Feature Selection for Mixed-Effects Models

Aleksei Sholokhov

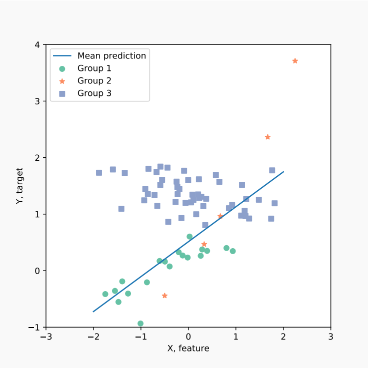
Thursday, 3^{rd} of December, 2020

Plan

Linear Mixed-Effect Models are a type of regression models for grouped data

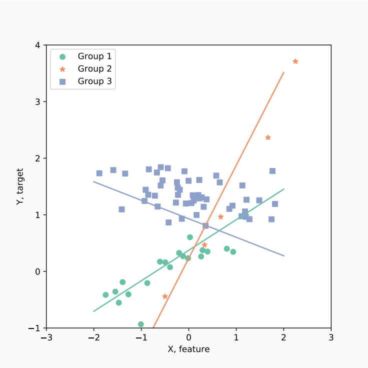
Dataset: m groups $(X_i, y_i), \quad i = 1, \dots m$

Standard Linear Regression:



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

Linear Mixed-Effect Model:



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

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

Notation

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

Notation

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

- $X_i \in \mathbb{R}^{n_i \times p}$ – features, or covariates,

Notation

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

- $X_i \in \mathbb{R}^{n_i \times p}$ – fixed features, or covariates, $Z_i \in \mathbb{R}^{n_i \times q}$ – random features
- p – number of fixed features, q – number of random effects.

Notation

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

- $X_i \in \mathbb{R}^{n_i \times p}$ – fixed features, or covariates, $Z_i \in R^{n_i \times q}$ – random features
- p – number of fixed features, q – number of random effects.
- $\beta \in \mathbb{R}^p$ – fixed effects, or mean effects

Notation

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

- $X_i \in \mathbb{R}^{n_i \times p}$ – fixed features, or covariates, $Z_i \in R^{n_i \times q}$ – random features
- 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

Notation

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

- $X_i \in \mathbb{R}^{n_i \times p}$ – fixed features, or covariates, $Z_i \in R^{n_i \times q}$ – random features
- 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

Notation

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

- $X_i \in \mathbb{R}^{n_i \times p}$ – fixed features, or covariates, $Z_i \in R^{n_i \times q}$ – random features
- 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
- $\varepsilon_i \in \mathbb{R}^{n_i}$ – observation noise

Notation

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

- $X_i \in \mathbb{R}^{n_i \times p}$ – fixed features, or covariates, $Z_i \in R^{n_i \times q}$ – random features
- 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
- $\varepsilon_i \in \mathbb{R}^{n_i}$ – observation noise
- $\Lambda_i \in R^{n_i \times n_i}$ – covariance matrix for noise

Notation

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

- $X_i \in \mathbb{R}^{n_i \times p}$ – fixed features, or covariates, $Z_i \in R^{n_i \times q}$ – random features
- 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
- $\varepsilon_i \in \mathbb{R}^{n_i}$ – observation noise
- $\Lambda_i \in R^{n_i \times n_i}$ – covariance matrix for noise
- Unknowns: β, u_i, Γ, sometimes Λ_i.

Notation

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

- $X_i \in \mathbb{R}^{n_i \times p}$ – fixed features, or covariates, $Z_i \in R^{n_i \times q}$ – random features
- 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
- $\varepsilon_i \in \mathbb{R}^{n_i}$ – observation noise
- $\Lambda_i \in R^{n_i \times n_i}$ – covariance matrix for noise
- Unknowns: β, u_i, Γ, sometimes Λ_i.

$$\Gamma = \mathrm{Diag}\,(\gamma),\ \gamma \in \mathbb{R}^q \quad \text{and } \Lambda_i \text{ are known.}$$

Likelihood for Mixed Models

Suppose we have a linear mixed effects model

$$\begin{aligned} Y_i &= X_i\beta + Z_i u_i + \varepsilon_i \\ u_i &\sim \mathcal{N}(0, \Gamma),\ \ \Gamma = \operatorname{Diag}\gamma \\ \varepsilon_i &\sim \mathcal{N}(0, \Lambda_i) \end{aligned} \tag{2}$$

where β, γ, and u_i are unknown model parameters, and λ_i are known.
Noticing that $Z_i u_i + \varepsilon_i \sim \mathcal{N}(0, Z_i \Gamma Z_i^T + \Lambda_i)$ we get the negative log-likelihood:

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

Feature Selection for Mixed Models

Consider the following feature selection setup:

$$\begin{aligned} &\min_{\beta,\gamma} \mathcal{L}(\beta,\gamma) \\ \text{s.t. } &\|\beta\|_0 \le k \\ &\|\gamma\|_0 \le j \\ &\gamma \ge 0 \end{aligned} \tag{4}$$

where k and j is the size of the sparse support for the fixed and random effects respectively.
This is a very hard problem to solve directly.

Methods for Feature Selection

There are several ways to address this setup:

(0) Exhaustive Search:

$$\begin{aligned} &\min_{\beta,\gamma} \mathcal{L}(\beta,\gamma) \\ \text{s.t. } &\|\beta\|_0 \le k \\ &\|\gamma\|_0 \le j \\ &\gamma \ge 0 \end{aligned}$$

AIC [VB05]
BIC [Jon11]

Methods for Feature Selection

There are several ways to address this setup:

(0) Exhaustive Search:

$$\begin{aligned} &\min_{\beta,\gamma} \mathcal{L}(\beta,\gamma) \\ \text{s.t.} \ & \|\beta\|_0 \leq k \\ & \|\gamma\|_0 \leq j \\ & \gamma \geq 0 \end{aligned}$$

AIC [VB05]
BIC [Jon11]

(1) ℓ_1-relaxation:

$$\begin{aligned} &\min_{\beta,\gamma} \mathcal{L}(\beta,\gamma) \\ \text{s.t.} \ & \|\beta\|_1 \leq \lambda_\beta \\ & \|\gamma\|_1 \leq \lambda_\gamma \\ & \gamma \geq 0 \end{aligned}$$

M-ALASSO [BKG10]
ALASSO [LPJ13]

Methods for Feature Selection

There are several ways to address this setup:

(0) Exhaustive Search:

$$\min_{\beta,\gamma} \mathcal{L}(\beta,\gamma)$$
$$\text{s.t. } \|\beta\|_0 \le k$$
$$\|\gamma\|_0 \le j$$
$$\gamma \ge 0$$

AIC [VB05]
BIC [Jon11]

(1) ℓ_1-relaxation:

$$\min_{\beta,\gamma} \mathcal{L}(\beta,\gamma)$$
$$\text{s.t. } \|\beta\|_1 \le \lambda_\beta$$
$$\|\gamma\|_1 \le \lambda_\gamma$$
$$\gamma \ge 0$$

M-ALASSO [BKG10]
ALASSO [LPJ13]

(3) Other types of sparsity-inducing penalties (ℓ_p, *SCAD*).

SCAD [FL01]
SCAD-P [FL12]

Methods for Feature Selection

There are several ways to address this setup:

(0) Exhaustive Search:

$$\min_{\beta,\gamma} \mathcal{L}(\beta,\gamma)$$
$$\text{s.t. } \|\beta\|_0 \le k$$
$$\|\gamma\|_0 \le j$$
$$\gamma \ge 0$$

(1) ℓ_1-relaxation:

$$\min_{\beta,\gamma} \mathcal{L}(\beta,\gamma)$$
$$\text{s.t. } \|\beta\|_1 \le \lambda_\beta$$
$$\|\gamma\|_1 \le \lambda_\gamma$$
$$\gamma \ge 0$$

(3) Other types of sparsity-inducing penalties (ℓ_p, *SCAD*).

(2) SR3-type Relaxation [ZAB+19]

$$\begin{aligned} \min_{\beta,\gamma,\hat\beta,\hat\gamma} \ & \mathcal{L}(\beta,\gamma) + \frac{\lambda_\beta}{2}\|\beta-\hat\beta\|_2^2 + \frac{\lambda_\gamma}{2}\|\gamma-\hat\gamma\|_2^2 \\ \text{s.t. } & \|\hat\beta\|_0 \le k \\ & \|\hat\gamma\|_0 \le j, \ \gamma \ge 0 \end{aligned} \tag{5}$$

Relax-and-Split from Optimization Perspective

$$
\begin{aligned}
\min_{\beta,\gamma,\hat\beta,\hat\gamma} & \mathcal{L}(\beta,\gamma) + \frac{\lambda_\beta}{2}\|\beta-\hat\beta\|_2^2 + \frac{\lambda_\gamma}{2}\|\gamma-\hat\gamma\|_2^2 \\
\text{s.t. } & \|\hat\beta\|_0 \le k,\ \|\hat\gamma\|_0 \le j \\
& \gamma \ge 0,\ \hat\gamma \ge 0
\end{aligned} \tag{6}
$$

We'll find β^*, γ^*, $\hat\beta^*$, and $\hat\gamma^*$ via alternating partial minimization:

Relax-and-Split from Optimization Perspective

$$
\begin{aligned}
\min_{\beta,\gamma,\hat{\beta},\hat{\gamma}} \mathcal{L}(\beta,\gamma) + \frac{\lambda_\beta}{2}\|\beta - \hat{\beta}\|_2^2 + \frac{\lambda_\gamma}{2}\|\gamma - \hat{\gamma}\|_2^2 \\
\text{s.t. } \|\hat{\beta}\|_0 \le k, \ \|\hat{\gamma}\|_0 \le j \\
\gamma \ge 0, \ \hat{\gamma} \ge 0
\end{aligned}
\tag{6}
$$

We'll find β^*, γ^*, $\hat{\beta}^*$, and $\hat{\gamma}^*$ via alternating partial minimization:

- Minimization w.r.t. β and γ is done via **Interior Point Method**.

Relax-and-Split from Optimization Perspective

$$
\begin{aligned}
\min_{\beta,\gamma,\hat{\beta},\hat{\gamma}} \mathcal{L}(\beta,\gamma) + \frac{\lambda_\beta}{2}\|\beta - \hat{\beta}\|_2^2 + \frac{\lambda_\gamma}{2}\|\gamma - \hat{\gamma}\|_2^2 \\
\text{s.t. } \|\hat{\beta}\|_0 \le k, \ \|\hat{\gamma}\|_0 \le j \\
\gamma \ge 0, \ \hat{\gamma} \ge 0
\end{aligned} \tag{6}
$$

We'll find β^*, γ^*, $\hat{\beta}^*$, and $\hat{\gamma}^*$ via alternating partial minimization:

- Minimization w.r.t. β and γ is done via **Interior Point Method**.
- Minimization w.r.t. $\hat{\beta}$ and $\hat{\gamma}$ is **projection** of β and γ onto ℓ_0-balls.

$$
\begin{aligned}
\hat{\beta}^{(k+1)} \leftarrow \operatorname*{Proj}_{\|\beta\|_0 \le k}(\beta^{(k)}) \quad &\Longrightarrow \quad \text{Take max k from } \beta\text{, rest to 0} \\
\hat{\gamma}^{(k+1)} \leftarrow \operatorname*{Proj}_{\|\gamma\|_0 \le s}(\gamma^{(k)}) \quad &\Longrightarrow \quad \text{Take max j from } \gamma\text{, rest to 0}
\end{aligned} \tag{7}
$$

Relax-and-Split from Optimization Perspective

$$\begin{aligned} \min_{\beta,\gamma,\hat\beta,\hat\gamma} & \mathcal{L}(\beta,\gamma) + \frac{\lambda_\beta}{2}\|\beta-\hat\beta\|_2^2 + \frac{\lambda_\gamma}{2}\|\gamma-\hat\gamma\|_2^2 \\ \text{s.t. } & \|\hat\beta\|_0 \le k, \ \|\hat\gamma\|_0 \le j \\ & \gamma \ge 0, \ \hat\gamma \ge 0 \end{aligned} \tag{6}$$

We'll find β^*, γ^*, $\hat\beta^*$, and $\hat\gamma^*$ via alternating partial minimization:

- Minimization w.r.t. β and γ is done via **Interior Point Method**.
- Minimization w.r.t. $\hat\beta$ and $\hat\gamma$ is **projection** of β and γ onto ℓ_0-balls.
- λ_β and λ_γ are **iteratively increased** until $\hat\beta \approx \beta$ and $\hat\gamma \approx \gamma$.

Relax-and-Split from Optimization Perspective

$$\begin{aligned} \min_{\beta,\gamma,\hat{\beta},\hat{\gamma}} & \mathcal{L}(\beta,\gamma) + \frac{\lambda_\beta}{2}\|\beta-\hat{\beta}\|_2^2 + \frac{\lambda_\gamma}{2}\|\gamma-\hat{\gamma}\|_2^2 \\ \text{s.t. } & \|\hat{\beta}\|_0 \le k,\ \|\hat{\gamma}\|_0 \le j \\ & \gamma \ge 0,\ \hat{\gamma} \ge 0 \end{aligned} \tag{6}$$

We'll find β^*, γ^*, $\hat{\beta}^*$, and $\hat{\gamma}^*$ via alternating partial minimization:

- Minimization w.r.t. β and γ is done via **Interior Point Method**.
- Minimization w.r.t. $\hat{\beta}$ and $\hat{\gamma}$ is **projection** of β and γ onto ℓ_0-balls.
- λ_β and λ_γ are **iteratively increased** until $\hat{\beta} \approx \beta$ and $\hat{\gamma} \approx \gamma$.

Advantages:

1. Interpretable hyperparameters.
2. No model bias when the $k-$subspace is chosen.

R&S-Mixed

The pseudocode of the proposed algorithm:

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

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:

$$\Gamma = \begin{bmatrix} 9 & 4.8 & 0.6 & 0 \\ 4.8 & 4 & 1 & 0 \\ 0.6 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{7}$$

- **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
$n = 30,\ n_i = 5$	**R&S-Mixed**	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	-	-
$n = 60,\ n_i = 10$	**R&S-Mixed**	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.

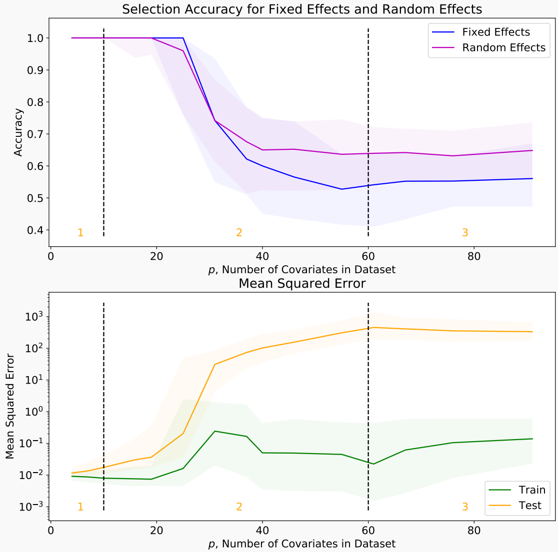
Scalability Experiment

- $n = 60, \; n_i = 10$
- $p = q \in [4, 7, 10, \ldots, 90]$, 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

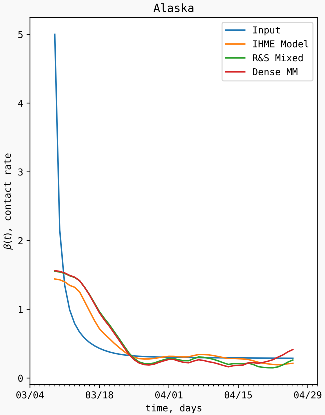
Scalability Experiment



Contact Rate Modeling for COVID-19 Forecasting

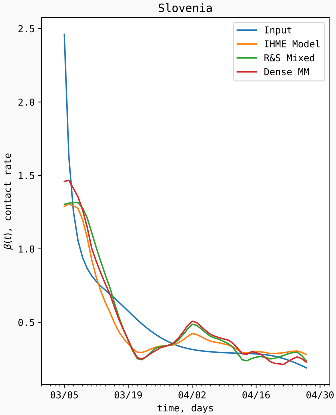
- $n = 60$ groups (countries and US states), $n_i \approx 50$
- Y_i – contact rate for COVID SEIIR Model
- $p = q = 5$ covariates related to temperature, mobility, population, testing; plus intercept.

Contact Rate Modeling for COVID-19 Forecasting



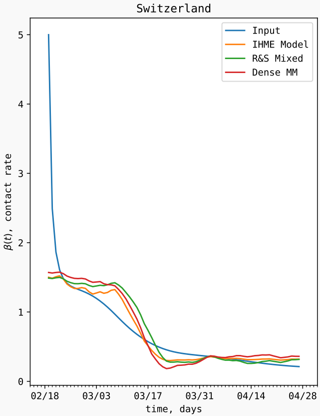

```
RMSE:
  IHME      : 3.85e+00
  Dense MM  : 3.98e+00   +3%
  R&S Mixed : 4.00e+00   +4%

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



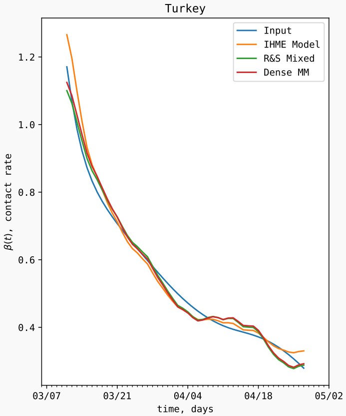

```
RMSE:
  IHME      : 1.46e+00
  Dense MM  : 1.40e+00   -4%
  R&S Mixed : 1.56e+00   +7%

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

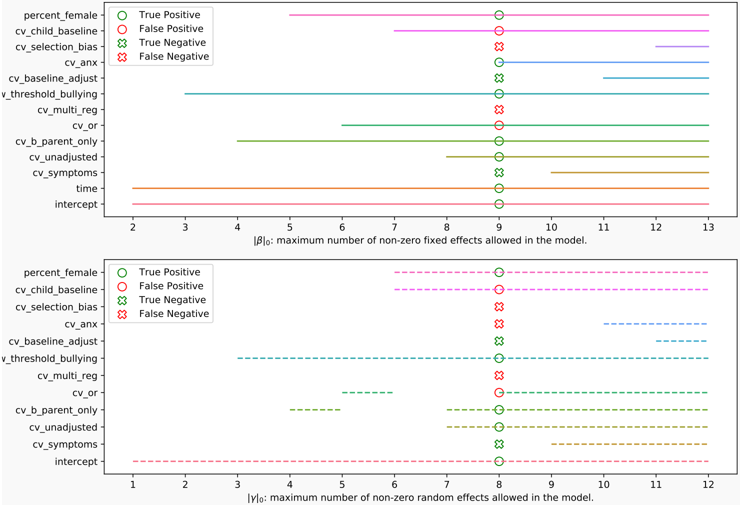



```
RMSE:
  IHME      : 2.67e-01
  Dense MM  : 1.85e-01   -31%
  R&S Mixed : 1.71e-01   -36%

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

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* $(\hat{\beta}^*, \hat{\gamma}^*)$ *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.

[ZAB+19] Peng Zheng, Travis Askham, Steven L. Brunton, J. Nathan Kutz, and Aleksandr Y. Aravkin. A Unified Framework for Sparse Relaxed Regularized Regression: SR3. IEEE Access, 7:1404–1423, 2019.