Trimmed Constrained Mixed Effects Models: Formulations and Algorithms

Peng Zheng and Aleksandr Aravkin*
Department of Applied Mathematics, University of Washington
and
Ryan Barber, Reed Sorensen, and Christopher Murray†
Institute for Health Metrics and Evaluation, University of Washington

September 25, 2019

Abstract

Mixed effects (ME) models inform a vast array of problems in the physical and social sciences, and are pervasive in meta-analysis. We consider ME models where the random effects component is linear. We then develop an efficient approach for a vast problem class that allows nonlinear measurements, priors, and constraints, and finds robust estimates in all of these cases using trimming in the associated marginal likelihood.

We illustrate the efficacy of the approach on a range of applications for meta-analysis of global health data. Constraints and priors are used to impose monotonicity, convexity and other characteristics on dose-response relationships, while nonlinear observations enable new epidemiological analyses in place of approximations. Robust extensions ensure that spurious studies do not drive our understanding of between-study heterogeneity. The software accompanying this paper is disseminated as an open-source Python package called limeTR.

Keywords: Mixed effects models, trimming, nonsmooth nonconvex optimization, meta-analysis

1 Introduction

Linear mixed effects (LME) models play a central role in a wide range of analyses [Bates et al. 2015]. Examples include longitudinal analysis [Laird et al. 1982], meta-analysis [DerSimonian and Laird 1986], and numerous domain-specific applications [Zuur et al. 2009].

*The authors gratefully acknowledge the Washington Research Foundation Data Science Professorship.
†This work was funded by the Bill & Melinda Gates Foundation.
The problem class we consider here lies strictly between LME models and fully general nonlinear mixed effects models. We allow nonlinear measurements, priors, and constraints, but require that the random effects enter the model in a linear way. This gives a tractable approach for a broad problem class, enabling a number of extensions. The key technical innovation is a trimmed extension for the marginal likelihood problem associated to these ME models, along with a specialized algorithm and convergence analysis that applies to the full class.

Robust LME models are typically obtained by using heavy tailed error models for random effects. The Student’s t distribution [Pinheiro et al. 2001], as well as weighting functions [Koller 2016] have been used. The resulting formulations are fit either by EM methods, estimating equations, or by MCMC [Rosa et al. 2003]. In this paper, we take a different tack, and extend the least trimmed squares (LTS) method to the ME setting. LTS has recently found wide use in modern applications particularly in machine learning [Aravkin and Davis 2019] and high-dimensional inference [Yang and Lozano, 2015, Yang et al., 2018b]. Trimming the ME likelihood extends prior art because it does not fall into the problem class of Aravkin and Davis [2019].

Table 1: Comparison with currently available robust mixed effects packages.

|                  | limeTR | metafor | robumeta | metaplus | robustlmm | clme | INLA |
|------------------|--------|---------|----------|----------|-----------|------|------|
| Robust option    | ✓      | x       | ✓        | ✓        | ✓         | x    | ✓    |
| Allows for known | ✓      | ✓       | ✓        |          | x         | x    | ✓    |
| observation variance | ✓      | x       | x        | ✓        | ✓         |      | x    |
| Covariates in random | ✓      | x       | x        | ✓        | ✓         |      | x    |
| effects variance | Nonlinear observations | ✓ | x | ✓ | ✓ | x | ✓ |
| Linear constraints | ✓ | x | x | ✓ | ✓ | x |
| Nonlinear constraints | ✓ | x | x | ✓ | x | x |

**Contributions.** We formulate and solve the trimming problem required for the trimmed ME approach. Our second contribution is to incorporate nonlinear measurements, constraints, and priors into the trimmed ME class. These extensions are essential for a range of problems, particularly those that use splines to model dose-response relationships. We show how splines capture
such nonlinear relationships, and how constraints can be used to directly control their shape.

The main code to perform the inference is published as open source Python package called limeTR (Linear Mixed Effects with Trimming, pronounced lime tree). All synthetic experiments using limeTR have been submitted for review as supplementary material with this paper. The limeTR package allows functionality that is not available through other available open source tools. The functionality of limeTR is summarized in Table 1.

The paper proceeds as follows. In Section 2.1 we describe the problem class of ME models and derive the marginal maximum likelihood (ML) estimator. In Section 2.2 we describe how constraints and priors are imposed on parameters. In Section 2.3 we review trimming approaches and develop a new trimming extensions for the ML approach. In Section 2.4 we present a customized algorithm based on variable projection, along with a convergence analysis. In Section 2.5 we discuss spline models for dose-response relationships and give examples of shape-constrained trimmed spline models. Section 3 shows the efficacy on the methods for synthetic and empirical data. In Section 3.1 we validate the ability of the method to detect outliers when working with heterogeneous longitudinal data, and compare with other packages. In Section 3.2 we apply the method to analyze empirical data sets for both linear and nonlinear relationships using trimmed constrained MEs. This section highlights new capability of limeTR that is not available in other packages.

2 Methods

2.1 Problem Class

We consider the following mixed effects model:

\[ y_i = F_i(\beta) + Z_i u_i + \epsilon_i \]

\[ u_i \sim N(0, \Gamma), \quad \Gamma = \text{diag}(\gamma), \quad \epsilon_i \sim N(0, \Lambda), \]

where \( y_i \in \mathbb{R}^{n_i} \) is the vector of observations from the \( i \)th group, \( \epsilon_i \in \mathbb{R}^{n_i} \) are measurement errors with covariance \( \Lambda \), \( u_i \in \mathbb{R}^{k_\gamma} \) are independent random effects, and \( Z_i \in \mathbb{R}^{n_i \times k_\gamma} \) is a linear map, and \( \beta \) are regression coefficients. The models \( F_i \) may be nonlinear, but we restrict the random effects to enter in a linear way through the term \( Z_i u_i \).
A range of assumptions may be placed on $\Lambda$. In longitudinal analysis, $\Lambda$ is often a diagonal or block-diagonal matrix, parametrized by set of shared unknown terms. In meta-analysis, $\Lambda$ is a known diagonal matrix whose entries are variances for each input datum.

The joint likelihood for the fixed effects ($\beta, \gamma, \Lambda$) and random effects $u$ is given by

$$p(\beta, \tau, \Lambda, u|y) \propto \prod_{i=1}^{m} \|y_i - F_i(\beta) - Z_i u\|^2_{\Lambda^{-1}} \|u\|^2_{\Gamma^{-1}} \det(\Lambda^{-1}) \det(\Gamma)^{-1}$$

Integrating out the random effects $u$ from (2) and taking the negative logarithm gives the associated objective to a minimization problem:

$$L_{ML}(\beta, \gamma, \Lambda) = -\ln \left( \int p(\beta, \gamma, \Lambda, u|y) du \right)$$

$$= \sum_{i=1}^{m} \frac{1}{2} (y_i - F_i(\beta))^\top (Z_i \Gamma Z_i^\top + \Lambda_i)^{-1} (y_i - F_i(\beta)) + \frac{1}{2} \ln |Z_i \Gamma Z_i^\top + \Lambda_i|.$$  

Problem (3) is equivalent to a maximum likelihood formulation arising from a Gaussian model with correlated errors:

$$y_i = F_i(\beta) + \omega, \quad \omega \sim N(0, Z_i \Gamma Z_i^\top + \Lambda_i).$$

The structure of this objective depends on the structural assumptions on $\Lambda$. We restrict our numerical experiments to two particular classes: (1) $\Lambda = \sigma^2 I$ with $\sigma^2$ unknown, used in standard longitudinal analysis, and (2) $\Lambda = \Sigma$, a known matrix of observation covariances, used in meta-analysis. limeTR allows other structure options as well, for example group-specific unknown $\sigma_i^2$ that extends case (1), but we restrict to a simple set of synthetic results to help focus on robust capabilities.

### 2.2 Constraints and Priors

The ML estimate (3) can be extended to incorporate linear and nonlinear inequality constraints

$$C(\theta) \leq c,$$

where $\theta$ are any parameters of interest. Constraints play a key role in Section 2.5 when we use polynomial splines to model nonlinear relationships. The trimming approach developed in the next section is applicable to both constrained and unconstrained ML estimates.
In many applications it is essential to allow priors on parameters of interest $\theta$. We assume that priors are given by a functional form

$$\theta \sim \exp(-\rho(\theta))$$

where $\rho$ is smooth (but may be nonlinear and nonconvex). The likelihood problem is then augmented by adding the term $\rho(\theta)$ to the ML objective.

In the next section we describe trimmed estimators, and extend them to the ME setting.

### 2.3 Trimming in Mixed Effect Models

Least trimmed squares (LTS) is a robust estimator proposed by Rousseeuw [1985], Rousseeuw and Croux [1993] for the standard regression problem. Given the problem

$$\min_{\beta} \sum_{i=1}^{n} \frac{1}{2} (y_i - \langle x_i, \beta \rangle)^2,$$  \hspace{1cm} (4)

the LTS estimator minimizes the sum of smallest $h$ residuals rather than all residuals. These estimators were initially introduced to develop linear regression estimators that have a high breakdown point (in this case 50%) and good statistical efficiency (in this case $n^{-1/2}$).\footnote{Breakdown refers to the percentage of outlying points which can be added to a dataset before the resulting M-estimator can change in an unbounded way. Here, outliers can affect both the outcomes and training data.} LTS estimators are robust against outliers, and arbitrarily large deviations that are trimmed do not affect the final $\hat{\beta}$.

Rather than writing the objective in terms of order statistics, it is simpler to extend the likelihood using an auxiliary variable $w$:

$$\min_{\beta, w} \sum_{i=1}^{n} w_i \left( \frac{1}{2} (y_i - \langle x_i, \beta \rangle)^2 \right) \text{ s.t. } 1^\top w = h, \quad 0 \leq w \leq 1. \hspace{1cm} (5)$$

The set

$$\Delta_h := \left\{ w : 1^\top w = h, \quad 0 \leq w \leq 1 \right\} \hspace{1cm} (6)$$

is known as the capped simplex, since it is the intersection of the $h$-simplex with the unit box (see e.g. Aravkin and Davis [2019] for details). For a fixed $\beta$, the optimal solution of (5) with respect to $w$ assigns weight 1 to each of the smallest $h$ residuals, and 0 to the rest. Problem (5) is solved jointly in $(\beta, w)$, simultaneously finding the regression estimate and classifying the observations.
into inliers and outliers. This joint strategy makes LTS different from post hoc analysis, where a model is first fit with all data, and then outliers are detected using that estimate.

Several approaches for finding LTS and other trimmed M-estimators have been developed, including FAST-LTS \cite{Rousseeuw2006}, exact algorithms with exponential complexity \cite{Mount2014}. The LTS approach (5) does not depend on the form of the least squares function, and this insight has been used to extend LTS to a broad range of estimation problems, including generalized linear models \cite{Neykov2003}, high dimensional sparse regression \cite{Alfons2013}, and graphical lasso \cite{Yang2015,Yang2018a}. The most general problem class to date, presented by \cite{Aravkin2019}, is formulated as

$$\min_{\beta, w} \sum_{i=1}^{n} w_i f_i(\beta) + R(\beta) \quad \text{s.t.} \quad 1^T w = h, \quad 0 \leq w \leq 1.$$  \hspace{1cm} (7)

where $f_i$ are continuously differentiable (possibly nonconvex) functions and $R$ describes any regularizers and constraints (which may also be nonconvex).

Critically, the general class (7) does not capture estimator (3). Problem (7) only applies to the very special problem of detecting entire outlying groups:

$$\min_{\beta, \gamma, \Lambda, w} \sum_{i=1}^{m} w_i \left( \frac{1}{2}(y_i - F_i(\beta))^T(Z_i \Gamma Z_i^T + \Lambda_i)^{-1}(y_i - F_i(\beta)) + \frac{1}{2} \ln |Z_i \Gamma Z_i^T + \Lambda_i| \right)$$

$$\text{s.t.} \quad 1^T w = h, \quad 0 \leq w \leq 1.$$  \hspace{0.5cm} (8)

This is limiting, since we want to differentiate measurements within groups. We solve the problem by using a new trimming formulation that goes outside (7).

To explain the approach we focus on trimming a single group term from the ML likelihood (3):

$$\left( \frac{1}{2}(y_i - F_i(\beta))^T(Z_i \Gamma Z_i^T + \Lambda_i)^{-1}(y_i - F_i(\beta)) + \frac{1}{2} \ln |Z_i \Gamma Z_i^T + \Lambda_i| \right)$$

Here, $y_i \in \mathbb{R}^{n_i}$, where $n_i$ is the number of observations in the $i$th group. To trim observations within the group, we introduce auxiliary variables $w_i \in \mathbb{R}^{n_i}$, and define

$$r_i := y_i - F_i(\beta), \quad W_i := \text{diag}(w_i), \quad \sqrt{W_i} := \text{diag}(\sqrt{w_i}).$$

We now form the objective

$$\frac{1}{2} r_i^T \sqrt{W_i} \left( \sqrt{W_i} Z_i \Gamma Z_i^T \sqrt{W_i} + \Lambda_i \odot w_i \right)^{-1} \sqrt{W_i} r_i + \frac{1}{2} \ln \left| \sqrt{W_i} Z_i \Gamma Z_i^T \sqrt{W_i} + \Lambda_i \odot w_i \right|,$$  \hspace{0.5cm} (9)
where $\odot$ denotes the elementwise power operation:

$$
\Lambda_i^{\odot w_i} := \begin{bmatrix}
(\lambda_{1j})^{w_{i1}} & 0 & \ldots & 0 \\
0 & \ddots & \ddots & \vdots \\
0 & \ldots & 0 & (\lambda_{in_i})^{w_{ini}}.
\end{bmatrix}
$$

(10)

When $w_{ij} = 1$, we recover the contribution of the $ij$th observation to the original likelihood. As $w_{ij} \downarrow 0$, the $ij$th contribution to the residual is correctly eliminated by $\sqrt{w_{ij}} \downarrow 0$. The $j$th row and column of $\sqrt{W_i}Z_i\Gamma Z_i^\top \sqrt{W_i}$ both go to 0, while the $j$th entry of $\Lambda_i^{\odot w_i}$ goes to 1, which effectively removes all impact of the $j$th point on the covariance matrix.

Combining trimmed ML with priors and constraints, we obtain the following estimator:

$$
\min_{\beta, \gamma, \Lambda, w} L(\beta, \gamma, \Lambda, w) := \sum_{i=1}^{m} \frac{1}{2} r_i^\top \sqrt{W_i} \left( \sqrt{W_i}Z_i\Gamma Z_i^\top \sqrt{W_i} + \Lambda_i^{\odot w_i} \right)^{-1} \sqrt{W_i}r_i + \\
\frac{1}{2} \ln \left| \sqrt{W_i}Z_i\Gamma Z_i^\top \sqrt{W_i} + \Lambda_i^{\odot w_i} \right| + \rho(\beta, \gamma, \Lambda)
$$

s.t. $r_i = y_i - F_i(\beta)$, $1^\top w = h$, $0 \leq w \leq 1$, $C \begin{pmatrix} \beta \\ \gamma \\ \Lambda \end{pmatrix} \leq c.$

(11)

The estimator (11) has not been previously considered in the literature. The fit is obtained using iterative techniques. Problem (11) is nonsmooth, so care must be taken when developing and analyzing the optimization algorithm. We present a specialized algorithms and its convergence theory in the next section.

### 2.4 Fitting Trimmed Constrained MEs: Algorithm and Analysis

Estimator (11) is nonsmooth and nonconvex. The key to algorithm design and analysis is to decouple this structure, and reduce the estimator to solving a smooth nonconvex value function over a convex set. This allows an efficient approach that combines classic nonlinear programming with first-order approaches for optimizing nonsmooth nonconvex problems. We partially minimize with respect to $(\beta, \gamma, \Lambda)$ using an interior point method, and then optimize the resulting value function with respect to $w$ using a first-order method. The approach leverages ideas from variable projection [Golub and Pereyra, 1973, 2003] [Aravkin and Van Leeuwen, 2012] [Aravkin et al., 2018].
We define $\theta = (\beta, \gamma, \Lambda)$, the implicit solution $\theta(w)$ and value function $v(w)$ as follows:

$$
\theta(w) := \arg \min_{\theta} L(\theta, w) \quad \text{s.t.} \quad C(\theta) \leq c
$$
$$
v(w) := \min_{\theta} L(\theta, w) \quad \text{s.t.} \quad C(\theta) \leq c
$$

(12)

where $L(\theta, w)$ is given in (11). The value function in (12) has first and second order derivatives under simple conditions that allow the implicit function theorem to be invoked [Bell and Burke, 2008; Aravkin et al., 2016, 2018]. We state the precise theorem below.

**Theorem 1** (Smoothness of the value function). Consider the function $v(w)$ in (12). Suppose that for any $\theta(w)$, we have

$$
\begin{bmatrix}
\nabla^2_{\theta} L|_{\theta(w), w} & \nabla C_{\theta(w)}^\top \\

\nabla C|_{\theta(w)} & 0
\end{bmatrix}
$$

is invertible. Then $v(w)$ is continuously differentiable by the implicit function theorem, with gradient given by

$$
\nabla v(w) = -\partial_w L(\theta, w)|_{(\theta(w), w)}.
$$

Partially minimizing over $\theta$ reduces the optimization problem (11) to

$$
\min_w v(w) \quad \text{s.t.} \quad 1^\top w = h, \quad 0 \leq w \leq 1,
$$

(13)

where $v(w)$ is a continuously differentiable nonconvex function, and the constrained set is the (convex) capped simplex $\Delta_h$ introduced in the trimming section. The high-level optimization over $w$ is implemented using projected gradient descent:

$$
\begin{aligned}
\mathbf{w}^+ &= \text{proj}_{\Delta_h}(\mathbf{w} - \alpha \nabla v(w)). \\
\end{aligned}
$$

(14)

Each update to $\mathbf{w}$ requires computing the gradient $\nabla v$, which in turn requires solving for $\theta$; see (12). The explicit implementation equivalent to (14) is summarized in Algorithm 1.

Step 5 of Algorithm 1 requires solving the constrained likelihood problem (11) with $\mathbf{w}$ held fixed. We solve this problem using IPOpt [Wächter and Biegler, 2006], a robust interior point solver that allows both simple box and functional constraints. While one could solve the entire problem using IPOpt, treating $\theta$ and $\mathbf{w}$ differently is key to efficient performance. Typically $\theta$ is small compared to $\mathbf{w}$, which is the size of the data. On the other hand the constrained likelihood problem in $\theta$ is difficult while constrained value function optimization over $\mathbf{w}$ can be solved with projected gradient.
Algorithm 1 Projected gradient descent on the Value Function $v$ of (12)

1: Input: $w_0, \lambda_w.$

2: Initialize: $\nu = 0$

3: while not converged do

4: $\nu \leftarrow \nu + 1$

5: $\theta^{\nu+1} \leftarrow \min_{\theta} \mathcal{L}(\theta, w^{\nu}) \text{ s.t } C(\theta) \leq c$

6: $w^{\nu+1} \leftarrow \text{proj}_{\Delta_h}(w - \lambda_w \partial_w \mathcal{L}(w^{\nu}, \theta^{\nu+1}))$

7: Output: $w_{\nu}, \theta_{\nu}$

2.5 Nonlinear Relationships using Constrained Splines

In this section we discuss using spline models to capture nonlinear relationships. The relationships most interesting to us are dose-response relationships, which allow us to analyze adverse effects of risk factor exposure (e.g. smoking, BMI, dietary consumption) on health outcomes. For an in-depth look at splines and spline regression see [De Boor et al. 1978] and [Friedman et al. 1991].

Constraints can be used to capture expert knowledge on the shape of such risk curves, particularly in segments informed by sparse data.

2.5.1 B-splines and bases

A spline basis is a set of piecewise polynomial functions with designated degree and domain. If we denote polynomial order by $p$, and the number of knots by $k$, we need $p+k$ basis elements $s_j^p$, which can be generated recursively as illustrated in Figure 1.

Given such a basis, we can represent any nonlinear curve as the linear combination of the spline basis elements, with coefficients $\beta \in \mathbb{R}^{p+k}$:

$$f(t) = \sum_{j=1}^{p+k} \beta_j s_j^p(t). \quad (15)$$

These coefficients are inferred by limeTR analysis. A more standard explicit representation of (15) is obtained by building a design matrix $X$. Given a set of $t$ values at which we have data,
the \( j \)th column of \( X \) is given by the expression
\[
X_{\cdot j} = \begin{bmatrix}
s^p_{j}(t_0) \\
\vdots \\
s^p_{j}(t_k)
\end{bmatrix}.
\]

The model for observed data coming from (15) can now be written compactly as
\[
y = X\beta + Z_iu_i + \epsilon_i,
\]
which is a special case of the main problem class (1).

2.5.2 Shape constraints

We can impose shape constraints such as monotonicity, concavity, and convexity on splines. This approach was proposed by Pya and Wood [2015], who used re-formulations using exponential representations of parameters to capture non-negativity. The development in this section uses explicit constraints, which is simple to encode and extends to more general cases, including functional inequality constraints \( C(\theta) \leq c \).
Monotonicity. Spline monotonicity across the domain of interest follows from monotonicity of the spline coefficients [De Boor et al., 1978]. Given spline coefficients $\beta$, the curve $f(t)$ in (15) is monotonically nondecreasing when

$$\beta_1 \leq \beta_2 \leq \cdots \leq \beta_n$$

and monotonically non-increasing if

$$\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n.$$  

The relationship $\beta_1 \leq \beta_2$ can be written as $\beta_1 - \beta_2 \leq 0$. Stacking these inequality constraints for each pair $(\beta_i, \beta_{i+1})$ we can write all constraints simultaneously as

$$\begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_n \end{bmatrix} \leq \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$  

These linear constraints are a special case of the general estimator (11).

2.5.3 Convexity and Concavity

For any $C^2$ (twice continuously differentiable) function $f : \mathbb{R} \to \mathbb{R}$, convexity and concavity are captured by the signs of the second derivative. Specifically, $f$ is convex if $f''(t) \geq 0$ is everywhere, and concave if $f''(t) \leq 0$ everywhere. We can compute $f''(t)$ for each interval, and impose linear inequality constraints on these expressions. We can therefore easily pick any of the eight shape combinations given in [Pya and Wood, 2015, Table 1], as well as imposing any other constraints on $\beta$ (including bounds) through the interface of limeTR.

2.5.4 Nonlinear measurements

Some of the studies in Section 3.2 use nonlinear observation mechanisms. In particular, given a dose-response curve

$$f(t) = \sum_{j=1}^{p+k} \beta_j^p \delta_j^p(t).$$
studies often report odds ratio of an outcome between exposed and unexposed groups that are defined across two intervals on the underlying curve:

\[ y_i = \frac{1}{a_1-a_0} \int_{a_0}^{a_1} f(t)dt - \frac{1}{b_1-b_0} \int_{b_0}^{b_1} f(t)dt. \]

When \( f(t) \) is represented using a spline, each integral is a linear function of \( \beta \). If we take the observations to be the log of the relative risk, this is given by

\[ y_i = \ln(\langle x_1^i, \beta \rangle) - \ln(\langle x_2^i, \beta \rangle) := F_i(\beta), \]

a particularly useful example of the general nonlinear term \( F_i(\beta) \) in problem class [1].

### 2.6 Variance Estimation

The limeTR package modifies the parametric bootstrap [Efron and Tibshirani 1994] to estimate the variance of the fitting procedure. This strategy is necessary when constraints are present, and standard Fisher-based strategies for posterior variance selection do not apply [Cox 2005].

The modified parametric bootstrap is similar to the standard bootstrap, but can be used more effectively for sparse data, e.g., when different studies sample sparsely across a dose-response curve. The approach can be used with any estimator [11].

In the linear Gaussian case, the standard bootstrap is equivalent to bootstrapping empirical residuals, since every datapoint can be reconstructed this way. When the original data is sparse, we modify the parametric procedure to sample modeled residuals. Having obtained the estimate \((\hat{\beta}, \hat{\Lambda}, \hat{\gamma})\), we sample model-based errors to get new bootstrap realizations \( \bar{y} \):

\[ \bar{y} = X\hat{\beta} + Z\bar{u} + \bar{\epsilon}, \]

where \( \epsilon_i \sim N(0, \hat{\Lambda}) \) and \( u_i \sim N(0, \hat{\gamma}) \). These realizations have the same structure as the input data. For each realization \( \bar{y} \), we re-run the fit, and obtain \( N \) estimates \( \{\hat{\beta}, \hat{\Lambda}, \hat{\gamma}\}_{1:N} \). This set of estimates is used to approximate the variance of the fitting procedure along with any confidence bounds. The procedure can be applied in sparse and complex cases but depends on the initial fit \((\hat{\beta}, \hat{\Lambda}, \hat{\gamma})\). Its exact theoretical properties are outside the scope of the current paper and are a topic of ongoing research.
3 Verifications

In this section we validate limeTR on synthetic and empirical datasets. In Section [3.1] we show how limeTR compares to existing robust packages on simple problems that all packages can solve; see Table [1]. We focus on robustness of the estimates to outliers, which is a key technical contribution of the paper.

In Section [3.2] we use the advanced features of limeTR to analyze multiple datasets in public health, where robustness to outliers, and information communicated through constraints and nonlinear measurements all play an important role.

3.1 Validation Using Synthetic Data

Here we consider two common mixed effects models. First we look at meta-analysis, where the $\Lambda$ term is known while $(\beta, \gamma)$ are unknown. Then we look at a simple longitudinal case, where all three parameters are unknown, and $\Lambda$ is modeled as $\sigma^2 I$ with unknown scalar $\sigma^2$. In both of these cases, we compare the performance of limeTR against several available packages. The simulated data is the same for both examples; only the model is different.

For the experiments, we generated 30 synthetic datasets with 10 observations in each of 10 studies ($n = 100$). The underlying true distribution is defined by $\beta_0 = 0$, $\beta_1 = 5$, $\gamma = 6$, and $\sigma = 4$, where $\gamma$ is the standard deviation of the between-study heterogeneity and $\sigma$ is the standard deviation of the measurement error. For the meta-analysis simulation, we assigned each observation a standard error of 4. The domain of the covariate $x_1$ is $[0, 10]$. To create outliers, we randomly chose 15 data points in the sub-domain $[6, 10]$ and offset them according to: $y'_i = y_i - 30 - |N(0, 80^2)|$.

3.1.1 Meta-analysis.

We compared limeTR to three widely-used packages that have some robust functionality: metafor, robumeta, and metaplus. The functionality developed in these packages differs from that of limeTR. The metafor and robumeta packages refer to robustness in the context of the sandwich variance estimator, which makes the uncertainty around predictions robust to correlation of observations within groups. metaplus uses heavy-tailed distributions to model random effects, which
potentially allows one to account for outlying studies but not measurements within studies.

Nonetheless, it is useful to see how a new package compares with competing alternatives on simple examples. We compared the packages to limeTR in terms of error incurred when estimating ground truth parameters \((\beta_0, \beta_1, \gamma)\), computation time, the true positive fraction (TPF) of outliers detected and the false positive fraction (FPF) of inliers incorrectly identified. If a threshold of 0.8 inliers is given to limeTR, then outliers are exactly data points with an estimated weight \(w_i\) of zero, and those correspond to the largest absolute model residuals. To compare with other packages in terms of TPF and FPF, we identified the 20 data points with the highest residuals according to each packages’ fit. Table 2 and Figure 2 show the results of the meta-analysis simulation. The metrics are averages of 30 estimates from models fit on the synthetic datasets. limeTR had lower absolute error in \((\beta_0, \beta_1, \gamma)\), higher TPF, lower FPF and faster computation time than the alternatives.

| Package    | \(\beta_0\) | \(\beta_1\) | \(\gamma\) | TPF  | FPF  | Seconds |
|-----------|-------------|-------------|------------|------|------|---------|
| Truth     | 0           | 5           | 6          | —    | —    | —       |
| limeTR    | 0.61        | 4.95        | 4.86       | 1.00 | 0.06 | 0.35    |
| robumeta  | 10.92       | 0.03        | 37.2       | 0.68 | 0.12 | 6.54    |
| metaplus  | 10.92       | 0.03        | 35.1       | 0.68 | 0.12 | 42.4    |
| metafor   | 10.92       | 0.03        | 35.4       | 0.68 | 0.12 | 1.51    |

Table 2: Results of Meta-Analysis Comparison. True values are \(\beta_0 = 0, \beta_1 = 5, \gamma = 6\). Results show average estimates across realizations. The limeTR package has much smaller absolute error in \(\beta_0, \beta_1\) and in \(\gamma\) than other packages.

3.1.2 Longitudinal Example.

Here we compare limeTR to R packages for fitting robust mixed effects models. Rather than assuming that errors are distributed as Gaussian, the packages use Huberized likelihoods (robustlmm) and Student’s t distributions (heavy) to model contamination by outliers. limeTR identifies outliers through the weights \(w_i\) in the likelihood estimate \[(3)\] that now captures simple longitudinal analysis. Specifically, \(\Lambda\) is no longer specified as in the meta-analysis case, but is instead parameterized through a single unknown error \(\Lambda = \sigma^2 I\) common to all observations. We use the same
Figure 2: A representative instance of the experiment summarized in Table 2. True mechanism is shown using solid line; the true model is successfully inferred by the limeTR package despite the outliers (red points).

simulation structure as in Section 3.1.1 now replacing observation-specific standard errors with random errors generated according to $N(0,\sigma^2)$. Since $\sigma^2$ is now also an unknown parameter, we check how well it is estimated by all packages and report this error in Table ??.

3.2 Real-World Case Studies

In this section we look at three real-data cases, all using meta-analysis. Across these examples, we show how trimming, dose-response relationships, and non-linear observation mechanisms come together to help understand complex and heterogeneous data.

3.2.1 Simple Example: Vitamin A vs. Diarrheal Disease
Figure 3: A representative instance of the experiment summarized in Table ?? . robustlmm and heavy packages both estimate $\beta$ better than lme4, likely because they use distributions with heavier tails. limeTR outperforms all three alternatives.

The first example is a simple linear model that aims to quantify the effect of vitamin A supplementation on diarrheal disease. This is an important topic in global health, and we refer the interested reader to the Cochrane systematic review of the topic [Imdad et al., 2017]. In this example, we examine the influence outliers can have on inferences from a model, but we do not discuss in detail how to interpret the findings.

In this example, the dependent variable is the natural log of relative risk, and no covariates are used. The model we consider is

$$y_i = 1\beta_0 + 1u_i + \epsilon_i,$$

where $y_i$ are data sets reported from each study, $\epsilon_i \sim N(0, \Sigma_i)$ are associated standard errors, and $u_i \sim N(0, \gamma^2)$ is a study-specific random effect, with $\gamma^2$ attempting to account for between-study heterogeneity.
Figure 4: Results for Vitamin A vs. diarrheal disease, presented using the standard funnel plot (effect size vs. reported standard error). The left panel shows model without trimming, while the right panel shows the model with trimming. Trimming 10% of the studies shifts the effect closer to 0, making it less plausible that Vitamin A is protective, and identifies potential outliers (right panel, shown in red), which are able to hide more easily in the left panel.

Figure 4 shows the results without trimming and with 10% trimming. We included data from 12 randomized control trials in the models, with a few studies having multiple observations. Without trimming, the estimated effect size is $\beta_0 = -0.15$; with trimming it is three time smaller, $\beta_0 = -0.05$. In the trimmed model, we observe all preserved points inside the funnel, indicating that there is no expected between-study heterogeneity. This is confirmed by the estimates – without trimming, between-study heterogeneity $\gamma^2$ is estimated to be 0.0435, and after trimming two studies it is reduced to $7.14 \times 10^{-9}$, nearly 0. Moreover, the potential outliers are unusual: while all other studies deal with ages 1 year or younger, the trimmed studies are among older ages, up to 10 years in one case and 3-6 years in the other. One of the trimmed studies was also conducted in slums, a non-representative population.

### 3.2.2 Spline Example: Smoking vs. Lung Cancer.

The correlation between smoking and lung cancer is indisputable [Gandini et al., 2008, Lee et al., 2012]. The exact nature of the relationship and its uncertainty requires accounting for the dose-response relationship between the amount smoked (typically measured in pack-years) and odds of lung cancer. We expect a nonlinear relationship between smoking and lung cancer, and the spline
The methodology described in Section 2.5.1 can be used.

The outcome here is the natural log of relative risk (compared to nonsmokers). The effect of interest is a function of a continuous exposure, measured in pack-years smoked. All studies compare different levels of exposure to non-smokers, and so the random effect for each study modifies the slope relative to non-smokers:

\[ y_{ij} = \left( \frac{\log(\text{risk}_{ij})}{\text{exposure}_{ij}} - \log(\text{risk}_0) + u_i \right) (\text{exposure}_{ij} - \text{exposure}_0) + \epsilon_{ij} \]

If we normalize risk of nonsmokers to 1, and use a spline to model the nonlinear risk curve, we have the explicit expression

\[ y_{ij} = \log(\langle x_{ij}, \beta \rangle) + \text{exposure}_{ij}u_i + \epsilon_{ij} \]

with \( x_{ij} \) computed using a spline basis matrix for \( \text{exposure}_{ij} \) (see Section 2.5.1), \( u_i \) the random effect for study \( i \), and \( \epsilon_{ij} \sim N(0, \sigma^2_{ij}) \) the variance reported by the \( i \)th study for its \( j \)th point.

To obtain the results in Figure 5 using limeTR, interior knots were set at the 10th, 50th, and 80th percentiles of the exposure values observed in the data, corresponding to pack-year levels of 10, 30, and 55, respectively. We included 199 datapoints across 25 studies in the analysis, and again trimmed 10% of the data. Trimming in this case removes datapoints that are far away from the group (even considering between-study heterogeneity), as well as points that are closer to the mean but over-confident; these types of outliers are specific to meta-analysis.

We also used multiple priors and constraints. First, we enforced that at an exposure of 0, the log relative risk must be equal to 1 (baseline: non-smokers) by adding direct constraint on \( \beta_0 \). To control changes in slope in data sparse segments, we included a Gaussian prior of \( N(0, 0.01) \) on the highest derivative in each segment. Additionally, the segment between the penultimate and exterior knots on the right side is particularly data sparse, and is more prone to implausible behavior due to its location at the terminus, and we force that spline segment to be linear.

We show the unconstrained cubic spline in the left panel of Figure 5, and the constrained analysis that uses monotonicity and concavity of the curve in the right panel the same figure. The mean relationship looks more regular when using constraints, but the model cannot explain the data as well and so the estimate of heterogeneity is higher. On the other hand, the more flexible model has higher fixed effects uncertainty, and a lower estimate for between-study heterogeneity. The point sets selected for trimming are slightly different as well between the two experiments.
Figure 5: Modeling dose-response relationship between exposure (pack years) and log-odds of lung cancer. Left figure shows a cubic spline, while right figure shows a cubic spline with monotonically increasing risk and concavity constraints. The uncertainty of the mean is shown in dark grey, while the additional uncertainty due to heterogeneity is shown in light gray. Constraints regularize the shape and decrease fixed effect uncertainty, but have higher estimated heterogeneity, whereas the more flexible model explains the data and thus has higher fixed effect uncertainty and lower heterogeneity. 10% trimming removes points that are far away from the mean dose-response relationship, as well as those moderately away from the mean but with very low reported standard deviation. Point radii on the graphs are inversely proportional to the reported standard deviations.

3.2.3 Indirect nonlinear observations: red meat vs. breast cancer.

The effect of red meat on various health outcomes is an ongoing debate. In this section we briefly consider the relationship between breast cancer and red meat consumption, which has been systematically studied [Anderson et al., 2018]. In this section, we use data from available studies on breast cancer and red meat to show two more features of the limeTR package: nonlinear observation mechanisms and monotonicity constraints.

The smoking example in Section 3.2.2 uses a direct observation model, since all measurement are comparisons to the baseline non-smoker group. This is not the case for other risk-outcome pairs. When considering the effect of red meat consumption, studies typically report multiple comparisons between groups that consume various amounts of meat. In particular, all datapoints across studies are given to us a tuple: odds ratio for group $[a, b]$ vs. group $[c, d]$. These datapoints
are thus not measuring the spline directly, but are average slopes between points in log-derivative space. Explicitly, the observation model in this case is as described in Section 2.5.4

\[
y_i = \frac{1}{a_1-a_0} \int_{a_0}^{a_1} f(t)dt \\
\frac{1}{b_1-b_0} \int_{b_0}^{b_1} f(t)dt
\]

(16)

As in the smoking example, we model the random effects on the average slopes of the log-relative risks, use a spline to represent the risk curve, and constrain the risk at zero exposure to be 1. The integral across any interval of a spline curve is a linear function of the coefficients, so we have

\[
\ln(y_{ij}) = \left( \ln(\langle x_{ij}^1, \beta \rangle) - \ln(\langle x_{i}^2, \beta \rangle) + u_i \right) (\text{alternative}_{ij} - \text{reference}_{i}) + \epsilon_{ij}
\]

\[
= \ln(\langle x_{ij}^1, \beta \rangle) - \ln(\langle x_{i}^2, \beta \rangle) + (\text{alternative}_{ij} - \text{reference}_{i})u_i + \epsilon_{ij}
\]

(17)

with \(x_{ij}^1\) and \(x_{i}^2\) computed using spline basis matrix for alternative exposure\(_{ij}\) and reference exposure\(_i\), (see Section 2.5.1), \(u_i\) the random effect for study \(i\), and \(\epsilon_{ij} \sim N(0, \sigma_{ij}^2)\) the variance reported by the \(i\)th study for its \(j\)th point.

\texttt{limeTR} is the only package that can infer the nonlinear dose-response relationship in this example using heterogeneous observations of ratios of average risks across variable exposures. The
meta-analysis to obtain the risk curve in Figure 6 integrates results from 14 prospective cohort studies. An ensemble of spline curves was used rather than a single choice of knot placement. The cone of uncertainty coming from the spline coefficients is shown in dark gray in the left panel, with additional uncertainty from random effects heterogeneity shown in light gray.

The right panel of Figure 6 shows the data fit. Each point represents a ratio between risks integrated across two intervals, so in effect on four points that define these intervals. We choose to plot the point at the midpoint defined by these intervals for the visualization. As in other plots, the radius of each point is inversely proportional to the standard deviation reported for it by the study. The derivatives of each spline curve fit are plotted in green so that they can be compared to these average ratios on log-derivative space.

After studying the relationship and the data in Figure 6, it is hard to justify forcing a monotonic increase in risk. However, to test the functionality of limeTR, we add this constraint and show the result in Figure 7. Our conclusions about the relationship and its strength would change in this example if we impose such shape constraints, unlike the example that compares smoking and lung cancer.

\[^{2}\text{Variation in risk here is much lower than for smoking, so we show relative risk instead of log relative risk.}\]
4 Conclusion

We have developed a new methodology for robust mixed effects models, and implemented it using the limeTR package. The package extends the trimming concept widely used in other robust statistical models to mixed effects. It solves the resulting problem using a new method that combines a standalone optimizer IPopt with a customized solver for the value function over the trimming parameters introduced in the reformulation. Synthetic examples show that limeTR is significantly more robust to outliers than available packages for meta-analysis, and also improves on the performance of packages for robust mixed effects regression/longitudinal analysis.

In addition to its robust functionality, limeTR includes additional features that are not available in other packages, including arbitrary nonlinear functions of fixed effects $F_i(\beta)$, as well as linear and nonlinear constraints. In the section that uses empirical data, we have shown how these features can be used to do standard meta-analysis as well as to infer nonlinear dose-response relationships from direct and indirect observations of complex nonlinear relationships.

References

Andreas Alfons, Christophe Croux, Sarah Gelper, et al. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, 7(1):226–248, 2013.

Jana J Anderson, Narisa DM Darwis, Daniel F Mackay, Carlos A Celis-Morales, Donald M Lyall, Naveed Sattar, Jason MR Gill, and Jill P Pell. Red and processed meat consumption and breast cancer: Uk biobank cohort study and meta-analysis. *European Journal of Cancer*, 90:73–82, 2018.

Aleksandr Y. Aravkin and Damek Davis. Trimmed statistical estimation via variance reduction. *Mathematics of Optimization Research*, 2019.

Aleksandr Y Aravkin and Tristan Van Leeuwen. Estimating nuisance parameters in inverse problems. *Inverse Problems*, 28(11):115016, 2012.

Aleksandr Y. Aravkin, Dmitriy Drusvyatskiy, and Tristan van Leeuwen. Variable projection without smoothness. *arXiv preprint arXiv:1601.05011*, 2016.
Aleksandr Y Aravkin, Dmitriy Drusvyatskiy, and Tristan van Leeuwen. Efficient quadratic penalization through the partial minimization technique. *IEEE Transactions on Automatic Control*, 63(7):2131–2138, 2018.

Douglas Bates, Martin Mchler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. ISSN 1548-7660. URL https://doi.org/article/7f279483412348928f01507440b0360d.

Bradley M Bell and James V Burke. Algorithmic differentiation of implicit functions and optimal values. In *Advances in Automatic Differentiation*, pages 67–77. Springer, 2008.

C Cox. Delta method. *Encyclopedia of biostatistics*, 2, 2005.

Carl De Boor, Carl De Boor, Etats-Unis Mathématicien, Carl De Boor, and Carl De Boor. *A practical guide to splines*, volume 27. springer-verlag New York, 1978.

Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188, 1986.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Jerome H Friedman et al. Multivariate adaptive regression splines. *The annals of statistics*, 19(1):1–67, 1991.

Sara Gandini, Edoardo Botteri, Simona Iodice, Mathieu Boniol, Albert B Lowenfels, Patrick Maisonneuve, and Peter Boyle. Tobacco smoking and cancer: A meta-analysis. *International journal of cancer*, 122(1):155–164, 2008.

Gene Golub and Victor Pereyra. Separable nonlinear least squares: the variable projection method and its applications. *Inverse problems*, 19(2):R1, 2003.

Gene H Golub and Victor Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on numerical analysis*, 10(2):413–432, 1973.
Aamer Imdad, Evan Mayo-Wilson, Kurt Herzer, and Zulfiqar A Bhutta. Vitamin a supplementation for preventing morbidity and mortality in children from six months to five years of age. *Cochrane Database of Systematic Reviews*, (3), 2017.

Manuel Koller. robustlmm: an r package for robust estimation of linear mixed-effects models. *Journal of statistical software*, 75(6):1–24, 2016.

Nan M Laird, James H Ware, et al. Random-effects models for longitudinal data. *Biometrics*, 38 (4):963–974, 1982.

Peter N Lee, Barbara A Forey, and Katharine J Coombs. Systematic review with meta-analysis of the epidemiological evidence in the 1900s relating smoking to lung cancer. *BMC cancer*, 12 (1):385, 2012.

David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. On the Least Trimmed Squares Estimator. *Algorithmica*, 69(1):148–183, 2014.

Neyko M Neykov and Christine H Müller. Breakdown Point and Computation of Trimmed Likelihood Estimators in Generalized Linear Models. In *Developments in robust statistics*, pages 277–286. Springer, 2003.

José C Pinheiro, Chuanhai Liu, and Ying Nian Wu. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics*, 10(2):249–276, 2001.

Natalya Pya and Simon N Wood. Shape constrained additive models. *Statistics and Computing*, 25(3):543–559, 2015.

GJM Rosa, Carlos R Padovani, and Daniel Gianola. Robust linear mixed models with normal/independent distributions and bayesian mcmc implementation. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 45(5):573–590, 2003.

Peter J Rousseeuw. Multivariate Estimation with High Breakdown Point. *Mathematical statistics and applications*, 8:283–297, 1985.
Peter J Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424):1273–1283, 1993.

Peter J Rousseeuw and Katrien Van Driessen. Computing LTS Regression for Large Data Sets. *Data mining and knowledge discovery*, 12(1):29–45, 2006.

Andreas Wächter and Lorenz T Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming*, 106(1):25–57, 2006.

E. Yang and A. Lozano. Robust Gaussian Graphical Modeling with the Trimmed Graphical Lasso. In *Advances in Neural Information Processing Systems*, pages 2602–2610, 2015.

Eunho Yang, Aurélie C Lozano, and Aleksandr Aravkin. A general family of trimmed estimators for robust high-dimensional data analysis. *Electronic Journal of Statistics*, 12(2):3519–3553, 2018a.

Eunho Yang, Aurélie C Lozano, and Aleksandr Y. Aravkin. A general family of trimmed estimators for robust high-dimensional data analysis. *Electronic Journal of Statistics*, 12(2):3519–3553, 2018b.

Alain Zuur, Elena N Ieno, Neil Walker, Anatoly A Saveliev, and Graham M Smith. *Mixed effects models and extensions in ecology with R*. Springer Science & Business Media, 2009.