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

January 11, 2020

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 of 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 using an open-source python code `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 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 computationally challenging; they are fit either by EM methods, or by estimating equation modifications, or by MCMC [Rosa et al., 2003]. In this paper, we take a very different tack, and extend the least trimmed squares (LTS) method to the ME setting.

Least trimmed squares, which has many advantages for basic regression, 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` `rlmer`	`clme`	`INLA`
Robust option	✓	✗	✓	✓	✗	✓
Allows for known observation variance	✓	✓	✓	✗	✗	✓
Covariates in random effects variance	✓	✗	✗	✓	✓	✗
Nonlinear observations	✓	✗	✗	✗	✗	✓
Linear constraints	✓	✗	✗	✗	✓	✗
Nonlinear constraints	✓	✗	✗	✗	✗	✗

Contributions. We formulate and solve the trimming problem required for the trimmed ME ap-

proach. 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 leverage the constrained extension to control their shape, particularly in regions where data is sparse.

The main code to perform the inference is published using an open source python package `LimeTR` (pronounced 'lime tree'). All synthetic experiments using `LimeTR` have been submitted for review as supplementary material with this paper. `LimeTR` is a significant contribution in its own right because it 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 3, 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 2.5, we present a customized algorithm based on variable projection, along with a convergence analysis. In Section 2.6, 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 real 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 real 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:

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{F}_i(\boldsymbol{\beta}) + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{\epsilon}_i \\ \boldsymbol{u}_i &\sim N(\boldsymbol{0}, \boldsymbol{\Gamma}), \quad \boldsymbol{\Gamma} = \operatorname{diag}(\boldsymbol{\gamma}), \quad \boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{\Lambda}), \end{aligned} \tag{1}$$

where $\boldsymbol{y}_i \in \mathbb{R}^{n_i}$ is the vector of observations from the ith group, $\boldsymbol{\epsilon}_i \in \mathbb{R}^{n_i}$ are measurement errors with covariance $\boldsymbol{\Lambda}$, $\boldsymbol{u}_i \in \mathbb{R}^{k_{\boldsymbol{\gamma}}}$ are independent random effects, and $\boldsymbol{Z}_i \in \mathbb{R}^{n_i \times k_{\boldsymbol{\gamma}}}$ is a linear map,

and $\boldsymbol{\beta}$ 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 $\boldsymbol{Z}_i\boldsymbol{u}_i$.

A range of assumptions may be placed on $\boldsymbol{\Lambda}$. In longitudinal analysis, $\boldsymbol{\Lambda}$ is often a diagonal or block-diagonal matrix, parametrized by set of shared unknown terms. In meta-regression and meta-analysis, $\boldsymbol{\Lambda}$ is a known diagonal matrix whose entries are variances for each input datum.

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

$$p(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\Lambda}, \boldsymbol{u}|\boldsymbol{y}) \propto \prod_{i=1}^{m} \|\boldsymbol{y}_i - \boldsymbol{F}_i(\boldsymbol{\beta}) - \boldsymbol{Z}_i\boldsymbol{u}\|^2_{\boldsymbol{\Lambda}^{-1}} \|\boldsymbol{u}\|^2_{\boldsymbol{\Gamma}^{-1}} \det(\boldsymbol{\Lambda}^{-1}) \det(\boldsymbol{\Gamma})^{-1} \tag{2}$$

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

$$\begin{aligned} \mathcal{L}_{ML}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Lambda}) &= -\ln\left(\int p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Lambda}, \boldsymbol{u}|\boldsymbol{y}) d\boldsymbol{u}\right) \\ &= \sum_{i=1}^{m} \frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{F}_i(\boldsymbol{\beta}))^\top (\boldsymbol{Z}_i\boldsymbol{\Gamma}\boldsymbol{Z}_i^\top + \boldsymbol{\Lambda}_i)^{-1}(\boldsymbol{y}_i - \boldsymbol{F}_i(\boldsymbol{\beta})) + \frac{1}{2}\ln|\boldsymbol{Z}_i\boldsymbol{\Gamma}\boldsymbol{Z}_i^\top + \boldsymbol{\Lambda}_i|. \end{aligned} \tag{3}$$

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

$$\boldsymbol{y}_i = \boldsymbol{F}_i(\boldsymbol{\beta}) + \boldsymbol{\omega}, \quad \boldsymbol{\omega} \sim N(\boldsymbol{0}, \boldsymbol{Z}_i\boldsymbol{\Gamma}\boldsymbol{Z}_i^\top + \boldsymbol{\Lambda}_i).$$

The structure of this objective depends on the structural assumptinos on $\boldsymbol{\Lambda}$. We restrict our numerical experiments to two particular classes: (1) $\boldsymbol{\Lambda} = \sigma^2\boldsymbol{I}$ with σ^2 unknown, used in standard longitudinal analysis, and (2) $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}_\epsilon$, a known matrix of observation covariances, used in meta-analysis and meta-regression.

2.2 Constraints and Priors

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

$$\boldsymbol{C}(\boldsymbol{\theta}) \leq \boldsymbol{c},$$

where $\boldsymbol{\theta}$ are any parameters of interest. Constraints play a key role in section 2.6, 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 $\boldsymbol{\theta}$. We assume that priors are given by a functional form

$$\boldsymbol{\theta} \sim \exp(-\rho(\boldsymbol{\theta}))$$

where ρ is smooth (but may be nonlinear and nonconvex). The likelihood problem is then augmented by adding the term $\rho(\boldsymbol{\theta})$ 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

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \frac{1}{2}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle)^2, \tag{4}$$

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 $n^{-1/2}$).[1] LTS estimators are robust against outliers, and arbitrarily large deviations that are trimmed do not affect the final $\hat{\boldsymbol{\beta}}$.

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

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

The set

$$\Delta_h := \left\{ \boldsymbol{w} : \mathbf{1}^\top \boldsymbol{w} = h, \quad \mathbf{0} \leq \boldsymbol{w} \leq \mathbf{1} \right\} \tag{6}$$

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 $\boldsymbol{\beta}$, the optimal solution of (5) with respect to $\boldsymbol{w}$ assigns weight 1 to each of the smallest h residuals, and 0 to the rest. Problem (5) is solved *jointly* in $(\boldsymbol{\beta}, \boldsymbol{w})$, simultaneously finding the regression estimate and classifying the observations

[1] 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 (features).

into inliers and outliers. This joint strategy makes LTS different from post-hoc analysis, where a model a fit first with all data, and then outliers are detected using that estimate.

Several approaches for finding LTS and other trimmed M-estimators have been developed. Rousseeuw and Van Driessen [2006] developed the FAST-LTS algorithm, which was able to find LTS estimators faster than existing algorithms for LMS estimations. Later, Mount et al. [2014] introduced an exact algorithm for computing LTS, which suffered from exponential complexity in higher dimensional problems. Moreover, the LTS strategy (5) does not depend on the form of the least squares function. We can replace each $\left(\frac{1}{2}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)^2\right)$ by an abstract data term $f_i(\boldsymbol{\beta})$. This insight has been used to extend LTS to a broad range of estimation problems, including generalized linear models [Neykov and Müller, 2003], high dimensional sparse regression [Alfons et al., 2013], and graphical lasso [Yang and Lozano, 2015, Yang et al., 2018a]. The most general problem class to date, presented by Aravkin and Davis [2019], is formulated as

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

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*:

$$\begin{aligned} \min_{\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\Lambda},\boldsymbol{w}} & \sum_{i=1}^{m} w_i \left(\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{F}_i(\boldsymbol{\beta}))^\top (\boldsymbol{Z}_i \boldsymbol{\Gamma} \boldsymbol{Z}_i^\top + \boldsymbol{\Lambda}_i)^{-1} (\boldsymbol{y}_i - \boldsymbol{F}_i(\boldsymbol{\beta})) + \frac{1}{2} \ln |\boldsymbol{Z}_i \boldsymbol{\Gamma} \boldsymbol{Z}_i^\top + \boldsymbol{\Lambda}_i| \right) \\ \text{s.t.} \quad & \mathbf{1}^\top \boldsymbol{w} = h, \quad \mathbf{0} \leq \boldsymbol{w} \leq \mathbf{1}. \end{aligned} \tag{8}$$

This is severely 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):

$$\left(\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{F}_i(\boldsymbol{\beta}))^\top (\boldsymbol{Z}_i \boldsymbol{\Gamma} \boldsymbol{Z}_i^\top + \boldsymbol{\Lambda}_i)^{-1} (\boldsymbol{y}_i - \boldsymbol{F}_i(\boldsymbol{\beta})) + \frac{1}{2} \ln |\boldsymbol{Z}_i \boldsymbol{\Gamma} \boldsymbol{Z}_i^\top + \boldsymbol{\Lambda}_i| \right)$$

Here, $\boldsymbol{y}_i \in \mathbb{R}^{n_i}$, where n_i is the number of observations in the ith group. To trim observations within the group, we introduce auxiliary variables $\boldsymbol{w}_i \in \mathbb{R}^{n_i}$, and define

$$\boldsymbol{r}_i := \boldsymbol{y}_i - \boldsymbol{F}_i(\boldsymbol{\beta}), \quad \boldsymbol{W}_i := \operatorname{diag}(\boldsymbol{w}_i), \quad \sqrt{\boldsymbol{W}_i} := \operatorname{diag}(\sqrt{\boldsymbol{w}_i}).$$

We now form the objective

$$\frac{1}{2}\boldsymbol{r}_i^\top \sqrt{\boldsymbol{W}_i}\left(\sqrt{\boldsymbol{W}_i}\boldsymbol{Z}_i\boldsymbol{\Gamma}\boldsymbol{Z}_i^\top\sqrt{\boldsymbol{W}_i}+\boldsymbol{\Lambda}_i^{\odot \boldsymbol{w}_i}\right)^{-1}\sqrt{\boldsymbol{W}_i}\boldsymbol{r}_i+\frac{1}{2}\ln\left|\sqrt{\boldsymbol{W}_i}\boldsymbol{Z}_i\boldsymbol{\Gamma}\boldsymbol{Z}_i^\top\sqrt{\boldsymbol{W}_i}+\boldsymbol{\Lambda}_i^{\odot \boldsymbol{w}_i}\right|, \tag{9}$$

where $^{\odot}$ denotes the elementwise power operation:

$$\boldsymbol{\Lambda}_i^{\odot \boldsymbol{w}_i} := \begin{bmatrix} (\lambda_{1j})^{w_{i1}} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & (\lambda_{in_i})^{w_{in_i}}. \end{bmatrix} \tag{10}$$

When $w_{ij}=1$, we recover the contribution of the ijth observation to the original likelihood. As $w_{ij}\downarrow 0$, The ijth contribution to the residual is correctly eliminated by $\sqrt{w_{ij}}\downarrow 0$. The jth row and column of $\sqrt{\boldsymbol{W}_i}\boldsymbol{Z}_i\boldsymbol{\Gamma}\boldsymbol{Z}_i^\top\sqrt{\boldsymbol{W}_i}$ both go to 0, while the jth entry of $\boldsymbol{\Lambda}_i^{\odot \boldsymbol{w}_i}$ goes to 1, which effectively removes all impact of the jth point on the covariance matrix.

2.4 General trimmed estimators for MEs.

Putting together the trimmed ML with priors and constraints, we arrive at the following estimator. The trimmed constrained regularized ML estimator is obtained by solving

$$\begin{aligned} \min_{\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\Lambda},\boldsymbol{w}} \mathcal{L}(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\Lambda},\boldsymbol{w}) := & \sum_{i=1}^m \frac{1}{2}\boldsymbol{r}_i^\top\sqrt{\boldsymbol{W}_i}\left(\sqrt{\boldsymbol{W}_i}\boldsymbol{Z}_i\boldsymbol{\Gamma}\boldsymbol{Z}_i^\top\sqrt{\boldsymbol{W}_i}+\boldsymbol{\Lambda}_i^{\odot \boldsymbol{w}_i}\right)^{-1}\sqrt{\boldsymbol{W}_i}\boldsymbol{r}_i + \\ & \frac{1}{2}\ln\left|\sqrt{\boldsymbol{W}_i}\boldsymbol{Z}_i\boldsymbol{\Gamma}\boldsymbol{Z}_i^\top\sqrt{\boldsymbol{W}_i}+\boldsymbol{\Lambda}_i^{\odot \boldsymbol{w}_i}\right| + \rho(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\Lambda}) \\ & \text{s.t.} \quad \boldsymbol{r}_i = \boldsymbol{y}_i - \boldsymbol{F}_i(\boldsymbol{\beta}), \quad \mathbf{1}^\top\boldsymbol{w}=h, \quad \mathbf{0}\le\boldsymbol{w}\le\mathbf{1}, \quad \boldsymbol{C}\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\\ \boldsymbol{\Lambda}\end{pmatrix}\le\boldsymbol{c}. \end{aligned} \tag{11}$$

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.

2.5 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 $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Lambda})$ using an interior point method, and then optimize the resulting value function with respect to $\boldsymbol{w}$ 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 $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Lambda})$, the implicit solution $\boldsymbol{\theta}(\boldsymbol{w})$ and value function $v(\boldsymbol{w})$ as follows:

$$\begin{aligned} \boldsymbol{\theta}(\boldsymbol{w}) &:= \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{w}) \quad \text{s.t.} \quad \boldsymbol{C}(\boldsymbol{\theta}) \leq \boldsymbol{c} \\ v(\boldsymbol{w}) &:= \min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{w}) \quad \text{s.t.} \quad \boldsymbol{C}(\boldsymbol{\theta}) \leq \boldsymbol{c} \end{aligned} \tag{12}$$

where $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{w})$ 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 $v(\boldsymbol{w})$ in (12). Suppose that for any $\boldsymbol{\theta}(\boldsymbol{w})$, we have*

$$\begin{bmatrix} \nabla^2_{\boldsymbol{\theta}} \mathcal{L}|_{\boldsymbol{\theta}(\boldsymbol{w}), \boldsymbol{w}} & \nabla \boldsymbol{C}^\top_{\boldsymbol{\theta}(\boldsymbol{w})} \\ \nabla \boldsymbol{C}|_{\boldsymbol{\theta}(\boldsymbol{w})} & 0 \end{bmatrix}$$

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

$$\nabla v(\boldsymbol{w}) = -\partial_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{w})|_{(\boldsymbol{\theta}(\boldsymbol{w}), \boldsymbol{w})},$$

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

$$\min_{\boldsymbol{w}} v(\boldsymbol{w}) \quad \text{s.t.} \quad \mathbf{1}^\top \boldsymbol{w} = h, \quad \mathbf{0} \leq \boldsymbol{w} \leq \mathbf{1}, \tag{13}$$

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

$$\boldsymbol{w}^+ = \text{proj}_{\Delta_h}(w - \alpha \nabla v(\boldsymbol{w})). \tag{14}$$

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

Algorithm 1 Projected gradient descent on the Value Function v of (12)

1: **Input:** $\boldsymbol{w}_0, \lambda_{\boldsymbol{w}}$.

2: **Initialize:** $\nu = 0$

3: **while** not converged **do**

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

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

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

7: **Output:** $\boldsymbol{w}_{\nu}, \boldsymbol{\theta}_{\nu}$

Step 5 of Algorithm 1 requires solving the constrained likelihood problem (11) with $\boldsymbol{w}$ 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 $\boldsymbol{\theta}$ and $\boldsymbol{w}$ differently is key to efficient performance. Typically $\boldsymbol{\theta}$ is small compared to $\boldsymbol{w}$, which is the size of the data. On the other hand the constrained likelihood problem in $\boldsymbol{\theta}$ is difficult while constrained value function optimization over $\boldsymbol{w}$ can be solved with projected gradient.

2.6 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, that allow us to analyze effects of risks with exposure (e.g. smoking, BMI, consumption) on adverse outcomes. For an in-depth look at splines and spline regression see De Boor et al. [1978] and Friedman et al. [1991].

The use of constraints is essential in this setting to capture expert knowledge on the shape of such risk curves, particularly in segments informed by sparse data.

2.6.1 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

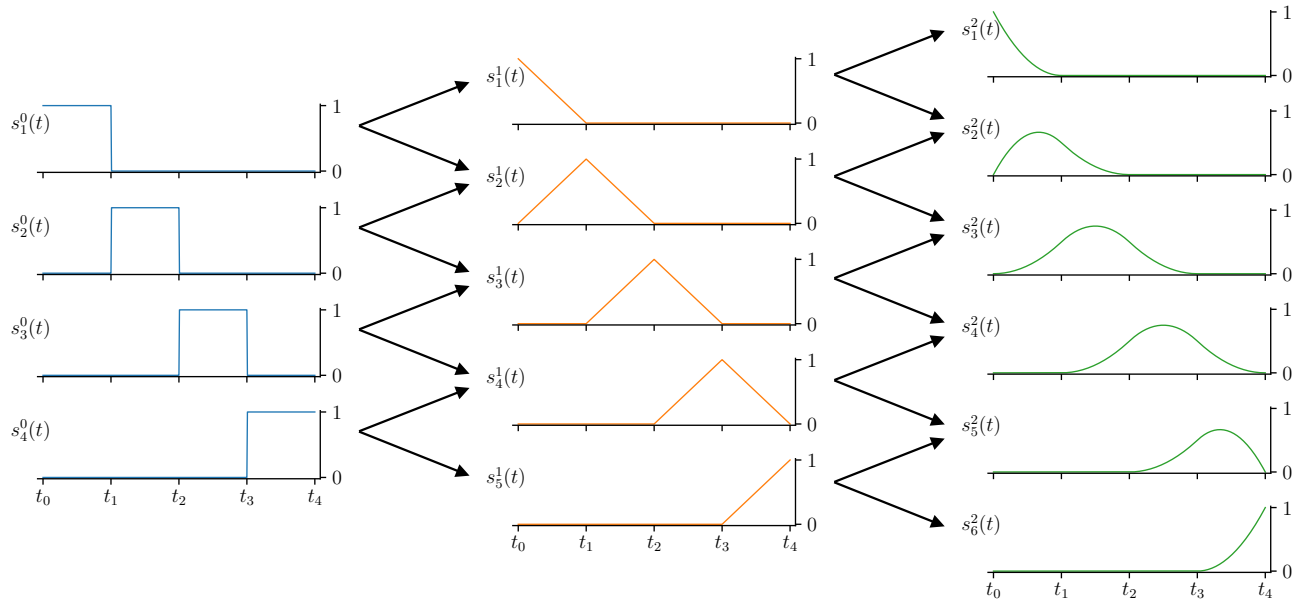

Figure 1: Recursive generation of bspline basis elements (orders 0, 1, 2).

basis elements, with coefficients $\boldsymbol{\beta} \in \mathbb{R}^{p+k}$:

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

These coefficients are inferred by `LimeTR` analysis. A more standard explicit representation of (15) is obtained by building a design matrix $\boldsymbol{X}$. Given a set of t values at which we have data, the jth column of $\boldsymbol{X}$ is given by the expression

$$\boldsymbol{X}_{\cdot,j} = \begin{bmatrix} s_j^p(t_0) \\ \vdots \\ s_j^p(t_k) \end{bmatrix}.$$

The model for observed data coming from (15) can now be written compactly as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{\epsilon}_i,$$

which is a special case of the main problem class (1).

2.6.2 Shape constraints

We can impose shape constraints such as monotonicity, concavity, and convexity on splines. Constraints on splines have been developed in the past, see e.g. [Pya and Wood, 2015]. However,

the authors took significant pains to avoid using explicit constraints, opting to re-formulate the problem using exponentials. The development in this section uses simple and explicit constrained formulations.

Monotonicity. Spline monotonicity across the domain of interest follows from monotonicity of the spline coefficients De Boor et al. [1978]. Given coefficients

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix},$$

we know the curve $f(t)$ in (15) is monotonically nondecreasing when

$$\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_n$$

and *monotonically non-increasing* if

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n.$$

The relationship $\alpha_1 \leq \alpha_2$ can be written as $\alpha_1 - \alpha_2 \leq 0$. Stacking these inequality constraints for each pair (α_i, α_{i+1}) we can write all constraints simultaneously as

$$\underbrace{\begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 1 & -1 \end{bmatrix}}_{\boldsymbol{C}} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_n \end{bmatrix} \leq \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

These linear constraints are a special case of the general estimator (11) that allows $\boldsymbol{C}(\boldsymbol{\beta}) \leq \boldsymbol{c}_{\boldsymbol{\beta}}$.

2.6.3 Convexity and Concavity

For any $\mathcal{C}^2$ (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, an 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 $\boldsymbol{\beta}$ (including bounds) through the interface of `limeTR`.

2.6.4 Nonlinear measurements

Some of the studies in the real data verifications use nonlinear observation mechanisms. For example, given a dose-response curve

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

studies often report odds of an outcome between exposed and unexposed groups that are defined across two intervals on the underlying curve:

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

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

$$y_i = \ln(\langle \boldsymbol{x}_i^1, \boldsymbol{\beta} \rangle) - \ln(\langle \boldsymbol{x}_i^2, \boldsymbol{\beta} \rangle) := F_i(\boldsymbol{\beta}),$$

a particularly useful example of the general nonlinear term $\boldsymbol{F}_i(\boldsymbol{\beta})$ in problem class (1).

2.7 Variance Estimation

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

The 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, the empirical bootstrap can be applied to sample *modeled* residuals. Having obtained the estimate $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\gamma}})$, we can sample model-based errors and get new bootstrap realizations $\bar{\boldsymbol{y}}$ as follows:

$$\bar{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\bar{\boldsymbol{u}} + \bar{\boldsymbol{\epsilon}},$$

where $\boldsymbol{\epsilon}_i \sim N(0, \hat{\Lambda})$ and $\boldsymbol{u}_i \sim N(0, \hat{\gamma})$. These realizations have the same structure as the input data, and reflect the uncertainty from the estimated variance parameters. For each realization

$\bar{\boldsymbol{y}}$, we then re-run the fit, and obtain N estimates $\{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\gamma}})\}_{1:N}$. This set of estimates is used to estimate the variance of the fitting procedure along with any confidence bounds.

3 Verifications

In this section we validate `limeTR` on synthetic and real 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. In particular 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 we need to consider shape constraints and nonlinear measurements, in addition to outlier robustness.

3.1 Validation Using Synthetic Data

3.2 Real-World Case Studies

3.2.1 Ratio model for any outcome

4 Conclusion

SUPPLEMENTAL MATERIALS

Package: Python package `LimeTR` that contains code to perform the analyses in the article.

Comparison Examples: R-code used to perform the comparisons in the validation section.

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.

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 Mächler, 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://doaj.org/article/7f27948341234892f01507440b0360d`.

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.

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.

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.

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.