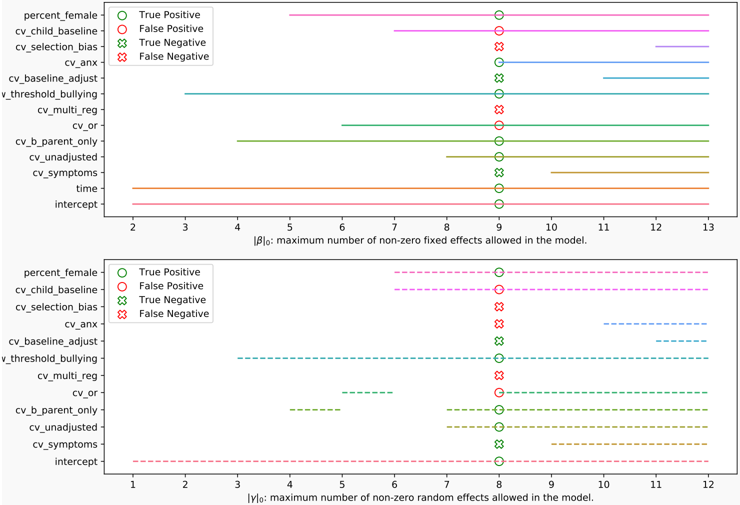



```
RMSE:
  IHME      : 2.67e-01
  Dense MM  : 1.85e-01   -31%
  R&S Mixed : 1.71e-01   -36%

Full MM Coefficients:
name                       local      mean        RE       Var
  intercept             1.28e+01  1.34e+01 -6.56e-01  5.66e-01
  temperature          -6.75e+02 -6.75e+02  0.00e+00  2.19e-20
  mobility_lift         3.05e+01  6.13e+01 -3.08e+01  1.83e+03
  proportion_over_1k   -7.14e+00 -7.14e+00  0.00e+00  9.61e-21
  testing_reference     1.49e+00 -2.21e+00  3.70e+00  5.09e+02

R&S Mixed Coefficients:
name                       local      mean        RE       Var
  intercept             1.26e+01  1.43e+01 -1.69e+00  7.95e+01
  temperature          -6.75e+02 -6.75e+02  0.00e+00  0.00e+00
  mobility_lift         2.88e+01  6.00e+01 -3.12e+01  1.83e+03
  proportion_over_1k    0.00e+00  0.00e+00  0.00e+00  0.00e+00
  testing_reference     0.00e+00  0.00e+00  0.00e+00  0.00e+00

Legend:
  Both Fixed and Random
  Fixed Only
  Excluded
```

Burden of Anxiety and Depression as Result of Bullying

- $m = 10$ cohort studies, $n = 77$, highly unbalanced
- $p = 13$, $q = 12$ (`time` was preselected fixed-only)
- Covariates are related to studies' designs.

Burden of Anxiety and Depression as Result of Bullying



Future Work: Theory

Theorem (Conditions for Convergence to True Estimator)

Under certain conditions the method converges in a finite number of iterations to $(\hat{\beta}, \hat{\gamma})$ *which projections* $(\tilde{\beta}, \tilde{\gamma})$ *belong to a k- and j-subspaces respectively that contain the true minimum* (β^*, γ^*).

Theorem (Consistency of Estimator)

There exists a local minimizer $(\hat{\beta}, \hat{\gamma})$ *for the proposed loss function, such that it is asymptotically consistent with true minimum* (β^*, γ^*).

Theorem (Consistency in Zeros)

If some coordinates of the true minimizer (β^*, γ^*) *are zero, then it is also zero in* $(\hat{\beta}, \hat{\gamma})$, *given that the later is sufficiently close to the former.*

Theorem (Asymptotic Normality)

The proposed estimator $(\hat{\beta}, \hat{\gamma})$ *asymptotically normally distributed around true minimizer* (β^*, γ^*) *in its true non-zero* $k + j$*-subspace.*

Future work: Algorithm

Question: Will exponential smoothing of projection improve the accuracy?

$\lambda_\beta = 0$; $\lambda_\gamma = 0$
repeat
  $\lambda_\beta \leftarrow 2(1 + \lambda_\beta)$
  $\lambda_\gamma \leftarrow 2(1 + \lambda_\gamma)$
  repeat
    $\tilde{\beta}^{(k+1)} \leftarrow \delta \, \text{Proj}_{\|\beta\|_0 \le k}(\beta^{(k)}) + (1-\delta)\tilde{\beta}^{(k)}$
    $\tilde{\gamma}^{(k+1)} \leftarrow \delta \, \text{Proj}_{\|\gamma\|_0 \le s}(\gamma^{(k)}) + (1-\delta)\tilde{\gamma}^{(k)}$
    $\beta^{(k+1)}, \gamma^{(k+1)} \leftarrow \text{argmin}_{\gamma \ge 0, \beta} \, \mathcal{L}(\beta, \gamma) + \frac{\lambda_\beta}{2}\|\beta - \tilde{\beta}^{(k)}\|_2^2 + \frac{\lambda_\gamma}{2}\|\gamma - \tilde{\gamma}^{(k)}\|_2^2$
  until converges;
until $\tilde{\beta} \approx \beta, \tilde{\gamma} \approx \gamma$;

Future Work: Implementation

Can we increase λ_β and λ_γ in a more careful way to avoid potential stacking? The approach can be based on the theorem:

Theorem (Distance Between Minima)

For a fixed dataset (X_i, Y_i) *and relaxation parameters* λ_β, λ_γ *the distance between* (β^*, γ^*), *the unconstrained minimizers of relaxed problem, and their projections* $(\tilde{\beta}^*, \tilde{\gamma}^*)$ *is bounded by a constant* M *depending on* (X_i, Y_i) *and the relaxation parameters.*

The End

Thank you for your attention!

References I

[BKG10] 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.

[FL01] 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.

[FL12] Yingying Fan and Runze Li. Variable selection in linear mixed effects models. The Annals of Statistics, 40(4):2043–2068, aug 2012.

[Jon11] Richard H. Jones. Bayesian information criterion for longitudinal and clustered data. Statistics in Medicine, 30(25):3050–3056, nov 2011.

[LPJ13] 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.

[VB05] Florin Vaida and Suzette Blanchard. Conditional Akaike information for mixed-effects models. Biometrika, 92(2):351–370, jun 2005.