Covariate Selection Techniques for Linear Mixed-Effects Models with Applications to Population Health Problems

Aleksei Sholokhov

Committee
Aleksandr Aravkin (chair),
James Burke,
Nathan Kutz,
Jonathan Wakefield (GSR)

A thesis proposal submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

Department of Applied Mathematics

University of Washington

Last modified: Sunday 21st February, 2021

Abstract

Linear Mixed-Effects models are widely used technique for modeling clustered data, such as cohort studies, longitudinal data analysis, and meta-research. In this work we address an important practical problem of simultaneous selection of fixed and random effects in this type of models. Our approach is based on the Relax-and-Split methodology, which uses a relaxation technique together with partial minimization to efficiently solve non-convex problems. We use it to relax the constraints on the maximal number of included covariates and then apply an interior point method to obtain a solution. The performance of the resulting estimator is studied on synthetic data, as well as on two real-data applications: covariates selection for modeling the contact rate for COVID-19, and the estimation of burden of anxiety and depression disorders as results of bullying. These experiments show that the proposed method is advantageous in terms of execution time and selection quality. In addition, several theoretical questions are discussed as a part of future work that would contribute to understanding of convergence conditions of Relax-and-Split based methods in general.

Contents

1 Introduction **5**

1.1 Linear Mixed-Effects Models 5

1.1.1 EM algorithm for Mixed Models 8

1.1.2 Newton Methods for Mixed Models 10

1.2 Feature Selection for Mixed Models 12

2 Relax-and-Split for Mixed-Effects Selection **16**

2.1 Introduction 16

2.2 Proposed Algorithm 17

2.2.1 Implementation Details 21

2.3 Experiments on Synthetic Data 22

2.3.1 Comparison to Other Algorithms 22

2.3.2 Scalability Experiments 24

2.4 Experiments on Real Data 29

2.4.1 Meta-Research on Internet Bullying 29

2.4.2 COVID-19 Contact Rate Forecasting Data 29

2.5 Software Implementation 32

3 Discussion and Future work **33**

3.1 Better Control over Relaxation Parameters 33

3.2 Asymptotic Properties 34

3.3 Exponentially Smoothed Projection 35

3.4 Correlated Features 35

A Derivatives of Marginalized Log-likelihood for Linear Mixed Models **41**

B Description of Datasets **43**

B.1 GBD Bullying Data 43

B.2 COVID-19 Contact Rate Forecasting Data 44

Notation

$$
\begin{aligned}
n &\text{ – Total number of objects (observations) in a dataset} \\
m &\text{ – Total number of groups (clusters, studies) in a dataset} \\
n_i &\text{ – Number of objects (observations) in the group } i. \\
p &\text{ – Total number of fixed features (covariates, predictors)} \\
q &\text{ – Total number of random features (covariates, predictors)} \\
X_i &\text{ – Data (design) matrix of fixed features of the group } i,\ X_i \in \mathbb{R}^{n_i \times p} \\
Z_i &\text{ – Data (design) matrix of random features of the group } i,\ Z_i \in \mathbb{R}^{n_i \times q} \\
Y_i &\text{ – Target variable for the group } i \\
\beta &\text{ – Vector of fixed (mean) coefficients in a model, } \beta \in \mathbb{R}^p \\
\Gamma &\text{ – Covariance matrix of random effects in a model, } \Gamma \in \mathbb{R}^{q \times q} \\
\gamma &\text{ – Vector of variances of random effects when random effects are independent: } \Gamma = \operatorname{Diag}(\gamma),\ \gamma \in \mathbb{R}^q \\
u_i &\text{ – Random effects for the group } i,\ u_i \in \mathbb{R}^q,\ u_i \sim \mathcal{N}(0, \Gamma) \\
\Lambda_i &\text{ – Covariance matrix for measurement (observation) errors for the group } i \\
\sigma_i &\text{ – Vector of variances for measurement (observation) errors for the group } i \text{ when } \Lambda_i = \operatorname{Diag}(\sigma_i) \\
\varepsilon_i &\text{ – Vector of measurement (observation) errors for the group } i,\ \varepsilon_i \sim \mathcal{N}(0, \Lambda_i) \\
\Omega_i &\text{ – Covariance matrix of within-cluster observations, } \Omega_i = \Omega_i(\Gamma) = Z_i \Gamma Z_i^T + \Lambda_i,\ \Omega_i \in \mathbb{R}^{n_i \times n_i} \\
k &\text{ – Number of non-zero mean coefficients allowed in a model} \\
s &\text{ – Number of non-zero variances of random effects allowed in a model} \\
\tilde{\beta} &\text{ – Sparse counterpart of } \beta,\ \tilde{\beta} \in \mathbb{R}^p,\ \|\tilde{\beta}\|_0 = k \\
\tilde{\gamma} &\text{ – Sparse counterpart of } \gamma,\ \tilde{\gamma} \in \mathbb{R}^q,\ \|\tilde{\gamma}\|_0 = s \\
\mathcal{L}(u_i \mid \beta, \gamma) &\text{ – Negative log-likelihood of random effects of a linear mixed model} \\
\mathcal{L}_{ML}(\beta, \gamma) &\text{ – Negative marginalized log-likelihood of parameters of a linear mixed model} \\
\mathcal{L}_{REML}(\beta, \gamma) &\text{ – Negative marginalized restricted log-likelihood (REML) of a linear mixed model} \\
\mathcal{L}_r(\beta, \gamma) &\text{ – Relaxed marginalized log-likelihood of a linear mixed model} \\
a \circ b &\text{ – Element-wise (Hadamard) product of } a \text{ and } b
\end{aligned}
\tag{0.1}
$$

1 Introduction

Linear Mixed-Effects Models are widely used technique for modeling clustered data, such as cohort studies, longitudinal data analysis, and meta-research. A key step in these types of works is feature selection, which goal is choosing the most useful predictors from a large set of possible covariates. In the next section we introduce the reader to the concept of linear mixed-effect models and provide a literature review on the most commonly used optimization methods for finding estimators for them, such as Newton method, EM-algorithm, and Interior Point Method. Section 1.2 gives reader an overview of existing covariates selection methods, such as subset selection, shrinkage penalties, and AIC/BIC-based methods. Section 2 introduces Relax-and-Split approach to feature selection, outlines the proposed algorithm, and discusses optimization subroutines involved in it. In the Section 2.3 we test our approach in two simulation setups, with the goals to compare the proposed algorithm to its competitors (Section 2.3.1) and to show its applicability conditions and current limitations (Section 2.3.2). Section 2.4 shows the application of our approach to two real-world problems: the estimation of burden of anxiety and depression disorders as results of bullying (Section 2.4.1), and the covariates selection for modeling the contact rate for COVID-19 (Section 2.4.2). Section 3 discusses shortcomings of the current version of the algorithm and outlines theoretical questions and practical improvements which we plan address as a part of the thesis work.

1.1 Linear Mixed-Effects Models

Mixed-effect regression models are a particular kind of regression models for clustered data. They describe the relationship between an outcome variable and its predictors in a setting where the observations are grouped according to one or many criteria. These models are used in a wide range of applications, such as longitudinal studies for clinical trials, population health problems, risk-outcome studies, and meta-research.

Mathematically, let there be m groups of observations indexed by i, with sizes n_i respectively, and the total number of observations to be $n = n_1 + n_2 + \cdots + n_m$. Depending on the domain of application, these groups are sometimes referred to as studies or clusters. For each group, we have design matrices $X_i \in \mathbb{R}^{n_i \times p}$, which are called the matrices of fixed features, and $Z_i \in \mathbb{R}^{n_i \times q}$, called matrices of random features, and we observe vectors of outcomes $Y_i \in \mathbb{R}^{n_i}$. Typically, columns of Z_i are a subset of columns X_i but it does not have to be the case in general. Let $X = [X_1, X_2, \dots, \mathbb{X}_m]^T$ and $Z = [Z_1, Z_2, \dots, Z_m]^T$. According to [Patterson and Thompson, 1971, Pinheiro and Bates, 2000], a linear mixed-effects model (LME) is defined as:

$$\begin{aligned} & Y_i = X_i\beta + Z_i u_i + \varepsilon_i, \quad i = 1 \dots m \\ & u_i \sim \mathcal{N}(0, \Gamma), \quad \Gamma \in \mathbb{R}^{q\times q}_{++} \\ & \varepsilon_i \sim \mathcal{N}(0, \Lambda_i), \quad \Lambda_i \in \mathbb{R}^{n_i \times n_i}_{++} \end{aligned} \tag{1.1}$$

where $\beta \in \mathbb{R}^p$ is a vector of fixed (mean) covariates and $u_i \in \mathbb{R}^q$ are the vectors of unobservable random effects, which are assumed to be distributed normally with zero mean and the unknown covariance matrix

Γ. The observation error's variances Λ_i are often considered to be known, or to have a parametric form, such as $\Lambda_i = \sigma I$ where σ is unknown. In this work, we consider Λ_i to be known, unless stated otherwise.

If one considers $\omega_i = Z_i u_i + \varepsilon_i$ to be an unknown per-cluster noise vectors, then the model 1.1 can also be viewed as a correlated noise model:

$$Y_i = X_i\beta + \omega_i, \quad \omega_i \sim \mathcal{N}(0, \Omega_i(\Gamma)), \quad \Omega_i(\Gamma) = Z_i \Gamma Z_i^T + \Lambda_i \tag{1.2}$$

For the sake of conciseness, since Ω_i is always a function of Γ, $\Omega_i(\Gamma)$ will further be spelled as just Ω_i, without the parentheses. In fact, Ω_i will be the only terms depending on Γ in the majority of expressions including the mixed model's likelihood and its derivatives described below.

When fitting an LME, the goal is to get both the estimation of fixed covariance β and the estimation of variance parameters of random effects Γ. One of the possible approaches is integrating random effects u_i out to get a marginalized posterior distribution which depends only on β and Γ, but not on u_i. According to Bayes Theorem:

$$f_{\beta,\Gamma|X=\hat{X},Y=\hat{Y}}(\beta,\Gamma) = \frac{f_{\beta,\Gamma}(\beta,\Gamma)}{f_{X,Y}(\hat{X},\hat{Y})} \int_v f_{X,Y|\beta,u=v}(\hat{X},\hat{Y}) f_{u|\Gamma}(v) dv \tag{1.3}$$

where $(\hat{X},\hat{Y})$ is a particular implementation of random variables (X,Y); $f_{\beta,\Gamma|X=\hat{X},Y=\hat{Y}}(\beta,\Gamma)$ is a posterior distribution of the model's parameters given data $(\hat{X},\hat{Y})$, also known as the likelihood of the model's parameters; $f_{\beta,\Gamma}(\beta,\Gamma)$ is a joint prior distribution of the model's parameters; $f_{X,Y|\beta,u}(\hat{X},\hat{Y})$ is likelihood of the data $(\hat{X},\hat{Y})$ for the model (1.1) given β and u_i are known; and $f_{u|\Gamma}(u)$ is a conditional distribution of random effects u_i given that Γ is known.

Omitting the parameters-independent multiplier $f_{X,Y}(\hat{X},\hat{Y})$ and taking the negative log-likelihood of (1.3), we obtain $\mathcal{L}_{ML}$:

$$\begin{aligned}
&\mathcal{L}_{ML}(\beta,\Gamma) \equiv -\log(f_{\beta,\Gamma|X=\hat{X},Y=\hat{Y}}(\beta,\Gamma)) \sim \\
&\sim \sum_{i=1}^m \frac{1}{2}(y_i - X_i\beta)^T (Z_i \Gamma^{-1} Z_i^T + \Lambda_i)^{-1} (y_i - X_i\beta) + \frac{1}{2}\log\det(Z_i\Gamma^{-1}Z_i^T + \Lambda_i) - \log(f(\beta,\Gamma)) = \\
&= \sum_{i=1}^m \frac{1}{2}(y_i - X_i\beta)^T \Omega_i^{-1}(y_i - X_i\beta) + \frac{1}{2}\log\det\Omega_i - \log(f_{\beta,\Gamma}(\beta,\Gamma))
\end{aligned} \tag{1.4}$$

This expression $\mathcal{L}_{ML}(\beta,\Gamma)$ is known as the negative log likelihood function of a linear mixed-effects model. The term $-\log(f_{\beta,\Gamma}(\beta,\Gamma))$ represents prior knowledge about β and Γ, and is often referred to as a regularization term. As an example, setting $-\log(f_{\beta,\Gamma}(\beta,\Gamma)) = \frac{1}{2\sigma_\beta^2}\|\beta\|_2^2 + \frac{1}{2\sigma_\Gamma^2}\|\Gamma\|_2^2$ would represent our prior belief that $\beta \sim \mathcal{N}(0,\sigma_\beta^2)$ and $\Gamma \sim \mathcal{N}(0,\sigma_\gamma^2)$.

The unknown coefficients β and Γ are to be obtained via solving the following optimization problem:

$$\begin{aligned} &\min_{\beta,\Gamma} \; \mathcal{L}_{ML}(\beta, \Gamma) \\ &s.t.\ \Gamma \in \mathbb{R}^{s\times s}_{++} \end{aligned} \tag{1.5}$$

In a simpler case where $\Gamma = \mathrm{Diag}(\gamma)$ is diagonal (often referred to as *the diagonal setup*), the positive definite constraint from (1.5) transforms into a box constraint:

$$\begin{aligned} &\min_{\beta,\gamma} \; \mathcal{L}_{ML}(\beta, \mathrm{Diag}(\gamma)) \\ &s.t.\ \gamma > 0 \end{aligned} \tag{1.6}$$

Use of Γ versus γ. Starting from Chapter 2 this work is focused on the case where the matrix Γ is diagonal: $\Gamma = \mathrm{Diag}(\gamma)$ (the diagonal setup (1.6)), so the unknown covariance parameters of the problem are often referred to as just γ. This simpler case corresponds to the situation when random effects u_{ij} corresponding to different mean effects β_j are independent of each other. In this case the positive definiteness constraint is guaranteed to satisfy whenever $\gamma > 0$. Authors made the decision to start with the diagonal case because it covers their immediate needs in applications, and they see the extensions of the algorithm proposed in Chapter 2.2 to the non-diagonal case to be a part of future work. This chapter, however, discusses both cases of γ and Γ to make connections between the general and the particular cases, and to preserve the notation consistency with the referenced works.

The expected value of the posterior mean effects β given Γ has a closed form representation which immediately follows from (1.2):

$$\beta^{(k+1)} = \operatorname*{argmin}_{\beta} \mathcal{L}_{ML}(\beta, \Gamma^{(k)}) = \left(\sum_{i=1}^{m} X_i^T \Omega_i^{-1} X_i\right)^{-1} \sum_{i=1}^{m} X_i^T \Omega_i^{-1} y_i \tag{1.7}$$

There is, however, no closed-form solution for Γ in general case, so it has to be obtained via an optimization subroutine. There are three aspects that make this problem challenging from the optimization point of view:

1. **Non-convexity**: the loss function $\mathcal{L}_{ML}$ is not convex with respect to Γ, which can prevent an optimization method from obtaining a global solution.
2. **Overfitting**: the matrix Γ contains as many as $\frac{q(q-1)}{2}$ free parameters, overfitting of which may result in near-singularities, especially when the sample size is small [Lange and Laird, 1989]. The

diagonal setup $\Gamma = \mathrm{Diag}(\gamma)$ is still prone to overfitting, although has less freedom (q free parameters) to do so.

3. **Positive Constraint**: the requirement of keeping the matrix Γ to be positive definite limits the scope of methods which can be used for optimization. The diagonal setup has a box constraint $\gamma > 0$ which also has to be accommodated by the optimization subroutine.

When β^* and Γ^*, the minimizers for (1.5), are obtained, the individual random effects u_i can be found through the minimization of conditional likelihood of random effects:

$$\min_{u_i} \mathcal{L}(u_i|\beta, \gamma) = \min_{u_i} \left(\frac{1}{2} u_i^T \Gamma^{-1} u_i + \frac{1}{2}(Z_i u_i - Y_i + F_i(\beta))^T \Lambda_i^{-1}(Z_i u_i - Y_i + F_i(\beta)) \right), \tag{1.8}$$

This problem has a closed form solution which was first derived by [Harville, 1976] as an extension to Gauss-Markov theorem:

$$u_i = (\Gamma^{-1} + Z_i^T \Lambda_i^{-1} Z_i)^{-1} Z_i^T \Lambda_i^{-1}(Y_i - X_i\beta). \tag{1.9}$$

As pointed by [Harville, 1974], Γ^* – the minimizer for the maximal likelihood formulation 1.5 – is known to be biased downwards because it does not take into account the loss of k degrees of freedom from estimating β^* within the same inference procedure. To overcome this effect they proposed a Restricted Maximal Likelihood (REML) formulation where the loss function is based on $n - k$ linearly independent error contrasts:

$$\mathcal{L}_{REML}(\beta, \Gamma) = -\frac{1}{2} \log\det \left(\sum_{i=1}^{m} X_i^T (Z_i \Gamma^{-1} Z_i^T + \Lambda_i)^{-1} X_i \right) + \mathcal{L}_{ML}(\beta, \Gamma) \tag{1.10}$$

Although different from ML, the REML formulation does not introduce more complexity to optimization: all the solution methods for ML formulation discussed below also have an adaptation to REML.

Two main optimization approaches to solving mixed effects likelihood formulation are used: EM algorithm [Dempster et al., 1977] and Newton-Raphson [Lindstrom and Bates, 1988].

1.1.1 EM algorithm for Mixed Models

One more way of thinking about mixed-effects models is considering $\{Y_i\}_{i=1}^m$ to be incomplete data and $\{Y_i, u_i\}_{i=1}^m$ to be complete data, where u_i is the unobservable part. In this setting, the problem of fitting a mixed model can also be classified as a statistical inference problem in the presence of missing data, and so the methods for solving this kind of setup might be applicable. One example is EM-algorithm – a general approach to estimating parameters in problems with missing data developed by [Dempster et al., 1977]. As later discussed by [Laird and Ware, 1982], one can view EM-algorithm as optimizing the approximating function $Q(\beta, \Gamma|\hat{\beta}, \hat{\Gamma})$ – the likelihood of (β, Γ) given that the missing data u_i comes from a normal distribution $\mathcal{N}(\hat{\beta}, Z_i\hat{\Gamma}Z_i^T + \Lambda_i)$. EM-algorithm, in its general form, consists of two steps: Expectation step and Minimization step:

$$\begin{aligned} \text{E-step:} \quad & Q^{(k)}(\beta, \Gamma|\hat{\beta}^{(k)}, \hat{\Gamma}^{(k)}) := p\left(\beta, \Gamma|\hat{u_i}^{(k)} = \mathbb{E}\left[u_i|\hat{\beta}^{(k)}, \hat{\Gamma}^{(k)}\right]\right) \\ \text{M-step:} \quad & \hat{\beta}^{(k+1)}, \hat{\Gamma}^{(k)} := \underset{\beta, \Gamma}{\text{argmax}} \;\; Q(\beta, \Gamma^{k+1}|\hat{\beta}^{(k)}, \hat{\Gamma}^{(k)}) \end{aligned} \tag{1.11}$$

On the E-step the function $Q(\beta, \Gamma|\hat{\beta}, \hat{\Gamma})$ is constructed using the expected values of missing data given the previous-iteration parameters $\hat{\beta}, \hat{\Gamma}$, and on M-step Q-function is minimized, and its minimizers $\hat{\beta}^{*}, \hat{\Gamma}^{*}$ are set to be next-iteration parameters. The process continues until convergence.

As discussed in [Laird et al., 1987], one can express the EM-algorithm's iterations in the linear mixed models setup as:

$$\begin{aligned} \beta^{(k+1)} &= \underset{\beta}{\text{argmin}}\, \mathcal{L}(\beta, \Gamma^{(k)}) = \left(\sum_{i=1}^{m} X_i^T \Omega_i^{-1}(\Gamma^{(k)}) X_i\right)^{-1} \sum_{i=1}^{m} X_i^T \Omega_i^{-1}(\Gamma^{(k)}) Y_i \\ u_i^{(k+1)} &= ((\Gamma^{(k)})^{-1} + Z_i^T \Lambda_i^{-1} Z_i)^{-1} Z_i^T \Lambda_i^{-1}(Y_i - X_i\beta^{(k)}) \\ \Gamma^{(k+1)} &= \frac{1}{m}\sum_{i=1}^{m} u_i^{(k)}(u_i^{(k)})^T + \Gamma^{(k)}(I - Z_i\Omega_i^{-1}(\Gamma^{(k)})Z_i^T\Gamma^{(k)}) \end{aligned} \tag{1.12}$$

EM-algorithm has two main advantages: simplicity of iterations and the absence of constraints in optimization subroutines. The later is especially important: EM-algorithm guarantees that if the initial guess $\Gamma^{(0)}$ was positive definite then all $\Gamma^{(k)}$, including, under certain conditions, the final result Γ^*, will be positive definite. This removes the necessity to handle the constraint $\Gamma \in \mathbb{R}^{s \times s}_{++}$ separately, e.g. via re-parametrization using Cholesky decomposition, which allows to keep the iterations fast and simple. As an apparent consequence, EM algorithm also correctly handles the box constraint in the diagonal setup (1.6).

The main drawback of EM-algorithm is its slow convergence speed. In the form above, [Dempster et al., 1977] showed that the algorithm monotonically converges, with geometric rate, as established by [Wu, 1983], to a local maximal likelihood estimator. The geometric rate is due to the fact that EM-algorithm can be equivalently viewed as a proximal gradient descent on KL-divergence based likelihood. There have been wide effort on both discovering more statistical properties of the algorithm, as well as on improving its convergence rate. For instance, in his recent work [Balakrishnan et al., 2017] provided a quantitative characterization of a region such that the EM-algorithm, initialized within that region, is guaranteed to converge to a "sufficiently good", fixed point: the one within statistical precision of the global optimum.

There are also works which are empirically improving the convergence speed of EM in terms of computations and time via implementation-level improvements, like [Liu and Rubin, 1994, Liu, 1998], to name a few.

1.1.2 Newton Methods for Mixed Models

Newton method as a second-order optimization method for estimating parameters of mixed models was proposed by [Jennrich and Schluchter, 1986] and later improved by [Lindstrom and Bates, 1988]. The core idea of this approach is to consider $\hat{\beta}(\Gamma)$ as a function of Γ (see Eq. 1.7), and apply Newton method to the resulting formulation:

$$\begin{aligned} \min_{\Gamma} \ & \mathcal{L}_{ML}(\beta(\Gamma), \Gamma) \\ \text{s.t. } & \Gamma \in \mathbb{R}_{+}^{s \times s} \end{aligned} \tag{1.13}$$

In a general setup when Γ is not diagonal, one can treat the positive-definiteness constraint by reparametrizing the problem via modified Cholesky decomposition, as suggested by [Chen and Dunson, 2003]:

$$\Gamma = DLL^T D \tag{1.14}$$

where D is an $s \times s$ diagonal matrix, and L is an $s \times s$ lower-triangular matrix with ones on the main diagonal. Applying this variables transformation to the equation (1.13) we get an unconstrained optimization problem on diagonal elements of D and on the off-diagonal elements of L:

$$\begin{aligned} \min_{D,L} \ & \mathcal{L}_{ML}(\hat{\beta}(DLL^T D), DLL^T D) \\ \text{s.t. } & D_{ij} = 0, \ i \neq j \\ & L_{ii} = 1, \ i = 1 \ldots s \end{aligned} \tag{1.15}$$

In a simpler case of Γ being diagonal ($\Gamma = \operatorname{Diag} \gamma$), the setup (1.13) simplifies to

$$\min_{\gamma > 0} \mathcal{L}(\hat{\beta}(\gamma), \gamma) \tag{1.16}$$

On each iteration, Newton-like family of methods is looking for $\gamma^{(k+1)}$ – a minimizer for a quadratic approximation of the problem at the point $\gamma^{(k)}$:

$$\begin{aligned} \gamma^{(k+1)} = \operatorname*{argmin}_{\Delta\gamma} & \mathcal{L}_{ML}(\gamma^{(k)}) + \nabla_{\gamma} \mathcal{L}_{ML}(\gamma^{(k)})^T \Delta\gamma + \Delta\gamma^T \nabla^2_{\gamma\gamma} \mathcal{L}_{ML}(\gamma^{(k)}) \Delta\gamma \\ \text{s.t. } & \gamma^{(k)} + \Delta\gamma \geq 0 \end{aligned} \tag{1.17}$$

The positivity constraint narrows the scope of optimization techniques which can be used. There is, however, still wide variety of methods to choose from like Truncated Newton Methods with Barriers [Nash, 2000], Projection Methods [Bertsekas, 1982], or Interior Point methods [Potra and Wright, 2000], one can refer to [Tru,] for a detailed overview of these algorithms. The Interior Point method shall be discussed in more details since it's used in Chapter 2.2 of this work.

Interior Point (IP) methods [Kojima et al., 1991, Nesterov and Nemirovskii, 1994, Wright, 1997] target the Karush-Kuhn-Tucker optimality conditions for the problem (1.18) directly by introducing "margins" – non-negative axillary variables which transform inequality constraints into equality constraints:

$$\begin{aligned} &\min_{\gamma, d} \mathcal{L}(\gamma) \\ &\text{s.t. } \gamma - d = 0 \\ &\qquad d \geq 0 \end{aligned} \tag{1.18}$$

The positivity constraint for d is then handled via a logarithmic barrier. The Lagrangian and the KKT system for this problem are:

$$\begin{aligned} F_\mu(\gamma, d, v) &= \mathcal{L}(\gamma) + v^T(\gamma - d) - \mu \sum_{i=1}^{s} \log(d_i) \\ \nabla F = \begin{bmatrix} \nabla_\gamma F_\mu \\ \nabla_v F_\mu \\ \nabla_d F_\mu \end{bmatrix} &= \begin{bmatrix} \nabla \mathcal{L}(\gamma) + v \\ \gamma - d \\ v \circ d - \mu \circ 1 \end{bmatrix} = 0 \end{aligned} \tag{1.19}$$

where v is a vector of dual variables, and μ is a positive hyper parameter which define the steepness of the logarithmic barrier, and $v \circ d$ denotes the element-wise product. Application of Newton root-finding method to this KKT system gives us the Interior Point method. In particular, the search direction $[\Delta\gamma, \Delta v, \Delta d]$ as the solution of the following system:

$$\nabla^2 F_\mu \begin{bmatrix} \Delta\gamma \\ \Delta v \\ \Delta d \end{bmatrix} = -\nabla F_\mu \begin{bmatrix} \Delta\gamma \\ \Delta v \\ \Delta d \end{bmatrix} \tag{1.20}$$

The iteration then looks like

$$\begin{aligned} \gamma^{(k+1)} &= \gamma^{(k)} + \alpha\Delta\gamma \\ v^{(k+1)} &= v^{(k)} + \alpha\Delta v \\ d^{(k+1)} &= d^{(k)} + \alpha\Delta d \end{aligned} \tag{1.21}$$

α is chosen via line search so all $\gamma^{(k+1)}$, $v^{(k+1)}$, and $d^{(k+1)}$ are positive, and a sufficient descent condition is satisfied:

$$\|\nabla F_\mu(\gamma^{(k+1)}, v^{(k+1)}, d^{(k+1)})\| \leq 0.99\|\nabla F_\mu(\gamma^{(k)}, v^{(k)}, d^{(k)})\| \tag{1.22}$$

Since this is a root finding procedure for the equation (1.19), the natural stopping criterion is

$$\|\nabla F_\mu(\gamma^{(k)}, v^{(k)}, d^{(k)})\| \leq \varepsilon \tag{1.23}$$

for some sufficiently small ε.

Interior Point method is a second-order method, so, under certain assumptions, it has a quadratic rate convergence region around the optimal point. This leads to better performance in terms of the number of iterations comparing to EM-algorithm. The performance in terms of time, however, strongly depends on the problem's size. It's due to the fact that working with Hessian information requires solving a linear system, which is not needed for first order methods like EM-algorithm.

1.2 Feature Selection for Mixed Models

In statistical learning setting, feature selection is defined as a process of selecting, or ranking, the most important predictors in a dataset. The goal is to get an interpretable model without sacrificing prediction quality. Selection process often happens simultaneously with fitting the model, where a predictor is considered to be selected if its respective coefficient in the model is non-zero. If the desired number of coefficients k is given, then the feature selection problem can be formulated as minimization of a loss function $f(\theta)$, which is often proportionate to the negative log-likelihood $\log p(\theta|X, Y)$, with respect to its parameters θ subject to a zero-norm constraint:

$$\begin{aligned} &\min_\theta f(\theta) \\ &\text{s.t. } \|\theta\|_0 \leq k \end{aligned} \tag{1.24}$$

In this form the problem is equivalent to the Knapsack problem, which is NP-complete. Many techniques based on exhaustive search, also known as subset selection process, were developed, and showed to be effective when the total number of predictors is small (see [Müller et al., 2013]). The same setup appears to be intractable in a high-dimension setting due to the exponential growth of the number of subsets to check when $n \to \infty$, however, it's possible to get an approximate solution via relaxation techniques discussed below.

One way to work with zero-norm constraint in the Eq.1.24 is to relax it to a first norm constraint. An example of such approach is LASSO, or least absolute square shrinkage operator, studied by [Tibshirani, 1996] for the least squares model:

$$\min_\beta ||y - X\theta||_2^2 + \lambda||\theta||_1 \tag{1.25}$$

It combines the small bias of the least square estimator and the interpretability of the subset selection method, producing, however, biased estimates for the large coefficients, according to [Zou, 2006]. This approach was later extended to exhibit other desirable properties, such as lesser bias for larger coefficients (SCAD from [Fan and Li, 2001]) or simultaneous selection for highly correlated predictors (Elastic Net from [Zou and Hastie, 2005]), and to accommodate other model setups, including classification (ℓ_1-SVM from [Bradley and Mangasarian, 1998, Sha, 2011, Belyy and Sholokhov, 2018]), clustering ([Jajuga, 1991]), and reinforcement learning ([Kim and Paik, 2019]), to name a few.

Feature selection for linear mixed models is generally a more challenging problem than feature selection for linear regression problem. First, in linear regression setting the observation are independent, whereas in mixed-effects setup they're typically dependent. Second, one needs to do simultaneous selection of mean effect variables (β) and the ones related to their random effects' covariance structure (Γ), and these groups can have different relative importance. Finally, selecting random features implies that Γ will have zero columns and rows, which prevents it being positive-definite. Consequently, selecting variance parameters leads to boundary issues which make both theoretical analysis and numerical computations more challenging.

The shrinkage operator approach was first adapted to the mixed-effect models setting by [Lan, 2006] to perform fixed effects selection. Selection of random effects is generally more difficult than selection of fixed effects alone because removing random effect from the model requires elimination of the entire respective row and column of Γ. This extension was provided by [Bondell et al., 2010], who performed simultaneous fixed and random effects selection by combining three components: the modified Cholesky decomposition $\Gamma = DLL^TD$, as suggested by [Chen and Dunson, 2003], the Adaptive LASSO penalty, as in [Lan, 2006], and EM algorithm ([Dempster et al., 1977]). Their penalized Q-function for EM-algorithm is defined as:

$$\begin{aligned} Q(\beta, D, L|\beta^{(k)}, D^{(k)}, L^{(k)}) =& \mathbb{E}_{u_i|\beta^{(k)}, D^{(k)}, L^{(k)}} \left[\sum_{i=1}^m \|Y_i - X_i\beta - Z_iDLu_i\|^2 \right] + \\ & + \lambda \left(\sum_{j=1}^k \frac{|\beta_j|}{|\hat{\beta}_j|} + \sum_{j=1}^s \frac{|D_{ii}|}{|\bar{D}_{ii}|} \right) \end{aligned} \tag{1.26}$$

They suggest tuning the regularization parameter λ using a BIC criterion. A similar formulation was used by [Lin et al., 2013] who proposed a joint selection with two-stage ALASSO. On stage 1 they approximate the shrinkage penalty with local quadratic approximation and use Newton method to select random effects, and then, on stage 2 they select fixed effects via penalized log-likelihood where the covariance variables are fixed.

Another type of regularization which is used for feature selection is SCAD [Fan and Li, 2001] which acts on the first derivative of the penalty function $p_\lambda(\theta)$:

$$p'_\lambda(|\theta|) = \lambda \left\{ \mathbb{I}_{[\theta \leq \lambda]} + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} \mathbb{I}_{[\theta - \lambda]} \right\} \tag{1.27}$$

The adaptation of SCAD penalty to selecting both fixed and random features in linear mixed model setup was developed by [Fan and Li, 2012].

In high-dimensional setting marginal likelihood is often replaced with a hierarchical approximation of it known as h-likelihood ([Lee and Nelder, 1996]), which, for a given data $(\hat{X}, \hat{Y})$ is defined as

$$\mathcal{L}_H(\beta, \Gamma, u) = -\log f_{X,Y|\beta,\Gamma,u}(\hat{X}, \hat{Y}) - \log f_{u|\Gamma}(u, \Gamma) \tag{1.28}$$

Where $f_{X,Y|\beta,\Gamma,u}(\hat{X}, \hat{Y})$ is a likelihood of data given β, Γ, and u, and $f_{u|\Gamma}(u, \Gamma)$ is a conditional density of random effects u given Γ. This formulation allows to avoid integration over random effects which is required when the classical marginal likelihood is used. Practically, the use of h-Likelihood eliminates the need to compute and to invert the covariance matrices Ω_i or their Cholesky factors, which is often the most expensive operation when computing the marginal likelihood numerically. When derived using h-Likelihood, the estimators of fixed effects β and random effects u_i are, as shown by [Lee and Nelder, 1996], asymptotically best unbiased predictors. In order to estimate variance components simultaneously with fixed effects an Adjusted Profile likelihood, which is an hierarchical generalization of REML, is used.

Similarly to marginal likelihood above, various feature selection approaches were adapted to be used with h-Likelihood. Similarly to what [Bondell et al., 2010] proposed for marginal likelihood based selection, [Xu et al., 2015] proposed to penalize the fixed effects and random effects variance components Ω_i in h-likelihood with LASSO-like family of penalties. They also showed, that, under certain assumptions, the proposed estimator possesses asymptotic normality property. The most recent development in this trend to date was provided by [Xie et al., 2020] who utilized multi-armed bandits approach for selecting mean effects in high-dimensional, h-Likelihood, setting.

Shrinkage penalties and norm-based penalties are not the only kind of relaxations which are used in feature selection. Recently [Ghosh and Thoresen, 2018] studied SCAD regularization for selecting mean effects in a high-dimensional setting motivated by genomics problems. Other approaches to feature selection are based on information criteria such as AICs and BICs [Jones, 2011]. According to [Müller et al., 2013], the most widely used AIC criterion is what [Vaida and Blanchard, 2005] call the marginal AIC criterion:

$$AIC = 2\,\mathcal{L}(\hat{\theta}) - 2\alpha_n(p+q) \tag{1.29}$$

where $\hat{\theta}$ includes all the estimated parameters in a mixed-effect model (β, Γ), and $\alpha_n = n(n-p-q-1)$ for a finite sample case [Sugiura, 1978]. As discussed by [Fang, 2011], AIC is asymptotically equivalent to leave-one-out cross-validation, so it can be used for choosing between a finite number of models.

One can refer to [Müller et al., 2013] for a detailed overview of different feature selection approaches.

Variable projection is another approach to tackle the zero-norm constraint in (1.24). Relax-and-Split [Zheng and Aravkin, 2020] technique is one of implementations of this approach. It's been successfully applied in basis pursuit denoise problems [Baraldi et al., 2019], compressed sensing [Zheng et al., 2019], and sparse models identification [Champion et al., 2020]. The key idea of this method is to use a set of auxiliary variables to relax the constraints:

$$\min_{\|\beta\|_0 \le k} f(\beta) \quad \to \quad \min_{\beta;\ \|w\|_0 \le k} f(\beta) + \frac{1}{2\nu}\|w - \beta\|_2^2 \tag{1.30}$$

and then to apply an alternating minimization algorithm to solve the relaxation. This approach will be discussed in details in the next chapter.

2 Relax-and-Split for Mixed-Effects Selection

2.1 Introduction

Recently [Zheng and Aravkin, 2020] proposed to use variable projection technique called Relax-and-Split for nonconvex-composite optimization problems of the form:

$$\min_{\beta} f(\beta) = h(A\beta) + g(\beta) \tag{2.1}$$

where $\beta \in \mathbb{R}^k$ are the decision variables, $A \in \mathbb{R}^{n\times k}$, $h(\beta)$ is non-smooth, non-convex, and separable, and $g(\beta)$ is convex. The key innovation of their work is to relax the original variable β with a set of auxiliary variables $\tilde{\beta}$, and then use partial minimization techniques to develop effective algorithms:

$$\min_{x} f(\beta) = h(A\beta) + g(\beta) \quad \rightarrow \quad \min_{\beta,\tilde{\beta}} h(\tilde{\beta}) + \frac{1}{2\theta}\|A\beta - \tilde{\beta}\|_2^2 + g(\beta) \tag{2.2}$$

The same approach can be used for relaxing constraints in non-convex optimization problems. This approach was discussed by [Zheng et al., 2019] as a part of SR3 framework:

$$\min_{\beta\in\Delta} f(\beta) \quad \rightarrow \quad \min_{\tilde{\beta}\in\Delta;\ \beta} f(\beta) + \frac{1}{2\nu}\|\beta - \tilde{\beta}\|_2^2 \tag{2.3}$$

As an example, the zero-norm constraint from (1.24) can be relaxed as follows:

$$\min_{\beta} f(\beta) \ \text{ s.t. } \|\beta\|_0 \leq k \quad \rightarrow \quad \min_{\beta,\tilde{\beta}} f(\beta) + \lambda\|\beta - \tilde{\beta}\|_2^2 \ \text{ s.t. } \|\tilde{\beta}\|_0 \leq k \tag{2.4}$$

The second norm links β, an unconstrained variable, and its constrained counterpart $\tilde{\beta}$. The coefficient λ represents the strength of this link and can be viewed as a Gaussian prior for the distance of β from the "wall" – a k-dimensional subspace. The larger λ is – the higher penalty is for β violating the zero-norm constraint.

This relaxation admits effective application of alternating minimization methods:

$$\begin{aligned}\beta^{(t+1)} &= \operatorname*{argmin}_{\beta} \left[f(\beta) + \lambda\|\beta - \tilde{\beta}^{(t)}\|_2^2\right] \\ \tilde{\beta}^{(t+1)} &= \operatorname*{argmin}_{\tilde{\beta}} \|\beta^{(t)} - \tilde{\beta}\|_2^2 \\ &\text{s.t. } \|\tilde{\beta}\|_0 \leq k\end{aligned} \tag{2.5}$$

The first optimization subroutine is an unconstrained optimization problem. The second subroutine is a Euclidean projection of $\beta^{(t)}$ onto a k-dimensional subspace. This operation can be efficiently implemented since it only involves taking k largest in absolute value elements of $\beta^{(t)}$. The coefficient λ is iteratively increased until $\|\beta - \tilde{\beta}\| \leq \varepsilon$ for some small ε.

This relaxation has certain advantages over ℓ_1-type relaxations:

- **Interpretability of hyper-parameters.** Notice that this relaxation does not change the meaning of the only hyper-parameter k. In both formulations in Eq. (2.4) the parameter k denotes the number of non-zero coefficients in the resulting estimator β^* whereas in the formulation (1.25) k is replaced with λ, which denotes the variance of Laplacian prior for β^*. There is no data-independent mapping between λ and k which would allow researcher to make an informed decision on the value of λ before calling the LASSO fit, which leads to multiple trial-launches resulting in a overall production pipeline slowdown.

- **No model bias given the subspace.** Notice that in (2.4) the local minimizer for the relaxation when $\lambda \to \infty$ will also be a local minimizer for a non-relaxed setup (1.24). The solution for the formulation (1.25) is, in a general case, not going to be a local minimizer for (1.24) since its non-zero coefficients are biased by the implicit Laplacian prior ([Zou, 2006] provides a comprehensive discussion on the nature of this bias). Because of that, feature selection via LASSO is normally executed as a two-stage process, when, after the optimal subset of features is being selected on the stage one, the optimization is repeated on that features only without ℓ_1 regularization to partially eliminate the model bias. There is, however, no need for the second stage when the formulation (2.4) is used.

The main disadvantage of the formulation (2.4) comparing to LASSO (1.25) is that the latter only needs to be solved once, whereas the former should be solved iteratively for increasing λ until the relaxation is eliminated, which naturally takes longer. This, however, can be mitigated by applying warm-start technique, when the optimization after increasing λ starts from the previous optimal β^*. This approach is especially efficient when the second order optimization subroutine is used for the step one of the algorithm (2.5): the subsequent problem differs from the previous one by a quadratic term, so if the method previously reached the quadratic convergence region then increasing λ won't change it.

2.2 Proposed Algorithm

Consider a mixed-effect model setting described in Eq. (1.1) with Γ being diagonal: $\Gamma = \mathrm{Diag}(\gamma)$. We wish to find a minimizer the following loss function:

$$\begin{aligned} \min_{\beta,\gamma} \ & \mathcal{L}_{ML}(\beta, \mathrm{Diag}\,(\gamma)) \\ \text{s.t. } & \|\beta\|_0 \leq k \\ & \|\gamma\|_0 \leq j \\ & \gamma \geq 0 \end{aligned} \tag{2.6}$$

where $\mathcal{L}_{ML}(\beta, \mathrm{Diag}\,(\gamma))$ is a marginalized log-likelihood of a mixed model defined in Equation (1.4).

In this form, the problem can have as many as C^n_k local minimizers (at least one per k-subspace), and finding a global solution is equivalent in complexity to solving a weighted knapsack problem. However, finding a "sufficiently good" local minimizer is usually enough for practical purposes.

We relax the original problem using Relax-and-Split approach ([Zheng and Aravkin, 2020]) in order to both automate choosing a good subspace and to do optimization in the whole space instead of doing optimization in subspaces. Namely, we introduce two constrained auxiliary variables $\tilde{\beta}$ and $\tilde{\gamma}$ and treat them as projections of now unconstrained variables β and γ on k- and j-subspaces respectively. The loss function $\mathcal{L}_r(\beta, \gamma, \tilde{\beta}, \tilde{\gamma})$ of for the relaxed setup is defined as:

$$\mathcal{L}_r(\beta, \gamma, \tilde{\beta}, \tilde{\gamma}) \equiv \mathcal{L}_{ML}(\beta, \gamma) + \frac{\lambda_\beta}{2}\|\beta - \tilde{\beta}\|_2^2 + \frac{\lambda_\gamma}{2}\|\gamma - \tilde{\gamma}\|_2^2 \tag{2.7}$$

where $\mathcal{L}_{ML}(\beta, \operatorname{Diag}(\gamma))$ is defined in (1.4):

$$\mathcal{L}_{ML}(\beta, \gamma) = \sum_{i=1}^{m} \frac{1}{2}(Y_i - X_i\beta)^T \Omega_i^{-1}(\gamma)(Y_i - X_i\beta) + \frac{1}{2}\log\det\Omega_i(\gamma) \tag{2.8}$$

and $\Omega_i(\gamma)$ is a covariance matrix for the i'th group observation's noise:

$$\Omega_i(\gamma) = Z_i \operatorname{Diag}(\gamma) Z_i^T + \Lambda_i \tag{2.9}$$

In order to solve a relaxed setup for given relaxation parameters λ_β and λ_γ we need to find a solution $(\beta^*, \gamma^*, \tilde{\beta}^*, \tilde{\gamma}^*)$ for the following problem:

$$\begin{aligned} \min_{\beta,\gamma,\tilde{\beta},\tilde{\gamma}} \; & \mathcal{L}_r(\beta, \gamma, \tilde{\beta}, \tilde{\gamma}) \\ \text{s.t. } & \|\tilde{\beta}\|_0 \leq k \\ & \|\tilde{\gamma}\|_0 \leq j \\ & \gamma \geq 0 \\ & \tilde{\gamma} \geq 0 \end{aligned} \tag{2.10}$$

We apply variable projection technique to solve the relaxed problem (2.10). This approach outlines the Algorithm 1, henceforth referred as R&S-Mixed , as an iterative procedure with four projection steps within each iteration. In order to guarantee that $(\tilde{\beta}, \tilde{\gamma})$ from (2.10) is a solution for (2.6), we iteratively increase λ_β and λ_γ until $\beta = \tilde{\beta}$ and $\gamma = \tilde{\gamma}$.

The algorithm has three alternating minimization steps in its inner cycle. We shall discuss all three steps in details below.

```
1  $\lambda_\beta = 0;\ \lambda_\gamma = 0$
2  repeat
3      $\lambda_\beta \leftarrow 2(1+\lambda_\beta)$
4      $\lambda_\gamma \leftarrow 2(1+\lambda_\gamma)$
5      repeat
6          $\tilde{\beta}^{(k+1)} \leftarrow \mathrm{Proj}_{\|\beta\|_0 \le k}(\beta^{(k)})$ ;      /* Take max k from $\beta$, rest to 0 */
7          $\tilde{\gamma}^{(k+1)} \leftarrow \mathrm{Proj}_{\|\gamma\|_0 \le s}(\gamma^{(k)})$ ;      /* Take max j from $\gamma$, rest to 0 */
8          $\beta^{(k+1)}, \gamma^{(k+1)} \leftarrow \mathrm{argmin}_{\gamma \ge 0, \beta}\, \mathcal{L}(\beta,\gamma) + \frac{\lambda_\beta}{2}\|\beta - \tilde{\beta}^{(k)}\|_2^2 + \frac{\lambda_\gamma}{2}\|\gamma - \tilde{\gamma}^{(k)}\|_2^2$ ; /*  IP Method */
9      until converges;
10 until $\tilde{\beta} \approx \beta, \tilde{\gamma} \approx \gamma$;
```

Algorithm 1: R&S-Mixed : Relax-and-Split-based feature selection algorithm for linear mixed-effects models.

Algorithm 1, lines 6 and 7: Projections on a ℓ_0 box. When β and γ are fixed the partial minimization solution for $\tilde{\beta}$ and $\tilde{\gamma}$ is just projections of β and γ on respective ℓ_0-boxes:

$$\begin{aligned}
\tilde{\beta}^{(k+1)} &= \underset{\tilde{\beta}}{\mathrm{argmin}}\, \|\tilde{\beta} - \beta^{(k)}\|_2^2 \text{ s.t. } \|\tilde{\beta}\|_0 \le k \quad &\Longrightarrow \quad \tilde{\beta}^{(k+1)} &= \text{ max } k \text{ components from } |\beta_1|, \dots, |\beta_p| \\
\tilde{\gamma}^{(k+1)} &= \underset{\tilde{\gamma}}{\mathrm{argmin}}\, \|\tilde{\gamma} - \gamma^{(k)}\|_2^2 \text{ s.t. } \|\tilde{\gamma}\|_0 \le j \quad &\Longrightarrow \quad \tilde{\gamma}^{(k+1)} &= \text{ max } j \text{ components from } |\gamma_1|, \dots, |\gamma_q| \\
& & &\text{... and the rest of components set to 0}
\end{aligned} \tag{2.11}$$

Algorithm 1, line 8: Interior Point Method. As described in Section 1.1.2, Interior Point (IP) Methods are a class of second-order methods which come from barrier relaxation of constraints followed by direct application of Newton method to KKT optimality conditions. Following the same idea, we transform the inequality constraints to equality constraints and then relax these constraints via logarithmic barriers:

$$\begin{aligned}
&\min_{\beta,\gamma,d} \mathcal{L}(\beta,\gamma) + \frac{\lambda_\beta}{2}\|\beta - \tilde{\beta}^{(k)}\|_2^2 + \frac{\lambda_\gamma}{2}\|\gamma - \tilde{\gamma}^{(k)}\|_2^2 + \mu\sum_{i=1}^{q}\log(d_i) \\
&\text{s.t. } \gamma - d = 0
\end{aligned} \tag{2.12}$$

where $d \in R^q$ are "offset" variables, and μ is a relaxation parameter: the smaller μ corresponds to "sharper" barrier approximation. Lagrangian for this optimization problem can be written as:

$$F_\mu(v,d,\beta,\gamma) = \mathcal{L}(\beta,\gamma) + \frac{\lambda_\beta}{2}\|\beta - \tilde{\beta}^{(k)}\|_2^2 + \frac{\lambda_\gamma}{2}\|\gamma - \tilde{\gamma}^{(k)}\|_2^2 + \mu\sum_{i=1}^{q}\log(d_i) + v^T(\gamma - d) \tag{2.13}$$

From the necessary optimality conditions on d we have that

$$\nabla_d F_\mu = \begin{bmatrix} v_1 - \mu/d_1 \\ \dots \\ v_q - \mu/d_q \end{bmatrix} = \begin{bmatrix} 0 \\ \dots \\ 0 \end{bmatrix} \implies v_i d_i = \mu \text{ for all } i \tag{2.14}$$

which can be written in a matrix form as

$$\operatorname{Diag}(v)d - \mu \circ 1 = 0 \tag{2.15}$$

where "$\circ$" denotes an element-wise product.

Using $\gamma - d = 0$ to express F_μ as a function of v, β, and γ alone, we can write down the KKT system:

$$\nabla F_\mu(v, \beta, \gamma) = \begin{bmatrix} \nabla_v F_\mu \\ \nabla_\beta F_\mu \\ \nabla_\gamma F_\mu \end{bmatrix} = \begin{bmatrix} v \circ \gamma - \mu \circ 1 \\ \nabla_\beta \mathcal{L}(\beta, \gamma) + \lambda_\beta(\beta - \tilde{\beta}^{(k)}) \\ \nabla_\gamma \mathcal{L}(\beta, \gamma) + \lambda_\gamma(\gamma - \tilde{\gamma}^{(k)}) - v \end{bmatrix} = 0 \tag{2.16}$$

Following the the approach of (1.20), on each iteration we choose the search direction $[\Delta v, \Delta\beta, \Delta\gamma]$ as the solution of the following system:

$$\nabla^2 F_\mu \begin{bmatrix} \Delta v \\ \Delta\beta \\ \Delta\gamma \end{bmatrix} = -\nabla F_\mu \begin{bmatrix} \Delta v \\ \Delta\beta \\ \Delta\gamma \end{bmatrix} \tag{2.17}$$

Once the search direction $[\Delta v, \Delta\beta, \Delta\gamma]$, is established the method performs a step

$$\begin{aligned} v^{(k+1)} &= v^{(k)} + \alpha_k \Delta v \\ \gamma^{(k+1)} &= \gamma^{(k)} + \alpha_k \Delta\gamma \\ \beta^{(k+1)} &= \beta^{(k)} + \alpha_k \Delta\beta \end{aligned} \tag{2.18}$$

where the step length α_k is chosen to satisfy positivity and sufficient descent conditions:

$$\begin{aligned} &\textit{Positivity:} && \gamma \geq 0,\ v \geq 0 \\ &\textit{Sufficient Descent:} && \|\nabla F_\mu(v^{(k+1)}, \beta^{(k+1)}, \gamma^{(k+1)})\| \leq 0.99 \|\nabla F_\mu(v^{(k)}, \beta^{(k)}, \gamma^{(k)})\| \end{aligned} \tag{2.19}$$

On each iteration the relaxation parameter μ is adjusted as $\mu^{(k+1)} = \frac{v^{(k)T}\gamma^{(k)}}{q}$. The method continues to iterate until the stopping criterion $\|\nabla F_\mu(v^{(k+1)}, \beta^{(k+1)}, \gamma^{(k+1)})\| \leq \texttt{tol}$ is met.

The Hessian matrix $\nabla^2 F_\mu$ in Eq. (2.17) is defined as

$$\nabla^2 F_\mu = \begin{bmatrix} \operatorname{Diag}(\gamma) & 0 & \operatorname{Diag}(v) \\ 0 & \nabla^2_{\beta\beta}\mathcal{L} + \lambda_\beta I & \nabla^2_{\beta\gamma}\mathcal{L} \\ -I & \nabla^2_{\gamma\beta}\mathcal{L} & \nabla^2_{\gamma\gamma}\mathcal{L} + \lambda_\gamma I \end{bmatrix} \tag{2.20}$$

The precise form of the derivatives $\nabla \mathcal{L}$ will depend on which likelihood is used: standard maximal likelihood $\mathcal{L}_{ML}$ or restricted maximum likelihood $\mathcal{L}_{REML}$. For instance, if the former is used then the derivatives are defined as:

$$\begin{aligned}
\nabla_\beta \mathcal{L}(\beta, \gamma) &= -\sum_{i=1}^m X_i^T \Omega_i^{-1}(Y_i - X_i\beta) \\
\nabla_\gamma \mathcal{L}(\beta, \gamma) &= \sum_{i=1}^m \operatorname{Diag} {Z_i}^T \Omega_i^{-1} Z_i - ({Z_i}^T \Omega_i^{-T}(Y_i - X_i\beta))^{\circ 2} \\
\nabla^2_{\beta\beta} \mathcal{L}(\beta, \gamma) &= \frac{1}{2}\sum_{i=1}^m X_i^T \Omega_i^{-1} X_i \\
\nabla^2_{\beta\gamma} \mathcal{L}(\beta, \gamma) &= -\frac{1}{2}\sum_{i=1}^m X_i^T \Omega_i^{-1} Z_i Z_i^T \Omega_i^{-1}(Y_i - X_i\beta) \\
\nabla^2_{\gamma\gamma} \mathcal{L}(\beta, \gamma) &= \frac{1}{2}\sum_{i=1}^m -(Z_i^T \Omega_i^{-T} Z_i)^{\circ 2} + 2 Z_i^T \Omega_i^{-T}(Y_i - X_i\beta)((Y_i - X_i\beta)^T \Omega^{-T} Z_i)^T \circ (Z_i^T \Omega^{-1} Z_i)
\end{aligned} \tag{2.21}$$

For the sake of completeness we provide the derivation of these formulas in Appendix A, although they can also be found in [Lindstrom and Bates, 1988].

2.2.1 Implementation Details

Hessians through Cholesky Decomposition Computing the derivatives in (2.23) are the most computationally intensive part of Algorithm 1. Here we provide the way to efficiently compute them via Cholesky decomposition.

Let's define L_i to be a Cholesky factor of the matrix Ω_i:

$$\Omega_i = L_i L_i^T \tag{2.22}$$

Having a Cholesky factor L_i for Ω_i one can calculate a Cholesky factor of the inverse matrix Ω_i^{-1} in $\mathcal{O}(max(m, n)^2)$ operations[1] Let's also define $\xi_i = Y_i - X_i\beta$. Then we can notice that all terms in formulae

[1] See the implementation of LAPACK's *dtrtri* as an example.

(2.23) can be build out of three building blocks: $L_i^{-1}Z_i$, $L_i^{-1}X_i$, and $L_i^{-1}\xi_i$. Namely:

$$
\begin{aligned}
\nabla_\beta \mathcal{L}(\beta,\gamma) &= -\sum_{i=1}^m (L_i^{-1}X_i)^T(L_i^{-1}\xi_i) \\
\nabla_\gamma \mathcal{L}(\beta,\gamma) &= \sum_{i=1}^m \operatorname{Diag}\left((L_iZ_i)^T(L_iZ_i)\right) - ((L_i^{-1}Z_i)^T(L_i^{-1}\xi_i))^{\circ 2} \\
\nabla^2_{\beta\beta} \mathcal{L}(\beta,\gamma) &= \frac{1}{2}\sum_{i=1}^m (L_i^{-1}X_i)^T(L_i^{-1}X_i) \\
\nabla^2_{\beta\gamma} \mathcal{L}(\beta,\gamma) &= -\frac{1}{2}\sum_{i=1}^m (L_i^{-1}X_i)^T(L_i^{-1}Z_i)(L_i^{-1}Z_i)^T(L_i\xi_i) \\
\nabla^2_{\gamma\gamma} \mathcal{L}(\beta,\gamma) &= \frac{1}{2}\sum_{i=1}^m -((L_i^{-1}Z_i)^T(L_i^{-1}Z_i))^{\circ 2} + 2(L_i^{-1}Z_i)^T(L_i^{-1}\xi_i)((L_i^{-1}\xi_i)^T(L_i^{-1}Z_i))^T \circ ((L_i^{-1}Z_i)^T(L_i^{-1}Z_i))
\end{aligned}
\tag{2.23}
$$

Apparently, one should calculate bigger building blocks like $(L_iZ_i)^TL_iZ_i$ only once and then reuse it for efficient performance.

Positive Approximation of Hessian Since the problem is non-convex with respect to γ the Hessian is not in general positive definite. Negative definite Hessians are known to hamper the convergence of second-order methods [Nocedal and Wright, 2006]. In order to prevent this we use a positive upper bound as an approximation of the true Hessian. Notice that true Hessian consists of two terms: positive definite and negative definite:

$$
\nabla^2_{\gamma\gamma} \mathcal{L}(\beta,\gamma) = \frac{1}{2}\sum_{i=1}^m \underbrace{-(Z_i^T\Omega_i^{-T}Z_i)^{\circ 2}}_{\text{negative definite}} + \underbrace{2Z_i^T\Omega_i^{-T}(Y_i - X_i\beta)((Y_i - X_i\beta)^T\Omega^{-T}Z_i)^T \circ (Z_i^T\Omega^{-1}Z_i)}_{\text{positive definite}} \tag{2.24}
$$

We approximate Hessian with its positive definite part $H(\beta,\gamma)$:

$$
\nabla^2_{\gamma\gamma} \mathcal{L}(\beta,\gamma) \prec H(\beta,\gamma) \equiv 2Z_i^T\Omega_i^{-T}(Y_i - X_i\beta)((Y_i - X_i\beta)^T\Omega^{-T}Z_i)^T \circ (Z_i^T\Omega^{-1}Z_i) \tag{2.25}
$$

2.3 Experiments on Synthetic Data

2.3.1 Comparison to Other Algorithms

In the first set of synthetic tests we replicate the experimental setup from [Bondell et al., 2010], which was used to estimate the performance of M-ALASSO. The same setup has been used for evaluating the performance of SCAD-P [Fan and Li, 2012] and two-stage adaptive LASSO of [Lin et al., 2013], which allows us to compare the performance of R&S-Mixed relatively to the performance of its competitors. We also include rPQL of [Hui et al., 2017] who applied shrinkage operators to PQL quasi-likelihood to

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 2.1: 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.

facilitate feature selection in Generalized Linear Mixed Models (GLMMs) for the sake of completeness, although rPQL is designed to address a more general setup then R&S-Mixed .

Two scenarios were considered. In each of them 200 datasets were generated from the multivariate normal distribution:

$$y_i \sim \mathcal{N}(X_i\beta,\ Z_i\Gamma Z_i^T + I) \tag{2.26}$$

where:

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

$$\Gamma = \begin{bmatrix} 9 & 4.8 & 0.6 & 0 \\ 4.8 & 4 & 1 & 0 \\ 0.6 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.27}$$

  Columns of X_i and Z_i were generated as i.i.d. vectors from a uniform distribution on $(-2, 2)$, and observation errors were generated from a standard normal distribution.

- **Scenario 2**: everything as in Scenario 1, but $n = 60$ and $n_i = 10$ to estimate performance on a larger dataset.

The aspect which makes this setup particularly challenging for R&S-Mixed is that the matrix Γ is not diagonal, so the true parameters (β^*, Γ^*) are out of the model's parameters scope. This inevitably pulls the performance down as the model is unable to take the covariance structure of random effects into account properly, and so the respective model bias is introduced to both estimates of β and u_i. It's still, however, possible to compare the performance of feature selection by comparing the subset of selected random covariates which is defined by non-zero elements of γ.

The evaluation results are listed in Table 2.1. Data for all algorithms except R&S-Mixed was borrowed from [Hui et al., 2017] who aggregated it from the original works. From this table it's clear that the performance of R&S-Mixed on Setup 1 is comparable to performance of other algorithms in terms of selecting correct fixed or random effects separately, but is worse in terms of making a correct selection simultaneously. This is expected given that for small datasets the finite sample error adds up to the error from restricting γ to be diagonal creating a significant gap in performance. This difference, however, disappears when the algorithms are compared on a bigger Setup 2, where R&S-Mixed performs on par with its strongest competitors. This behavior suggests that the diagonal γ restriction is not as restrictive as it might seem, and our algorithm can enjoy both optimization stability and high feature selection accuracy given a sufficiently large sample size even if true random effects are correlated. Besides, R&S-Mixed demonstrated lower mean-squared prediction error and four orders of magnitude shorter execution time comparing to rPQL[2].

2.3.2 Scalability Experiments

In the second set of synthetic experiments we want to examine how the feature selection quality of R&S-Mixed changes when the number of features grows and the number of observations stays fixed.

The setting is similar to **Scenario 2** from Chapter 2.3.1: $n = 60$, $n_i = 10$, q was set equal to p, and also all X_i were equal to Z_i for all i. The number of features p was iteratively increasing by blocks of three, where the blocks of new columns X_i and Z_i were sampled i.i.d. from a multivariate normal distribution $\mathcal{N}(0, \Psi)$ with the variance-covariance matrix Ψ:

$$\Psi = \begin{bmatrix} 9 & 4.8 & 0.6 \\ 4.8 & 4 & 1 \\ 0.6 & 1 & 1 \end{bmatrix} \tag{2.28}$$

For each experiment, random half of all features were active fixed features with 70% of those active fixed features also having a random effect. To be precise, β had a random 50% subset of its coordinates equal to 1 with zeros elsewhere. Random 70% out of those non-zero coordinates in β had 1s in respective positions in γ ($\Gamma := \mathrm{Diag}(\gamma)$), with the rest coordinates of γ to be 0.

The results are presented on Figure 2.1. One can immediately notice that the behavior of plots splits the picture into three different zones:

1. **Zone 1: "tall" datasets.** This zone corresponds to the situation when $n_i \geq p$ for all i, so every group has more observations than there are covariates in the dataset. The performance of the

[2] Authors of other above mentioned algorithms did not report their execution times.

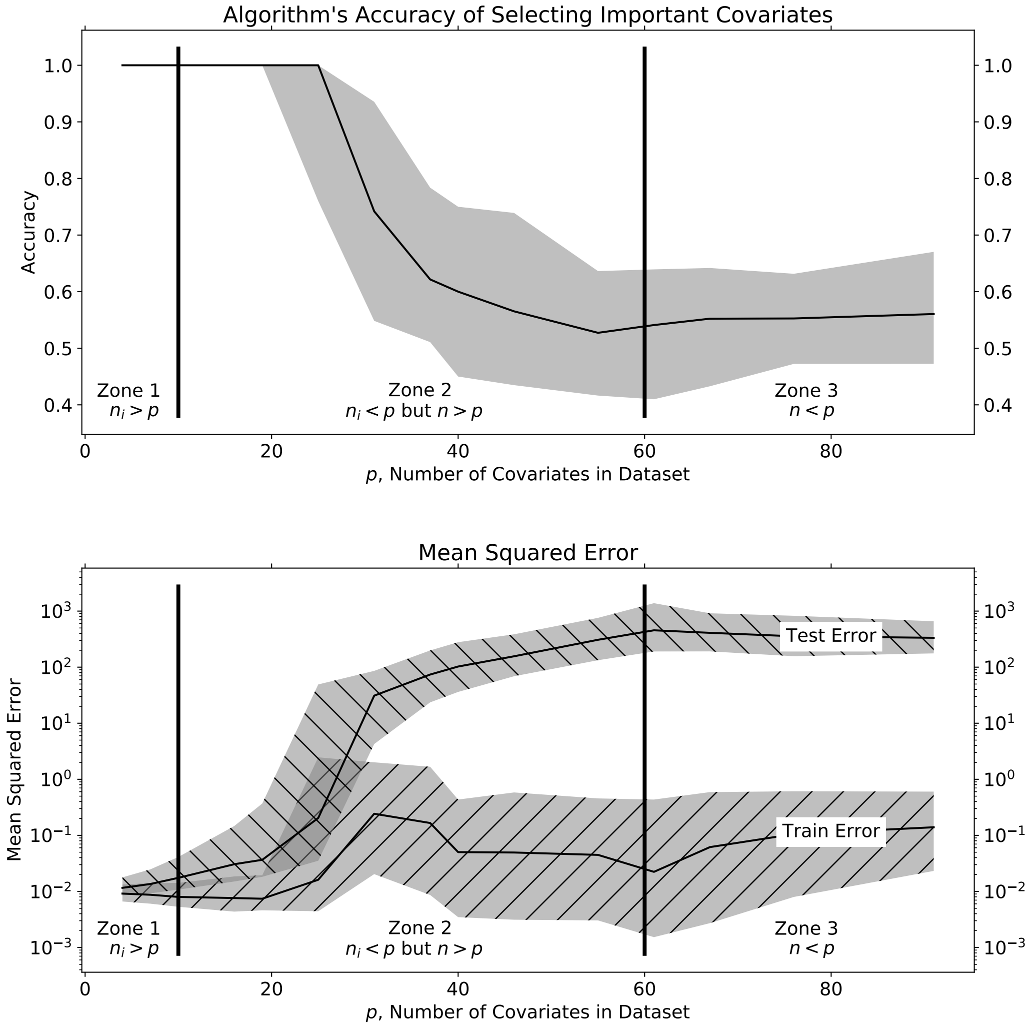


Figure 2.1: This is a scalability experiment that examines the performance of R&S-Mixed on synthetic data. We looked at how algorithm's performance varies when the number of observations is fixed and the number of features changes. In each synthetic dataset a random half of covariates were assigned to be significant and were used to generate observations, the rest was not used. The top figure shows the accuracy of selecting true important covariates depending on the number of covariates available. The bottom figure shows mean squared error of the model's prediction on train and test data. Each synthetic experiment was repeated 300 times, the shaded areas indicate 95% confidence intervals of the empirical distribution of outcomes of those experiments. Depending on the relation between the number of observations per group (n_i), total number of observations ($n = \sum_i n_i$), and the number of features p, we see three distinctive "performance zones". In Zone 1, which corresponds to "tall" datasets ($n_i \geq p$ for all i), algorithm identifies important covariates correctly in 100% cases. In Zone 2 ("square" datasets: $n_i < p$ but $n > p$) the performance falls around $p = n/2$ to the performance of random guessing, keeping this poor performance in Zone 3 ("wide" datasets: $p > n$). Given that the train error is significantly lower than the test error in Zones 2 and 3, we conclude that the resulting models overfit the data when number of covariates is bigger than the number of observations.

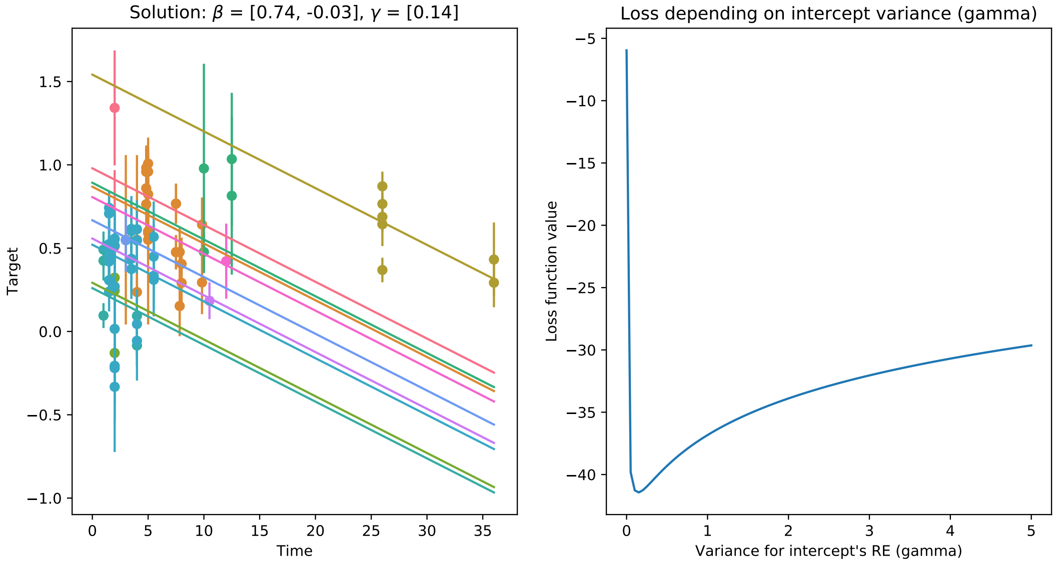


Figure 2.2: Model fit for Internet Bullying Data from GBD 2020 using only intercept and time. We see that the groups are not separable, but the fit of a mixed model is reliable: there is a distinctive minimum of the loss function around $\gamma = 0.17$.

algorithm in this zone is nearly perfect, with all correct covariates to be selected. The train error is nearly the same to the test error which evidences that the model generalizes the data properly.

2. **Zone 2: "square" datasets.** This zone corresponds to the situation when $n_i < p$ but $n = \sum_i n_i > p$, so a problem of fitting individual regression models would lead to an underdetermined system for each group, but the problem of fitting a linear mixed regression as a whole is still well-defined. Although for $p \leq n/2$ the performance is still nearly perfect, there is a clear drop in performance in terms of both selection accuracy and RMSE on test data once $p \geq n/2$. This is a strong sign that the model overfits the data which leads to generalization errors.

3. **Zone 3: "wide" datasets.** This zone corresponds to solving underdetermined systems where $p > n$. We see that the performance of the method is roughly equal to performance of random guessing, which means that R&S-Mixed is not suitable for learning from "wide" datasets at this point.

This analysis shows that the method can be successfully applied to the problems up until $p > n/2$, with "wide" dataset setups to be an apparent space for improvement.

Figure 2.3: Fixed Feature Selection for Bullying Data from GBD 2020. The algorithm picks five historically significant covariates and two historically insignificant. See the Appendix B.1 for covariates description and assessment of significance.

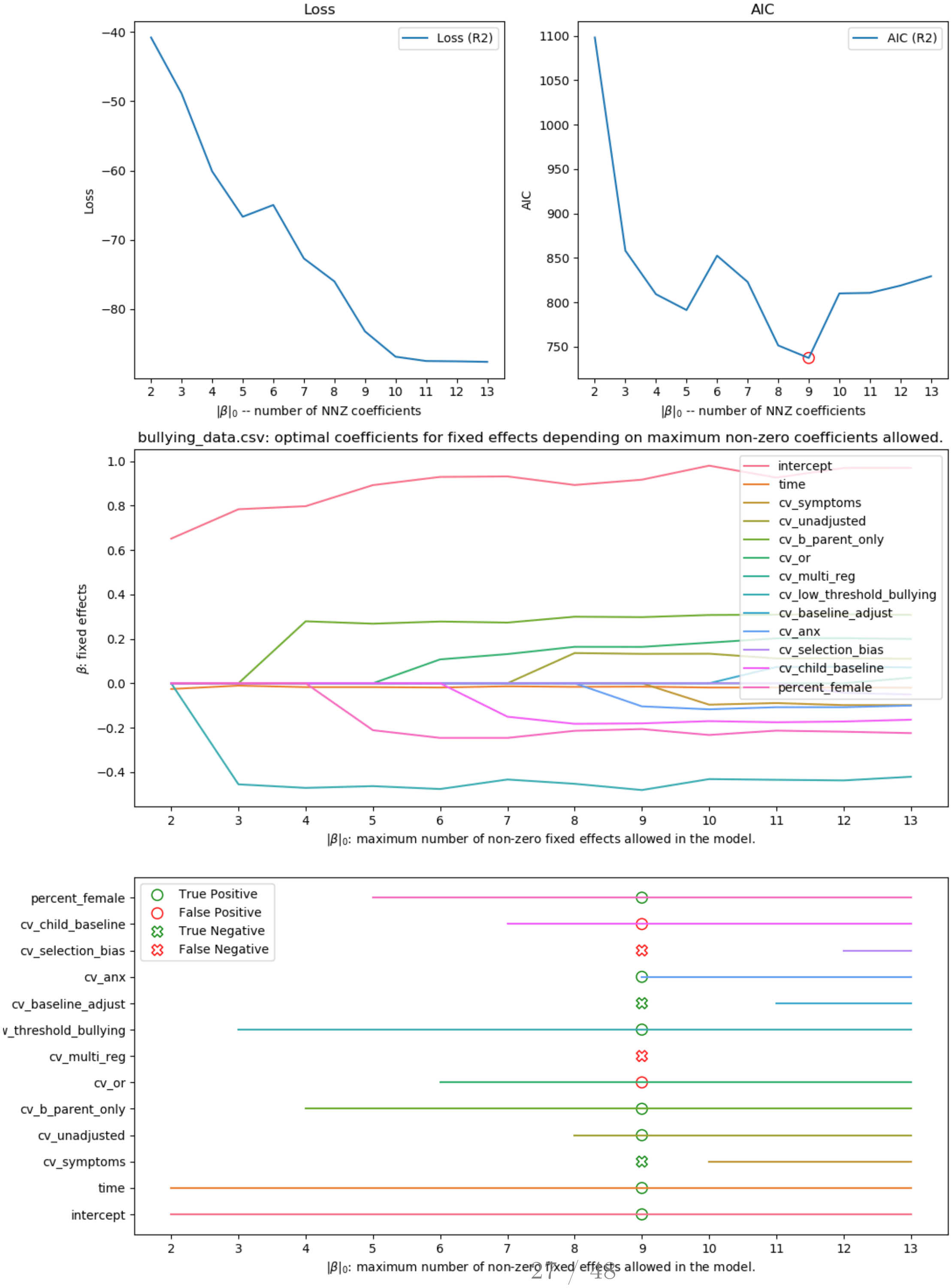
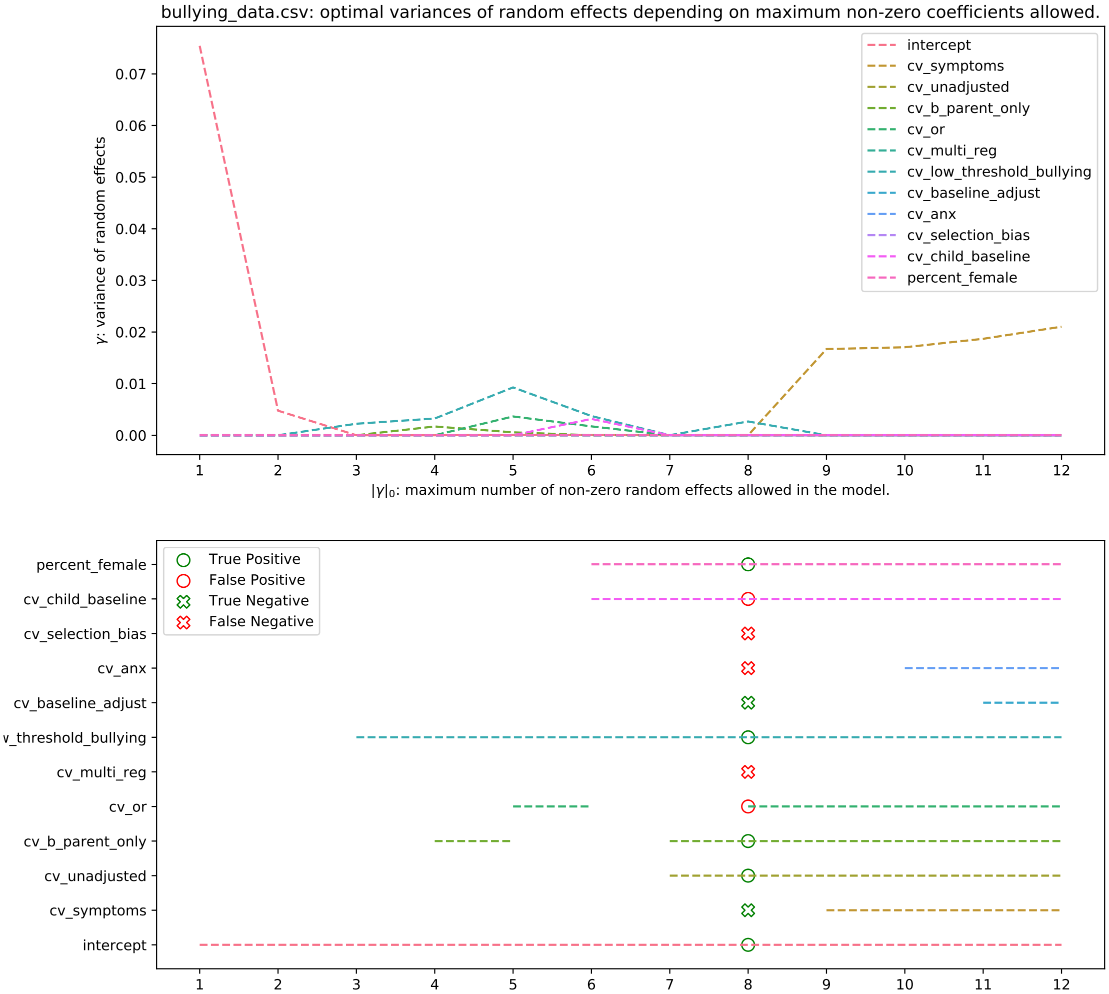


Figure 2.4: Random Feature Selection for Bullying Data from GBD 2020. The algorithm picks five historically significant covariates and two historically insignificant. See the Appendix B.1 for covariates description and assessment of significance.



*the time is constrained to be always included as a fixed effect, so the procedure asks to choose i fixed effects and $i-1$ random effects (out of those i fixed effects).

2.4 Experiments on Real Data

2.4.1 Meta-Research on Internet Bullying

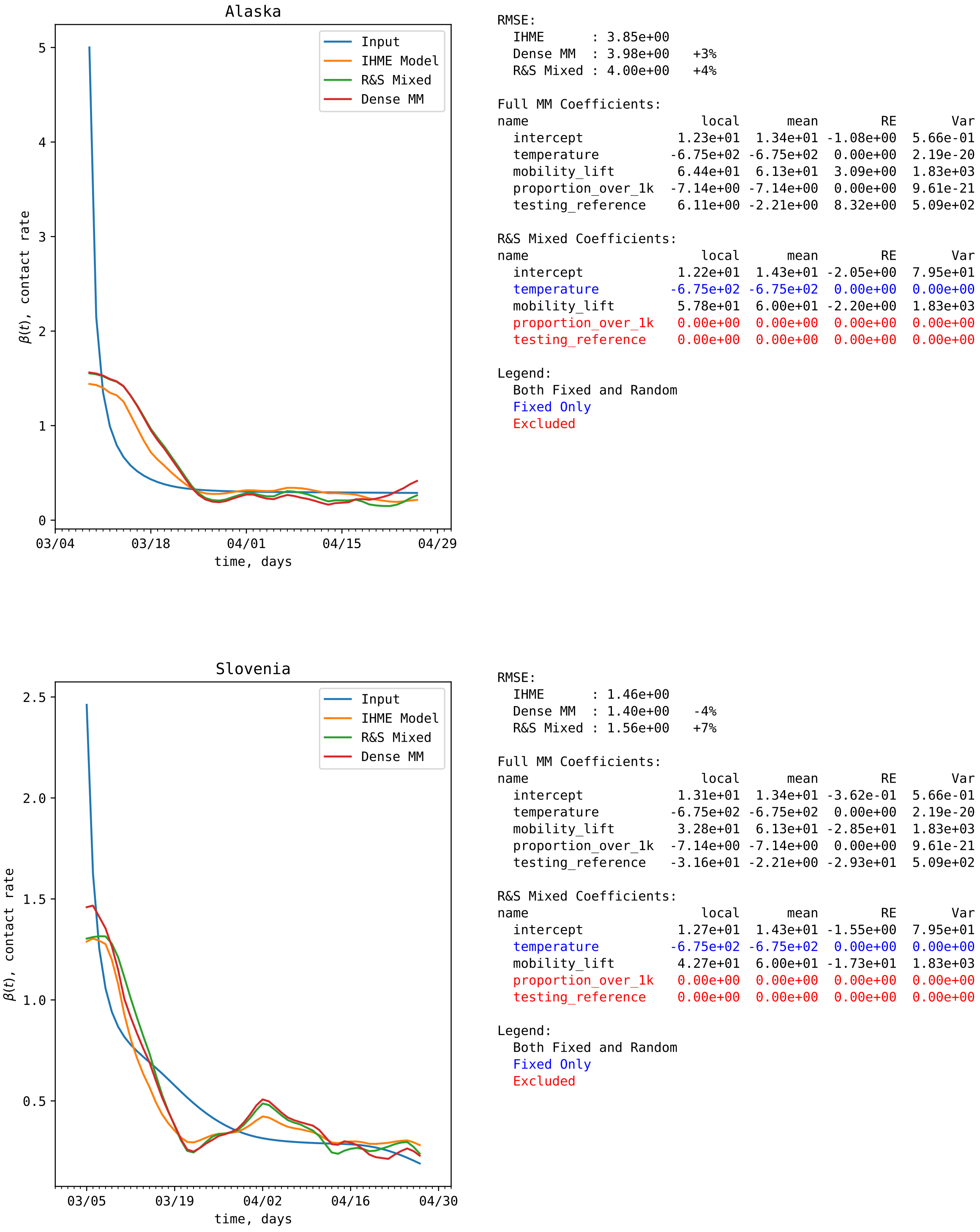
In this section we apply R&S-Mixed to identify the most important covariates in research on relative risk of anxiety and depression disorders depending on the exposure to bullying in young age[3]. This research has been a part of Global Burden of Diseases study for the last several years. The end goal for this work was to estimate the burden (DALYs) of major depressive disorder (MDD) and anxiety disorders that caused by bullying. For this risk factor, the exposure is primarily concentrated in childhood and adolescents, but the risk for MDD and anxiety disorders is anticipated to continue well into adulthood. This elevated risk is, however, expected to decrease with time as other risk factors come into play in adulthood (unemployment, relationship issues, etc.). To accommodate this, the research team uses the models which estimate the relative risk (RR) of MDD and anxiety disorders among persons exposed to bullying depending on how many years it has been since the first exposure. Studies informing the model were sourced from a systematic review and consist of longitudinal cohort studies. They measure exposure to bullying at baseline, and then follow up years later and assess them for MDD or anxiety disorders. The detailed description of the covariates can be found in Appendix B.1 .

The feature selection process is illustrated on Figures 2.3 and 2.4. Since there was no apparent prior on k – the number of features to keep in the model – an AIC criterion from [Vaida and Blanchard, 2005] was used to make a choice (see the plot on the first row, column two on Figure 2.3). We see that there is a minimum around $k = 9$, although the minimum is shallow so $k = 8$ can also be considered. Five extra covariates (the ones outside `intercept` and `time`) were selected by R&S-Mixed : `cv_unadjisted`, `cv_threshold_bullying`,`cv_b_parent_only`, `cv_anx` and `percent_female` (have been important in the previous years of GBD). The covariates `cv_or` and `cv_child_baseline` has not been important previously but also got selected.

2.4.2 COVID-19 Contact Rate Forecasting Data

In this section we apply our method to a COVID-19 Contact Rate Forecasting problem. The global pandemic created an unprecedented need for robust and accurate disease transmission forecasting. Since the beginning of the pandemic Institute for Health Metrics and Evaluation has been providing guidance to local authorities across the world with their pioneering COVID-19 Projections tool [IHME, 2020]. The key methodology which was essential for success in their forecasting of the disease's dynamics was transforming the death data into the contact rate time series data, and then relating this contact rate to available covariates such as temperature, social mobility, population, and others [IHME COVID-19 Forecasting Team, 2020]. All these covariates were collected in real time using limited human resources, and identifying the most important covariates was crucial for distributing these resources effectively. In the example below we show how R&S-Mixed can be used for making covariate selection on IHME data.

[3] The data is available at ...



```
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
```

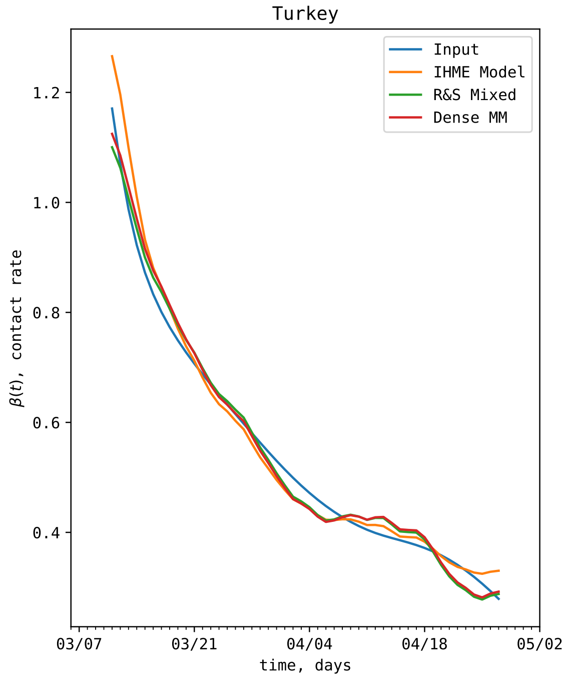
```
RMSE:
  IHME      : 1.46e+00
  Dense MM  : 1.40e+00   -4%
  R&S Mixed : 1.56e+00   +7%

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
```

Figure 2.5: Comparison of fits of two different models (fully dense linear mixed model and R&S-Mixed) to the original IHME Projections for Alaska and Slovenia. The quality (RMSE) achieved by a sparse fit is within 10% from a quality of both dense models which is evidences that the model picked up informative features.

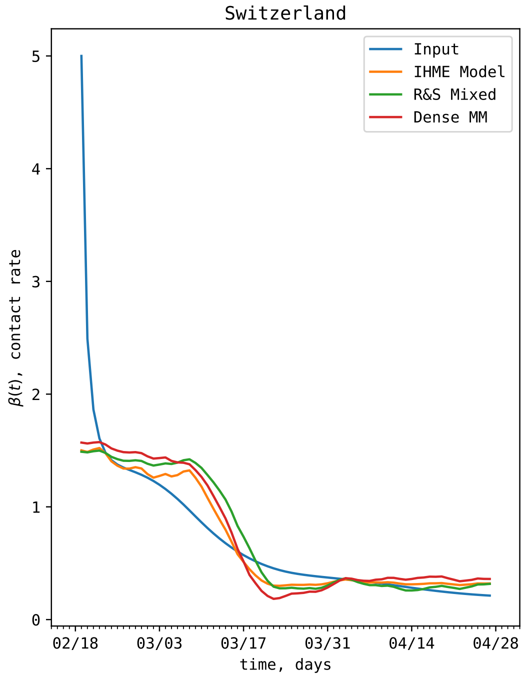


```
RMSE:
  IHME      : 2.67e-01
  Dense MM  : 1.85e-01   -31%
  R&S Mixed : 1.71e-01   -36%

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
```



```
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
```

Figure 2.6: Comparison of fits of two different models (fully dense linear mixed model and R&S-Mixed) to the original IHME Projections for Turkey and Switzerland. The quality (RMSE) achieved by a sparse fit is within 10% from a quality of both dense models which is evidences that the model picked up informative features.

The dataset consists of $m = 60$ groups (countries or states), the detailed description of groups sizes (n_i) and time spans is provided in the Table B.1 in the Appendix. The target variable Y_i was the contact rate – the coefficient $\beta(t)$ from an SEIR model (not to be confused with β – vector of fixed effects). The covariates – columns of X_i and Z_i – were: `intercept` –a column of ones, `temperature` – average air temperature in degrees Fahrenheit, `mobility_lift` – social mobility, `proportion_over_1k` – population size threshold, and `testing_reference` – testing. The observation error's variance σ_i was set to be 0.1.

Feature selection results for four particular locations (Alaska, Slovenia, Turkey, and Switherland) are presented on Figures 2.5 and 2.6, with coefficients for the rest of the locations attached in supplementary materials. R&S-Mixed was charged with a task to produce a fit using only three covariates, one of which had to be mean-only (no random effects). On the left we see the original data (`Data`, in blue), and predictions of three models: IHME Projections (`IHME`, in orange), a linear mixed model fit with no selection (`Dense MM`, in red), and a sparse fit of R&S-Mixed (R&S-Mixed in green). The R&S-Mixed model has chosen to use `intercept`, `mobility_lift`, and `temperature` covariates, with the later only as a fixed covariate; `testing_reference` and `proportion_over_1k` were left out. On the plots to the left we see that the exclusion of two covariates did not significantly affect the quality of predictions. The residual errors (RMSE) to the right support this statement: the difference in RSME is within 10% of RSME of a "dense" model which uses all the covariates, as well as from IHME's original model which fit all the locations via sequence of independent regressions. This choice also seems reasonable given the timespan: the proliferation of testing during the spring was not yet significant, so it did not inform predictions in a major way. The exclusion of `proportion_over_1k` could have been due to the scale of grouping: locations were grouped on the level of states and countries, not on the level of individual counties where the influence of population-based covariates could have been more significant.

2.5 Software Implementation

In order to ensure reproducibility of our research, the proposed algorithm (R&S-Mixed , Algorithm 1) has been implemented as a part of `skmixed`[4] library. This library implements functionality for fitting linear mixed models and selecting covariates. The user interface was designed to be fully compliant with the standards[5] of `sklearn` library to minimize learning time.

[4] Available at https://github.com/aksholokhov/skmixed

[5] https://scikit-learn.org/stable/developers/develop.html

3 Discussion and Future work

Although the proposed algorithm performs on par with its competitors in its current form (Algorithm 1), there are several improvements which the author plans to work on in future. In this section we discuss these improvements and outline possible paths to them. We plan to test and deploy these methods on global health applications through our collaborations with researchers at the Institute for Health Metrics and Evaluation.

3.1 Better Control over Relaxation Parameters

Both Relax-and-Split and Interior Point methods introduce relaxations to the problem. In particular, Relax-and-Split approach relaxes ℓ_0-constraints using auxiliary variables, which gives us relaxation parameters λ_β and λ_γ, whereas IP method relaxes equality constraints in (2.12) introducing another parameter μ. All three parameters λ_β, λ_γ, and μ need to be tightened iteratively (λ_β, $\lambda_\gamma \to \infty$ and $\mu \to 0$) to guarantee that the variables β and γ satisfy the original constraints from (2.6). One can see this triplet as one big relaxation for the original problem and tune them all at the same place.

For the IP method μ is set to via a complementary slackness condition:

$$\mu^{(k)} := \sum_{i=1}^{m} \frac{v_i^{(k)} \gamma_i^{(k)}}{m} \tag{3.1}$$

that come from necessary optimality conditions on a KKT-system (2.15). At the same time, λ_β and λ_γ are increased in a static and rather arbitrary fashion: $\lambda := 2(1+\lambda)$. This approach is effective but not efficient: [Zheng and Aravkin, 2020] showed that the convergence is guaranteed, but one needs to iteratively solve the IP relaxation in full every time λ_β and λ_γ are increased. Adjusting them in the same place where μ is adjusted does not work because the steady exponential growth in λ drives $\|F_\mu\|$ up faster than Newton method drives it down over iterations, which breaks the stopping criterion of IP method $\|F_\mu\| \leq \varepsilon$.

The proposed approach is to increase λ_β and λ_γ based on the current step length α_k. The logic behind this choice is that if the relaxation parameters are too large then the method will struggle with making a step away from the projections $\tilde{\beta}$ and $\tilde{\gamma}$, and so no further increase is required. The key theorem which lays a theoretical foundation behind it would be:

Theorem 1 (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.*

There are two main reasons why the joint relaxation is more attractive than the current version. First, it will simplify the Algorithm 1 by removing the nested loop. This will accelerate the algorithm in terms of time and the number of iterations as both relaxations will be tuned within the same cycle with no further need to solve one relaxation per each step of another relaxation. Second, these theoretical findings may lead to developing a universal tuning technique for Relax-and-Split methods and might be used in other applications outside of mixed models setting.

3.2 Asymptotic Properties

Although R&S-Mixed can be classified as a penalized likelihood based method, regularity conditions from [Fan and Li, 2001], which are being widely used by competitor algorithms, are not applicable here directly because the penalty is not of the form $P_{\lambda,n}(|\beta|, |\gamma|)$. Hence, a separate analysis of statistical properties of the proposed estimator is needed.

The first theorem establishes the foundation for proving three consequent theorems by outlining the conditions under which the proposed method is guaranteed to find the right k-subspace:

Theorem 2 (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^*).*

The key challenge is to find the set of conditions which satisfy the theorem and at the same time are sufficiently general to cover real-world cases. For example, the conditions of X_i and Z_i to be sets of orthonormal vectors are known to be sufficient due to [Tibshirani, 1996], but without costly preprocessing the real world datasets are rarely orthonormal matrices.

The next step is to prove a set of statements which is normally discussed in this type of works (see, for example [Bondell et al., 2010, Hui et al., 2017]) using Theorem 1.

Theorem 3 (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^*).*

Proof Outline. The proof is based on the fact that the set of stationary points for the original formulation (2.6) matches to the set of stationary points of the relaxation $\mathcal{L}_r$ (2.10) when $\lambda_\beta, \lambda_\gamma \to \infty$. Hence, the first step would be providing a concentration bounds on how far a stationary point $(\hat{\beta}, \hat{\gamma})$ of the relaxation $\mathcal{L}_r(\beta, \gamma)$ can be from a stationary point (β^*, γ^*) of the original setup $\mathcal{L}(\beta, \gamma)$ given a fixed $\lambda_\beta \gg 1$ and $\lambda_\gamma \gg 1$. This bound *should be* n-independent meaning that the minima will match for any finite number of samples n.

Theorem 4 (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.*

Proof Outline. This is similar to the previous statement: the fact that the limiting minima of the relaxation $\mathcal{L}_r(\beta, \gamma)$ match the minima of the original setup $\mathcal{L}_{ML}(\beta, \gamma)$ implies that their zero coordinates also match.

Theorem 5 (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.*

These proofs would not only contribute to this particular work, but would also be a valuable development of Relax-and-Split approach in general. This is ongoing work which should be done by the time of Thesis Defense.

3.3 Exponentially Smoothed Projection

One of the aspects of Algorithm 1 which drags the selection accuracy downwards is premature, all-in, guessing of the sparse subspaces. Namely, $\tilde{\beta}$ and $\tilde{\gamma}$ are assigned to be the direct projections of β and γ, which may not identify the right subspace correctly during the first iterations when the problem is significantly relaxed. At the same time, the attraction to $\tilde{\beta}$ and $\tilde{\gamma}$ chosen on the previous iteration biases the next β and γ. One way of mitigating this effect is using exponential smoothing. In particular, lines 6 and 7 in the Algorithm 1 can be changed to the weighted exponential average. The proposed modification is outlined in Algorithm 2.

```
1  $\lambda_\beta = 0$; $\lambda_\gamma = 0$
2  repeat
3      $\lambda_\beta \leftarrow 2(1 + \lambda_\beta)$
4      $\lambda_\gamma \leftarrow 2(1 + \lambda_\gamma)$
5      repeat
6          $\tilde{\beta}^{(k+1)} \leftarrow \delta \operatorname{Proj}_{\|\beta\|_0 \leq k}(\beta^{(k)}) + (1-\delta)\tilde{\beta}^{(k)}$
7          $\tilde{\gamma}^{(k+1)} \leftarrow \delta \operatorname{Proj}_{\|\gamma\|_0 \leq s}(\gamma^{(k)}) + (1-\delta)\tilde{\gamma}^{(k)}$
8          $\beta^{(k+1)}, \gamma^{(k+1)} \leftarrow \operatorname{argmin}_{\gamma \geq 0, \beta} \mathcal{L}(\beta, \gamma) + \frac{\lambda_\beta}{2}\|\beta - \tilde{\beta}^{(k)}\|_2^2 + \frac{\lambda_\gamma}{2}\|\gamma - \tilde{\gamma}^{(k)}\|_2^2$
9      until $\underline{\text{converges}}$;
10 until $\underline{\tilde{\beta} \approx \beta, \tilde{\gamma} \approx \gamma}$;
```

Algorithm 2: A prospective version of R&S-Mixed with exponential smoothing for projected variables. This modification should increase accuracy of covariates selection.

The hyper parameter δ can be assigned via cross-validation tuning. Alternatively, it's possible to analyze this scheme via Nesterov's Accelerated Methods framework [Nesterov, 2004] or Polyak's Heavy Ball framework [Polyak, 1964] in hope that a proper choice of δ will increase the convergence rate of the algorithm.

3.4 Correlated Features

In Algorithm 1, when we project β onto a k-subspace constraint to get $\tilde{\beta}$ we take k maximum elements in their absolute value from β and put them into $\tilde{\beta}$, setting the rest to 0. This can also be viewed as a minimization of a local quadratic model with a (trivial) approximation of Hessian with an identity matrix.

Namely, for the iteration j:

$$\begin{aligned}\tilde{\beta}^* := \operatorname*{argmin}_{\tilde{\beta}} & \mathcal{L}_r(\beta^{(j)}, \gamma^{(j)}) + \nabla_\beta \mathcal{L}_r(\beta^{(j)}, \gamma^{(j)})^T(\tilde{\beta} - \beta^{(j)}) + \frac{1}{2}(\tilde{\beta} - \beta^{(j)})^T I(\tilde{\beta} - \beta^{(j)}) \\ \text{s.t. } & \|\tilde{\beta}\|_0 \leq k\end{aligned} \tag{3.2}$$

Since $(\beta^{(j)}, \gamma^{(j)})$ is a partial minimizer of $\mathcal{L}_r(\beta, \gamma, \tilde{\beta}, \tilde{\gamma})$ for the iteration j, we have that the second term equal to 0. Notice $\tilde{\beta}^*$ would be a correct minimizer of the quadratic approximation of the loss function *given that Hessian can be approximated by an identity*, which is not always the case. As we have from (2.23):

$$\nabla^2_{\beta\beta} \mathcal{L}(\beta, \gamma) = \frac{1}{2}\sum_{i=1}^{m} X_i^T \Omega_i^{-1} X_i \tag{3.3}$$

which is equal to $\frac{1}{2}I$ in, for instance, the case of "orthonormal design": when all columns of X_i are orthonormal for all I. This is not the case in general when fixed features are correlated, which suggests that R&S-Mixed might not work as well in a setup with highly correlated features. This formulation also suggests an extension:

$$\begin{aligned}\tilde{\beta}^* := \operatorname*{argmin}_{\tilde{\beta}} & \mathcal{L}_r(\beta^{(j)}, \gamma^{(j)}) + \nabla_\beta \mathcal{L}_r(\beta^{(j)}, \gamma^{(j)})^T(\tilde{\beta} - \beta^{(j)}) + \frac{1}{2}(\tilde{\beta} - \beta^{(j)})^T (I + \nabla^2_{\beta\beta}\mathcal{L}(\beta^{(j)}, \gamma^{(j)}))(\tilde{\beta} - \beta^{(j)}) \\ \text{s.t. } & \|\tilde{\beta}\|_0 \leq k\end{aligned} \tag{3.4}$$

This is a “quadratic knapsack” problem which is NP-complete. Nevertheless, there are effective approximation methods for solving this problem. The hope is that this version of the algorithm will perform ultimately better in the setup where covariates are highly correlated.

References

[Tru,] Part III Trust-Region Methods for Constrained Optimization with Convex Constraints, pages 439–439.

[Sha, 2011] (2011). Pegasos: Primal estimated sub-gradient solver for SVM. Mathematical Programming, 127(1):3–30.

[Balakrishnan et al., 2017] Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. The Annals of Statistics, 45(1):77–120.

[Baraldi et al., 2019] Baraldi, R., Kumar, R., and Aravkin, A. (2019). Basis Pursuit Denoise With Nonsmooth Constraints. IEEE Transactions on Signal Processing, 67(22):5811–5823.

[Belyy and Sholokhov, 2018] Belyy, A. and Sholokhov, A. (2018). MEMOIR: Multi-class Extreme Classification with Inexact Margin.

[Bertsekas, 1982] Bertsekas, D. P. (1982). Projected Newton Methods for Optimization Problems with Simple Constraints. SIAM Journal on Control and Optimization, 20(2):221–246.

[Bondell et al., 2010] Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models. Biometrics, 66(4):1069–1077.

[Bradley and Mangasarian, 1998] Bradley, P. and Mangasarian, O. (1998). Feature Selection via Concave Minimization and Support Vector Machines. Proceedings of the International Conference on Machine Learning, (6):82–90.

[Champion et al., 2020] Champion, K., Zheng, P., Aravkin, A. Y., Brunton, S. L., and Kutz, J. N. (2020). A Unified Sparse Optimization Framework to Learn Parsimonious Physics-Informed Models From Data. IEEE Access, 8:169259–169271.

[Chen and Dunson, 2003] Chen, Z. and Dunson, D. B. (2003). Random Effects Selection in Linear Mixed Models. Biometrics, 59(4):762–769.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22.

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

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

[Fang, 2011] Fang, Y. (2011). Asymptotic Equivalence between Cross-Validations and Akaike Information Criteria in Mixed-Effects Models. Journal of Data Science, 9:15–21.

[Ghosh and Thoresen, 2018] Ghosh, A. and Thoresen, M. (2018). Non-concave penalization in linear mixed-effect models and regularized selection of fixed effects. AStA Advances in Statistical Analysis, 102(2):179–210.

[Harville, 1976] Harville, D. (1976). Extension of the Gauss-Markov Theorem to Include the Estimation of Random Effects. The Annals of Statistics, 4(2):384–395.

[Harville, 1974] Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. Biometrika, 61(2):383–385.

[Hui et al., 2017] Hui, F. K., Müller, S., and Welsh, A. H. (2017). Joint Selection in Mixed Models using Regularized PQL. Journal of the American Statistical Association, 112(519):1323–1333.

[IHME, 2020] IHME (2020). IHME COVID-19 Projections.

[IHME COVID-19 Forecasting Team, 2020] IHME COVID-19 Forecasting Team (2020). Modeling COVID-19 scenarios for the United States. Nature Medicine.

[Jajuga, 1991] Jajuga, K. (1991). L1-norm based fuzzy clustering. Fuzzy Sets and Systems, 39(1):43–50.

[Jennrich and Schluchter, 1986] Jennrich, R. I. and Schluchter, M. D. (1986). Unbalanced Repeated-Measures Models with Structured Covariance Matrices. Biometrics, 42(4):805.

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

[Kim and Paik, 2019] Kim, G. S. and Paik, M. C. (2019). Doubly-robust lasso bandit. Advances in Neural Information Processing Systems, 32(NeurIPS).

[Kojima et al., 1991] Kojima, M., Megiddo, N., Noma, T., and Yoshise, A. (1991). A unified approach to interior point algorithms for linear complementarity problems: A summary. Operations Research Letters, 10(5):247–254.

[Laird et al., 1987] Laird, N., Lange, N., and Stram, D. (1987). Maximum likelihood computations with repeated measures: Application of the EM algorithm. Journal of the American Statistical Association, 82(397):97–105.

[Laird and Ware, 1982] Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. Biometrics, 38(4):963.

[Lan, 2006] Lan, L. (2006). Variable Selection in Linear Mixed Model for Longitudinal Data. PhD thesis.

[Lange and Laird, 1989] Lange, N. and Laird, N. M. (1989). The Effect of Covariance Structure on Variance Estimation in Balanced Growth-Curve Models with Random Parameters. Journal of the American Statistical Association, 84(405):241–247.

[Lee and Nelder, 1996] Lee, Y. and Nelder, J. A. (1996). Hierarchical Generalized Linear Models. Journal of the Royal Statistical Society: Series B (Methodological), 58(4):619–656.

[Lin et al., 2013] Lin, B., Pang, Z., and Jiang, J. (2013). Fixed and random effects selection by REML and pathwise coordinate optimization. Journal of Computational and Graphical Statistics, 22(2):341–355.

[Lindstrom and Bates, 1988] Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data. Journal of the American Statistical Association, 83(404):1014.

[Liu, 1998] Liu, C. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm. Biometrika, 85(4):755–770.

[Liu and Rubin, 1994] Liu, C. and Rubin, D. B. (1994). The ECME Algorithm: A Simple Extension of EM and ECM with Faster Monotone Convergence. Biometrika, 81(4):633.

[Müller et al., 2013] Müller, S., Scealy, J. L., and Welsh, A. H. (2013). Model selection in linear mixed models. Statistical Science, 28(2):135–167.

[Nash, 2000] Nash, S. G. (2000). A survey of truncated-Newton methods. Journal of Computational and Applied Mathematics, 124(1-2):45–59.

[Nesterov, 2004] Nesterov, Y. (2004). Introductory Lectures on Convex Optimization, volume 87 of Applied Optimization. Springer US, Boston, MA.

[Nesterov and Nemirovskii, 1994] Nesterov, Y. and Nemirovskii, A. (1994). Interior-Point Polynomial Algorithms in Convex Programming. Society for Industrial and Applied Mathematics.

[Nocedal and Wright, 2006] Nocedal, J. and Wright, S. (2006). Numerical optimization. Springer Science & Business Media.

[Patterson and Thompson, 1971] Patterson, H. D. and Thompson, R. (1971). Recovery of Inter-Block Information when Block Sizes are Unequal. Biometrika, 58(3):545.

[Pinheiro and Bates, 2000] Pinheiro, J. C. and Bates, D. M. (2000). Mixed-Effects Models in Sand S-PLUS, volume 96 of Statistics and Computing. Springer New York, New York, NY.

[Polyak, 1964] Polyak, B. (1964). Some methods of speeding up the convergence of iteration methods. USSR Computational Mathematics and Mathematical Physics, 4(5):1–17.

[Potra and Wright, 2000] Potra, F. A. and Wright, S. J. (2000). Interior-point methods. Journal of Computational and Applied Mathematics, 124(1-2):281–302.

[Sugiura, 1978] Sugiura, N. (1978). Further analysts of the data by akaike' s information criterion and the finite corrections. Communications in Statistics - Theory and Methods, 7(1):13–26.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288.

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

[Wright, 1997] Wright, S. J. (1997). Primal-Dual Interior-Point Methods. Society for Industrial and Applied Mathematics.

[Wu, 1983] Wu, C. F. J. (1983). On the Convergence Properties of the EM Algorithm. The Annals of Statistics, 11(1):95–103.

[Xie et al., 2020] Xie, Y., Li, Y., Xia, Z., Yan, R., and Luan, D. (2020). A Penalized h-Likelihood Variable Selection Algorithm for Generalized Linear Regression Models with Random Effects. Complexity, 2020:1–13.

[Xu et al., 2015] Xu, P., Wang, T., Zhu, H., and Zhu, L. (2015). Double Penalized H-Likelihood for Selection of Fixed and Random Effects in Mixed Effects Models. Statistics in Biosciences, 7(1):108–128.

[Zheng and Aravkin, 2020] Zheng, P. and Aravkin, A. (2020). Relax-and-split method for nonconvex inverse problems. Inverse Problems, 36(9).

[Zheng et al., 2019] Zheng, P., Askham, T., Brunton, S. L., Kutz, J. N., and Aravkin, A. Y. (2019). A Unified Framework for Sparse Relaxed Regularized Regression: SR3. IEEE Access, 7:1404–1423.

[Zou, 2006] Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101(476):1418–1429.

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320.

Acknowledgement

I am extremely grateful to my advisor, Sasha Aravkin, for providing support, guidance, and enthusiasm during my work on this project in the Department of Applied Mathematics and in Institute for Health Metrics and Evaluation.
I would also like to thank Jim Burke for his detailed review of my work which would hopefully lead to fruitful collaboration.
Last, but not least: I am grateful to Damian Santomauro, who introduced me to his work on the consequences of bullying, provided with a dataset, and with a valuable feedback on the performance of the proposed method.

A Derivatives of Marginalized Log-likelihood for Linear Mixed Models

For conciseness, let us define the mismatch $\xi_i = Y_i - X_i\beta$. We also omit the dependence on β, as it's fixed at this point. The loss function 1.4 takes the form

$$\mathcal{L}(\gamma) = \sum_{i=1}^{m} \frac{1}{2}\xi_i^T(\Omega_i(\gamma))^{-1}\xi_i + \frac{1}{2}\log\det(\Omega_i(\gamma)). \tag{A.1}$$

The derivative of the objective w.r.t γ_j, the j'th diagonal element of the matrix Γ is

$$\frac{\partial \xi_i^T\Omega_i^{-1}\xi_i}{\partial\Gamma_{jj}} = \mathrm{Tr}\left[\left(\frac{\partial \xi_i^T\Omega_i^{-1}\xi_i}{\partial\Omega_i}\right)\frac{\partial\Omega}{\partial\Gamma_{jj}}\right] = \mathrm{Tr}\left[\left(-\Omega_i^{-T}\xi_i\xi_i^T\Omega_i^{-T}\right)^T Z_i\frac{\partial\Gamma}{\partial\Gamma_{jj}}Z_i^T\right] = \tag{A.2}$$

where $\frac{\partial\Gamma}{\partial\Gamma_{jj}}$ is a structure matrix, which, in a general case, is equal to a single-entry matrix J^{jj} with jj'th element is equal to 1 and zeroes elsewhere. Substituting this back we get

$$= \mathrm{Tr}\left[\left(-\Omega_i^{-T}\xi_i\xi_i^T\Omega_i^{-T}\right)^T Z_i^j {Z_i^j}^T\right] = \tag{A.3}$$

where Z_i^j is a j'th column of the matrix Z_i. Making a circular swap we end up with

$$= \mathrm{Tr}\left[-{Z_i^j}^T\Omega_i^{-T}\xi_i\xi_i^T\Omega_i^{-T}Z_i^j\right] = -({Z_i^j}^T\Omega_i^{-T}\xi_i)^2 \tag{A.4}$$

Similarly,

$$\frac{\partial\log\det\Omega_i}{\partial\Gamma_{jj}} = \mathrm{Tr}\left[\left(\frac{\partial\log\det\Omega_i}{\partial\Omega_i}\right)\frac{\partial\Omega_i}{\partial\Gamma_{jj}}\right] = \mathrm{Tr}\left[\Omega_i^{-1}Z_i^j{Z_i^j}^T\right] = {Z_i^j}^T\Omega_i^{-1}Z_i^j \tag{A.5}$$

Taking into account that Ω_i is symmetric, we have

$$\left[\nabla_\gamma \mathcal{L}(\beta, \gamma)\right]_j = \sum_{i=1}^{m} -({Z_i^j}^T \Omega_i^{-T} \xi_i)^2 + {Z_i^j}^T \Omega_i^{-1} Z_i^j = \tag{A.6}$$

or, in vector form

$$= \sum_{i=1}^{m} \operatorname{Diag} {Z_i}^T \Omega_i^{-1} Z_i - ({Z_i}^T \Omega_i^{-T} \xi_i)^{\circ 2} = \tag{A.7}$$

where $\circ$ denotes the Hadamard (element-wise) product. Using the Cholesky decomposition $\Omega_i = L_i L_i^T$ we can calculate it more effectively, using only one triangular matrix inversion:

$$= \sum_{i=1}^{m} \left[\sum_{\text{rows}} \left(L_i^{-1} Z_i\right)^{\circ 2} - [(L_i^{-1} Z_i)^T (L_i^{-1} \xi_i)]^{\circ 2} \right] \tag{A.8}$$

Notice, that the loss function (1.4) and the optimal β solution (1.7) can also be effectively computed using Cholesky:

$$\mathcal{L}(\gamma) = \sum_{i=1}^{m} \frac{1}{2} \xi_i^T (\Omega_i(\gamma))^{-1} \xi_i + \frac{1}{2} \log \det(\Omega_i(\gamma)) = \sum_{i=1}^{m} \frac{1}{2} \, \|L_i^{-1} \xi_i\|^2 - \sum_{j=1}^{k} \log \left[L_i^{-1}\right]_{jj} \tag{A.9}$$

$$\begin{aligned} \beta_{k+1} = \operatorname*{argmin}_{\beta} \mathcal{L}(\beta, \gamma_k) &= \left(\sum_{i=1}^{m} X_i^T \Omega_i^{-1} X_i \right)^{-1} \sum_{i=1}^{m} X_i^T \Omega_i^{-1} y_i = \\ &= \left(\sum_{i=1}^{m} (L_i^{-1} X_i)^T L_i^{-1} X_i \right)^{-1} \sum_{i=1}^{m} (L_i^{-1} X_i)^T L_i^{-1} y_i \end{aligned} \tag{A.10}$$

The Hessian w.r.t. γ also can be found:

$$\begin{aligned} \frac{\partial^2 \mathcal{L}(\beta, \gamma)}{\partial \gamma_j^2} &= \sum_{i=1}^{m} -2({Z_i^j}^T \Omega_i^{-T} \xi_i) \operatorname{Tr} \left[\frac{\partial {Z_i^j}^T \Omega_i^{-T} \xi_i}{\partial \Omega_i} \frac{\partial \Omega_i}{\partial \Gamma_{jj}} \right] + \operatorname{Tr} \left[\frac{\partial {Z_i^j}^T \Omega_i^{-1} Z_i^j}{\partial \Omega_i} \frac{\partial \Omega_i}{\partial \Gamma_{jj}} \right] = \\ &= \sum_{i=1}^{m} 2({Z_i^j}^T \Omega_i^{-T} \xi_i) \operatorname{Tr} \left[\Omega_i^{-1} Z_i^j {\xi_i}^T \Omega_i^{-1} Z_i^j {Z_i^j}^T \right] - ({Z_i^j}^T \Omega_i^{-T} Z_i^j)^2 = \\ &= \sum_{i=1}^{m} 2({Z_i^j}^T \Omega_i^{-T} \xi_i)({Z_i^j}^T \Omega_i^{-1} Z_i^j)({\xi_i}^T \Omega_i^{-1} Z_i^j) - ({Z_i^j}^T \Omega_i^{-T} Z_i^j)^2 \end{aligned} \tag{A.11}$$

$$
\begin{aligned}
\frac{\partial^2 \mathcal{L}(\beta,\gamma)}{\partial\gamma_j\partial\gamma_k} &= \sum_{i=1}^m -2({Z_i^j}^T\Omega_i^{-T}\xi_i)\operatorname{Tr}\left[\frac{\partial {Z_i^j}^T\Omega_i^{-T}\xi_i}{\partial\Omega_i}\frac{\partial\Omega_i}{\partial\Gamma_{kk}}\right] + \operatorname{Tr}\left[\frac{\partial {Z_i^j}^T\Omega_i^{-1}Z_i^j}{\partial\Omega_i}\frac{\partial\Omega_i}{\partial\Gamma_{kk}}\right] = \\
&= \sum_{i=1}^m 2({Z_i^j}^T\Omega_i^{-T}\xi_i)\operatorname{Tr}\left[\Omega_i^{-1}Z_i^j{\xi_i}^T\Omega_i^{-1}Z_i^k{Z_i^k}^T\right] - ({Z_i^j}^T\Omega_i^{-T}Z_i^k)^2 = \\
&= \sum_{i=1}^m 2(\xi_i^T\Omega_i^{-T}Z_i^j)({Z_i^j}^T\Omega_i^{-1}Z_i^k)({Z_i^k}^T\Omega_i^{-1}\xi_i) - ({Z_i^j}^T\Omega_i^{-T}Z_i^k)^2
\end{aligned}
\tag{A.12}
$$

$$
\begin{aligned}
\nabla^2_\gamma \mathcal{L}(\beta,\gamma) = \frac{1}{2}\sum_{i=1}^m -(Z_i^T\Omega_i^{-T}Z_i)^{\circ 2} + 2\operatorname{Diag}(Z_i^T\Omega_i^{-T}\xi_i)(Z_i^T\Omega^{-1}Z_i)\operatorname{Diag}(\xi_i^T\Omega^{-T}Z_i) = \\
= \frac{1}{2}\sum_{i=1}^m -(Z_i^T\Omega_i^{-T}Z_i)^{\circ 2} + 2(Z_i^T\Omega_i^{-T}\xi_i)(\xi_i^T\Omega^{-T}Z_i)^T \circ (Z_i^T\Omega^{-1}Z_i)
\end{aligned}
\tag{A.13}
$$

B Description of Datasets

B.1 GBD Bullying Data

The author acknowledges his colleague and collaborator Damian Santomauro[6] for providing the dataset, the description of its covariates, and the expert assessment of their historical importance in different rounds of GBD study below.

1. `cv_symptoms`
   - 0 = study assesses participants for MDD or anxiety disorders via a diagnostic interview to determine whether they have a diagnosis.
   - 1 = study uses a symptom scale (e.g., Beck Depression Inventory) and uses an established cut-off on that scale to determine caseness.
   - Has not historically been significant.
2. `cv_unadjusted`
   - 0 = RR is adjusted for potential confounders (e.g., SES, etc.)
   - 1 = RR is not adjusted for potential confounders
   - Has been significant in the past.
3. `cv_b_parent_only`

[6] d.santomauro@uq.edu.au, Affiliate Assistant Professor of Health Metrics Sciences, Institute for Health Metrics and Evaluation, University of Washington

- 0 = Child is involved in reporting their own exposure to bullying.
- 1 = Only parent is involved in reporting the child's exposure to bullying
- This covariate has recently started becoming significant (but not consistently).

4. `cv_or`

    - 0 = estimate is a RR
    - 1 = estimate is an odds ratio (OR)
    - ORs are always larger than RRs. However the magnitude may be very small / insignificant.

5. `cv_multi_reg`

    - 0 = RR is the ratio of the rate of the outcome in persons exposed vs all persons unexposed (including persons exposed to low-threshold bullying victimization)
    - 1 = RRs are estimated via a logistic regression where exposure represented by 3 categories: 1) No exposure, 2) Occasional exposure, 3) Frequent exposure. The RR for occasional exposure will exclude participants with frequent exposure, and the RR for frequent exposure will exclude participants with occasional exposure.
    - Is expected to be significant.

6. `cv_low_threshold_bullying`

    - 0 = uses a 'frequent' exposure frequency threshold for classing someone as exposed to bullying.
    - 1 = uses an 'occasional' exposure frequency threshold for classing someone as exposed to bullying.
    - Has been consistently significant with a strong magnitude.

7. `cv_anx`

    - 0 = estimate represents risk for MDD
    - 1 = estimate represents risk for anxiety disorders

8. `cv_selection_bias`

    - 0 = $<$ 15% attrition at followup
    - 1 = $\geqslant$ 15% attrition at followup
    - Has been significant in the past

9. `Percent_female`

    - Indicates % of sample in estimate that are female.

10. `cv_child_baseline`

    - Has not been significant in the past.

B.2 COVID-19 Contact Rate Forecasting Data

Table B.1: List of locations, number of observations, start and end date for each location for COVID-19 Contact Rate Fococasting data

Location	Obs	Start	End
Malaysia	60	2020-02-27	2020-04-26
Philippines	67	2020-02-21	2020-04-27
Bulgaria	50	2020-03-09	2020-04-27
Croatia	50	2020-03-08	2020-04-26
Czechia	54	2020-03-05	2020-04-27
Hungary	55	2020-03-04	2020-04-27
Poland	56	2020-03-03	2020-04-27
Romania	56	2020-03-03	2020-04-27
Serbia	55	2020-03-04	2020-04-27
Slovakia	32	2020-03-26	2020-04-26
Slovenia	54	2020-03-05	2020-04-27
Estonia	48	2020-03-10	2020-04-26
Latvia	26	2020-04-01	2020-04-26
Lithuania	53	2020-03-05	2020-04-26
Republic of Moldova	48	2020-03-11	2020-04-27
Ukraine	53	2020-03-06	2020-04-27
Japan	68	2020-02-20	2020-04-27
Republic of Korea	85	2020-02-02	2020-04-26
Austria	62	2020-02-26	2020-04-27
Belgium	65	2020-02-23	2020-04-27
Cyprus	49	2020-03-09	2020-04-26
Denmark	61	2020-02-27	2020-04-27
Finland	53	2020-03-06	2020-04-27
France	63	2020-02-24	2020-04-26
Greece	62	2020-02-26	2020-04-27
Iceland	43	2020-03-15	2020-04-26
Ireland	58	2020-03-01	2020-04-27
Israel	56	2020-03-03	2020-04-27
Luxembourg	58	2020-02-29	2020-04-26
Netherlands	61	2020-02-27	2020-04-27
Norway	62	2020-02-26	2020-04-27
Portugal	58	2020-03-01	2020-04-27
Sweden	63	2020-02-25	2020-04-27
Switzerland	69	2020-02-19	2020-04-27

Continued on next page

Table B.1: List of locations, number of observations, start and end date for each location for COVID-19 Contact Rate Fococasting data

Location	Obs	Start	End
United Kingdom	70	2020-02-18	2020-04-27
Argentina	56	2020-03-03	2020-04-27
Chile	54	2020-03-05	2020-04-27
Dominican Republic	58	2020-03-01	2020-04-27
Ecuador	50	2020-03-01	2020-04-19
Peru	55	2020-03-04	2020-04-27
Colombia	55	2020-03-04	2020-04-27
Panama	50	2020-03-09	2020-04-27
Egypt	68	2020-02-20	2020-04-27
Iran (Islamic Republic of)	69	2020-02-19	2020-04-27
Turkey	48	2020-03-11	2020-04-27
Puerto Rico	45	2020-03-14	2020-04-27
Alabama	48	2020-03-11	2020-04-27
Alaska	49	2020-03-09	2020-04-26
Arizona	55	2020-03-04	2020-04-27
Arkansas	52	2020-03-07	2020-04-27
California	67	2020-02-21	2020-04-27
Colorado	54	2020-03-05	2020-04-27
Connecticut	52	2020-03-07	2020-04-27
Delaware	51	2020-03-08	2020-04-27
District of Columbia	54	2020-03-05	2020-04-27
Florida	57	2020-03-02	2020-04-27
Georgia	56	2020-03-03	2020-04-27
Hawaii	43	2020-03-15	2020-04-26
Idaho	50	2020-03-08	2020-04-26
Illinois	55	2020-03-04	2020-04-27

Table B.2: List of location-specific coefficients for the R&S-Mixed model fit, as well as RMSEs for three models discussed in the respective chapterCoefficient for `temperature` was set to -674.86. Coefficients for `proportion_over_1k` and `testing_reference` were set to 0.

Location	Intercept	Mobility	RMSE_IHME	RMSE_Dense	RMSE_Sparse
Malaysia	13.94	52.15	5.17	5.00	5.00
Philippines	13.60	29.45	4.20	4.16	4.16

Continued on next page

Table B.2: List of location-specific coefficients for the R&S-Mixed model fit, as well as RMSEs for three models discussed in the respective chapterCoefficient for `temperature` was set to -674.86. Coefficients for `proportion_over_1k` and `testing_reference` were set to 0.

Location	Intercept	Mobility	RMSE_IHME	RMSE_Dense	RMSE_Sparse
Bulgaria	14.24	123.94	3.42	3.20	3.20
Croatia	13.28	56.84	3.70	3.67	3.67
Czechia	13.77	103.24	2.86	3.07	3.12
Hungary	12.51	28.59	0.73	0.70	0.72
Poland	12.66	39.47	0.64	0.58	0.62
Romania	13.63	80.40	3.59	3.89	4.01
Serbia	13.16	56.92	3.79	3.80	3.84
Slovakia	11.27	-43.53	5.27	4.50	5.01
Slovenia	12.74	42.75	1.46	1.40	1.56
Estonia	14.17	155.26	2.21	2.55	2.73
Latvia	11.82	-14.61	4.94	4.41	4.68
Lithuania	12.96	66.04	3.89	3.86	3.86
Republic of Moldova	15.15	153.32	3.05	2.44	2.44
Ukraine	13.54	103.82	3.06	3.29	3.30
Japan	12.47	35.51	4.21	4.22	4.22
Republic of Korea	12.62	99.62	4.84	4.67	4.70
Austria	13.07	64.96	3.97	3.90	3.93
Belgium	13.44	51.77	3.47	3.39	3.39
Cyprus	12.65	25.21	1.84	0.64	0.66
Denmark	12.79	47.14	1.79	1.87	1.94
Finland	12.87	73.08	3.53	3.63	3.68
France	12.95	32.79	1.57	1.64	1.66
Greece	12.74	29.97	1.42	1.56	1.56
Iceland	16.35	226.87	5.77	3.09	3.17
Ireland	13.54	57.98	3.98	4.00	4.00
Israel	13.83	66.97	3.46	3.71	3.71
Luxembourg	12.51	21.39	1.47	1.53	1.73
Netherlands	12.90	52.99	1.35	1.46	1.47
Norway	12.22	34.27	1.70	1.64	1.65
Portugal	13.51	52.51	2.47	2.55	2.55
Sweden	12.95	99.37	3.93	3.80	3.90
Switzerland	12.68	66.51	3.79	3.88	3.98
United Kingdom	13.28	51.22	4.70	4.71	4.72
Argentina	13.29	36.93	1.45	1.63	1.67

Continued on next page

Table B.2: List of location-specific coefficients for the R&S-Mixed model fit, as well as RMSEs for three models discussed in the respective chapterCoefficient for `temperature` was set to -674.86. Coefficients for `proportion_over_1k` and `testing_reference` were set to 0.

Location	Intercept	Mobility	RMSE_IHME	RMSE_Dense	RMSE_Sparse
Chile	13.64	73.48	3.51	3.62	3.63
Dominican Republic	13.78	40.14	2.21	2.23	2.23
Ecuador	15.97	128.02	8.21	8.07	8.07
Peru	12.97	16.00	1.08	0.96	1.18
Colombia	13.88	47.20	3.63	3.63	3.63
Panama	13.12	10.67	0.25	0.27	0.27
Egypt	13.11	41.80	3.89	3.95	3.96
Iran (Islamic Republic of)	12.27	15.23	0.93	1.03	1.05
Turkey	12.60	28.83	0.27	0.18	0.17
Puerto Rico	13.76	45.18	0.68	0.34	0.35
Alabama	12.81	27.02	0.70	0.65	0.84
Alaska	12.24	57.82	3.85	3.98	4.00
Arizona	13.40	66.31	4.00	3.85	4.03
Arkansas	13.32	91.92	4.02	3.87	3.88
California	13.14	54.89	3.81	3.86	3.88
Colorado	12.45	37.13	0.75	1.01	1.03
Connecticut	13.28	69.51	0.59	0.77	0.92
Delaware	13.34	68.28	3.86	3.77	3.83
District of Columbia	13.27	49.45	3.42	3.55	3.56
Florida	13.34	30.01	0.29	0.34	0.36
Georgia	12.74	19.19	0.45	0.34	0.66
Hawaii	14.97	134.74	3.74	3.34	3.39
Idaho	12.75	92.74	3.81	3.84	3.89
Illinois	12.73	32.57	0.64	0.56	0.70