Trimmed Constrained Mixed Effects Models: Formulations and Algorithms

Peng Zheng[†], Ryan Barber[†], Reed Sorensen[†],
Christopher Murray[†], and Aleksandr Aravkin[°†] *

Institute for Health Metrics and Evaluation, University of Washington[†]

Department of Applied Mathematics, University of Washington[°]

December 2, 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 broad 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.

The software accompanying this paper is disseminated as an open-source Python package called `LimeTr` . `LimeTr` is able to recover results more accurately in the presence of outliers compared to available packages for both standard longitudinal analysis and meta-analysis, and is also more computationally efficient than competing robust alternatives. Supplementary materials that reproduce the simulations, as well as run `LimeTr` and third party code are available online. We also present analyses of global health data, where we use advanced functionality of `LimeTr` , including constraints to impose monotonicity and concavity for dose-response relationships. Nonlinear observation models allow new analyses in place of classic approximations, such as log-linear models. Robust extensions in all analyses ensure that spurious data points do not drive our understanding of either mean relationships or between-study heterogeneity.

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

*This work was funded by the Bill & Melinda Gates Foundation. The authors also gratefully acknowledge the Washington Research Foundation Data Science Professorship. The authors are very grateful to the referees for insightful questions and suggestions that have improved exposition and extended the scope of the paper.

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].

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 track, and extend the least trimmed squares (LTS) method to the ME setting. While LTS has found wide use in a range of applications [Aravkin and Davis, 2019, Yang and Lozano, 2015, Yang et al., 2018], trimming the ME likelihood extends prior work.

Contributions. In this paper, we consider a subclass of nonlinear mixed effects models. We allow nonlinear measurements, priors, and constraints, but require that the random effects enter the model in a linear way. We call this class *partially nonlinear* ME models, and it covers a broad class of problems while allowing tractable algorithms. We develop new conditions that guarantee the existence of estimators for partially nonlinear models, a trimming approach that robustifies any linear or partially nonlinear model against outliers, and algorithms for solving the nonconvex optimization problems required to find estimates with standard guarantees (convergence to stationary points). We also show splines (and associated shape constraints) can be used to capture key nonlinear relationships, and illustrate the full modeling capability on real-data examples based on dose-response relationships.

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 2.2, we describe the problem class of ME models and derive the marginal maximum likelihood (ML) estimator. In Section 2.3, we describe how constraints and priors are imposed on parameters. In Section 2.4, we review trimming approaches and develop a new trimming extension for the ML approach. In Section 2.5, we present a cus-

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

	`LimeTr`	`metafor`	`robumeta` `metaplus`	`robustlmm` `heavy`	`clme`	`INLA`
Robust option	✓	✗	✓	✓	✗	✓
Allows for known observation variance	✓	✓	✓	✗	✗	✓
Covariates in random effects variance	✓	✗	✗	✓	✓	✗
Nonlinear observations	✓	✗	✗	✗	✗	✓
Linear constraints	✓	✗	✗	✗	✓	✗
Nonlinear constraints	✓	✗	✗	✗	✗	✗

tomized 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 of 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 Notation and Modeling Concepts

In this section, we define notation and concepts used throughout the paper. Additional definitions and notation are introduced in the analysis section. We use lower case letters to denote scalars, e.g. β, and scalar-valued functions, e.g. $f(\beta)$, and bold letters represent vectors, e.g. $\boldsymbol{\beta}$, and vector-valued functions, e.g. $\boldsymbol{f}_i(\boldsymbol{\beta})$. We use capital bold letters to represent matrices, e.g. $\boldsymbol{X}$. All variables and vectors are real, i.e. in $\mathbb{R}^k$, with k indicating dimension. For a smooth function $f : \mathbb{R}^k \to \mathbb{R}$, we denote the vector of first derivatives, or **gradient**, by ∇f, and the matrix of second

derivatives, or **Hessian**, by $\nabla^2 f$. We use $\mathrm{diag}(\boldsymbol{x})$ to denote a diagonal matrix whose diagonal is an input vector $\boldsymbol{x}$. We use $\boldsymbol{M}^{-1}$ to denote the inverse of a matrix, and denote weighted norms (Mahalanobis distances) by

$$\|\boldsymbol{x}\|_{\boldsymbol{M}}^2 := \boldsymbol{x}^T \boldsymbol{M} \boldsymbol{x},$$

and $|\mathbf{M}|$ to denote the determinant of $\mathbf{M}$. We use the $\odot$ notation to denote the Hadamard product or operation, so in particular

$$\boldsymbol{x} \odot \boldsymbol{y} = \begin{bmatrix} x_1 y_1 \\ \vdots \\ x_n y_n \end{bmatrix}, \quad \boldsymbol{x}^{\odot \boldsymbol{y}} = \begin{bmatrix} x_1^{y_1} \\ \vdots \\ x_n^{y_n} \end{bmatrix}.$$

A **likelihood** maps parameters to an associated density function for observed data. In the mixed effects context, these parameters can be separated into **fixed effects** (e.g. population mean) and **random effects** (e.g. study-specific random intercept). A **marginal** likelihood function refers to the likelihood obtained by integrating out random effects from the joint likelihood.

We incorporate additional information about parameters using statistical priors, and restrict parameter domains using **constraints**. When constraints are present we use the term **constrained likelihood**. We use **trimming** (Section 2.4) to robustify a (marginal) likelihood, and use the term **trimmed likelihood** to describe such likelihoods.

The goal of an inference problem is to maximize the (marginal) likelihood or modified likelihood, or equivalently to minimize the negative logarithm of such a function, called an **objective function** in optimization. The optimization problem specification includes constraints as well as the objective. We use min to refer to the minimum value of an optimization problem, and arg min to refer to the minimizer, which corresponds to the **estimator** in this setting.

We define projection of a point $\boldsymbol{x}$ onto a closed set C by

$$\mathrm{proj}_C(\boldsymbol{x}) := \arg\min_{\boldsymbol{y} \in C} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

2.2 Problem Class

We consider the following mixed effects model:

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

where m is the number of groups, $\boldsymbol{y}_i \in \mathbb{R}^{n_i}$ is the vector of observations from the ith group, and $n = \sum_i n_i$ is the total number of observations. Measurement errors are denoted by $\boldsymbol{\epsilon}_i \in \mathbb{R}^{n_i}$, with covariance $\boldsymbol{\Lambda}_i \in \mathbb{R}^{n_i \times n_i}$, and we denote by $\boldsymbol{\Lambda} \in \mathbb{R}^{n \times n}$ the full block diagonal measurement error covariance matrix. Regression coefficients are denoted by $\boldsymbol{\beta} \in \mathbb{R}^{k_\beta}$. Random effects are denoted by $\boldsymbol{u}_i \in \mathbb{R}^{k_\gamma}$, where $\boldsymbol{Z}_i \in \mathbb{R}^{n_i \times k_\gamma}$ are linear maps. The functions $\boldsymbol{f}_i$: $\mathbb{R}^{k_\beta} \to \mathbb{R}^{n_i}$ may be nonlinear, but we restrict the random effects to enter in a linear way through the linear maps $\boldsymbol{Z}_i : \mathbb{R}^{k_\gamma} \to \mathbb{R}^{n_i}$.

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 a small set of parameters, with the simplest example $\boldsymbol{\Lambda} = \sigma^2 \boldsymbol{I}$ where σ^2 is unknown. In meta-analysis, $\boldsymbol{\Lambda}$ is a known diagonal matrix whose entries are variances for each input datum. We do not restrict the term 'meta-analysis' to a single observation per study, since many analyses include multiple observations, such as summary results for quartiles based on exposure. The distinguishing feature of meta-analysis is the specification of a known $\boldsymbol{\Lambda}$ matrix.

For convenience, we denote by $\boldsymbol{\theta}$ the tuple of fixed parameters:

$$\boldsymbol{\theta} := (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Lambda}).$$

The joint likelihood corresponding to model (1) for $\boldsymbol{\theta}$ and random effects $\boldsymbol{u}$ is given by

$$p(\boldsymbol{\theta}, \boldsymbol{u} | \boldsymbol{y}) = \frac{\exp\left(-\frac{1}{2}\|\boldsymbol{u}\|^2_{\boldsymbol{\Gamma}^{-1}}\right)}{\sqrt{|2\pi\boldsymbol{\Gamma}|}} \prod_{i=1}^{m} \frac{\exp\left(-\frac{1}{2}\|\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\beta}) - \boldsymbol{Z}_i \boldsymbol{u}\|^2_{\boldsymbol{\Lambda}^{-1}}\right)}{\sqrt{|2\pi\boldsymbol{\Lambda}_i|})} \tag{2}$$

Maximizing (2) with respect to both fixed and random parameters is problematic, as the number of random parameters $\boldsymbol{u}_i$ grows with the number of groups. In the extreme case of one observation per group, there are more unknowns than datapoints. Standard practice is to marginalize random effects, integrating (2) with respect to all $\boldsymbol{u}_i$. The numerical estimation is then accomplished by taking the negative logarithm (a simplifying transformation) and minimizing the result, which is

an equivalent problem to maximizing the marginal likelihood:

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

Problem (3) is equivalent to a maximum likelihood formulation arising from a 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 assumptions on $\boldsymbol{\Lambda}$. In the scope of this paper we always assume $\boldsymbol{\Lambda}$ is a diagonal matrix, namely all the measurement error are independent with each other. 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) the measurements variance is provided and used in meta-analysis. `LimeTr` allows other structure options as well, for example group-specific unknown σ_i^2 that extends case (1), but we consider a simple set of synthetic results to help focus on robust capabilities.

2.3 Constraints and Priors

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

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

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 follow a distribution defined by the density function

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

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 (3). The most common use case is $\rho(\cdot)=\frac{1}{2\sigma_p^2}\|\cdot\|^2$, for some user-defined σ_p^2.

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

2.4 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. Starting from a standard least squares estimator,

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

the LTS approach modifies (4) to minimize 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 estimate.

The explicit LTS extension to (4) is formed by introducing auxiliary variables $\boldsymbol{w}$:

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

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 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 [Rousseeuw and Van Driessen, 2006], and exact algorithms with exponential complexity [Mount et al., 2014]. 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 [Neykov and Müller, 2003], high dimensional sparse regression [Alfons et al., 2013], and graphical lasso [Yang and Lozano, 2015, Yang et al., 2018].

[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.

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).

Problem (3) is not of the form (7), except for the special case where we want to detect *entire outlying groups*. This is limiting, since we want to differentiate measurements within groups. We solve the problem by using a new trimming formulation.

To explain the approach we focus on trimming a single group term from the 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

$$\frac{1}{2} \boldsymbol{r}_i^\top \sqrt{\boldsymbol{W}_i} \left(\sqrt{\boldsymbol{W}_i} \boldsymbol{Z}_i \boldsymbol{\Gamma} \boldsymbol{Z}_i^\top \sqrt{\boldsymbol{W}_i} + \boldsymbol{\Lambda}_i^{\odot \boldsymbol{w}_i}\right)^{-1} \sqrt{\boldsymbol{W}_i} \boldsymbol{r}_i + \frac{1}{2} \ln \left| \sqrt{\boldsymbol{W}_i} \boldsymbol{Z}_i \boldsymbol{\Gamma} \boldsymbol{Z}_i^\top \sqrt{\boldsymbol{W}_i} + \boldsymbol{\Lambda}_i^{\odot \boldsymbol{w}_i} \right|, \tag{8}$$

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

$$\boldsymbol{\Lambda}_i^{\odot \boldsymbol{w}_i} := \begin{bmatrix} (\lambda_{i1})^{w_{i1}} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & (\lambda_{in_i})^{w_{in_i}}. \end{bmatrix} \tag{9}$$

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 removes all impact of the jth point. Specifically, the matrix $\sqrt{\boldsymbol{W}_i} \boldsymbol{Z}_i \boldsymbol{\Gamma} \boldsymbol{Z}_i^\top \sqrt{\boldsymbol{W}_i} + \boldsymbol{\Lambda}_i^{\odot \boldsymbol{w}_i}$ with $w_{ij} = 0$ and all remaining $w_{i\cdot} = 1$ is the same as the matrix $\boldsymbol{Z}_i \boldsymbol{\Gamma} \boldsymbol{Z}_i^\top + \boldsymbol{\Lambda}_i$ obtained after deleting the ij-th point.

Combining trimmed ML with priors and constraints, we obtain the following modified log-likelihood:

$$\begin{aligned}\min_{\boldsymbol{\theta},\boldsymbol{w}} \mathcal{L}(\boldsymbol{\theta},\boldsymbol{w}) := & \sum_{i=1}^{m} \frac{1}{2}\boldsymbol{r}_i^\top \sqrt{\boldsymbol{W}_i}\left(\sqrt{\boldsymbol{W}_i}\boldsymbol{Z}_i\boldsymbol{\Gamma}\boldsymbol{Z}_i^\top\sqrt{\boldsymbol{W}_i}+\boldsymbol{\Lambda}_i^{\odot \boldsymbol{w}_i}\right)^{-1}\sqrt{\boldsymbol{W}_i}\boldsymbol{r}_i + \\ & \frac{1}{2}\ln\left|\sqrt{\boldsymbol{W}_i}\boldsymbol{Z}_i\boldsymbol{\Gamma}\boldsymbol{Z}_i^\top\sqrt{\boldsymbol{W}_i}+\boldsymbol{\Lambda}_i^{\odot \boldsymbol{w}_i}\right| + \rho(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\Lambda}) \\ & \text{s.t.} \quad \boldsymbol{r}_i = \boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\beta}), \quad \mathbf{1}^\top \boldsymbol{w} = h, \quad \mathbf{0} \le \boldsymbol{w} \le \mathbf{1}, \quad \boldsymbol{C}\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\\ \boldsymbol{\Lambda}\end{pmatrix} \le \boldsymbol{c}.\end{aligned} \tag{10}$$

Problem (10) has not been previously considered in the literature. We present a specialized algorithm and analysis in the next section.

2.5 Fitting Trimmed Constrained MEs: Algorithm and Analysis

Problem (10) 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:

$$\begin{aligned}\boldsymbol{\theta}(\boldsymbol{w}) &:= \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta},\boldsymbol{w}) \quad \text{s.t.} \quad \boldsymbol{C}(\boldsymbol{\theta}) \le \boldsymbol{c} \\ v(\boldsymbol{w}) &:= \min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta},\boldsymbol{w}) \quad \text{s.t.} \quad \boldsymbol{C}(\boldsymbol{\theta}) \le \boldsymbol{c}\end{aligned} \tag{11}$$

where $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{w})$ is given in (10). The term $\boldsymbol{\theta}(\boldsymbol{w})$ refers to the entire set of minimizers for a given w, which may not necessarily be a singleton.

We first develop conditions and theory to guarantee the existence of minimizers $\boldsymbol{\theta}(\boldsymbol{w})$. While there are results in the literature for particular classes of linear mixed effects models, conditions for the existence of minimizers for the partial nonlinear case (10) have not been derived. Both Harville [2018] and Davidian and Giltinan [1995] analyze the Aitken model, where $\boldsymbol{\Lambda} = \sigma^2\boldsymbol{H}$ for a nonsingular $\boldsymbol{H}$, by essentially deriving the closed form estimates for $\boldsymbol{\theta}$ in this case, where the conditions

that the residual is not exactly 0 guarantees a positive estimator for σ^2. In the nonlinear case (10), this is not possible, and to guarantee existence of minimizers we have to obtain some conditions for model validity.

Let $\mathcal{S}_{++}$ denote the set of positive definite matrices, and for $\boldsymbol{M} \in \mathcal{S}_{++}$ consider the function

$$f(\boldsymbol{M}, \boldsymbol{r}) = \boldsymbol{r}^T \boldsymbol{M}^{-1} \boldsymbol{r} + \ln |\boldsymbol{M}|. \tag{12}$$

To connect the general functional form (12) with the problem (10), we specify functional domains for feasible $\boldsymbol{r}$ amd $\boldsymbol{M}$.

Definition 1 (Domains)**.** *We define domains $\mathcal{D}_r$ for $\boldsymbol{r}$ and $\mathcal{D}_M$ for $\boldsymbol{M}$ as follows:*

$$\mathcal{D}_r := \{\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\beta}) : \forall \boldsymbol{\beta}\},$$
$$\mathcal{D}_M := \{\boldsymbol{Z}\mathrm{Diag}(\boldsymbol{\gamma})\boldsymbol{Z}^\top + \boldsymbol{\Lambda}(\boldsymbol{\sigma}) : \forall \boldsymbol{\gamma} \geq 0, \forall \boldsymbol{\sigma} > 0\}.$$

Using these definitions, we can write our objective as

$$\min_{\boldsymbol{r} \in \mathcal{D}_r, \boldsymbol{M} \in \mathcal{D}_M} f(\boldsymbol{r}, \boldsymbol{M}) \tag{13}$$

for f in (12). We also need to define the **level set** of a function.

Definition 2 (Level set)**.** *The α-level set of a function $f : \mathbb{R}^k \to \mathbb{R}$, denoted $\mathcal{L}_{f,\alpha}$, is defined by*

$$\mathcal{L}_{f,\alpha} := \{\boldsymbol{\theta} : f(\boldsymbol{\theta}) \leq \alpha\}$$

To guarantee the existence of estimators ($\boldsymbol{\theta}$) we make two assumptions.

Assumption 1. *We assume that the image of $\boldsymbol{f}$ is closed. This implies that $\mathcal{D}_r$ is closed.*

Assumption 2. *Denote the eigenvalue decomposition for any $\boldsymbol{M} \in \mathcal{D}_m$ by $\boldsymbol{M} = \boldsymbol{X}_M Diag(\boldsymbol{\lambda}_M)\boldsymbol{X}_M^\top$. For $\forall \boldsymbol{r} \in \mathcal{D}_r$ and $\forall \boldsymbol{M} \in \mathcal{D}_M$, we assume there exist $\alpha > 0$ such that*

$$\min_i \max(|\tilde{r}_i|, \lambda_i) \geq \alpha \tag{14}$$

where $\tilde{\boldsymbol{r}} = X_M^\top \boldsymbol{r}$ and λ_i is the corresponding eigenvalue.

Assumption 2 tells us that, after being pre-whitened by the covariance matrix, either the residual is bounded away from 0 or else the corresponding eigenvalue is bounded away from zero. These assumptions are necessary and sufficient for the existence of minimizers. Before we proceed, we make a simple remark about the meta-analysis case.

Remark 1. *Assumption 2 is always satisfied for the case of meta-analysis, as long as all reported covariance matrices are positive definite.*

The remark follows because in the meta-analysis case, $\mathbf{\Lambda}$ is block diagonal, with blocks precisely the covariance matrices reported by studies. For many functions $\mathbf{f}$, Assumption 1 is satisfied, including all linear and piecewise linear-quadratic functions. Functions that may violate these assumptions include logarithms and fractional functions, and complex models may therefore require further analysis. We can now prove the following lemma.

Theorem 1. *Under Assumption 1 and Assumption 2, all level sets $\mathcal{L}_{f,\alpha}$ for f in* (13) *are bounded. Moreover, all level sets $\mathcal{L}_{f,\alpha}$ for f in* (13) *are closed.*

Proof: We first prove that

$$\|\boldsymbol{r}\|_2^2 + \|\boldsymbol{M}\|_F^2 \to \infty \quad \Rightarrow \quad f(\boldsymbol{r}, \boldsymbol{M}) \to \infty,$$

which implies bounded level sets. From (13) and the eigenvalue decomposition, we have

$$\begin{aligned} f(\boldsymbol{r}, \boldsymbol{M}) =& \frac{1}{2}\sum_{i=1}^{m}\left[\frac{\tilde{r}_i^2}{\lambda_i} + \ln(\lambda_i)\right] \\ \geq& \frac{1}{2}\sum_{i=1}^{m}\max\left(\ln(\lambda_i), 1 + \ln(\tilde{r}_i^2)\right) \geq \frac{m}{2}\min(\ln(\alpha), 1 + 2\ln(\alpha)) \end{aligned}$$

When $\|\boldsymbol{r}\|_2^2 + \|\boldsymbol{M}\|_F^2 = \|\tilde{\boldsymbol{r}}\|_2^2 + \|\boldsymbol{\lambda}_M\|_2^2 > c$, we know that there at least exists one $\tilde{r}_i$ or λ_i such that $\tilde{r}_i^2 > c/(2m)$ or $\lambda_i^2 > c/(2m)$, and

$$f(\boldsymbol{r}, \boldsymbol{M}) \geq \frac{1}{2}\max\left(\ln\left(\sqrt{\frac{c}{2m}}\right), 1 + \ln\left(\frac{c}{2m}\right)\right) + \frac{m-1}{2}\min(\ln(\alpha), 1 + 2\ln(\alpha)) \tag{15}$$

And we know when $c \to \infty$, $f(\boldsymbol{r}, \boldsymbol{M}) \to \infty$.

Next we prove that $\mathcal{L}_{f,\alpha}$ are also closed. Since f is the continuous function and $\mathcal{D}_r$ is closed from Assumption 1, we only need to show that the intersection between any $\mathcal{L}_c$ and open boundary of $\mathcal{D}_M$, denoted as $\mathcal{B}_M$, is empty. And we know that $\mathcal{B}_M$ is contained in the space of positive indefinite matrices. For any $\overline{\boldsymbol{M}} \in \mathcal{B}_M$, and any sequence $\boldsymbol{M}^i \subset \mathcal{D}_M$ approaching $\overline{\boldsymbol{M}}$, without loss of generality, we assume $\overline{\lambda_m} = 0$, and we know that, $\lambda_m^i \to 0$ and the corresponding $|\tilde{r}_m^i|$ has to be stay above α (from Assumption 2). And from the inequality (15), we have

$$f(\boldsymbol{r}, \boldsymbol{M}^i) \geq \frac{1}{2}\left(\frac{\alpha^2}{\lambda_m^i} + \ln(\lambda_m^i)\right) + \frac{m-1}{2}\min(\ln(\alpha), 1 + 2\ln(\alpha))$$

And we know as $\lambda_m^i \to 0$, $f(\boldsymbol{r}, \boldsymbol{M}^i) \to \infty$, therefore there exists N such that $f(\boldsymbol{r}, \boldsymbol{M}^i) > c$ for all $i \geq N$, and $\overline{\boldsymbol{M}} \notin \mathcal{L}_c$.
We therefore have an immediate corollary.

Corollary 1 (Existence of minimizers)**.** *Under Assumptions 2 and 1, the set of minimizers f is nonempty.*

By Theorem 1, the level sets of f are closed and bounded, hence compact. A continuous function assumes its minimum on a compact set, therefore the set of minimizers is nonempty. We can also use Theorem 1 to prove specific results about existence of minimizers for (10) under specific parametrizations, including meta-analysis and more standard longitudinal assumptions. Linear constraints will not fundamentally change the situation, so long as they allow any feasible solution, as is stated in the next corollary.

Corollary 2 (Effects of constraints)**.** *As long as the intersection of domains $\mathcal{D}_r$ and $\mathcal{D}_M$ with the polyhedral set $\boldsymbol{C\theta} \leq \boldsymbol{c}$ is nonempty, minimizers exist under Assumptions 2 and 1. Moreover, constraints can be used to ensure Assumption 1, for specific cases, by ensuring that the intersection of the feasible region with $\mathcal{D}_r$ is closed.*

This follows immediately from the fact that the intersection of a compact set with any closed set remains compact.

Theorem 1 and following corollaries establish conditions for the existence of $\boldsymbol{\theta}(w)$, given assumptions 2 and 1. These assumptions must hold for all $\boldsymbol{w}$ in the capped simplex specified by the modeler through the h parameter, which roughly means that for any selection of h datapoints, the problem has to be well defined in the sense of assumptions 2 and 1.

The existence of global minimizers $\boldsymbol{\theta}(\boldsymbol{w})$ underpins the approach, which at a high level optimizers the value function $\boldsymbol{v}(\boldsymbol{w})$ over the capped simplex to detect outliers. The optimization problem (10) has a nonconvex objective and nonlinear constraints. Nevertheless, we can specify the conditions under which $\boldsymbol{v}(\boldsymbol{w})$ is a differentiable function, and we can find an explicit formula for its derivative. The result is an application of the Implicit Function Theorem to the Karush Kuhn Tucker conditions that characterize optimal solutions of (10). We summarize the relevant portion of this classic result presented by [Still, Theorem 4.4], and refer the reader to Craven [1984], Bonnans and Shapiro [2013] where these results are extended under weaker conditions.

Theorem 2 (Smoothness of the value function)**.** *For a given $\boldsymbol{w}$, consider $\boldsymbol{\theta}(\boldsymbol{w})$ and let $\mathcal{I}_0$ denote the set of active constraints:*

$$\mathcal{I}_0 := \{i : \boldsymbol{C}_i(\boldsymbol{\theta}(\boldsymbol{w})) = \boldsymbol{c}_i\}.$$

Consider the extended Lagrangian function restricted to this index set:

$$L(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{w}) = \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{w}) + \sum_{i \in I_0} \mu_i(\boldsymbol{C}_i(\boldsymbol{\theta}) - \boldsymbol{c}_i)$$

Suppose that the following conditions hold:

- *Stationarity of the Lagrangian: there exist multiplies $\boldsymbol{\mu}(\boldsymbol{w})$ with*

$$\nabla_{\boldsymbol{\theta}} L = \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}(\boldsymbol{w}), \boldsymbol{w}) + \sum_{i \in I_0} \mu_i(\boldsymbol{w}) \nabla_{\boldsymbol{\theta}} \boldsymbol{C}_i(\boldsymbol{\theta}(\boldsymbol{w})) = 0$$

- *Linear independence constraint qualification*

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{C}_i(\boldsymbol{\theta}(\boldsymbol{w})), i \in \mathcal{I}_0 \quad \textit{are linearly independent}$$

- *Either of the following conditions hold:*
  - *Strict complementarity: number of active constraints is equal to the number of elements in $\boldsymbol{\theta}$, and all $\mu_i(\boldsymbol{w}) > 0$.*
  - *Second order condition: $\nabla^2_{\boldsymbol{\theta}} L$ is positive definite when restricted to the tangent space $\mathcal{T}$ induced by the active constraints:*

$$\mathcal{T} = \{d : \nabla_{\boldsymbol{\theta}} C_i(\boldsymbol{w}) d = 0, i \in \mathcal{I}_0.\}$$

Then the value function $v(\boldsymbol{w})$ is differentiable, with derivative given by

$$\begin{aligned} \nabla v(\boldsymbol{w}) &= \nabla_{\boldsymbol{w}} L(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{w})|_{\boldsymbol{\theta}(\boldsymbol{w}), \boldsymbol{\mu}(\boldsymbol{w}), \boldsymbol{w}} \\ &= \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{w})|_{\boldsymbol{\theta}(\boldsymbol{w}), \boldsymbol{w}}. \end{aligned} \tag{16}$$

The second order condition above guarantees that $\boldsymbol{\theta}(\boldsymbol{w})$ is isolated (locally unique) minimizer [Still, Theorem 2.5].

Partially minimizing over $\boldsymbol{\theta}$ reduces problem (10) to

$$\min_{\boldsymbol{w}} v(\boldsymbol{w}) \quad \text{s.t.} \quad \mathbf{1}^\top \boldsymbol{w} = h, \quad \mathbf{0} \leq \boldsymbol{w} \leq \mathbf{1}, \tag{17}$$

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:

$$\boldsymbol{w}^+ = \text{proj}_{\Delta_h}(\boldsymbol{w} - \alpha \nabla v(\boldsymbol{w})), \tag{18}$$

where proj_{Δ_h} is the projection operator onto a set defined in the introduction. Since Δ_h is a convex set, the projection is unique.

Each update to $\boldsymbol{w}$ requires computing the gradient ∇v, which in turn requires solving for $\boldsymbol{\theta}$; see (11). The explicit implementation equivalent to (18) is summarized in Algorithm 1. Projected gradient descent is guaranteed to converge to a stationary point in a wide range of settings, for any semi-algebraic differentiable function v (Attouch et al. [2013]).

Algorithm 1 Projected gradient descent on the Value Function v of (11)

1: **Input:** $\boldsymbol{w}_0, \lambda_{\boldsymbol{w}}$.
2: **Initialize:** $\nu = 0$
3: **while** not converged **do**
4: $\quad \nu \leftarrow \nu + 1$
5: $\quad \boldsymbol{\theta}^{\nu+1} \leftarrow \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{w}^\nu) \quad \text{s.t} \quad \boldsymbol{C}(\boldsymbol{\theta}) \leq \boldsymbol{c}$
6: $\quad \boldsymbol{w}^{\nu+1} \leftarrow \text{proj}_{\Delta_h}(\boldsymbol{w} - \alpha \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{\theta}^{\nu+1}, \boldsymbol{w}))$
7: **Output:** $\boldsymbol{w}_\nu, \boldsymbol{\theta}_\nu$

Step 5 of Algorithm 1 requires solving the constrained likelihood problem (10) 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`, optimizing with simultaneously in $\boldsymbol{\theta}$ and $\boldsymbol{w}$ turns the problem into a high dimensional nonsmooth nonconvex problem. Instead, we treat them differently to exploit problem structure. 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. We therefore iteratively solve the difficult small-dimensional problem using `IPopt`, while handling the optimization over the larger $\boldsymbol{w}$ variable through value function optimization over the capped simplex, a convex set.

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, 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. Basic spline functionality is available in many tools and packages. The two main innovations here are (1) use of constraints to capture the shape of the relationship, explained in Section 2.6.2, and (2) nonlinear functions of splines, such as ratios or differences of logs, developed in Section 2.6.3. First we introduce basic concepts and notation for spline models.

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

$$f(t) = \sum_{j=1}^{p+k} \beta_j^p s_j^p(t). \tag{19}$$

These coefficients are inferred by `LimeTr` analysis. A more standard explicit representation of (19) is obtained by building a design matrix $\boldsymbol{X}$. Given a set of n_i range values $(t_1, \dots t_{n_i})$ from study i, the jth column of $\boldsymbol{X}_i \in \mathbb{R}^{n_i \times (p+k)}$ is given by the expression

$$\boldsymbol{X}_i(\cdot, j) = \begin{bmatrix} s_j^p(t_1) \\ \vdots \\ s_j^p(t_{n_i}) \end{bmatrix}. \tag{20}$$

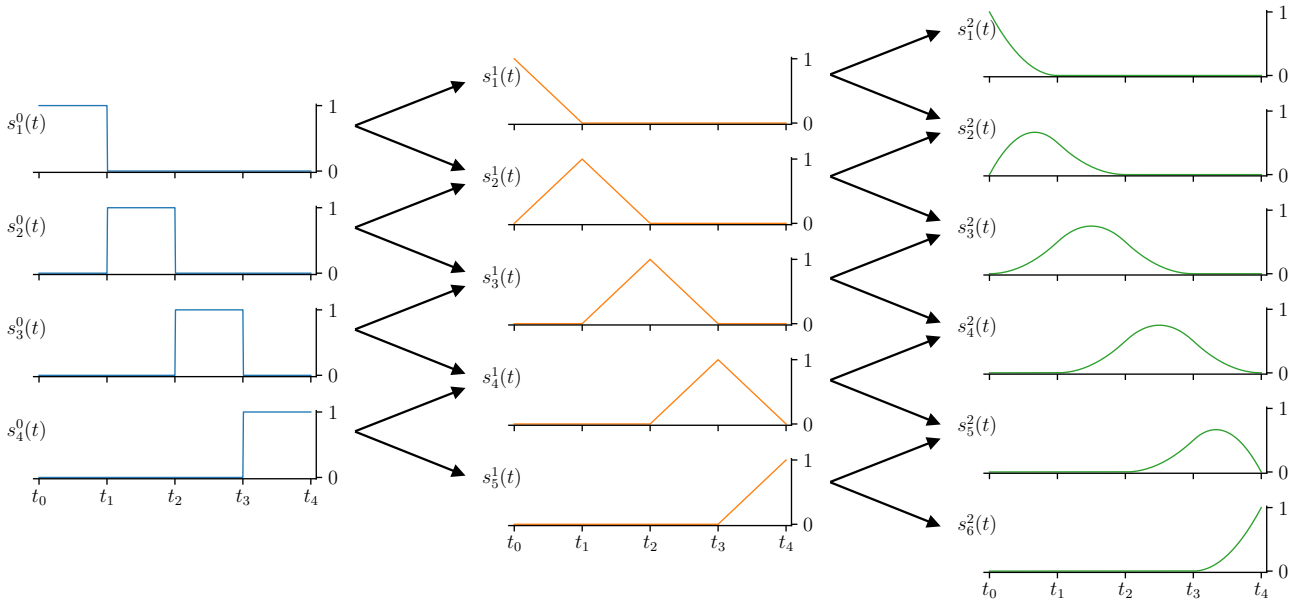

Figure 1: Recursive generation of bspline basis elements (orders 0, 1, 2). The points t_i in the figure are locations for knots, specified (in this case) to be equidistant. Once the spline bases are formed, they can then be evaluated at any points t in the domain, forming $\boldsymbol{X}$ as in (20).

The model for observed data coming from (19) can now be written compactly as

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

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

Monotonicity. Spline monotonicity across the domain of interest follows from monotonicity of the spline coefficients [De Boor et al., 1978]. Given spline coefficients $\boldsymbol{\beta}$, the curve $f(t)$ in (19) is

monotonically nondecreasing when

$$\beta_1 \leq \beta_2 \leq \cdots \leq \beta_{p+k}$$

and monotonically non-increasing if

$$\beta_1 \geq \beta_2 \geq \cdots \geq \beta_{p+k}.$$

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

$$\underbrace{\begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 1 & -1 \end{bmatrix}}_{\boldsymbol{C}} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_{p+k} \end{bmatrix} \leq \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

These linear constraints are a special case of the general estimator (10).

Convexity and Concavity For any 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 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` .

2.6.3 Nonlinear measurements

Some of the studies in Section 3.2 use nonlinear observation mechanisms. In particular, given a dose-response curve of the form (19), studies often report odds ratio of an outcome between exposed and unexposed groups that are defined across two intervals on the underlying curve:

$$\text{relative risk} = \frac{\frac{1}{a_1 - a_0} \int_{a_0}^{a_1} f(t)dt}{\frac{1}{b_1 - b_0} \int_{b_0}^{b_1} f(t)dt} \tag{21}$$

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 for observation j in study i, this is given by

$$y_{ij} := \log \text{relative risk for } ij \text{ observation} = \ln(\langle \boldsymbol{x}_{ij}^1, \boldsymbol{\beta} \rangle) - \ln(\langle \boldsymbol{x}_{ij}^2, \boldsymbol{\beta} \rangle) := f_{ij}(\boldsymbol{\beta}),$$

a particularly useful example of the general nonlinear term $\boldsymbol{f}_{ij}(\boldsymbol{\beta})$ in problem class (1). A set of examples of epidemiological models that arise from relative risk are discussed in Section 3.2.2.

2.7 Uncertainty Estimation

The `LimeTr` package modifies the parametric bootstrap [Efron and Tibshirani, 1994] to estimate the uncertainty 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 (10).

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 $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\gamma}})$, we sample model-based errors to get new bootstrap realizations $\bar{\boldsymbol{y}}$:

$$\bar{\boldsymbol{y}}_i = \boldsymbol{f}_i(\hat{\boldsymbol{\beta}}) + \boldsymbol{Z}_i \bar{\boldsymbol{u}}_i + \bar{\boldsymbol{\epsilon}}_i,$$

where $\bar{\boldsymbol{\epsilon}}_{\boldsymbol{i}} \sim N(0, \hat{\Lambda})$ and $\bar{\boldsymbol{u}}_{\boldsymbol{i}} \sim N(0, \hat{\gamma})$. These realizations have the same structure as the input data. 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 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 $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\gamma}})$. Its exact theoretical properties are outside the scope of the current paper and are a topic of ongoing research. We use N = 1000 in all of the numerical experiments in the next section. This significantly increases computational load, compared to a single fit. How to make the procedure more efficient (or develop alternatives) is another ongoing research topic.

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 can be solved by other available tools; 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. These examples illustrate advanced `LimeTr` functionality that is not available in other tools. All examples in the paper are available online, along with a growing library of additional examples (see the Supplementary Materials section).

3.1 Validation Using Synthetic Data

Here we consider two common mixed effects models. First we look at meta-analysis, where the $\boldsymbol{\Lambda}$ term is known while $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ are unknown. Then we look at a simple longitudinal case, where all three parameters are unknown, and $\boldsymbol{\Lambda}$ is modeled as $\sigma^2 \boldsymbol{I}$ with unknown scalar σ^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 γ is the standard deviation of the between-study heterogeneity and σ 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: $y_i' = y_i - 30 - |N(0, 80^2)|$. The setting of the simulation uses large outliers to show the robustness of the model to trimming. The larger the outliers, the larger the difference in performance between `LimeTr` and other tools. The behavior of trimming in the meta-analysis of real data is more clear in Section 3.2.

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	β_0	β_1	$\sqrt{\gamma}$	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 β_0, β_1 and in γ 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, $\mathbf{\Lambda}$ is no longer specified as in the meta-analysis case, but is instead parameterized through a single unknown error $\mathbf{\Lambda} = \sigma^2 \boldsymbol{I}$ common to all observations. We use the same simulation structure as in Section 3.1.1, now replacing observation-specific standard errors with

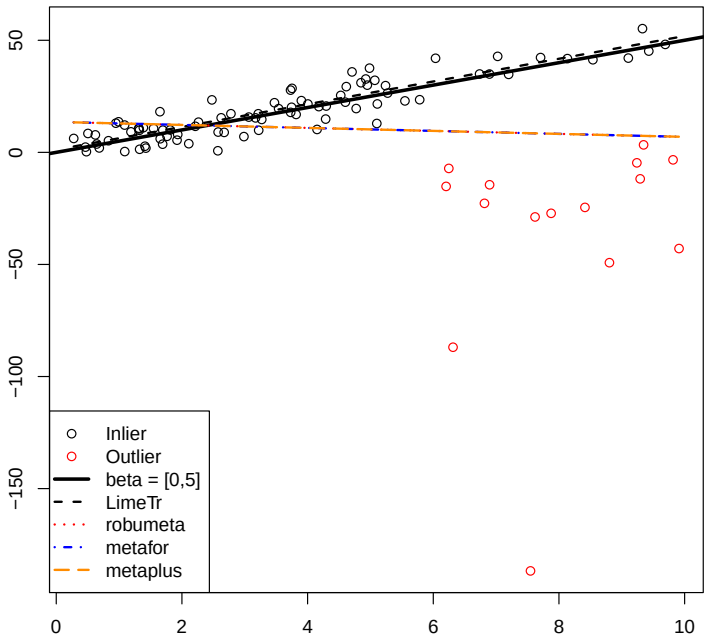

Figure 2: A representative instance of the experiment summarized in Table 2. True mechanism is shown using solid line; the true model is successfully inferred by the `LimeTr` package despite the outliers (red points).

random errors generated according to $N(0, \sigma^2)$. Since σ^2 is now also an unknown parameter, we check how well it is estimated by all packages and report this error in Table 3.

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

.

The first example is a simple linear model that aims to quantify the effect of vitamin A

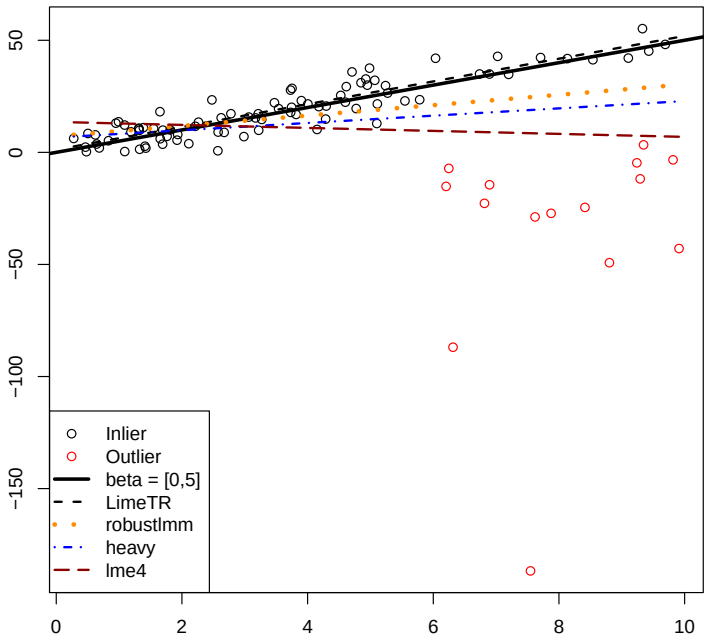

Figure 3: A representative instance of the experiment summarized in Table 3. `robustlmm` and `heavy` packages both estimate β better than `lme4`, likely because they use distributions with heavier tails. `LimeTr` outperforms all three alternatives.

Package	β_0	β_1	$\sqrt{\gamma}$	σ	TPF	FPF	Seconds
Truth	0	5	6	4	—	—	—
LimeTr	**0.64**	**4.95**	4.93	**3.61**	**1.000**	**0.06**	1.04
robustlmm	3.32	3.67	**5.14**	9.53	0.99	0.06	7.7
heavy	2.43	3.55	3.61	NA	0.95	0.07	0.16
lme4	10.6	0.09	4.89	31.0	0.68	0.11	**0.08**

Table 3: Results of Longitudinal Comparison. True values are $\beta_0 = 0, \beta_1 = 5, \gamma = 6, \sigma = 4$. Both `robustlmm` and the `heavy` package do better than the standard `lme4`. `LimeTr` more accurately estimates β_0, β_1, and σ. The `heavy` package estimates γ more accurately. However, estimated γ typically increases with a worse fit to the data, and heavy does not accurately estimate β.

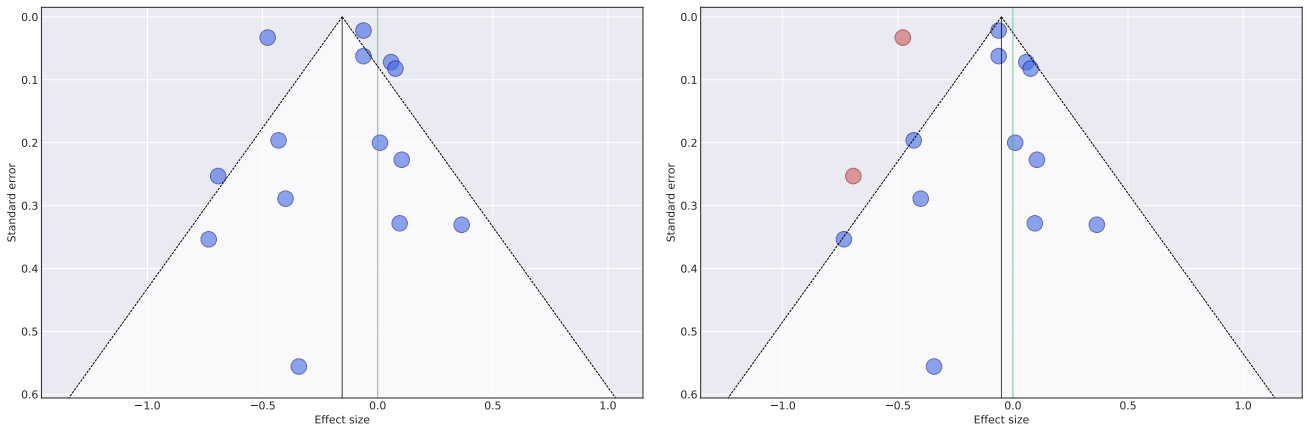

Figure 4: Results for Vitamin A vs. incidence and/or mortality rates for diarrheal disease, shown using the funnel plot (effect size vs. reported standard error). The left panel shows model without trimming, while the right panel shows the model with trimming. Trimming 10% of the studies shifts the effect 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.

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 of incidence or mortality from diarrheal disease from 12 studies. Two of the studies report on both incidence and mortality, giving 14 total datapoints. No covariates are used. The model we consider is

$$\boldsymbol{y}_i = \mathbf{1}_i \beta_0 + \mathbf{1}_i u_i + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{y}_i$ are data sets reported from each study, $\mathbf{1}_i$ is a vector of 1's of the same size as the number of observations for study i, $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_i)$ are associated standard errors, and $u_i \sim N(0, \gamma^2)$ is a study-specific random effect, with γ^2 accounting for between-study heterogeneity.

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 γ^2 is estimated to be 0.0435, and after trimming two studies it is reduced to 7.14×10^{-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.

log relative risk model. The log relative risk model is very common in the epidemiological context, and is often approximated by the log-linear relative risk model. A brief introduction to the nonlinear model was given in Section 2.6.3. Here we derive two models, and appropriate random effect specification, that can be used to analyze smoking and diet data.

In words, the log-relative risk model can be written as

$$\ln(\text{rel risk}) = \left(\frac{\ln(\text{risk at alt exposure}) - \ln(\text{risk at ref exposure})}{\text{alt exposure} - \text{ref exposure}} + \text{random effect} \right) \times$$
$$(\text{alt exposure} - \text{ref exposure}) + \text{measurement noise}$$

where the random effect is on the average slope and the measurement error in the log space. Instead of making the strong assumption that the log of the risk is a linear function of exposure, we use a spline to represent this relationship:

$$\ln(\text{risk at exposure } t) = \ln(\langle \boldsymbol{x}, \boldsymbol{\beta} \rangle)$$

where $\boldsymbol{x}$ is the design vector at exposure t and the spline is parametrized by $\boldsymbol{\beta}$, see Section 2.6.1.

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 methodology described in Section 2.6.1 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 (exposure = 0), and we assume there is no

risk for non-smokers (ln(risk at exposure 0) = 0). Since all datapoints share a common reference, we can simplify the relative risk model as follows:

$$\ln(\text{rel risk}) = \left(\frac{\ln(\text{risk at alt exposure})}{\text{alt exposure}} + \text{random effect}\right) \times \text{alt exposure} + \text{measurement noise}$$

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

$$y_{ij} = \log(\langle \boldsymbol{x}_{ij}, \boldsymbol{\beta} \rangle) + \text{exposure}_{ij} u_i + \epsilon_{ij}$$

with y_{ij} the log relative risk, $\boldsymbol{x}_{ij}$ computed using a spline basis matrix for exposure_{ij} (see Section 2.6.1), u_i the random effect for study i, and $\epsilon_{ij} \sim N(\mathbf{0}, \sigma_{ij}^2)$ the variance reported by the ith study for its jth 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 0 (baseline: non-smokers) by adding direct constraint on β_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.

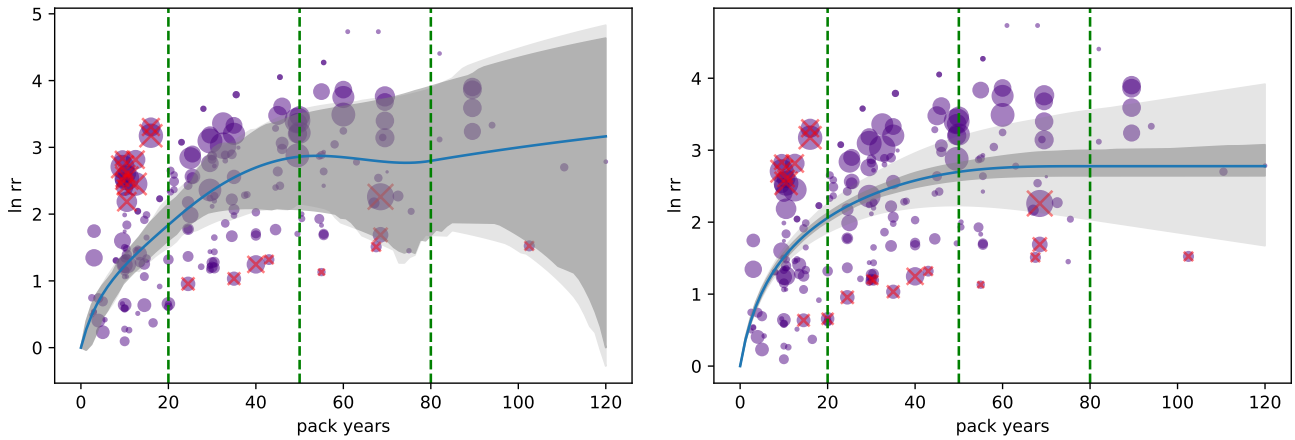

Figure 5: Modeling dose-response relationship between exposure (pack years) and log-odds of lung cancer. Left figure shows a cubic spline, while right figure shows a cubic spline with monotonically increasing risk and concavity constraints. The reference group is 'nonsmokers' in all studies, so taking nonsmoking risk at 1 we can plot the relative risk for each exposure group at its midpoint, and show the comparison to the estimated risk curve. The uncertainty of the mean is shown in dark grey, and 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 has higher fixed effect uncertainty but 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 a topic of 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. The observation model given by (21). As in the smoking example, we model the random

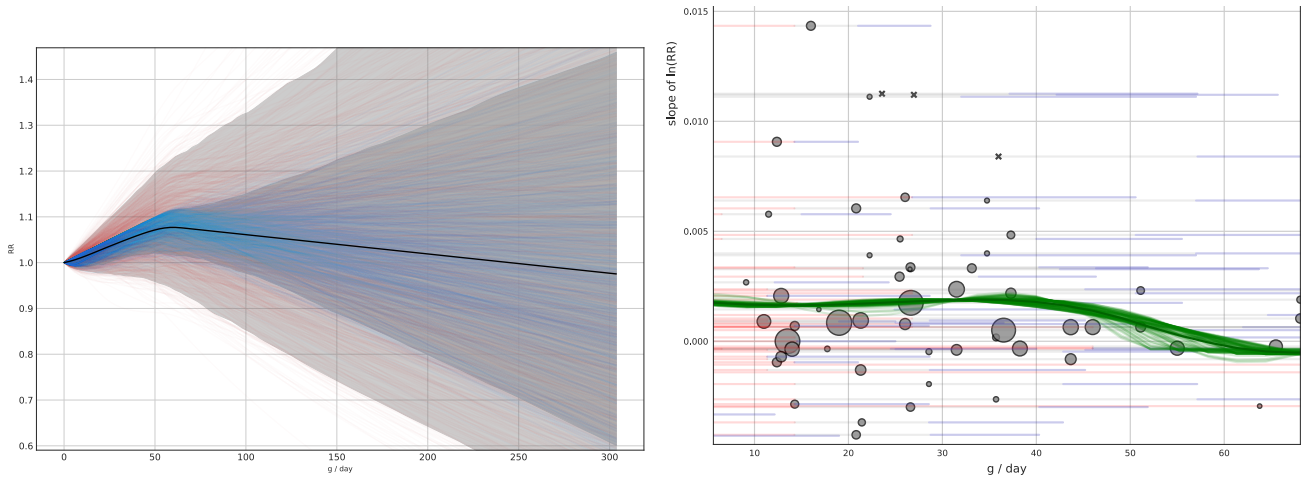

Figure 6: Spline fit obtained from observation model (22). The left panel shows the inferred dose response mechanism from indirect observations, across multiple spline knot placement options. Uncertainty from the spline coefficient fits is plotted in dark gray, with additional uncertainty coming from the random effects shown in light gray. Observations are shown in the right panel. The midpoint between the reference and alternative intervals is displayed as a point, with the reference and alternative exposures plotted in blue and red.

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. In contrast to the smoking case, here we must account for both reference and alternative group definitions, so the observation model is given by

$$\begin{aligned} y_{ij} &= \left(\frac{\ln(\langle \boldsymbol{x}_{ij}^1, \boldsymbol{\beta} \rangle) - \ln(\langle \boldsymbol{x}_{ij}^0, \boldsymbol{\beta} \rangle)}{\text{alt exposure}_{ij} - \text{ref exposure}_{ij}} + u_i \right) (\text{alt exposure}_{ij} - \text{ref exposure}_{ij}) + \epsilon_{ij} \\ &= \ln(\langle \boldsymbol{x}_{ij}^1, \boldsymbol{\beta} \rangle) - \ln(\langle \boldsymbol{x}_{ij}^0, \boldsymbol{\beta} \rangle) + (\text{alt exposure}_{ij} - \text{ref exposure}_{ij}) u_i + \epsilon_{ij}. \end{aligned} \tag{22}$$

with y_{ij} the log relative risk, $\boldsymbol{x}_{ij}^1$ and $\boldsymbol{x}_{ij}^0$ computed using spline basis matrix for alternative exposure_{ij} and reference exposure_i, (see Section 2.6.1), u_i the random effect for study i, and $\epsilon_{ij} \sim N(\mathbf{0}, \sigma_{ij}^2)$ the variance reported by the ith study for its jth point.

`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

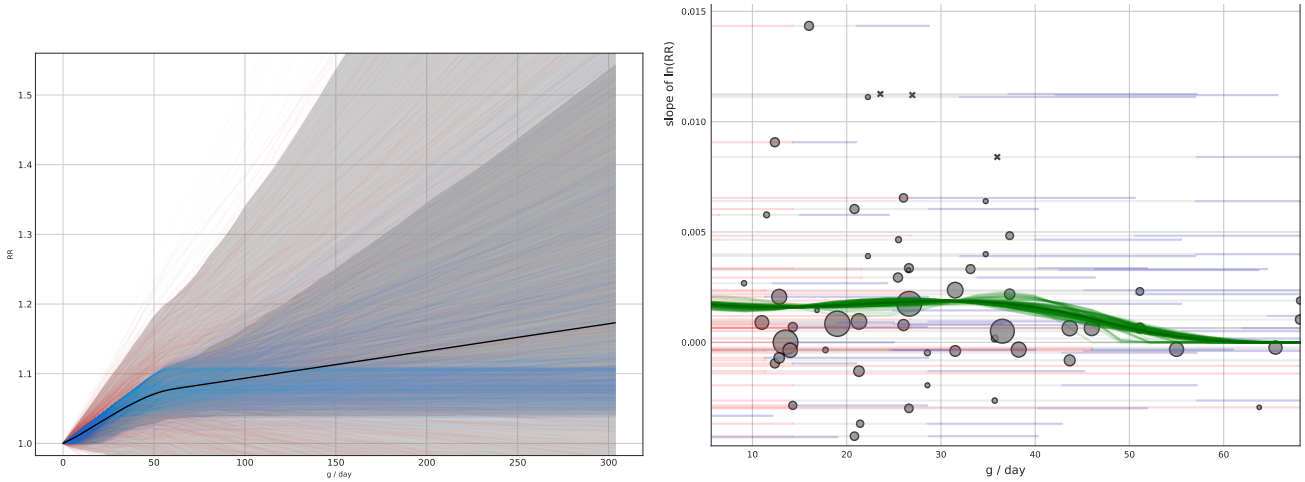

Figure 7: Adding monotonicity constraints to example in Figure 6.

meta-analysis to obtain the risk curve in Figure 6 integrates results from 14 prospective cohort studies[2]. 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]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 $\boldsymbol{f}_i(\boldsymbol{\beta})$, 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.

SUPPLEMENTAL MATERIALS

`LimeTr` package: Python package `LimeTr` containing code to perform robust estimation of mixed effects models, for both meta-analysis and simple longitudinal analysis. Available online through github: `https://github.com/zhengp0/limetr`

Experiments: Set of script files available online to produce simulated data and run `LimeTr` and third party code: `https://github.com/zhengp0/limetr/tree/paper/experiments`

- `Settings.R`: R code to Specifies folder structure and simulation parameters.
- `functions.R`: R code for auxiliary functions for simulating data and aggregating results.
- `0_create_sim_data.R`: R code to create simulated data.
- `1_limetr.py`: Python code to run `LimeTr` .
- `2_metafor.R`: R code to run `metafor` package
- `3_robumeta.R`: R code to run `robumeta` package
- `4_metaplus.R`: R code to run `metaplus` package

- `5_lme4.R`: R code to run `lme4` package
- `6_robustlmm.R`: R code to run `robustlmm` package
- `7_heavy.R`: R code to run `heavy` package

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. Efficient quadratic penalization through the partial minimization technique. *IEEE Transactions on Automatic Control*, 63(7):2131–2138, 2018.

Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

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/7f279483412348928f01507440b0360d`.

J Frédéric Bonnans and Alexander Shapiro. *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.

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

BD Craven. Non-linear parametric optimization (b. bank, j. guddat, d. klatte, b. kummer and k. tammer). *SIAM Review*, 26(4):594, 1984.

Marie Davidian and David M Giltinan. *Nonlinear models for repeated measurement data*, volume 62. CRC press, 1995.

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.

David A Harville. *Linear models and the relevant distributions and matrix algebra*. CRC Press, 2018.

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.

Georg Still. Lectures on parametric optimization: An introduction. *Optimization Online*.

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 Y. Aravkin. A general family of trimmed estimators for robust high-dimensional data analysis. *Electronic Journal of Statistics*, 12(2):3519–3553, 2018.

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.