A Relaxation Approach to Feature Selection for Linear Mixed Effects Models

Aleksei Sholokhov[*],

Department of Applied Mathematics, University of Washington,

James V. Burke[†]

Department of Mathematics, University of Washington,

Damian F. Santomauro[*],

School of Public Health, The University of Queensland,

Queensland Centre for Mental Health Research,

Institute of Health Metrics and Evaluation, University of Washington,

Peng Zheng[*],

Department of Health Metrics Sciences &

Institute of Health Metrics and Evaluation, University of Washington,

and

Aleksandr Aravkin[*],

Department of Applied Mathematics &

Institute of Health Metrics and Evaluation, University of Washington

Monday 23rd October, 2023

Abstract

Linear Mixed-Effects (LME) models are a fundamental tool for modeling correlated data, including cohort studies, longitudinal data analysis, and meta-analysis. Design and analysis of variable selection methods for LMEs is more difficult than for linear regression because LME models are nonlinear. In this work we propose a novel optimization strategy that enables a wide range of variable selection methods for LMEs using both convex and nonconvex regularizers, including ℓ_1, Adaptive-ℓ_1, SCAD, and ℓ_0. The computational framework only requires the proximal operator for each regularizer to be readily computable, and the implementation is available in an open source `python` package `pysr3`, consistent with the `sklearn` standard. The numerical results on simulated data sets indicate that the proposed strategy improves on the state of the art for both accuracy and compute time. The variable selection techniques are also validated on a real example using a data set on bullying victimization.

Keywords: Mixed effects models, feature selection, nonconvex optimization

1 Introduction

Linear mixed-effects (LME) models use covariates to explain the variability of target variables in a grouped data setting. For each group, the relationship between covariates and observations is modeled using group-specific coefficients that are linked by a common prior distribution across all groups, allowing LMEs to borrow strength across groups in order to estimate statistics for the common prior. LMEs are used in settings with insufficient data to resolve each group independently, making them fundamental tools for regression analysis in population health sciences (Reiner et al. (2020); Murray et al. (2020)), meta-analysis (DerSimonian and Laird (1986); Zheng et al. (2021)), life sciences, and as well as in many others domains (Zuur et al. (2009)).

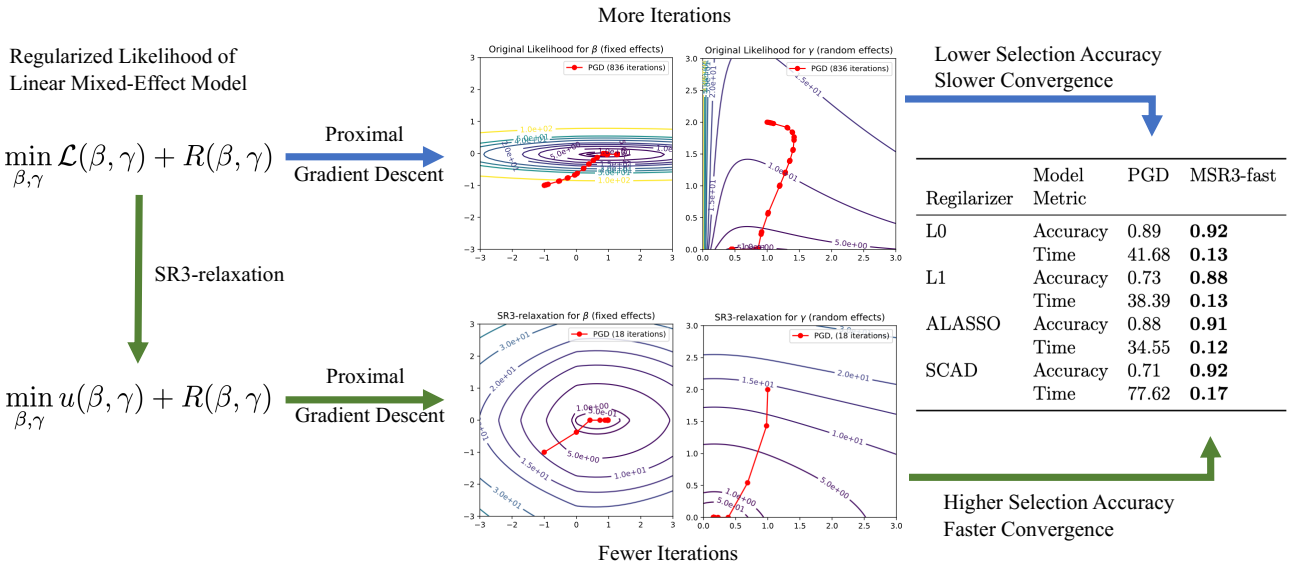
Variable selection is a fundamental problem in all regression settings. In linear regression, the LASSO method (Tibshirani, 1996a) and related extensions have been widely used. However,

*Bill and Melinda Gates Foundation
†U.S. NSF grant DMS-1908890

variable selection for LMEs is complicated by the nonlinear structure and relative sparsity of the within-group data. While standard methods and software are available for linear regression (see e.g. `glmnet` Friedman et al. (2010)), there are few open source libraries for variable selection for LMEs. Many covariates selection algorithms for LMEs have been proposed over the last 20 years (see the survey Buscemi and Plaia (2019)), but comparison of these strategies and practical application remains difficult. Approaches vary by choice of likelihood (e.g. marginal, restricted, or h- likelihood), regularizer (e.g. ℓ_1 (Bondell et al., 2010) or SCAD Ibrahim et al. (2011a)), and information criteria (Vaida and Blanchard, 2005; Ibrahim et al., 2011b). Implementations vary as well, typically using regularizer-specific local quadratic approximations to apply solution methods for smooth problems (Newton-Raphson, EM, sequential least squares) to fit the original nonsmooth model. All of these decisions make it difficult to compare and evaluate performance of available variable selection strategies and to determine which method is best suited for a given task. This challenge is exacerbated by the absence of standardized datasets and open source libraries for each method. Our main practical goal to fill this gap by developing a unified methodological framework that accommodates a wide variety of variable selection strategies based on a set of easily implementable regularizers, and made available in an open source library, `pysr3`[1] that is easy to use and to compare different methods. All experiments in the paper can be reproduced using `pysr3` and code in the reproducibility guide[2].

In this work we introduce a regularization-agnostic covariate selection strategy that (1) is fast and simple to implement, (2) provides robust models, and (3) is flexible enough to support most regularizers currently used in variable selection across different domains. The baseline approach uses the proximal gradient descent (PGD) method, which has been studied by the optimization community for over 40 years, but has not been widely used in LME covariate selection. In our initial numerical experiments, using a naive PGD approach indicated that, at best, the method yields only a marginal improvement over the equally unsatisfactory alternative methods in accurately determining the correct variables in our variable selection test problems. We conjecture that

[1] https://github.com/aksholokhov/pysr3
[2] https://github.com/aksholokhov/msr3-paper



Regilarizer	Model Metric	PGD	MSR3-fast
L0	Accuracy	0.89	**0.92**
	Time	41.68	**0.13**
L1	Accuracy	0.73	**0.88**
	Time	38.39	**0.13**
ALASSO	Accuracy	0.88	**0.91**
	Time	34.55	**0.12**
SCAD	Accuracy	0.71	**0.92**
	Time	77.62	**0.17**

Figure 1: Selection of fixed and random effects for LME likelihoods $\mathcal{L}$ using 'regularization-agnostic' framework and its SR3 extension using four regularizers. SR3 relaxation accelerates algorithmic converge (middle panel), and gives better robustness and improved performance on synthetic problems across regularizers (right panel)

.

the weakness of the naive PGD is a result of the first-order likelihood approximation used in PGD. To overcome this problem, we propose an alternative likelihood approximation with the goal of incorporating global variational properties of the likelihood. For this purpose, we extend the sparse relaxed regularized regression (SR3) framework (Zheng et al. (2019)) to the LME setting. This idea is supported by the success of SR3 in the context of linear regression where it accelerates and improves the performance of regularization strategies. This extension and its mathematical foundations in Aravkin et al. (2022) constitute the major innovations of this work. Here we introduce the modeling framework and its relaxation, discuss the resulting algorithms and their implementation details, and validate the method on both simulated and real data sets, while the mathematical foundations are presented in Aravkin et al. (2022). The SR3 framework introduces auxiliary variables x and w to decouple the likelihood $\mathcal{L}$ and sparsity regularizer R which are tied together by adding a multiple of the norm squared difference $\frac{\eta}{2}\|x-w\|^2$ to the objective. Then, fixing the variables w dedicated to the nonsmooth regularizer R, the smooth function $\mathcal{L}(x) + \frac{\eta}{2}\|x-w\|^2$ is globally optimized over the variables x to obtain an optimal value

function $u_\eta(w)$. We then show that $u_\eta(w)$ is smooth. This opens the door to the application of the PGD algorithm to minimizing $u_\eta + R$ where now u_η contains global variational information on the likelihood function $\mathcal{L}$. The main obstacle in the application of this approach is the evaluation of u_η and its gradient. In Section 3.2 we present a method for overcoming this difficulty using variable metric techniques and interior point technology.

All new methods are implemented in an open-source library called `pysr3`, which fills a gap for python mixed-models selection tools in `Python` (Buscemi and Plaia (2019), Table 3). Our algorithms are 1-2 orders of magnitude faster than available LASSO-based libraries for mixed effects selection in `R`, see Table 3. `pysr3` enables a standardized comparison of different methods in the LME setting, and makes both the PGD framework and its SR3 extension available to practitioners working with LME models.

We begin in Section 2 by giving a precise description of the LME model and set the notation for the remainder of the paper. This is followed by a brief discussion of prior work on LMEs. In Section 3 we present our algorithms for the LME model starting with the naive PGD algorithm. This is followed by a description of the variable splitting technique used in Zheng et al. (2019) to incorporate global variation information on the likelihood function into the direction finding subproblem for the PGD algorithm. Next we tackle the problem of how to approximate the resulting optimal value function u_η and its gradient where $\eta > 0$ is the decoupling parameter. As noted, this is done using variable metric techniques and and interior point technology. We conclude Section 3 with a discussion of the MSR3 and MSR3-fast algorithms. In Section 4 we discuss how the underlying algorithmic parameters are set and test the algorithm on both simulated problems and a problem with real data. The paper is concluded in Section 5 with a brief discussion of the contributions.

2 Linear Mixed-Effects Models: Notation and Fundamentals

Mixed-effect models describe the relationship between an outcome variable and its predictors when the observations are grouped, for example in studies or clusters. To set the notation, consider m groups of observations indexed by i, with sizes n_i, and the total number of observations equal to $n = n_1 + n_2 + \cdots + n_m$. For each group, we have design matrices for fixed features $X_i \in \mathbb{R}^{n_i \times p}$, and matrices of random features $Z_i \in \mathbb{R}^{n_i \times q}$, along with vectors of outcomes $Y_i \in \mathbb{R}^{n_i}$. Let $X = [X_1^T, X_2^T, \ldots, X_m^T]^T$ and $Z = [Z_1^T, Z_2^T, \ldots, Z_m^T]^T$. Following Patterson and Thompson (1971); Pinheiro and Bates (2000), we define a Linear Mixed-Effects (LME) model as

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

where $\beta \in \mathbb{R}^p$ is a vector of fixed (mean) covariates, $u_i \in \mathbb{R}^q$ are unobservable random effects assumed to be distributed normally with zero mean and the unknown covariance matrix Γ, and $\mathbb{S}_+^\nu$ and $\mathbb{S}_{++}^\nu$ are the sets of real symmetric $\nu \times \nu$ positive semi-definite and positive definite matrices, respectively. Matrices Z_i encode a wide variety of models, including random intercepts (Z_i are columns of 1's that add u_i to all datapoints from the ith study) and random slopes (Z_i also scale u_i according to the magnitude of a covariate), see e.g. Pinheiro and Bates (2006). In our study, we assume that the observation error covariance matrices Λ_i are given and that the random effects covariance matrix is an unknown diagonal matrix, i.e., $\Gamma = \mathrm{Diag}(\gamma)$, $\gamma \in \mathbb{R}_+^s$. This assumption corresponds to the meta-analysis and meta-regression branch of mixed effects problems, which is the primary focus of our applied collaborations (see e.g. Zheng et al. (2022); Lescinsky et al. (2022); Razo et al. (2022); Stanaway et al. (2022); Dai et al. (2022).) The theoretical developments in this work allow extensions to other types of repeated measure models, but practical implementation requires significant additional effort, and we leave these extensions to future work.

Defining group-specific error terms $\omega_i = Z_i u_i + \varepsilon_i$, we get a compact formulation that recasts (1)

as a correlated noise model:

$$Y_i = X_i\beta + \omega_i, \quad \omega_i \sim \mathcal{N}(0, \Omega_i(\Gamma)), \quad \Omega_i(\Gamma) = Z_i \Gamma Z_i^T + \Lambda_i. \tag{2}$$

For brevity, we refer to $\Omega_i(\Gamma)$ as just Ω_i. The reformulation (2) yields the following marginalized negative log-likelihood function of a linear mixed-effects model (Patterson and Thompson, 1971):

$$\mathcal{L}_{ML}(\beta, \Gamma) := \sum_{i=1}^{m} \frac{1}{2}(y_i - X_i\beta)^T \Omega_i^{-1}(y_i - X_i\beta) + \frac{1}{2}\ln\det\Omega_i. \tag{3}$$

Maximum likelihood estimates for β and Γ are obtained by solving the optimization problem

$$\min_{\beta,\Gamma} \ \mathcal{L}_{ML}(\beta, \Gamma) \quad \text{s.t.} \quad \Gamma \in \mathbb{S}_+^q \,. \tag{4}$$

In the discussion below, make use of basic concepts from Rockafellar and Wets (2009), defined in the Appendix.

The negative log likelihood (4) is nonlinear and nonconvex, and requires an iterative numerical solver. However, it is convex with respect to β, and weakly convex with respect to γ, with a weak convexity constant $\overline{\eta}$ computed in (Aravkin et al., 2022, Section 5.1) . The expected value of the posterior mode β given Γ has the closed form representation

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

By using the simplification $\Gamma = \operatorname{Diag}(\gamma)$, we obtain the problem

$$\min_{\beta\in\mathbb{R}^p, \gamma\in\mathbb{R}_+^q} \ \mathcal{L}(\beta, \gamma) := \mathcal{L}_{ML}(\beta, \operatorname{Diag}(\gamma)) \,. \tag{5}$$

In this setting, when an entry γ_j takes the value 0 the corresponding coordinates of all random effects u_{ij} are identically 0 for all i.

Verification of the existence to solutions to (5) and, more generally, (4) follows from the work of Zheng et al. (2021). Standalone proofs for the existence of minimizers are developed in (Aravkin et al., 2022, Theorem 1), and extended to the presence of regularizers in (Aravkin et al., 2022, Theorem 2).

This paper focuses the case where Γ is diagonal, (often referred to as *the diagonal setup*) and all Λ_i are known (see (5)), following the meta-analysis use-case (Zheng et al., 2021) that is widely employed in epidemiological studies Murray et al. (2020). While the proposed approach can be extended to the non-diagonal case, we leave it for future work, save for a brief discussion in Section 4.

2.1 Prior Work on Feature Selection for Mixed-Effects Models

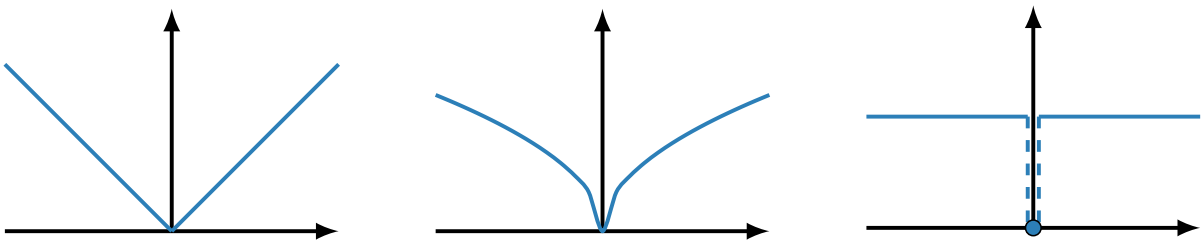
Variable (feature) selection models seeks to select or rank the most important predictors in a dataset in order to get a parsimonious model at a minimal cost to prediction quality. Feature selection may be performed both on β, to find the sparse set of covariates that best explains the mean, and on γ, to find the sparse set of covariates that best accounts for variation between groups. Both types of selection have been studied in the literature, and both are accessible using the methods developed here. If the desired number of coefficients k is given, then the feature selection problem can be formulated as the minimization of a loss function $f(\theta)$ (e.g. the negative log-likelihood) subject to a zero-norm constraint:

$$\min_{\theta} f(\theta) \quad \text{s.t.} \quad \|\theta\|_0 \leq k \tag{6}$$

where $\|\theta\|_0$ denotes the number of nonzero entries in θ, see panel (c) of Figure 2.

The constraint in (6) is combinatorial, and a common workaround is to relax it to a one-norm constraint, with $\|\theta\|_1$ equal to the sum of absolute values of the entries of θ. The best-known example of this approach is the least absolute square shrinkage operator (LASSO) studied by Tibshirani (1996b) for linear regression, see panel (a) of Figure 2.

Feature selection for LMEs is more difficult than for linear regression models. In linear regression



(a) ℓ_1 (b) $\ell_{1/2}$ (c) ℓ_0

Figure 2: Common convex and non-convex regularizers used for feature selection.

the observations are independent, whereas in mixed-effects setup they are generally correlated. In addition, LMEs have both mean effect variables β as well as random variance variables Γ. The shrinkage operator approach for linear regression (Tibshirani, 1996b) was first adapted to the problem of feature selection for the fixed effects in mixed-effect models by Lan (2006). The removal of a random effect from the model requires the elimination of an entire row and column from Γ. To make the problem more tractable, Chen and Dunson (2003) reparamtrized Γ through a modified Cholesky decomposition $\Gamma(D, L) := DLL^TD$, where D is a diagonal matrix and L is a lower-triangular matrix with ones on the main diagonal, and focused on selecting elements of D. Based on this idea, Bondell et al. (2010) extended the Adaptive LASSO regularizer (Lan (2006); Xu et al. (2015)) to mixed-effects setting using the objective $\mathcal{L}(\beta, \Gamma(D, L)) + \lambda \left(\sum_{i=1}^{p} \left| \frac{\beta_i}{\hat{\beta}_i} \right| + \sum_{j=1}^{q} \frac{D_{ii}}{\hat{D}_{ii}} \right)$, where $\hat{\beta}$ and $\hat{D}$ are the solution of a non-penalized maximum likelihood problem and λ is a tuning parameter for the weighted regularizer and is called the regularization parameter. Ibrahim et al. (2011b) use a similar approach, penalizing non-zero elements Γ_{ij} directly. Other methods that use Adaptive LASSO for simultaneous selection of fixed and random effects are Lin et al. (2013a); Fan et al. (2014); Pan and Shang (2018). Adaptive LASSO is available to practitioners via `R` packages `glmmLasso`[3] (Groll and Tutz (2014)) and `lmmLasso`[4](Schelldorfer et al. (2011)).

A popular nonconvex regularizer used for feature selection is smoothed clipped absolute deviation (SCAD) Fan and Li (2001). The adaptation of the SCAD penalty to select both fixed and random features in linear mixed models was developed by Fan and Li (2012). SCAD was also used by

[3] https://rdrr.io/cran/glmmLasso/man/glmmLasso.html
[4] https://rdrr.io/cran/lmmlasso/

Chen et al. (2015) for selecting fixed effects and establishing the existence of random effects in ANOVA-type models. Finally, Ghosh and Thoresen (2018) studied SCAD regularization for selecting mean effects in high-dimensional genomics problems.

To better compare methods, we need to consider the tuning of the regularization parameter λ. The output of a shrinkage model critically depends on the tuning parameter λ. The entire range of λ values is captured by the notion of a "λ-path in the model space", with the best parameter and the final model chosen using information criteria. According to Müller et al. (2013), the most widely used information criterion is the marginal AIC criterion (Vaida and Blanchard (2005)), $AIC := 2\,\mathcal{L}(\hat{\theta}) + 2\alpha_n(p+q)$, where $\hat{\theta}$ includes all the estimated parameters (β, Γ), and $\alpha_n := n(n-p-q-1)$ for the finite sample case (Sugiura (1978)). Alternatively, LASSO-type methods (Bondell et al. (2010); Ibrahim et al. (2011b)) use a BIC-type information criterion, $BIC := 2\,\mathcal{L}(\hat{\theta}) + \log(n)(p+q)$. BIC performs well in practice, but does not have theoretical guarantees (Schelldorfer et al. (2011)).

3 Algorithms for Feature Selection

We approach feature selection by adding a regularizer to model (5):

$$\min_{x} \mathcal{L}(x) + R(x) + \delta_{\mathcal{C}}(x), \tag{7}$$

where $x = (\beta, \gamma)$, $\mathcal{C} := \mathbb{R}^p \times \mathbb{R}^q_+$, $R : \mathbb{R}^P \times \mathbb{R}^q_+ \to \overline{\mathbb{R}}_+ := \mathbb{R}_+ \cup \{+\infty\}$ is a lower semi-continuous (lsc) regularization term, and $\delta_{\mathcal{C}}$ is the convex indicator function, where $\delta_{\mathcal{C}}(x) := 0$ for $x \in \mathcal{C}$ and $+\infty$ otherwise. By (Aravkin et al., 2022, Theorem 2), solutions to (7) always exist when R has compact lower level sets. The most common regularizers are separable taking the form

$$R(x) = \sum_{i=1}^{p} r_i(x_i), \tag{8}$$

with typical choices for the component functions r_i given in Table 1.

3.1 Variable Selection via Proximal Gradient Descent

Since $\mathcal{L}$ is differentiable on its domain and proximal operator for $\alpha R + \delta_{\mathcal{C}}$ is computationally tractable, the Proximal Gradient Descent (PGD) Algorithm (e.g. see Beck (2017)) offers a simple numerical strategy for estimating first-order stationary points for (7). The proximal operator for $\alpha R + \delta_{\mathcal{C}}$ is defined as the mapping $\text{prox}_{\alpha R+\delta_{\mathcal{C}}}(z) := \text{argmin}_{y\in\mathcal{C}}\ R(y) + \frac{1}{2\alpha}\|y - z\|_2^2$, and the PGD iteration is given by $x^+ = \text{prox}_{\alpha R+\delta_{\mathcal{C}}}(x - \alpha\nabla\mathcal{L}(x))$, where α is a stepsize. When $R(x)$ has the form given in (8), we have $\text{prox}_R(z) = (\text{prox}_r(z_1), \ldots, \text{prox}_r(z_q))$. Table 1 provides closed form expressions for the proximal operators of commonly used regularizers. For all of these cases, the following theorem gives closed form expressions for $\text{prox}_{\alpha R+\delta_{\mathcal{C}}}(z)$.

Regularizer	$r(x)$, $x \in \mathbb{R}$	$\text{prox}_{\alpha r}(z)$
LASSO (ℓ_1)	$\lvert x\rvert$	$\text{sign}(z)(\lvert z\rvert - \alpha)_+$
A-LASSO	$\bar{w}\lvert x\rvert$, $\bar{w} \geq 0$	$\text{sign}(z)(\lvert z\rvert - \alpha\bar{w})_+$
SCAD	$\begin{cases} \sigma\lvert x\rvert, & \lvert x\rvert \leq \sigma \\ \frac{-x^2+2\rho\sigma x-\sigma^2}{2(\rho-1)}, & \sigma < \lvert x\rvert < \rho\sigma \\ \frac{\sigma^2(\rho+1)}{2}, & \lvert x\rvert > \rho\sigma \end{cases}$	$\begin{cases} \text{sign}(z)(\lvert z\rvert - \sigma\alpha)_+, & \lvert z\rvert \leq \sigma(1+\alpha) \\ \frac{(\rho-1)z-\text{sign}(z)\rho\sigma\alpha}{\rho-1-\alpha}, & \sigma(1+\alpha) < \lvert z\rvert \leq \rho\sigma \\ z, & \lvert z\rvert > \max(\rho, 1+\alpha)\sigma \end{cases}$
$\delta_{\|x\|_0\leq k}$ (ℓ_0 ball)	$\begin{cases} 0, & \#\{\lvert x_i\rvert \neq 0\} \leq k \\ \infty, & \text{otherwise} \end{cases}$	keep k largest $\lvert x_i\rvert$, set the rest to 0

Table 1: Proximal operators for commonly used sparsity-promoting regularizers.

Theorem 1 (prox for bounded γ)**.** *We consider modified regularizers $r(\gamma)$ from the Table 1 that include an additional constraint on γ of the form $0 \leq \gamma \leq \bar{\gamma}$, for $\bar{\gamma} \in [0, +\infty]$. We have the following results.*

1. *For SCAD, we have for all i that* $\operatorname{prox}_{(\alpha r+\delta_{[0,\bar{\gamma}]})}(\gamma_i) = \begin{cases} \operatorname{prox}_{\alpha r}(\gamma_i), & 0 \le \gamma_i < \bar{\gamma} \\ \bar{\gamma}, & \gamma_i \ge \bar{\gamma} \\ 0, & \textit{otherwise} \end{cases}$.

2. *For LASSO, A-LASSO we have for all i that* $\operatorname{prox}_{(\alpha r+\delta_{[0,\bar{\gamma}]})}(\gamma_i) = \begin{cases} \operatorname{prox}_{\alpha r}(\gamma_i), & 0 \le \gamma_i < \bar{\gamma} + \alpha \\ \bar{\gamma}, & \gamma_i \ge \bar{\gamma} + \alpha \\ 0, & \textit{otherwise} \end{cases}$.

3. *For $R(\cdot) = \delta_{\operatorname{lev}_{\|\cdot\|_0}(k)}$ the $\operatorname{prox}_{\alpha R+\delta_{\mathcal{C}}}(\gamma)$ can be evaluated by taking k largest coordinates of γ such that $0 \le \gamma_i \le \bar{\gamma}$, and setting the remainder to 0.*

The proof of the Theorem 1 is provided in Appendix B.2. The PGD algorithm is detailed in Algorithm 1. The algorithm's step-size α depends on the Lipschitz constant; an upper-bound is given in Appendix B.3. In practice, α is computed using a line-search, since the available estimate for L is very conservative.

$x = x_0$, $\alpha < \frac{1}{L}$, where $\mathcal{L}$ is L-Lipschitz
while not converged **do**
  $x^+ = \operatorname{prox}_{\alpha R+\delta_{\mathcal{C}}}(x - \alpha \nabla \mathcal{L}(x))$;
end

Algorithm 1: Proximal Gradient Descent for Linear Mixed-Effect Models

The main advantages of Algorithm 1 are its simplicity and flexibility. The main loop needs only the gradient and prox operator, and the structure of the algorithm is independent of the choice of R. Algorithm 1 locates first-order stationary points under weak assumptions, in particular neither the objective nor the regularizer need be convex (Beck, 2017; Attouch et al., 2013).

3.2 Variable Selection via MSR3

To develop an approach that is both more efficient and accurate, we extend the SR3 regularization of Zheng et al. (2019) to LMEs. We call the extension MSR3, since we are focusing on mixed effects

models. Starting with the regularized likelihood (7) we introduce auxiliary parameters designed to discover the fixed and random features:

$$\min_{x,w} \mathcal{L}(x) + R(w) + \delta_{\mathcal{C}}(x) + \kappa_\eta(x - w), \tag{9}$$

where κ_η penalizes deviations between $x = (\beta, \gamma)$ and $w = (\hat{\beta}, \hat{\gamma})$, and also guarantees that the objective is convex with respect to the γ components of x for sufficiently large η: $\kappa_\eta(x - w) = \frac{\eta}{2}\|x - w\|^2 = \frac{\eta}{2}\|\beta - \hat{\beta}\|^2 + \frac{\eta}{2}\|\gamma - \hat{\gamma}\|^2$ with $\eta \geq \bar{\eta}$ where $\bar{\eta}$ is the weak convexity constant computed in (Aravkin et al., 2022, Section 5.1). As $\eta \uparrow \infty$, the extended objective (9) converges in an epigraphical sense to the original objective (7). However, feature selection accuracy does not require this continuation, indeed, we show that a fixed modest value such as $\eta = 1$ can be used (Zheng et al., 2019).

To understand the algorithm and logic behind the objective (9), we define an optimal value function $u_\eta(w)$ and the solution set $S_\eta(w)$:

$$\begin{aligned} u_\eta(w) &= \min_x \ \mathcal{L}(x) + \delta_{\mathcal{C}}(x) + \kappa_\eta(x - w) \\ S_\eta(w) &= \operatorname*{argmin}_x \mathcal{L}(x) + \delta_{\mathcal{C}}(x) + \kappa_\eta(x - w). \end{aligned} \tag{10}$$

Substituting (10) into (9) transforms (9) into

$$\min_w u_\eta(w) + R(w) \tag{11}$$

Here we have transformed the original regularized likelihood (7) through relaxation and partial minimization to obtain an equivalent problem (11) for w with the same regularizer. The value function u_η encapsulates global variational information on the function $\mathcal{L}(x) + \delta_{\mathcal{C}}(x)$ relative to w.

In the case of linear regression, the function u_η has a closed form solution Zheng et al. (2019). However, in both the linear regression context of Zheng et al. (2019) and in the LME context studied here, we need only compute $S_\eta(w)$ in order to optimize (11). Indeed, in (Aravkin et al.,

2022, Section 5) it is shown that there exists a computable $\bar{\eta} > 0$, which we have called the weak convexity constant, such that $\mathcal{L} + \delta_{\mathcal{C}} + \kappa_\eta(\cdot - w)$ is strongly convex for all $\eta > \bar{\eta}$ regardless of the choice of w. This allows us to show that u_η is well-defined, differentiable, and Lipschitz continuous, with

$$\nabla u_\eta(w) = \nabla_w k_\eta(x - w)|_{x=S_\eta(w)} = \eta(w - S_\eta(w)). \tag{12}$$

Our empirical studies indicate that (11) has advantages over (7) from an optimization perspective since u_η typically has nearly spherical level-sets while keeping the position of minima close to those of $\mathcal{L}(x)$. This effect is extensively studied and validated for a quadratic loss function in the original work of Zheng et al. (2019). In the center panel of Figure 1, we plot the level-sets of $\mathcal{L}(x) + \|x\|_1$ (left column) and $u_\eta + \|\cdot\|_1$ (right column) for the same mixed-effect problem. The more spherical geometry of the latter allows the Algorithm 2 (described below) to converge in 21 iterations, whereas Algorithm 1 takes 1284 iterations. The difference is most pronounced when the minimum sits on the boundary of the feasible set, which is always the case for the variable selection problems with sparse support.

We apply PGD to optimize the regularized value function u_η which yields the iteration

$$w^+ = \text{prox}_{\alpha^{-1}R}(w - \alpha \nabla u_\eta(w)) \tag{13}$$

The results in Aravkin et al. (2022) show that all components of the iteration (13) are well-defined. The equivalence of Algorithm 2 and (13) is established in the following lemma, which extends the relationship studied by Zheng et al. (2019) to the case of $x = (\beta, \gamma)$.

Lemma 2 (Equivalence of Algorithms)**.** *Algorithm 2 is equivalent to* (13)*.*

Proof. Substituting (12) into (13), we see that the iteration (13) is equivalent to the alternating minimization scheme outlined in the Algorithm 2. □

1 $w = w_0$

2 **while** not converged **do**

3 $\quad x^+ = \arg\min_x \mathcal{L}(x) + \delta_{\mathcal{C}}(x) + \kappa_\eta(x - w)$

4 $\quad w^+ = \text{prox}_{\alpha^{-1}R}(x^+)$

5 **end**

Algorithm 2: Proximal Gradient Descent for Value Function

In (Aravkin et al., 2022, Theorem 6), it is shown that for any sequence $\eta_k \uparrow \infty$ the associated optimal solutions (x^k, w^k) to (11) satisfy $\mathcal{L}(x^k)+R(w^k) \uparrow \inf_{x\in\mathbb{R}^p\times\mathbb{R}^q_+} \mathcal{L}(x)+R(x)$ with $\|x^k - w^k\| \to 0$. In particular, every cluster point of the sequences $\{x^k\}$ and $\{w^k\}$ are solutions to (5), where such cluster points exist whenever the function R is coercive, i.e. $\lim_{\|x\|\uparrow\infty} R(x) = +\infty$. Just how close w^k is to a solution to (5) remains an open question, however, our numerical studies in Section 4 show that η can be chosen surprisingly small. Indeed, we typically take $\eta = 1$.

In the linear regression setting of Zheng et al. (2019), Algorithm 2 can be implemented exactly. In the nonlinear case, evaluating x^+ requires an iterative algorithm. For this we use an interior point method which replaces the indicator function $\delta_{\mathcal{C}}$ by a smooth log-barrier term. This allows us to approximate both u_η and its gradient where the degree of the approximation is controlled by the convergence criteria of the interior point algorithm.

An Interior Point Method for Approximating u_η. In order to solve for the x^+ update in line 2 of Algorithm 2, we must optimize a convex loss with linear inequality constraints, that is, for a fixed $w = (\hat{\beta}, \hat{\gamma})$, we need to solve

$$\min_{\beta,\gamma} \mathcal{L}(\beta, \gamma) + \kappa_\eta(\beta - \hat{\beta}, \gamma - \hat{\gamma}) \quad \text{s.t.} \quad 0 \le \gamma. \tag{14}$$

This problem is well suited for an interior point approach (Kojima et al., 1991; Nesterov and Nemirovskii, 1994; Wright, 1997; Vanderbei and Shanno, 1999). First, the constraint $0 \le \gamma$ is

relaxed using a log-barrier penalty, obtaining a minimization problem for a relaxed objective $\mathcal{L}_{\mu,\eta}$:

$$\min_{\beta,\gamma}\left\{\mathcal{L}_{\mu,\eta}(\beta,\gamma) := \mathcal{L}(\beta,\gamma) + \kappa_\eta(\beta-\hat{\beta},\gamma-\hat{\gamma}) - \mu\sum_{i=1}^{q}\ln(\gamma_i)\right\}. \tag{15}$$

Here the log-barrier penalty approximates the indicator function to the positive orthant as μ decreases; indeed, the function $\gamma \mapsto \mu\ln(\gamma)$ epi-converges to the indicator function $\delta_{\mathbb{R}^n_+}(\gamma)$ as $\mu \downarrow 0$ (Rockafellar and Wets (2009)). The penalty (homotopy) parameter μ is progressively decreased to 0 as the algorithm proceeds as described below. The existence of solutions for the problem (15) for any positive μ is shown in (Aravkin et al., 2022, Theorem 5), and the convergence of solutions to the MSR3 solution as $\mu \downarrow 0$ is shown in (Aravkin et al., 2022, Theorem 7). Finally, (Aravkin et al., 2022, Theorem 6) shows that the MSR3 relaxation is consistent with respect to the barrier, so that as the MSR3 parameter $\eta \uparrow \infty$, limit points of global solutions to the former are global solutions to the latter. However, in the applications considered here, the empirical studies in Sections 3.3 and 4.2 indicate that one does not need to make η particularly large in order to accurately identify the correct sparsity pattern.

For $\gamma > 0$, the necessary optimality conditions for $\mathcal{L}_{\mu,\eta}$ in γ give us the relation

$$\nabla_\gamma \mathcal{L}_{\mu,\eta}(\beta,\gamma) = \nabla_\gamma \mathcal{L}(\beta,\gamma) + \eta(\gamma-\hat{\gamma}) - \mu\mathrm{Diag}\,(\gamma)^{-1}\,\mathbf{1} = 0, \tag{16}$$

where $\mathbf{1}$ is the vector of all ones of the appropriate dimension. By setting $v = \nabla_\gamma \mathcal{L}_{\mu,\eta}(\beta,\gamma)+\eta(\gamma-\hat{\gamma})$, we can rewrite this equation as

$$v \odot \gamma - \mu\mathbf{1} = 0, \tag{17}$$

where $\mathbf{1}$ is the vector of all ones of the appropriate dimension and "$\odot$" denotes the Hadamard (or simply element-wise) product. The complete set of optimality conditions for (15) can now be

written as

$$G_{\mu,\eta}(v,\beta,\gamma) := \begin{bmatrix} v \odot \gamma - \mu \mathbf{1} \\ \nabla_\beta \mathcal{L}(\beta,\gamma) + \eta(\beta - \hat{\beta}) \\ \nabla_\gamma \mathcal{L}(\beta,\gamma) + \eta(\gamma - \hat{\gamma}) - v \end{bmatrix} = 0. \tag{18}$$

We then apply Newton's method to (18), that is, in each iteration the search direction $[\Delta v, \Delta\beta, \Delta\gamma]$ solves the linear system

$$\nabla G_{\mu,\eta}(v,\beta,\gamma) \begin{bmatrix} \Delta v \\ \Delta\beta \\ \Delta\gamma \end{bmatrix} = -G_{\mu,\eta}(v,\beta,\gamma), \quad \nabla G_{\mu,\eta}(v,\beta,\gamma) = \begin{bmatrix} \mathrm{Diag}\,(\gamma) & 0 & \mathrm{Diag}\,(v) \\ 0 & \nabla^2_{\beta\beta}\mathcal{L} + \eta I & \nabla^2_{\beta\gamma}\mathcal{L} \\ -I & \nabla^2_{\gamma\beta}\mathcal{L} & \nabla^2_{\gamma\gamma}\mathcal{L} + (\eta + \overline{\lambda})I \end{bmatrix}$$

and we have used the fact that $v \odot \gamma = \mathrm{Diag}\,(v)\,\gamma = \mathrm{Diag}\,(\gamma)\,v$. The exact formulae for the derivatives of $\mathcal{L}$ are provided in the Appendix B.1.

The general structure of the algorithm is as follows. Given a search direction $[\Delta v^{(k)}, \Delta\beta^{(k)}, \Delta\gamma^{(k)}]$, choose a step of size $\alpha_k > 0$ so that the update

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

satisfies the conditions

$$\begin{aligned} \textit{Positivity:} \qquad & \gamma^{(k+1)} > 0, \ v^{(k+1)} > 0 \\ \textit{Sufficient Descent:} \qquad & \|G_{\eta,\mu}(v^{(k+1)}, \beta^{(k+1)}, \gamma^{(k+1)})\| \leq 0.99 \|G_{\eta,\mu}(v^{(k)}, \beta^{(k)}, \gamma^{(k)})\|, \end{aligned}$$

where the parameter 0.99 is used to bias toward the acceptance of a full Newton step. At each iteration the relaxation parameter μ is updated by the formula $\mu^{(k+1)} = {v^{(k)}}^T \gamma^{(k)}/q$, where ${v^{(k)}}^T \gamma^{(k)}$

is the duality gap at iteration k. The algorithm terminates when the criteria

$$\|G_{\mu,\eta}(v^{(k+1)}, \beta^{(k+1)}, \gamma^{(k+1)})\| \leq \texttt{tol}$$
$$\mu \leq \texttt{tol}$$

are both satisfied, so the interior point problem is nearly stationary, and closely approximates the original problem (14). MSR3 is summarized in Algorithm 3, which approximates Algorithm 2 as the tolerance goes to 0. In the numerical experiments, we use $\texttt{tol} = 10^{-5}$, and accuracy does not change as the tolerance parameter decreases.

$w = w_0$
while not converged **do**
  x^+ satisfies $\|G_{\mu,\eta}(v^+, x^+)\| \leq \texttt{tol},\ \mu \leq \texttt{tol}$
  $w^+ = \text{prox}_{\alpha^{-1}R}(x^+)$
end

Algorithm 3: MSR3

Positive Approximation of the Hessian For many datasets the weak convexity constant $\overline{\eta}$ can be extremely large and difficult to compute. However, if η is too small $\nabla^2_{\gamma\gamma}\,\mathcal{L}_{\mu,\eta}(\beta,\gamma)$ is negative-(semi)definite. Negative definite Hessians can hamper the convergence of second-order methods (e.g., see Nocedal and Wright (2006)). Therefore, one must take care in selecting η. For this, we recall from (Aravkin et al., 2022, Lemma 3) that

$$\nabla^2 \mathcal{L}\,(\beta,\gamma) = \sum_{i=1}^{m} S_i^T \begin{bmatrix} X_i^T \\ -Z_i^T \end{bmatrix} \Omega_i(\gamma)^{-1} \begin{bmatrix} X_i & -Z_i \end{bmatrix} S_i - \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{2}(Z_i^T \Omega_i(\gamma)^{-1} Z_i)^{\circ 2} \end{bmatrix}.$$

This implies that negative eigenvalues for the Hessian must arise from the Hessian with respect to γ, $\nabla^2_{\gamma\gamma}\,\mathcal{L}(\beta,\gamma)$, and more specifically, the term $(Z_i^T\Omega_i(\gamma)^{-1}Z_i)^{\circ 2}$. A positive semidefinite approximation to the Hessian is obtained by simply dropping this term.

3.3 Relaxation and Efficient Algorithms: MSR3 and MSR3-Fast

While algorithm (2) is modular, it requires solving a nonlinear optimization problem in $x = (\beta, \gamma)$ for each single update of $w = (\hat{\beta}, \hat{\gamma})$. To make the implementation as efficient as possible, we designed a more balanced updating scheme, that alternates Newton iterations as described in the interior point algorithm with w updates. We update w whenever we are sufficiently close to the 'central path' in the interior point method, a condition that can be checked rigorously using optimality conditions. This scheme is detailed in Algorithm 4.

In designing Algorithm 4, we chose a particular central path parameter, $\tau = 0.5$ in line 8, that controls how far the interior point method needs to proceed before we take a proximal gradient step. We explored the effect of this parameter on performance and timing in Appendix C, and found that it did not have any effect on either for values between 0.1 and 0.9. MSR3-fast was competitive with respect to time compared to PGD and PGD with line search (also as reported in Appendix C) for problems up to 1000 features.

1 `progress` $\leftarrow$ **True**; `iter = 0`;
2 $\beta^+, \tilde{\beta}^+ \leftarrow \beta_0$; $\gamma^+, \tilde{\gamma}^+ \leftarrow \gamma_0$; $v^+ \leftarrow 1 \in \mathbb{R}^q$; $\mu \leftarrow \frac{{v^+}^T \gamma^+}{10q}$
3 **while** `iter` $<$ `max_iter` and $\|G_{\eta,\mu}(\beta^+, \gamma^+, v^+)\| >$ `tol` and `progress` **do**
4 $\beta \leftarrow \beta^+$; $\gamma \leftarrow \gamma^+$; $\tilde{\beta} \leftarrow \tilde{\beta}^+$; $\tilde{\gamma} \leftarrow \tilde{\gamma}^+$
5 $[dv, d\beta, d\gamma] \leftarrow \nabla G_{\eta,\mu}((\beta, \gamma, v), (\tilde{\beta}, \tilde{\gamma}))^{-1} G_{\eta,\mu}((\beta, \gamma, v), (\tilde{\beta}, \tilde{\gamma}))$ `// Newton Iteration`
6 $\alpha \leftarrow 0.99 \times \min\left(1, -\frac{\gamma_i}{d\gamma_i}, \forall i : \ d\gamma_i < 0\right)$
7 $\beta^+ \leftarrow \beta + \alpha d\beta$; $\gamma^+ = \gamma + \alpha d\gamma$; $v^+ \leftarrow v + \alpha dv$
8 **if** $\|\gamma^+ \odot v^+ - q^{-1}{\gamma^+}^T v^+ \mathbf{1}\| > 0.5 q^{-1} {v^+}^T \gamma^+$ **then**
9 continue `// Keep doing Newton iterations`
10 **end**
11 **else**
12 $\tilde{\beta}^+ = \text{prox}_{\alpha R}(\beta^+)$; $\tilde{\gamma}^+ = \text{prox}_{\alpha R + \delta_{\mathbb{R}_+}}(\gamma^+)$; $\mu = \frac{1}{10}\frac{{v^+}^T \gamma^+}{q}$ `// Near central path`
13 **end**
14 `progress` $= (\|\beta^+ - \beta\| \geq \text{tol}$ or $\|\gamma^+ - \gamma\| \geq \text{tol}$ or $\|\tilde{\beta}^+ - \tilde{\beta}\| \geq \text{tol}$ or $\|\tilde{\gamma}^+ - \tilde{\gamma}\| \geq \text{tol})$
15 `iter += 1`
16 **end**
17 **return** $\tilde{\beta}^+$, $\tilde{\gamma}^+$

Algorithm 4: MSR3-fast (Optimized Proximal Gradient Descent for the Value function)

Regularizer	Model Metric	PGD	MSR3	MSR3-fast
L0	Accuracy	0.89	**0.92**	**0.92**
	Time	47.47	109.86	**0.36**
L1	Accuracy	0.73	**0.89**	0.88
	Time	43.02	13.74	**0.35**
ALASSO	Accuracy	0.88	**0.91**	**0.91**
	Time	38.68	81.52	**0.45**
SCAD	Accuracy	0.71	**0.92**	**0.92**
	Time	87.24	104.20	**0.45**

Table 2: Comparison of performance of algorithms measured as accuracy of selecting the correct covariates and run-time. The L0 strategy stands out over other standard regularizers. MSR3 improves performance significantly for all regularizers, while MSR3-fast improves convergence speed while preserving the accuracy of MSR3. More detailed results are in the Table 4 of Appendix C.1.

4 Verifications

4.1 MSR3 for Covariate Selection

In this section we compare the feature selection accuracy and the numerical efficiency of Algorithms 1 and 4 when using the LASSO, A-LASSO, SCAD, and L0 sparsity regularizers. We begin by describing how the data is generated for our numerical simulations followed by a description of how the regularization parameter λ and the coupling parameter η were chosen. Our experiments on real data are presented in Section 4.2.

Experimental Setup. The number of fixed effects p and random effects q are set at 20 with $\beta = \gamma = [\frac{1}{2}, \frac{2}{2}, \frac{3}{2}, \ldots, \frac{10}{2}, 0, 0, 0, \ldots, 0]$, i.e. the first 10 covariates are increasingly important and the last 10 covariates are not. The data is generated as

$$
\begin{aligned}
y_i &= X_i\beta + Z_i u_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, 0.3^2 I) \\
X_i &\sim \mathcal{N}(0, I)^p, \quad Z_i = X_i \\
u_i &\sim \mathcal{N}(0, \mathrm{Diag}\,(\gamma)),
\end{aligned}
$$

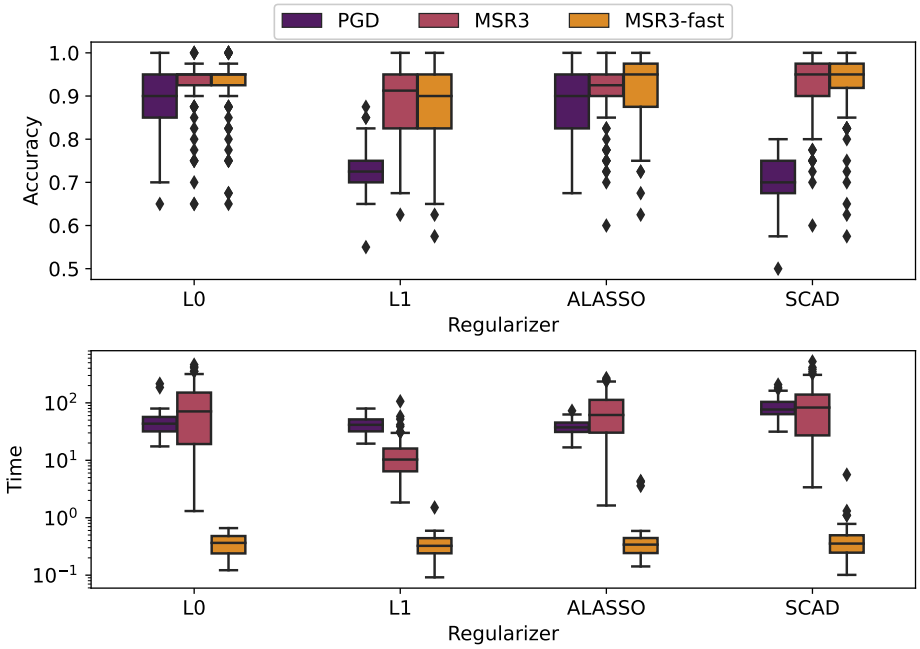


Figure 3: Feature selection accuracy and execution time in seconds for PGD (Algorithm 1), MSR3 (Algorithm 2), and MSR3-fast (Algorithm 4) with various regularizers. MSR3-Fast has the same accuracy as MSR3 and significantly decreases computation time.

with 9 groups of sizes $[10, 15, 4, 8, 3, 5, 18, 9, 6]$. The data generation is repeated 100 times in order to estimate the uncertainty bounds. The smallest non-zero components in the generated signals are just above the level of observation noise.

Parameter Selection. The regularization parameter λ multiplying R and the coupling parameter η restricting the difference between (β, γ) and $(\tilde{\beta}, \tilde{\gamma})$ are chosen to maximize a classic BIC criterion from Jones (2011). We set a log-uniform grid of 20 candidate values for the parameter $\eta \in [10^{-3}, 10^2]$. For each value of η, the BIC is optimized using a golden search in $\lambda \in [0, 10^5]$. The final values of η and λ are chosen to maximize the BIC criterion.

Figure 4 shows the dependence of accuracy on the values of η for the first data set generated in our test set. There are three distinct regions, corresponding to loose, moderate, and tight levels of coupling. When η is small the coupling term does not have sufficient strength and the training does not progress far from the initial point (a fully dense vector $\mathbf{1}$ in this case). When the coupling is tight, the level-sets and minimizers are closer to those of the the original problem. For the values in between, the coupling significantly improves the model's accuracy. These results are consistent with experiments in the sparse linear regression setting Zheng et al. (2019).

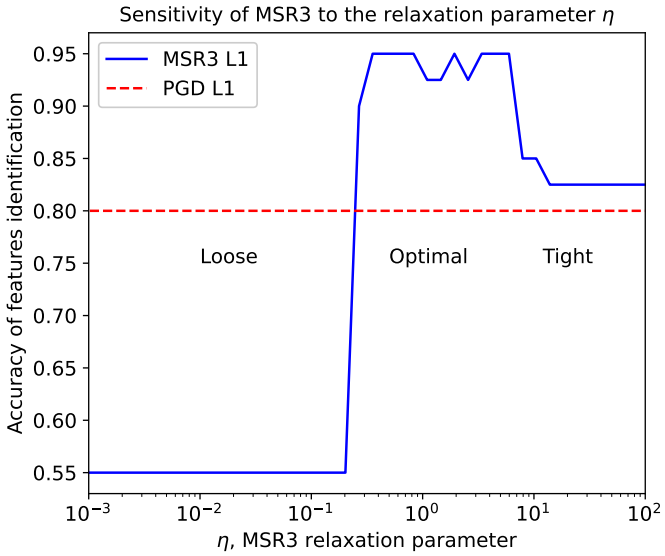


Figure 4: Dependence of model performance on the relaxation η for a sample problem.

Results. The experimental results are presented in the Table 2 and Figure 3. MSR3 improves the selection accuracy of most regularization techniques described in Table 1, showing a near-perfect performance, while converging two orders of magnitude faster in wall-clock time.

Comparison to `glmmLasso` and `lmmLasso`. We used (Buscemi and Plaia, 2019, Table 3) as a reference for feature selection libraries. Of the 17 entries mentioned, the four libraries that successfully ran on our synthetic data described above were packages `glmmLasso`[5] (Groll and Tutz (2014)), `lmmLasso`[6](Schelldorfer et al. (2011)), `fence`[7] (Jiang et al. (2008)) and `PCO` (Lin et al. (2013b)) libraries. `fence` caused a memory overflow on the experimental system during the performance evaluation on the datasets described above. We could not evaluate `PCO` because it did not support datasets where the total number of random effects mq exceeded the total number of observations n. We compare performance of MSR3 (available through the open source `pysr3` library) to the performance of the `R` packages `glmmLasso`[8] (Groll and Tutz (2014)) and `lmmLasso`[9]

[5] https://rdrr.io/cran/glmmLasso/man/glmmLasso.html
[6] https://rdrr.io/cran/lmmlasso/
[7] https://rdrr.io/cran/fence/
[8] https://rdrr.io/cran/glmmLasso/man/glmmLasso.html
[9] https://rdrr.io/cran/lmmlasso/

(Schelldorfer et al. (2011)) which are the functionally closest libraries available online. As of this writing, `glmmlasso` does not allow the user to specify Γ as a diagonal matrix. Since the diagonal specification simplifies the problem, this puts `glmmlasso` package at a disadvantage in our numerical comparison. We evaluate all algorithms' performance on the same set of problems as described above. We tuned the hyperparameters of `glmmLasso` and `lmmLasso` by minimizing the BIC scores provided by the libraries over $\lambda \in [0, 10^5]$. The results are presented in Table 3. Overall, MSR3 executes, on average, 5 times faster in wall-clock time than `glmmLasso` and 60 times faster than `lmmLasso` and shows much higher accuracy in selecting correct fixed and random effects simultaneously. The accuracy of `glmmLasso` is lower relative to the other libraries' scores likely due to its BIC selection criterion choosing dense models. The package `lmmLasso` supports the diagonal specification of Γ, thus allowing a direct comparison with the scores from `pysr3`. `lmmLasso` yields a competitive accuracy of selecting random effects but `lmmLasso` provides dense solutions for fixed effects β for chosen values of λ.

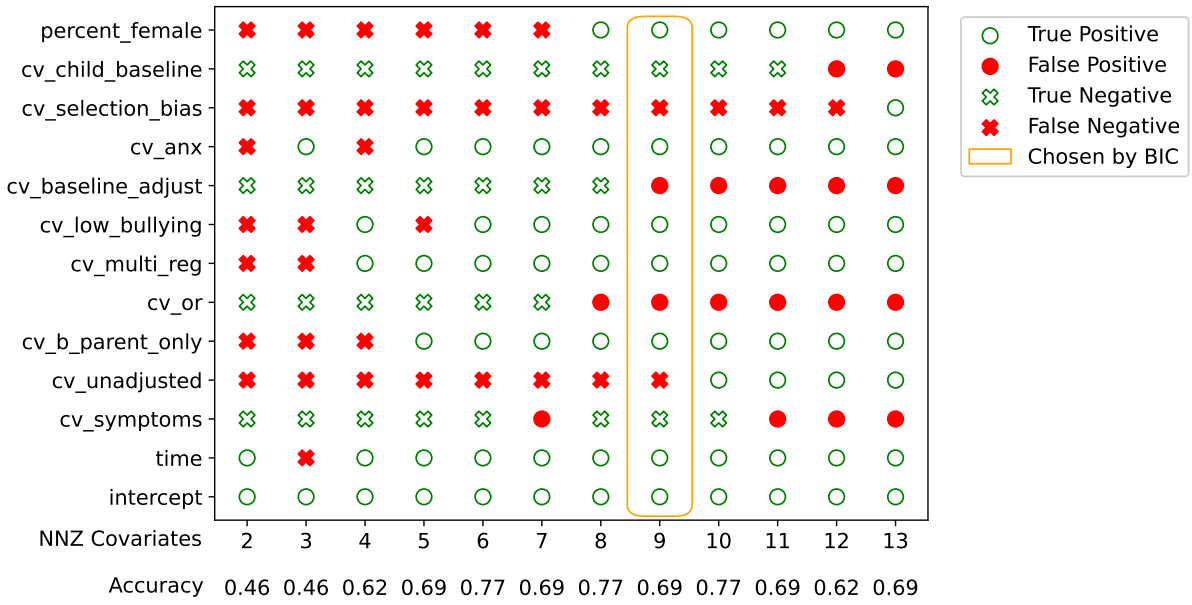
Algorithm	Units (perc. / 100 runs)	MSR3-Fast (ℓ_1)	glmmLasso	lmmLasso
Accuracy	% (5%-95%)	**88 (72-98)**	48 (42-55)	66 (55-73)
FE Accuracy	% (5%-95%)	**86 (64-100)**	52 (40-66)	47 (45-55)
RE Accuracy	% (5%-95%)	**91 (74-100)**	45 (45-45)	84 (55-100)
F1	% (5%-95%)	**89 (73-97)**	63 (60-66)	65 (0-77)
FE F1	% (5%-95%)	**88 (69-100)**	64 (57-70)	57 (0-64)
RE F1	% (5%-95%)	**90 (73-100)**	62 (62-62)	78 (0-100)
Time	sec. (5%-95%)	**0.19 (0.14-0.24)**	1.37 (0.78-1.89)	11.51 (5.35-23.66)
Iterations	num. (5%-95%)	**34 (28-45)**	50 (33-77)	-

Table 3: Comparison of performance of MSR3-Fast for ℓ_1 regularizer vs `glmmLasso`. MSR3-Fast executes 5 times faster in wall time and has higher accuracy of selecting correct covariates.

4.2 Experiments on Real Data

In this section we validate the MSR3-empowered ℓ_0-regularized mixed-effect model ($R(x) = \delta_{\|x\|_0 \le k}$ from Table 1) by using it to identify the most important covariates in real data on relative risk

Figure 5: Validation of Random Feature Selection for Bullying Data from GBD 2020. The panel evaluates each algorithm's choice against expert knowledge. The algorithm picks seven historically significant covariates and two historically insignificant, for the model selected using the BIC criteria. See the Appendix C.3 for covariates description and assessment of significance.



of anxiety and depressive disorders depending on the exposure to bullying in young age[10]. This research has been a part of Global Burden of Diseases (GBD) study for the last several years. The end goal is to estimate the burden through disability adjusted life years (DALYs) (Murray and Acharya, 1997) of major depressive disorder (MDD) and anxiety disorders that are 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

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

Appendix C.3.

The feature selection process is illustrated on Figure 5. Here, the BIC criterion from Jones (2011) was used to select k, which suggests $k = 9$. The selected covariates (`intercept`, `time`, `cv_low_bullying`, `cv_multi_reg` `cv_b_parent_only`, `cv_anx`, `percent_female`) are known as important and were used in the analysis in previous years of GBD. The algorithm also selects `cv_baseline_adjust` and `cv_or`, which were not used before. The `cv_or` variable describes whether the estimate is a relative risk or odds ratio; the selection of this variable suggests a closer look at the data reporting mechanisms across studies. For example, there is an active literature on converting estimates between relative risks and odds ratios Grant (2014); Wang (2013).

4.3 Software Implementation

To ensure reproducibility of this research, all new algorithms have been implemented as a part of the `pysr3`[11] 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[12] of `sklearn` library to minimize learning time.

5 Discussion

In this paper, we developed and implemented a variable selection framework for LMEs based on the PGD algorithm applied to an optimal value function associated with the likelihood function $\mathcal{L}$ which uses second-order information on $\mathcal{L}$. The method has the ability to handle both convex and nonconvex regularizers. Our numerical studies show that the MSR3 relaxation (11) improves the covariates selection accuracy of a wide group of popular sparsity-promoting regularizers. We introduce a modification of MSR3, MSR3-fast, to improve numerical efficiency while maintaining the improved accuracy of MSR3. As in Zheng et al. (2019), we found that SR3 formulations yield more accurate results than the original problem that the SR3 relaxes, likely because the auxiliary

[11] Available at https://github.com/aksholokhov/pysr3
[12] https://scikit-learn.org/stable/developers/develop.html

variables w help to estimate the sparse support. The experiments in this paper show that the phenomenon extends to LME models, and deserves further study.

Since the LME relaxation does not have a closed form, we used an interior method to evaluate the requisite value function. The more efficient version of the algorithm (MSR3-fast) interleaves the interior point iterations with updates of the auxiliary variables, and this method was chosen for the open source library `pysr3`. Numerical experiments on synthetic data showed that the MSR3 approach for variable selection extends regions of hyper-parameter values where the highest accuracy is achieved, making it easier for information criteria to select the best model. The variable selection library for the accelerated method MSR3-fast is much faster than currently available software, and allows the MSR3 approach to be easily applied to a range of regularizers that have computationally efficient prox operators.

The main analytic limitations of the proposed method stem from a lack of an analytical representation of the value function in the MSR3 relaxation for LMEs (11). However, the MSR3 framework (Algorithm 2) incorporates global variational information about the likelihood $\mathcal{L}$ into the PGD algorithm whereas the standard application of the PGD algorithm (Algorithm 1) only uses a local linear approximation to $\mathcal{L}$ at each iteration. This difference reveals itself in both the increased speed and accuracy of the MSR3 approach on this class of problems. In contrast to SR3 in linear regression settings, where the Conjugate Gradient (CG) method can be efficiently used to evaluate the value function (see e.g. Baraldi et al. (2019)), the nonlinear optimization problem required for LMEs is more difficult. Although the use of Hessian information makes each iteration computationally efficient, it limits the size of the problems to which the method can be applied. On the other hand, switching to first-order methods for the inner problem inside the relaxation may be prohibitively slow. A potential path to balance these limitations is to develop efficient upper-bounding models for the value function that can be evaluated more efficiently.

The suggested methodology can be expanded to a wider class of models. In particular, one can extend MSR3 to the setting of non-linear mixed-effect models or generalized linear mixed models, which are known to be challenging setups for covariate selection tasks. Both of these problem

classes require optimizing highly nonlinear objective functions that arise when we consider marginal likelihoods. The SR3 approach may allow new avenues for more efficient strategies, analogous to what was done here for LMEs.

References

Aravkin, A., J. Burke, A. Sholokhov, and P. Zheng (2022). Analysis of relaxation methods for feature selection in mixed effects models. https://arxiv.org/abs/2205.06925.

Attouch, H., J. Bolte, and B. F. Svaiter (2013). Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. Mathematical Programming 137(1), 91–129.

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

Beck, A. (2017). First-Order Methods in Optimization. MOS-SIAM Series on Optimization. SIAM.

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

Buscemi, S. and A. Plaia (2019). Model selection in linear mixed-effect models. AStA Advances in Statistical Analysis.

Chen, F., Z. Li, L. Shi, and L. Zhu (2015). Inference for mixed models of anova type with high-dimensional data. Journal of Multivariate Analysis 133, 382–401.

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

Dai, X., G. F. Gil, M. B. Reitsma, N. S. Ahmad, J. A. Anderson, C. Bisignano, S. Carr, R. Feldman, S. I. Hay, J. He, et al. (2022). Health effects associated with smoking: a burden of proof study. Nature Medicine 28(10), 2045–2055.

DerSimonian, R. and N. Laird (1986). Meta-analysis in clinical trials. Controlled clinical trials 7(3), 177–188.

Fan, J. (1997). Comments on "wavelets in statistics: A review" by a. antoniadis. Journal of the Italian Statistical Society 6(2), 131.

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

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

Fan, Y., G. Qin, and Z. Y. Zhu (2014). Robust variable selection in linear mixed models. Communications in Statistics-Theory and Methods 43(21), 4566–4581.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33(1), 1–22.

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

Grant, R. L. (2014). Converting an odds ratio to a range of plausible relative risks for better communication of research findings. Bmj 348.

Groll, A. and G. Tutz (2014). Variable selection for generalized linear mixed models by l 1-penalized estimation. Statistics and Computing 24(2), 137–154.

Ibrahim, J. G., H. Zhu, R. I. Garcia, and R. Guo (2011a). Fixed and random effects selection in mixed effects models. Biometrics 67(2), 495–503.

Ibrahim, J. G., H. Zhu, R. I. Garcia, and R. Guo (2011b, jun). Fixed and Random Effects Selection in Mixed Effects Models. Biometrics 67(2), 495–503.

Jiang, J., J. S. Rao, Z. Gu, and T. Nguyen (2008). Fence methods for mixed model selection. The Annals of Statistics 36(4), 1669–1692.

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

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

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

Lescinsky, H., A. Afshin, C. Ashbaugh, C. Bisignano, M. Brauer, G. Ferrara, S. I. Hay, J. He, V. Iannucci, L. B. Marczak, et al. (2022). Health effects associated with consumption of unprocessed red meat: a burden of proof study. Nature Medicine 28(10), 2075–2082.

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

Lin, B., Z. Pang, and J. Jiang (2013b). Fixed and random effects selection by reml and pathwise coordinate optimization. Journal of Computational and Graphical Statistics 22(2), 341–355.

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

Murray, C. J. and A. K. Acharya (1997). Understanding dalys. Journal of health economics 16(6), 703–730.

Murray, C. J., A. Y. Aravkin, P. Zheng, C. Abbafati, K. M. Abbas, M. Abbasi-Kangevari, F. Abd-Allah, A. Abdelalim, M. Abdollahi, I. Abdollahpour, et al. (2020). Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. The Lancet 396(10258), 1223–1249.

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

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

Pan, J. and J. Shang (2018). A simultaneous variable selection methodology for linear mixed models. Journal of Statistical Computation and Simulation 88(17), 3323–3337.

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

Pinheiro, J. and D. Bates (2006). Mixed-effects models in S and S-PLUS. Springer science & business media.

Pinheiro, J. C. and D. M. Bates (2000, sep). Mixed-Effects Models in Sand S-PLUS. Journal of the American Statistical Association 96(455), 1135–1136.

Razo, C., C. A. Welgan, C. O. Johnson, S. A. McLaughlin, V. Iannucci, A. Rodgers, N. Wang, K. E. LeGrand, R. J. Sorensen, J. He, et al. (2022). Effects of elevated systolic blood pressure on ischemic heart disease: a burden of proof study. Nature Medicine 28(10), 2056–2065.

Reiner, R. C., R. M. Barber, J. K. Collins, P. Zheng, S. I. Hay, S. S. Lim, C. J. L. Murray, and IHME COVID-19 Forecasting Team (2020). Modeling covid-19 scenarios for the United States. Nature medicine.

Rockafellar, R. T. and R. J.-B. Wets (2009). Variational analysis, Volume 317. Springer Science & Business Media.

Schelldorfer, J., P. Bühlmann, and S. V. DE GEER (2011). Estimation for high-dimensional linear mixed-effects models using l1-penalization. Scandinavian Journal of Statistics 38(2), 197–214.

Stanaway, J. D., A. Afshin, C. Ashbaugh, C. Bisignano, M. Brauer, G. Ferrara, V. Garcia, D. Haile, S. I. Hay, J. He, et al. (2022). Health effects associated with vegetable consumption: a burden of proof study. Nature Medicine 28(10), 2066–2074.

Sugiura, N. (1978, jan). 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, R. (1996a). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58(1), 267–288.

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

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

Vanderbei, R. and D. Shanno (1999). An interior-point algorithm for nonconvex nonlinear programming. Comp. Opt, and Appl. 13, 231–252.

Wang, Z. (2013). Converting odds ratio to relative risk in cohort studies with partial data information. Journal of Statistical Software 55, 1–11.

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

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

Zheng, P., A. Afshin, S. Biryukov, C. Bisignano, M. Brauer, D. Bryazka, K. Burkart, K. M. Cercy, L. Cornaby, X. Dai, et al. (2022). The burden of proof studies: assessing the evidence of risk. Nature Medicine 28(10), 2038–2044.

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

Zheng, P., R. Barber, R. J. Sorensen, C. J. Murray, and A. Y. Aravkin (2021). Trimmed constrained mixed effects models: formulations and algorithms. Journal of Computational and Graphical Statistics, 1–13.

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

Zuur, A., E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith (2009). Mixed effects models and extensions in ecology with R. Springer Science & Business Media.

A Definitions

Definition 1 (Epigraph and level sets)**.** *The epigraph of a function* $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *is defined as* $\operatorname{epi} f = \{(x, \alpha) : f(x) \leq \alpha\}$. *For a given* α*, the* α*-level set of* f *is defined as* $lev_\alpha f = \{x : f(x) \leq \alpha\}$.

Definition 2 (Lower semicontinuity and level-boundedness)**.** *A function* $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *is lower semicontinous (lsc) when* $\operatorname{epi} f$ *is closed, and level-bounded when all level sets* $lev_\alpha f$ *are bounded.*

Definition 3 (Convexity)**.** *A function* $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *is convex when* $\operatorname{epi} f$ *is a convex set. Equivalently,*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \quad \forall x, y \in \operatorname{dom} f,\ \lambda \in (0,1),$$

where $\operatorname{dom} f := \{x \in \mathbb{R}^n \mid f(x) < +\infty\}$.

Definition 4 (Weak convexity)**.** *A function* $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *is* η*-weakly convex* $f(\cdot) + \frac{\eta}{2}\|\cdot\|^2$ *is convex.*

B Additional derivations

B.1 Derivatives of Marginalized Log-likelihood for Linear Mixed Models

For conciseness, let us define the mismatch $\xi_i = Y_i - X_i\beta$. The loss function (3) takes the form

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

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

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

Similarly,

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

Using the symmetry of Ω_i, we have

$$\nabla_\gamma \mathcal{L}(\beta,\gamma) = \sum_{i=1}^{m}\mathrm{diag}\left(({Z_i}^T\Omega_i^{-1}Z_i)\right) - ({Z_i}^T\Omega_i^{-1}\xi_i)^{\circ 2}, \tag{19}$$

where $\circ$ denotes the Hadamard (element-wise) product and $\mathrm{diag}\,(\cdot)$ takes a square matrix to its diagonal. Using the Cholesky decomposition $\Omega_i = L_iL_i^T$ we can calculate (19) using only one triangular matrix inversion:

$$\nabla_\gamma \mathcal{L}(\beta,\gamma) = \sum_{i=1}^{m}\left[\sum_{\text{rows}}\left(L_i^{-1}Z_i\right)^{\circ 2} - [(L_i^{-1}Z_i)^T(L_i^{-1}\xi_i)]^{\circ 2}\right]$$

Notice, that the loss function (3) and the optimal β can also be effectively computed using Cholesky:

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

$$\beta_{k+1} = \underset{\beta}{\operatorname{argmin}}\,\mathcal{L}(\beta,\gamma_k) = \left(\sum_{i=1}^{m} X_i^T\Omega_i^{-1}X_i\right)^{-1}\sum_{i=1}^{m} X_i^T\Omega_i^{-1}y_i =$$

$$= \left(\sum_{i=1}^{m}(L_i^{-1}X_i)^T L_i^{-1}X_i\right)^{-1}\sum_{i=1}^{m}(L_i^{-1}X_i)^T L_i^{-1}y_i.$$

The Hessian w.r.t. γ is derived below:

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

$$= \sum_{i=1}^{m} 2({Z_i^j}^T\Omega_i^{-1}\xi_i)\operatorname{Tr}\left[\Omega_i^{-1}Z_i^j{\xi_i}^T\Omega_i^{-1}Z_i^j{Z_i^j}^T\right] - ({Z_i^j}^T\Omega_i^{-1}Z_i^j)^2$$

$$= \sum_{i=1}^{m} 2({Z_i^j}^T\Omega_i^{-1}\xi_i)({Z_i^j}^T\Omega_i^{-1}Z_i^j)({\xi_i}^T\Omega_i^{-1}Z_i^j) - ({Z_i^j}^T\Omega_i^{-1}Z_i^j)^2,$$

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

$$= \sum_{i=1}^{m} 2({Z_i^j}^T\Omega_i^{-1}\xi_i)\operatorname{Tr}\left[\Omega_i^{-1}Z_i^j{\xi_i}^T\Omega_i^{-1}Z_i^k{Z_i^k}^T\right] - ({Z_i^j}^T\Omega_i^{-1}Z_i^k)^2 =$$

$$= \sum_{i=1}^{m} 2(\xi_i^T\Omega_i^{-1}Z_i^j)({Z_i^j}^T\Omega_i^{-1}Z_i^k)({Z_i^k}^T\Omega_i^{-1}\xi_i) - ({Z_i^j}^T\Omega_i^{-1}Z_i^k)^2, \qquad \text{and so}$$

$$\nabla_\gamma^2\mathcal{L}(\beta,\gamma) = \frac{1}{2}\sum_{i=1}^{m} -(Z_i^T\Omega_i^{-1}Z_i)^{\circ 2} + 2\operatorname{Diag}\left(Z_i^T\Omega_i^{-1}\xi_i\right)(Z_i^T\Omega^{-1}Z_i)\operatorname{Diag}\left(\xi_i^T\Omega^{-1}Z_i\right) =$$

$$= \frac{1}{2}\sum_{i=1}^{m} -(Z_i^T\Omega_i^{-1}Z_i)^{\circ 2} + 2(Z_i^T\Omega_i^{-1}\xi_i)(\xi_i^T\Omega^{-1}Z_i)^T \circ (Z_i^T\Omega^{-1}Z_i).$$

B.2 Derivation of Selected Proximal Operators from Table 1

SCAD For a scalar variable $x \in \mathbb{R}$, SCAD-regularizer is defined as

$$r(x) = \begin{cases} \sigma|x|, & |x| \leq \sigma \\ \frac{-x^2+2\rho\sigma|x|-\sigma^2}{2(\rho-1)}, & \sigma < |x| < \rho\sigma \\ \frac{\sigma^2(\rho+1)}{2}, & |x| > \rho\sigma \end{cases}$$

To evaluate the $\text{prox}_{\alpha r}$ operator we need to solve the following minimization problem:

$$\min_x r(x) + \frac{1}{2\alpha}(x-z)^2$$

For $\alpha = 1$, the solution was derived by Fan (1997). Here we extend it for an arbitrary α. To identify the set of stationary points $\{x^*\}$ of a non-smooth function $f(x)$, we the optimality condition

$$0 \in \partial_x f(x^*)$$

where $\partial_x f(x)$ denotes a sub-differential set of f at the point x. For the prox problem, we get

$$0 \in \frac{1}{\alpha}(x^* - z) + \partial r(x)_{x=x^*}$$

Since $r(x)$ is piece-wise defined the precise value of $\partial r(x)_{x=x^*}$ will depend on x^*:

1. Let $0 < x^* \leq \sigma$, then we have $\partial r(x)_{x=x^*} = \{x^*\}$ and so $x = z - \sigma\alpha, z \in [\sigma\alpha, \sigma + \sigma\alpha]$.

2. Let $-\sigma\alpha \leq x^* < 0$, then we have $\partial r(x)_{x=x^*} = \{-x^*\}$ and so $x = z + \sigma\alpha, z \in [-\sigma - \sigma\alpha, -\sigma\alpha]$.

3. Let $x^* = 0$, then $\partial r(x)_{x=x^*} = [-1, 1]$, which yields $\frac{1}{\alpha}(x^* - z) \in -\sigma[-1, 1] \Rightarrow z \in [-\sigma\alpha, \sigma\alpha]$.

4. Let $\sigma < x^* < \rho\sigma$, then $r(x)_{x=x^*} = \frac{-{x^*}^2+2\rho\sigma x^*-\sigma^2}{2(\rho-1)}$, which gives us $\frac{1}{\alpha}(x^* - z) = \frac{x^*-\rho\sigma}{\rho-1}$. To ensure that the stationary point is indeed a minimizer, we need to ensure that $\frac{1}{\alpha} - \frac{1}{\rho-1} > 0 \Rightarrow \alpha < \rho - 1$. Rearranging the terms we get $x^* = \frac{(\rho-1)z-\lambda\rho\sigma}{\rho-1-\alpha} \Rightarrow z \in [\sigma + \alpha\sigma, \rho\sigma]$.

5. Let $-\rho\sigma < x^* < -\sigma$, then, similarly to the previous case, we get $\frac{1}{\alpha}(x^* - z) = \frac{x^*+\rho\sigma}{\rho-1}$. Rearranging the terms to express x in terms of z we get: $x^* = \frac{(\rho-1)z+\lambda\rho\sigma}{\rho-1-\alpha} \Rightarrow z \in [-\sigma-\alpha\sigma, -\sigma]$.

6. Finally, when $|x^*| \geq \sigma\rho$ we have $\partial r(x)_{x=x^*} = \{0\}$ and so $x^* = z, |z| \geq \sigma\rho$. Bundling all six cases together, we have

$$\text{prox}_{\alpha r}(z) = \begin{cases} \text{sign}(z)(|z| - \sigma\alpha)_+, & |z| \leq \sigma(1+\alpha) \\ \frac{(\rho-1)z - \text{sign}(z)\rho\sigma\alpha}{\rho-1-\alpha}, & \sigma(1+\alpha) < |z| \leq \max(\rho, 1+\alpha)\sigma \\ z, & |z| > \max(\rho, 1+\alpha)\sigma \end{cases} \tag{20}$$

The middle branch is active only when $\rho > 1 + \alpha$. One special case of this is when $\alpha = 1$, and then (20) recovers the classic result by Fan and Li (2001).

To get $\text{prox}_{\alpha r+\delta_{\mathbb{R}_+}}(z)$ from $\text{prox}_{\alpha r}(z)$ we only need to notice that (1) the minimizer x^* of

$$\min_x r(x) + \delta_{\mathbb{R}_+} + \frac{1}{\alpha}(x - z)^2$$

can never be negative, and that (2) when the minimizer x^* is exactly zero we get:

$$\frac{1}{\alpha}(x^* - z) \in -\partial(r(x)|_{x=x^*} + \delta_{\mathbb{R}_+}(x)|_{x=x^*}) \quad \Rightarrow \quad z \in [-\infty, \sigma\alpha]$$

A-LASSO A-LASSO regularizer is defined as $r(x) = w|x|$ where $w = 1/|\hat{x}|$ with $\hat{x}$ the solution of a non-regularized problem (Zou (2006)). The derivation of the proximal operator of A-LASSO nearly matches the steps 1, 2, and 3 that of SCAD above. We wish to evaluate $\min_x w|x| + \frac{1}{2\alpha}(x-z)^2$ as a function of z. The sub-differential optimality criterion yields $0 \in \frac{1}{\alpha}(x^* - z) + w\partial|x|$.

1. Let $0 < x^*$, then we have $\partial r(x)_{x=x^*} = \{x^*\}$ and so $x^* = z - \alpha w,\ z > \alpha w$.

2. Let $x^* < 0$, then we have $\partial r(x)_{x=x^*} = \{-x^*\}$ and so $x^* = z + \alpha w,\ z < -\alpha w$.

3. Let $x^* = 0$, then $\partial r(x)_{x=x^*} = [-1, 1]$, which yields $\frac{1}{\alpha}(x^* - z) \in [-w, w] \Rightarrow z \in [-\alpha w, \alpha w]$.

Combining all cases together we get $\text{prox}_{\alpha r}(z) = \text{sign}(z)(|z| - \alpha w)_+$. Finally, $\text{prox}_{\alpha r + \delta_{\mathbb{R}}}(z)$ can be derived by noticing that, in this case, (1) $x^* \geq 0$, and (2) when $x^* = 0$ the sub-differential changes due to the presence of the delta-function:

$$x^* = 0 \implies \frac{1}{\alpha}(x^* - z) \in -([-\alpha w, \alpha w] + [-\infty, 0]) = [-\alpha w, +\infty],$$

which gives us the condition $x^* = z, \quad z \in [-\infty, \alpha w]$.

LASSO LASSO is a particular case of A-LASSO above when $w = 1$.

ℓ_0-regularizer Comparing to its counterparts above, the regularizer $R(x) = \delta_{\|x\| \leq k}(x)$ is non-separable. However, the proximal operator of it can still be evaluated analytically:

$$[\text{prox}_{\alpha R}(z)]_i = \left[\underset{\|x\| \leq k}{\text{argmin}} \frac{1}{2\alpha} \|x - z\|^2 \right]_i = \begin{cases} z_i, & i \in \mathcal{I}_k \\ 0, & \text{otherwise} \end{cases},$$

where $\mathcal{I}_k$ is a set of k largest in their absolute value coordinates of z. To get $\text{prox}_{\alpha R + \delta_{\mathbb{R}_+}}$ we replace $\mathcal{I}_k$ with a set of k largest positive coordinates of z, and set the rest of the coordinates to 0.

B.3 Lipschitz-constant for Likelihood of a Linear Mixed-Effects Model

Recall that a function $\mathcal{L}(x)$ is called L-Lipschitz smooth when $\|\nabla \mathcal{L}(x) - \nabla \mathcal{L}(y)\|_2 \leq L\|x - y\|_2$. To find the Lipschitz-constant of the function $\mathcal{L}_{ML}$ (3) we will use the fact that $\mathcal{L}(x)$ is L-Lipschitz if and only if $\|\nabla^2 \mathcal{L}(x)\| \leq L$ for any x. Hence, to upper-bound L we need to upper-bound the norms of Hessians. Assume that $\|y_i - X_i\beta\| \leq \eta$ where $\eta > 0$. We get

$$\left\|\nabla^2 \mathcal{L}(x)\right\|_2 = \left\| \begin{bmatrix} \nabla^2_{\beta\beta} \mathcal{L}(\beta,\gamma) & \nabla^2_{\beta\gamma} \mathcal{L}(\beta,\gamma) \\ \nabla^2_{\gamma\beta} \mathcal{L}(\beta,\gamma) & \nabla^2_{\gamma\gamma} \mathcal{L}(\beta,\gamma) \end{bmatrix} \right\| \leq \sum_{i=1}^{m} \left\| \begin{bmatrix} \frac{\|X_i\|_2^2}{\|\Lambda_i\|_2} & \frac{\eta\|X_i\|_2\|Z_i\|_2^2}{\|\Lambda_i\|^2} \\ \frac{\eta\|X_i\|_2\|Z_i\|_2^2}{\|\Lambda_i\|^2} & \frac{\eta\|Z_i\|_2^4}{\|\Lambda_i\|_2^3} \end{bmatrix} \right\| \tag{21}$$
$$\leq \sum_{i=1}^{m} \max\left(\frac{\|X_i\|_2^2}{\|\Lambda_i\|_2}, \frac{\eta\|X_i\|_2\|Z_i\|_2^2}{\|\Lambda_i\|^2}, \frac{\eta\|X_i\|_2\|Z_i\|_2^2}{\|\Lambda_i\|^2}, \frac{\eta\|Z_i\|_2^4}{\|\Lambda_i\|_2^3} \right) = L$$

C Description of Datasets and Experiments

Table 4 below provides a more detailed overview of the relative performance of the algorithms from Table 2 in the main body.

C.1 Detailed Results from Simulation from Table 2

Model	Regularizer / Metric	L0	L1	ALASSO	SCAD
PGD	Accuracy	89 (75-95)	73 (68-82)	88 (72-98)	71 (62-78)
	FE Accuracy	88 (70-95)	56 (45-70)	84 (65-100)	53 (45-65)
	RE Accuracy	90 (75-100)	91 (80-100)	92 (80-100)	89 (75-100)
	F1	88 (74-95)	77 (71-83)	88 (74-97)	75 (68-80)
	FE F1	87 (72-95)	67 (62-75)	85 (70-100)	66 (62-72)
	RE F1	89 (74-100)	91 (78-100)	91 (78-100)	88 (74-100)
	Time	47.47 (20.22-78.43)	43.02 (23.02-67.01)	38.68 (20.52-58.26)	87.24 (40.73-160.34)
	Iterations	29662 (20985-43234)	31693 (22361-45603)	28912 (20915-39210)	41724 (26911-69881)
MSR3	Accuracy	92 (75-98)	89 (72-100)	91 (75-98)	92 (75-100)
	FE Accuracy	92 (70-100)	85 (60-100)	91 (70-100)	93 (70-100)
	RE Accuracy	91 (78-95)	92 (75-100)	91 (75-100)	92 (80-100)
	F1	91 (76-97)	89 (73-100)	91 (76-98)	92 (76-100)
	FE F1	92 (75-100)	87 (69-100)	92 (75-100)	93 (75-100)
	RE F1	90 (74-94)	91 (74-100)	90 (73-100)	91 (75-100)
	Time	109.86 (5.49-335.01)	13.74 (3.12-31.69)	81.52 (5.94-232.98)	104.20 (6.46-308.19)
	Iterations	1135 (27-3148)	126 (41-314)	895 (81-2262)	1182 (47-3146)
MSR3-fast	Accuracy	92 (75-100)	88 (68-100)	91 (75-98)	92 (75-100)
	FE Accuracy	92 (65-100)	85 (60-100)	91 (70-100)	94 (75-100)
	RE Accuracy	93 (85-100)	91 (75-100)	92 (75-100)	91 (70-100)
	F1	92 (76-100)	88 (71-100)	91 (75-97)	92 (74-100)
	FE F1	92 (72-100)	87 (69-100)	91 (75-100)	94 (78-100)
	RE F1	92 (82-100)	90 (74-100)	90 (74-100)	90 (71-100)
	Time	0.36 (0.15-0.57)	0.35 (0.15-0.56)	0.45 (0.18-0.55)	0.45 (0.16-0.77)
	Iterations	86 (41-119)	87 (43-123)	115 (45-119)	102 (49-145)

Table 4: Comparison of performance of algorithms

C.2 Scalability and Sensitivity Analysis

C.2.1 Scalability

We tested the scalability of the new approach (MSR3-fast) compared to proximal gradient descent and proximal gradient descent with line search. To do this, chose an initially small problem and we scaled the number of features in the data from 100 to 1000, while scaling the number of observations proportionally, and tested the time to completion of these three methods, averaged over 100 replicates. To get the problems of different sizes we assigned A to be 1, 2, 5, 10, 20, 50, and 100, and for each choice of A we generated 100 random problems. Each problem had 8 groups of $10A$ observations each, β and γ had $20A$ features equally split between 0 and 1. Since MSR3-fast has a relaxation parameter η, we evaluated MSR3-fast across different η values to also test the effect of η on timing. For each experiment, we also computed the accuracy of the feature selection, to make sure that there was no degradation in performance. The results are presented in Tables 5 and 6. In terms of timing, we see a superlinear increase in computational complexity with respect to the number of features. Nonetheless, MSR3-fast is competitive with the alternatives across the experiments, and the results are far more accurate. Larger problems could likely significantly benefit from iterated solvers within the interior point framework.

Algorithm	MSR3-Fast							PGD	PGD-LineSearch
η	0.01	0.05	0.10	0.50	1.00	5.00	10.00		
# Features									
100	0m 7s	0m 7s	0m 7s	0m 6s	0m 7s	0m 8s	0m 10s	2m 44s	4m 44s
200	0m 36s	0m 39s	0m 36s	0m 39s	0m 39s	0m 49s	1m 8s	7m 43s	11m 28s
400	5m 2s	4m 51s	4m 34s	4m 26s	5m 16s	7m 33s	10m 38s	47m 46s	12m 36s
1000	59m 10s	57m 12s	60m 30s	69m 57s	68m 55s	111m 31s	139m 47s	469m 16s	55m 8s

Table 5: Execution time for feature selection problems of varying sizes. Each cell shows total time, including grid-search with respect to the sparsity parameter λ. Each cell shows averaged value over 100 randomly-generated problems.

Algorithm	MSR3-Fast							PGD	PGD-LineSearch
η	0.01	0.05	0.10	0.50	1.00	5.00	10.00		
# Features									
100	0.94	0.94	0.95	0.94	0.91	0.86	0.84	0.77	0.77
200	0.99	0.99	0.99	0.98	0.98	0.97	0.95	0.78	0.82
400	0.99	0.99	0.99	0.99	0.99	1.00	1.00	0.80	0.84
1000	0.99	0.98	0.98	0.98	0.98	1.00	1.00	0.83	0.87

Table 6: Accuracy of feature selection problems of varying sizes. Each cell shows averaged value over 100 randomly-generated problems.

C.2.2 Closeness to the Central Path for IP

The τ parameter of MSR3-fast controls how close the interior point method gets to the central path before taking a prox-gradient step. This is a heuristic parameter in the algorithm, and to understand its impact we tested the sensitivity of the execution time and accuracy for a problem with 200 features for four selections of relaxation parameter η. The problems were identical to those from the second row of Table 5. The results are reported in Tables 7 and 8. Neither time nor accuracy were affected by τ across the levels of η.

Algorithm	MSR3-Fast			
η	0.01	0.10	1.00	10.00
τ				
0.1	0m 41s	0m 40s	0m 41s	1m 12s
0.3	0m 35s	0m 36s	0m 38s	1m 1s
0.5	0m 34s	0m 35s	0m 36s	0m 57s
0.7	0m 33s	0m 33s	0m 35s	0m 59s
0.9	0m 33s	0m 33s	0m 35s	0m 52s

Table 7: Execution time of MSR3-fast for different values of τ - a parameter that controls how close the IP needs to be to the central path before doing a projection step. Each cell shows total time, including grid-search with respect to the sparsity parameter λ. Each cell shows averaged value over 100 randomly-generated problems.

Algorithm	MSR3-Fast			
η	0.01	0.10	1.00	10.00
τ				
0.1	0.99	0.99	0.99	0.95
0.3	0.99	0.99	0.98	0.95
0.5	0.99	0.99	0.98	0.95
0.7	0.99	0.99	0.98	0.95
0.9	0.99	0.99	0.98	0.95

Table 8: Accuracy of MSR3-fast for different values of τ - a parameter that controls how close the IP needs to be to the central path before doing a projection step. Each cell shows averaged value over 100 randomly-generated problems.

C.3 GBD Bullying Data

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.
2. `cv_unadjusted`
    - 0 = RR is adjusted for potential confounders (e.g., SES, etc.)
    - 1 = RR is not adjusted for potential confounders
3. `cv_b_parent_only`
    - 0 = Child is involved in reporting their own exposure to bullying.
    - 1 = Only parent is involved in reporting the child's exposure to bullying
4. `cv_or`
    - 0 = estimate is a RR
    - 1 = estimate is an odds ratio (OR)

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.
6. `cv_low_threshold_bullying`
    - 0 = uses a 'frequent' exposure threshold for classing someone as exposed to bullying.
    - 1 = uses an 'occasional' exposure threshold for classing someone as exposed to bullying.
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
9. `Percent_female`
    - Indicates % of sample in estimate that are female.
10. `cv_child_baseline`
    - Indicates whether mid-age of sample is above or below 13.

Covariates 2, 3, 5, 6, 8 have been statistically significant in past models.