JDSSV

Journal of Data Science, Statistics, and Visualisation

November 2024, Volume IV, Issue VII. *doi: 10.52933/jdssv.v4i7.112*

A Rank-Based Estimation Method for Mixed Effects Models in the Presence of Outlying Data

Barbara Brune
TU Wien and
TÜV AUSTRIA

Irene Ortner
Digitas GmbH

Peter Filzmoser
TU Wien

Abstract

We present an approach to the rank-based estimation of mixed effects models, extending existing methods to random effects structures beyond random intercepts. The estimates obtained from our procedure are insensitive to outlying observations including leverage points, (almost) tuning-parameter free and can be computed very efficiently. Furthermore, the resulting estimates allow for model diagnostics, especially with respect to the identification of outlying observations or groups in the data. The properties of the proposed estimators, in particular, their robustness to different outlier types, are studied by means of simulation studies. The methodology is illustrated with applications to the sleep study data set and to data from accelerated aging experiments on photovoltaic (PV) modules.

Keywords: Non-parametric regression, mixed effects models, longitudinal data, rank-based regression, robust statistics.

1. Introduction

The mixed effects model is a popular extension of the standard linear regression model. It is able to deal with clustered error structures, e.g., in cases where multiple measurements are made on the same subject. In this case, the error terms are no longer independent and identically distributed.

In general, the model equation for a standard linear mixed effects model is given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\beta}$ are the so-called fixed effects, $\boldsymbol{b}$ are the individual- or group-specific random effects, and $\boldsymbol{\varepsilon}$ are the independent errors. $\boldsymbol{X}$ and $\boldsymbol{Z}$ are matrices of predictors and $\boldsymbol{y}$ is the response. If $\boldsymbol{Z}$ is the vector of ones, the associated random effect is called a random intercept, otherwise we speak about random slopes. In the classical parametric setting, it is assumed that both the errors and random effects are homoscedastic and follow a normal distribution. Based on this assumption, the standard approaches of estimation are maximum likelihood (ML) and restricted maximum likelihood (REML). Pinheiro and Bates (2000) provide a comprehensive overview of both approaches.

In practice, we are often confronted with data that do not comply with the normality assumption imposed by the standard estimation approaches mentioned above. The data may contain measurement errors or unusual observations (so-called *outliers*) and the error distributions might be non-symmetric or heavy-tailed. Such deviations can cause the standard parametric estimation procedures to break down, resulting in heavily distorted coefficient estimates. The contaminated estimates make it difficult to identify the outlying observations which can be masked by wrong fits.

The fields of non-parametric and robust statistics aim to develop methodology that does not rely (or relies far less) on distributional assumptions, and is not distorted by the presence of unusual measurements. Different approaches to estimate mixed effects models in a robust manner have been proposed in the literature. Various methods are based on down-weighing observations, i.e., M- or S-estimation (see e.g., Copt and Victoria-Feser 2006; Koller 2013; Agostinelli and Yohai 2016). These methods are often computationally very expensive. Other methods replace the normality assumption for the error term and the random effects with more heavy-tailed distributions such as the t-distribution (Pinheiro et al. 2001). A further line of robust and non-parametric estimation methods is provided in the framework of rank-based regression (see e.g., Kloke et al. 2009; Bilgic 2012; Jung and Ying 2003; Wang and Zhu 2006; Wang and Zhao 2008). These methods have the advantage of being computationally cheap, robust against outlying responses and efficient, even if the modeling assumptions for ML and REML as mentioned above are violated.

There are two main shortcomings of the existing rank-based methods for the estimation of linear mixed effects models: (1) In the setting of experiments that are not designed, some of the predictors included in $\boldsymbol{X}$ or $\boldsymbol{Z}$ might be unusual or outlying (e.g., due to measurement errors). Such leverage points can still distort the coefficient estimates obtained from the existing models. (2) To the best of our knowledge, there does not exist a rank-based estimation method that is able to deal with more complex random effects structures: All rank-based methods are designed for random intercepts, but cannot account for random slopes of any type. However, we often do not want to

limit our model estimates to random intercepts. Instead, we might be interested in the individual-specific response to a treatment, not only with regard to changes in the average response, but also regarding the slope in a certain covariate.

The present work extends the rank-based estimation approaches presented in Kloke et al. (2009) and Bilgic (2012) to allow for model structures that include random slopes. Furthermore, a weighting procedure is introduced that protects the estimates against the effects of leverage points. The results allow for model diagnostics both for the whole sample and on group-level. Unusual groups and observations can be identified, aiding interpretation and deeper understanding of the modeled data.

2. Preliminaries

2.1. Rank-based regression

Given a regular regression model

$$\boldsymbol{y} = \alpha \mathbf{1}_N + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{2}$$

with $\boldsymbol{y} \in \mathbb{R}^N$, $\boldsymbol{X} \in \mathbb{R}^{N \times p}$, intercept parameter $\alpha \in \mathbb{R}$, regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\mathbf{1}_N$ the N-dimensional all-ones vector, where N describes the number of observations. The error term $\boldsymbol{e} \in \mathbb{R}^N$ is assumed to be generated by a continuous distribution with a positive definite covariance matrix given by $\sigma_e^2 \boldsymbol{I}_N$ with variance parameter $0 < \sigma_e^2 < \infty$. $\boldsymbol{I}_N$ is the N-dimensional identity matrix.

A robust rank-based estimate of the regression coefficients $\boldsymbol{\beta}$ can be obtained by minimizing a pseudo norm of the form

$$||\boldsymbol{r}||_\varphi = \sum_{j=1}^{N} a[R(r_j)] r_j \tag{3}$$

in the residuals $\boldsymbol{r} = \boldsymbol{y} - \alpha \mathbf{1}_N - \boldsymbol{X}\boldsymbol{\beta} \in \mathbb{R}^N$. Here, $R(r_j)$ denotes the rank of the residual r_j among the components $r_1, \ldots, r_N$ of the residual vector $\boldsymbol{r}$. The scores are given by $a[t] = \varphi(t/(N+1))$ where $\varphi(u)$ is a non-decreasing, bounded and square-integrable function such that $\sum_t a[t] = 0$. Proposed in Jureckova (1971), the pseudo norm (3) is also called *Jaeckel's dispersion function* after Jaeckel (1972). The score function $\varphi(\cdot)$ can be specified in various ways. We use the *Wilcoxon score*

$$\varphi(u) = \sqrt{12}\left[u - \frac{1}{2}\right], \tag{4}$$

which provides a good trade-off between efficiency and robustness, if the error term $\boldsymbol{e}$ is not normally distributed (McKean 2004). With this score function, the norm can be written equivalently as a scaled sum of pairwise differences (see McKean 2004):

$$||\boldsymbol{r}||_\varphi = \frac{\sqrt{3}(N+1)}{2} \sum_{j=1}^{N} \sum_{k=1}^{N} |r_j - r_k|. \tag{5}$$

The ordinary rank-based regression estimate is then given by

$$\hat{\boldsymbol{\beta}}_\varphi = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_\varphi. \tag{6}$$

The criterion (6) does not allow to estimate the intercept parameter α directly, because the ranks of the residuals are invariant to shifts by a constant. The intercept can be estimated by applying a robust estimator of location $T : \mathbb{R}^N \to \mathbb{R}$ to the vector of residuals $\tilde{\boldsymbol{r}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\varphi$:

$$\hat{\boldsymbol{\alpha}} = \mathrm{T}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\varphi) = \mathrm{T}(\tilde{\boldsymbol{r}}). \tag{7}$$

The error variance σ_e^2 is estimated by applying a robust scale estimator, denoted by $S : \mathbb{R}^N \to \mathbb{R}^+$, to the resulting vector of residuals $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\varphi - \hat{\boldsymbol{\alpha}}$. We suggest using the *one-sample Hodges-Lehmann estimator* (Hodges and Lehmann 1963) for intercept estimation, and the Q_n *estimator of scale* (Rousseeuw and Croux 1993) to estimate the error variance. These two estimators provide a good trade-off between robustness and efficiency. Other possible choices for $T(\cdot)$ and $S(\cdot)$ are e.g., the *sample median* and the *median absolute deviation* (*MAD*).

There is no closed form solution for the minimizer $\hat{\boldsymbol{\beta}}_\varphi$ in (6). However, the optimization problem is convex and can be solved efficiently using numerical optimization techniques. Useful starting values for the optimization can be obtained from an initial ordinary least squares (OLS) fit or, if we suspect outlying values, from least trimmed squares (LTS, Rousseeuw and Van Driessen 2006).

A thorough overview of the theory regarding the rank-based regression model may be found in Chapter 3 of Hettmansperger and McKean (2011).

2.2. Terminology of mixed effects models

Given are multiple measurements from g independent groups. Then, for the n_i-dimensional response vector $\boldsymbol{y}_i$ from group i, $i = 1, \ldots, g$, we aim to fit a model of the form

$$\boldsymbol{y}_i = \alpha \mathbf{1}_{n_i} + \boldsymbol{X}_i\boldsymbol{\beta} + a_i\mathbf{1}_{n_i} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \tag{8}$$

where $N = \sum_{i=1}^g n_i$ is the total sample size.

We call the coefficients α and $\boldsymbol{\beta}$ that are shared by all groups the *fixed effects*. $\boldsymbol{X}_i \in \mathbb{R}^{n_i \times (p-1)}$ contains the predictors associated with the fixed effects.

The group-specific *random effects* $a_i \in \mathbb{R}$ and $\boldsymbol{b}_i \in \mathbb{R}^{k-1}$ are treated as realizations of a centered random variable $(a, \boldsymbol{b}')' \in \mathbb{R}^k$ with positive definite covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}) \in \mathbb{R}^{k \times k}$ that is parametrized by a coefficient vector $\boldsymbol{\theta}$. We assume $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is a diagonal matrix, and $\boldsymbol{\theta} \in \mathbb{R}^k$ contains the random effects' standard deviations. More generally, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ could also be specified as compound symmetric or unstructured with a $\boldsymbol{\theta}$ of suitable length. $\boldsymbol{Z}_i \in \mathbb{R}^{n_i \times (k-1)}$ is a matrix of predictors corresponding to the random slope(s) $\boldsymbol{b}_i$. The matrices $\boldsymbol{Z}_i$ can be subsets of the columns of the matrices $\boldsymbol{X}_i$, but also contain further external predictors.

The errors $\boldsymbol{\varepsilon}_i$ are assumed to follow a distribution with finite second moment and covariance matrix $\sigma_\varepsilon^2 \boldsymbol{I}_{n_i}$. In addition, the errors and the random effects are independent from each other. The coefficients σ_ε and $\boldsymbol{\theta}$ are also referred to as *variance components*.

Given estimators $\hat{\alpha}$, $\hat{\boldsymbol{\beta}}$ and $\hat{a}_i$, $\hat{\boldsymbol{b}}_i$, $i = 1, \ldots, g$, respectively, we denote the *marginal* and *conditional* residuals by

$$\boldsymbol{r}_{i,\text{marg}} = \boldsymbol{y}_i - \mathbf{1}_{n_i}\hat{\alpha} - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}, \text{ and} \tag{9}$$

$$\boldsymbol{r}_{i,\text{cond}} = \boldsymbol{y}_i - \mathbf{1}_{n_i}\hat{\alpha} - \boldsymbol{X}_i\hat{\boldsymbol{\beta}} - \mathbf{1}_{n_i}\hat{a}_i - \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i, \tag{10}$$

accordingly. For ML and REML estimation (Pinheiro and Bates 2000), the random effects and the errors are assumed to be centered, independent and normally distributed.

2.3. Rank-based regression for dependent data

Kloke et al. (2009) extend the methodology of rank-based regression with independent errors to estimate simple mixed effects models with random intercept terms. Their methodology is further extended by Bilgic (2012) to allow for nested random intercepts of arbitrary depth. Both methods rely on a two-step estimation procedure, where the fixed effects are estimated in the first step. The residuals from this first regression are then used to derive the random effects and the variance components. The authors give asymptotic properties of their estimators and use them to draw inference on the fixed effects.

Other rank-based estimation methods for repeated measurements data include the method proposed in Jung and Ying (2003), and its extensions in Wang and Zhu (2006) and Wang and Zhao (2008). Those methods follow a different model formulation and do not explicitly specify the dependency structure of the data beforehand. The authors consider a simple linear regression model for repeated measurements data, and propose estimation methods that are able to deal with possible dependencies and heteroscedasticity. Similarly, Abebe et al. (2016) present rank-based fits for generalized estimating equations.

All of the above mentioned papers focus on drawing inference on the fixed effects in the model. Contrarily, we focus on the accurate estimation of fixed and random effects in the presence of outliers. This especially allows for model diagnostics and identification of unusual (groups of) observations as well as interpretation of the realizations of the random effects. We do not consider inference for the parameters.

3. Methodology

In the following, we use the formulation of the mixed effects model as presented in Section 2.2. The proposed estimation procedure relies on the rank-based regression presented in Section 2.1.

3.1. Rank-based estimation of mixed effects models with random slopes

For simplicity of the exposition and without loss of generality, we drop the intercept terms α and a_i, $i = 1, \ldots, g$, from model (8). Details for the estimation with intercept can be found in Algorithm A.1.

By ignoring the dependency structure induced by the random effects, we can obtain an initial estimate $\hat{\boldsymbol{\beta}}^{(0)}$ of $\boldsymbol{\beta}$ from a rank-based regression of the stacked models (8)

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_g \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_g \end{pmatrix} \boldsymbol{\beta} + \tilde{\boldsymbol{e}}. \tag{11}$$

The random effect structure is absorbed into the error term $\tilde{\boldsymbol{e}} \in \mathbb{R}^N$. Then, we calculate the marginal residuals as

$$\boldsymbol{r}_{i,\text{marg}}^{(0)} = \boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}^{(0)} \approx \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{\varepsilon}_i. \tag{12}$$

for each $i = 1, \ldots, g$, This gives us g further regression problems that can be solved for the most likely realizations of the random effects $\boldsymbol{b}_i$:

$$\hat{\boldsymbol{b}}_i^{(0)} = \arg\min_{\boldsymbol{b}} ||\boldsymbol{r}_{i,\text{marg}}^{(0)} - \boldsymbol{Z}_i \boldsymbol{b}||_{\varphi}. \tag{13}$$

We obtain estimates of the variance components as $\hat{\sigma}_{\varepsilon}^{(0)}$ and $\hat{\boldsymbol{\theta}}^{(0)}$ by applying the Q_n estimator to the conditional residuals

$$\boldsymbol{r}_{i,\text{cond}}^{(0)} = \boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{Z}_i \hat{\boldsymbol{b}}_i^{(0)}, \tag{14}$$

and the estimated random effects $\hat{\boldsymbol{b}}_i^{(0)}$.

This set of initial estimates $\hat{\boldsymbol{\beta}}^{(0)}$, $\hat{\boldsymbol{b}}_i^{(0)}$, $\hat{\sigma}_{\varepsilon}^{(0)}$ and $\hat{\boldsymbol{\theta}}^{(0)}$ can now be improved using a reweighting scheme in the manner of iteratively reweighted least squares. Similar methodology for the random intercept model is proposed as the *generalized rank procedure* in Bilgic (2012).

The basic idea is the following: From the model equation (8), we can directly obtain the covariance of each vector $\boldsymbol{y}_i$, $i = 1, \ldots, g$. It is given by

$$\boldsymbol{\Sigma}_{\boldsymbol{y}_i} = \boldsymbol{\Sigma}_{\boldsymbol{y}_i}(\sigma_{\varepsilon}, \boldsymbol{\theta}) = \text{Cov}(\boldsymbol{y}_i \,|\, \boldsymbol{X}_i, \boldsymbol{Z}_i) = \sigma_{\varepsilon}^2 \boldsymbol{I}_{n_i} + \boldsymbol{Z}_i \boldsymbol{\Sigma}(\boldsymbol{\theta}) \boldsymbol{Z}_i' \tag{15}$$

and can be robustly estimated by plugging in the variance component estimates:

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(0)} = \boldsymbol{\Sigma}_{\boldsymbol{y}_i}(\hat{\sigma}_{\varepsilon}^{(0)}, \hat{\boldsymbol{\theta}}^{(0)}). \tag{16}$$

By multiplying the system of regression equations (11) with $\boldsymbol{\Sigma}_{\boldsymbol{y}_i}^{-1/2}$, the errors $\tilde{\boldsymbol{e}}$ are transformed to be homoscedastic and uncorrelated. One can then solve the rescaled equation system(s)

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(0)-1/2} \boldsymbol{y}_i = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(0)-1/2} \boldsymbol{X}_i \boldsymbol{\beta} + \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(0)-1/2} \tilde{\boldsymbol{e}}_i \tag{17}$$

to achieve a better, more efficient estimate of $\boldsymbol{\beta}$.

However, this rescaling step can impose problems in the presence of outlying predictors or responses. If observations do not follow the regression line, the rotation (17) can amplify their influence: Due to the linear combinations built in the multiplication with the estimated covariance matrix, one outlying observation or predictor in group i is able to contaminate the whole group.

This can be circumvented by a preceding reweighting step: Based on the conditional residuals (14), we define *outlyingness weights*

$$\nu_{ij}^{(0)} = \min\left\{1, \frac{c \cdot \hat{\sigma}_{\varepsilon}^{(0)}}{|r_{ij,\text{cond}}^{(0)}|}\right\}, \tag{18}$$

for $i = 1, \ldots, g,\ j = 1, \ldots, n_i$, where $r_{ij,\text{cond}}^{(0)}$ are the elements of the conditional residuals in (14). The tuning parameter c is the cutoff value that is used to classify the residuals as outlying. As it is a common choice in the literature we set $c = 2$. We reweigh the regression equations as follows,

$$\text{diag}\left(\boldsymbol{\nu}^{(0)}\right)\begin{pmatrix}\boldsymbol{y}_1\\ \vdots \\ \boldsymbol{y}_g\end{pmatrix} = \text{diag}\left(\boldsymbol{\nu}^{(0)}\right)\begin{pmatrix}\boldsymbol{X}_1\\ \vdots \\ \boldsymbol{X}_g\end{pmatrix}\boldsymbol{\beta} + \tilde{\tilde{\boldsymbol{e}}}, \tag{19}$$

where $\boldsymbol{\nu}^{(0)} = (\nu_{11}^{(0)}, \ldots, \nu_{gn_g}^{(0)})'$ are the stacked weights. We then proceed with the rescaling step as described in Equation (17) and re-estimate $\boldsymbol{\beta}$. The new estimate can then be used to recalculate the residuals $\boldsymbol{r}_{i,\text{marg}}$ as in (12), and improve the random effects estimates $\hat{\boldsymbol{b}}_i$ as in (13). We iterate between these steps until the estimates stabilize.

For error variance estimation, we suggest correcting for the degrees of freedom lost through the separate fitting of the g regression models that estimate the random effects. Thus, we apply the following finite-sample correction to the Q_n estimator of scale:

$$Q_{n,\text{corr}}(\boldsymbol{r}) = \sqrt{\frac{N}{N-(p+g)}}\, Q_n(\boldsymbol{r}). \tag{20}$$

This correction gives good results in simulation studies.

The model fitting procedure is summarized in Algorithm A.1 in the appendix. In all our simulation experiments, the algorithm converged in less than 5 iterations.

3.2. Further robustification against leverage points

The proposed method is robust against outlying values in the response space (*y-outliers*), but can be heavily affected by outliers in the predictor space (*X-outliers*, *leverage points*).

McKean (2004) suggests a method that additionally robustifies the rank-based regression against leverage points. This is achieved by introducing robustness weights to the Wilcoxon norm as given in (5). The weights are determined such that influential and outlying observations, either in terms of leverage (predictor space) or in terms of unusual responses, are downweighted.

To estimate mixed effects model (8) with a rank-based approach that is robust against both outliers in response and predictor space, we replace the two regression steps in Algorithm A.1 points 1. and 3. by weighted versions. A weighted estimator for the fixed effects $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_w = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})||_{\varphi}, \tag{21}$$

where $\boldsymbol{W} = \text{diag}(w_{11}, \ldots, w_{gn_g}) \in \mathbb{R}^{N \times N}$ is a diagonal matrix of weights w_{ij}, $i = 1, \ldots, g, j = 1, \ldots, n_i$. Analogously, the weighted estimates of the random effects are given by

$$\hat{\boldsymbol{b}}_{i,w} = \arg\min_{\boldsymbol{b}} ||\tilde{\boldsymbol{W}}_i(\boldsymbol{r}_{i,\text{marg}} - \boldsymbol{Z}_i\boldsymbol{b})||_\varphi, \tag{22}$$

where $\tilde{\boldsymbol{W}}_i = \text{diag}(\tilde{w}_{i1}, \ldots, \tilde{w}_{in_i})$ is a matrix of group specific robustness weights.

The weights w_{ij} can be specified in different ways. We robustify the estimates against leverage points and downweigh observations that are associated with influential predictors. The weights are determined based on a robust estimate of leverage. Let

$$w_{ik} = \min\left\{1, \frac{c_p}{\nu_{ik}}\right\}, \text{ where } \nu_{ik} = (\boldsymbol{x}_{ik} - \boldsymbol{v}_c)'\boldsymbol{V}^{-1}(\boldsymbol{x}_{ik} - \boldsymbol{v}_c), \tag{23}$$

for $i = 1, \ldots, g$ and $k = 1, \ldots, n_i$. The cutoff c_p is selected as the 95%-quantile of the $\chi^2(p-1)$-distribution, where $p-1$ is the number of columns of $\boldsymbol{X}$. $\boldsymbol{V}$ and $\boldsymbol{v}_c$ are robust estimates of the covariance and mean of the predictor matrix $\boldsymbol{X}$. We use the (fast) minimum covariance determinant estimator (MCD, Hubert et al. 2018). The robustification against leverage points using the MCD estimator assumes that the predictors are continuous and roughly elliptically distributed. In the case that the predictor matrices include dummy variables, an alternative way to construct weights is offered based on the hat matrix, see e.g., Cantoni and Ronchetti (2001). A short discussion can be found in Section S5 of the supplement.

For the group-specific weights in (22), we proceed analogously, and calculate the leverage based on the matrices $\boldsymbol{Z}_i$. Note that the matrix $\boldsymbol{W}$ needs to be recalculated based on the rescaled observations in every iteration. The group-specific weights in $\tilde{\boldsymbol{W}}_i$ only need to be determined once, as the matrices $\boldsymbol{Z}_i$ are not modified in the iteration steps. The detailed estimation procedure is given in Algorithm A.2 in the appendix.

The approach of downweighting data points with high leverage can cause a loss of efficiency, as "good" leverage points might receive the same weights as "bad" leverage points. To further improve estimation, McKean (2004) suggests using the so-called *Wilcoxon high-breakdown (HBR)* estimator as an alternative. Based on an initial *least trimmed squares* (LTS) fit, it determines weights that account for outlyingness in both predictor- and response space. The extension of our methodology to this estimator is straightforward.

4. Simulation

In the following, we assess the statistical properties and robustness of our two rank-based procedures in simulation studies. We refer to the estimator without additional leverage weights (Algorithm A.1) as *Rank*, and to the estimator with leverage weights (Algorithm A.2) as *Weighted Rank*. The proposed procedures are implemented in the package **rankLME** which is available on GitHub (Brune 2024). For comparison, we consider the classic REML estimator as implemented in the R-package **lme4** (Bates et al. 2015). Also, we compare with the estimator proposed by Koller (2013). This estimator is based on a different robustness concept, namely MM-estimation, and implemented in the R-package **robustlmm** (Koller 2016). It is denoted by SMDM in the following.

All simulations are carried out on a Linux server using R, version 3.6.3 (R Core Team 2020). Replication files for the simulation studies can be found in the GitHub repository (Brune 2024).

For the simulation study, we start by generating an uncontaminated data set according to model equation (8). The entries of the predictor matrices $\boldsymbol{X}_i \in \mathbb{R}^{n_i \times (p-1)}$ are drawn from the $\mathcal{N}(0, 4)$ distribution, and the random slope matrices $\boldsymbol{Z}_i \in \mathbb{R}^{n_i \times (k-1)}$ contain the first $(k-1)$ columns of $\boldsymbol{X}_i$. We add a fixed intercept α and set $(\alpha, \boldsymbol{\beta}')' = \mathbf{1}_p$; also random intercepts a_i, $i = 1, \ldots, g$ are included. The random effects are distributed as $(a, \boldsymbol{b}')' \sim \mathcal{N}(0, 0.5^2 \boldsymbol{I}_k)$. We limit the analysis to $p = 4$ and $k = 2$. Thus, we have regression coefficients $(\alpha, \boldsymbol{\beta}')' = (\alpha, \beta_1, \beta_2, \beta_3)$ and random effects (a, b_1) with scale parameters $\boldsymbol{\theta} = (\theta_0, \theta_1)$. The errors are drawn as $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \boldsymbol{I}_{n_i})$.

In the next step, up to 50% of the observations are contaminated with one of the following outlier types:

- *y-* or *response outliers*: Observations y_{ij} are modified as either $y_{ij} + s$ (*additive*) or $y_{ij} \cdot s$ (*multiplicative*) for different outlier sizes s
- *leverage points* or *outlying predictors*: Rows of the predictor matrices are modified as $\boldsymbol{x}_{ij} + \ell$ (*additive*) or $\boldsymbol{x}_{ij} \cdot \ell$ (*multiplicative*) where $\boldsymbol{x}_{ij} = (x_{ij,1}, \ldots, x_{ij,p-1}) \in \mathbb{R}^{p-1}$ denotes the j'th row of the matrix $\boldsymbol{X}_i$, and ℓ is the leverage coefficient. The corresponding rows of the matrices $\boldsymbol{Z}_i$ are adjusted accordingly.

The outlying observations can either be spread *randomly* over the g groups, or be located *sequentially*. In the sequential case, one observation at a time is replaced and we achieve groupwise contamination.

Each setting is repeated for $R = 200$ replications. We report the MSE of the estimates:

$$\text{MSE}(\hat{\beta}_i) = \frac{1}{R} \sum_{r=1}^{R} (\hat{\beta}_i^{(r)} - 1)^2, \tag{24}$$

where $\beta_i^{(r)}$ is the the estimate of the i'th component of the coefficient vector $\boldsymbol{\beta}$, $i = 0, \ldots, p-1$, obtained in the r'th simulation run, $r = 1, \ldots, R$. β_0 corresponds to the intercept parameter α.

In the interest of clarity, we only report results for randomly located multiplicative y-outliers and leverage points. Further simulation results for additive and sequentially located outliers can be found in the supplemental material.

4.1. Bias of the estimates

To explore the bias of the proposed estimators, we repeatedly draw data sets from the null model (i.e., without outliers), and fit the rank-based mixed effects model as described in Section 3.

The top row of Figure 1 shows boxplots of the coefficient estimates for different group sizes $n_i = \tilde{n}$, $i = 1, \ldots, g$, and numbers of groups g. The boxplots are symmetrical and centered around the true values of the coefficients $\alpha = \beta_1 = \beta_2 = \beta_3 = 1$. Even for small total sample sizes, the estimates give no evidence for a bias, the median estimate

resembles the true value. Increasing the number of groups g causes a decrease in the estimates' variance. The effect of increasing group sizes n_i is less pronounced. Having many groups seems to be more important than large groups. We observe a higher variance for the coefficients that are associated with random effects (α and β_1). This is the case as the same predictors are used for estimation of both the fixed and the random effects. Therefore, higher uncertainty is introduced to the estimation.

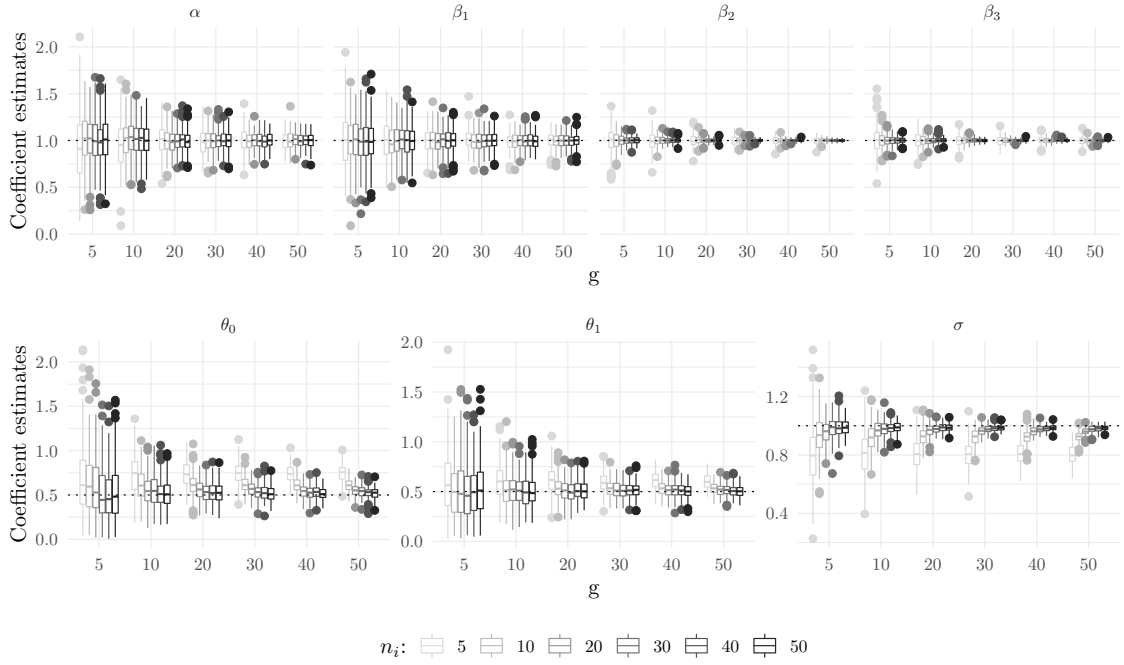

Figure 1: Boxplots of regression coefficient and variance component estimates obtained from the rank-based estimator for different combinations of group sizes n_i and number of groups g (total sample size: $N = g \cdot n_i$).

The respective estimates of the variance coefficients (or rather standard deviations) are displayed in the bottom row of Figure 1. The number of groups g and the number of observations per group n_i both have strong influence on the accuracy of the estimates. The number of groups g, guides the variation of the coefficient estimates. The larger g, the more random effects are available for variance estimation, and thus the variation of the estimates decreases. Our method overestimates θ_0 and θ_1 if the group sizes n_i are small. Due to the two-stage approach for estimation, our method cannot make use of the self-regularizing properties that evolve from direct likelihood-based estimation of the variance parameters (with REML). Especially for small groups, the ratio of observations to parameters for the regression is comparably small. Thus, we tend to overfit the groups and the variance of the estimated random effects $\boldsymbol{b}_i,\ i = 1, \ldots, g$, is overestimated. The contrary effect can be observed for the scale parameter of the errors, σ, which tends to be underestimated: The good fit in each group can reduce the variance of the conditional residuals. The larger the groups, the more degrees of freedom remain, and thus the better the variance is conserved.

However, our methodology yields accurate estimates of the random effects. These can help with the interpretation of the model and the identification of unusual observations. This is especially helpful if we want to run model diagnostics and interpret the realizations of the random effects (also see the applications in Section 5). As soon as n_i is large enough (this study indicates $n_i \geq 20$), all estimates seem to be consistent, regardless of the number of groups g.

The results for the weighted rank-based estimator are similar and provided in the supplemental material.

4.2. Validity of the updating step and efficiency

It is well-known that non-parametric methods are outperformed by parametric methods if the sample sizes are very small and the assumptions of the parametric approaches are valid.

Table 1 reports the relative efficiency (Serfling 1980) of our rank-based estimator as opposed to the baseline REML, the SMDM and the weighted rank-based estimator, as well as the efficiency of the rank-based estimator to that of its initial value (i.e., the estimate $\boldsymbol{\beta}^{(0)}$ obtained before iteration).

Table 1: Efficiency of the rank-based estimator vs. the initial value, REML, SMDM and Weighted Rank for $g, n_i \in \{5, 20, 50\}$.

g	n_i	initial value	REML	SMDM	Weighted Rank
5	5	0.858	1.159	1.112	0.978
5	20	0.730	1.048	1.046	0.994
5	50	0.712	1.085	1.048	1.003
20	5	0.821	1.134	1.084	0.966
20	20	0.728	1.041	1.017	0.987
20	50	0.726	1.080	1.013	0.988
50	5	0.796	1.163	1.097	0.969
50	20	0.718	1.068	1.031	0.985
50	50	0.768	1.084	1.021	1.000

The rank-based estimator is slightly less efficient than REML and SMDM, but its efficiency approaches that of REML (and SMDM) with increasing total sample size N. The weighted version of our estimator is slightly less efficient due to possible down-weighting of useful observations / leverage points. The updating procedure yields a considerable improvement over the initial value. Further simulation studies shown in the supplemental material confirm that the improvement induced by the updating step is even more pronounced in the presence of outliers.

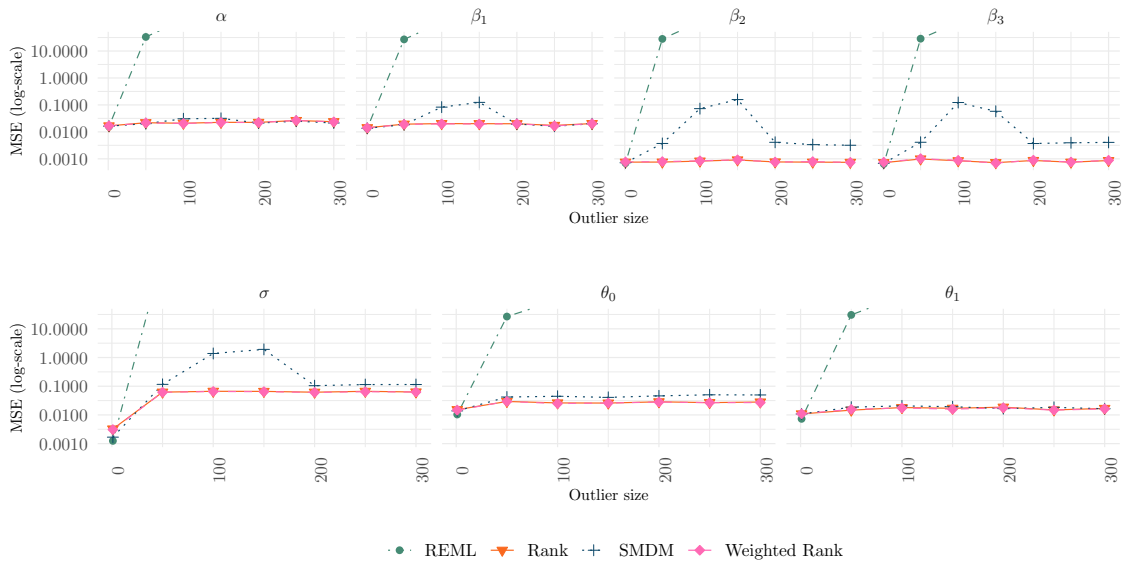

Figure 2: MSE (log-scale) of the SMDM, REML, Rank and Weighted Rank estimators for a 10% proportion of multiplicative response outliers of increasing size, $n_i = g = 20$. The curves for Rank and Weighted Rank are overplotted.

4.3. Behavior in the presence of outliers

In the following, we examine the behavior of the estimates under contamination with randomly located multiplicative response outliers and leverage points. We limit our analyses to the setting with $g = 20$ and $n_i = 20$. Further simulation results are reported in the supplemental material.

Response outliers

Figure 2 shows the MSEs of the four estimators for the situation of 10% response outliers of increasing size s. The MSEs of the two rank-based estimators and SMDM are fairly constant with increasing outlier size. However, SMDM seems to require a certain outlier size to make sure they are recognized and accounted for by the algorithm (explaining the higher error at medium outlier sizes. At the same time, the REML estimates are severely affected by even small contaminations. Thus, as soon as we suspect outliers to be present in the samples, it is definitely worth applying a robust estimator.

Figure S3 in the supplemental material examines how the estimator reacts to an increasing proportion of response outliers of size $s = 1000$. The estimates of α and $\boldsymbol{\beta}$ provided by the SMDM estimator break down at an outlier proportion larger than 20%. The estimates from the rank-based and weighted rank-based estimator stay stable longer (up to 30% contamination).

Outlying predictors

The MSEs for the coefficient estimates under 10% contamination with leverage points are reported in Figure 3. This setting clearly shows the advantage of the leverage

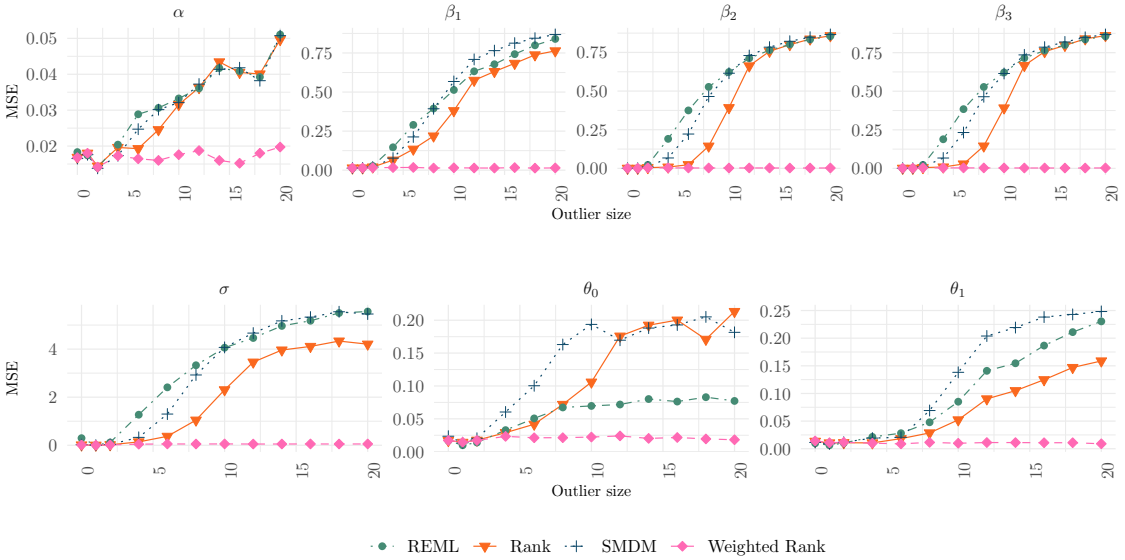

Figure 3: MSE of the SMDM, REML, Rank and Weighted Rank estimators for a 10% proportion of multiplicative leverage points of increasing size, $g = n_i = 20$

weighted rank-based estimator: While REML, SMDM and the non-weighted rank-estimator are severely distorted by the leverage points, the weighted rank-based estimator remains largely unaffected. Its coefficient estimates remain constant with increasing outlier size ℓ.

Figure S7 in the supplemental material examines how the estimator reacts to an increasing proportion of leverage points with $\ell = 100$. The estimates of the leverage weighted estimator are stable up to an outlier proportion of at least 20%, while the other three estimators are distorted even for small outlier proportions.

5. Applications

In the following section we illustrate the performance of the (weighted) rank-based estimation method in three different applications, parts of which are mixtures of real data analysis and introduction of artificial outliers. Especially, we illustrate different diagnostic tools that may help with identifying outliers or outlying / unusual groups in the data. Replication files for the data analyses can be found in the GitHub repository (Brune 2024).

5.1. Sleep study data

We analyze the *sleepstudy* dataset (Belenky et al. 2003) as included in the **lme4** R-package (Bates et al. 2015). It consists of data from 18 subjects that were exposed to sleep deprivation and had to complete a reaction test each day. Thus, we expect the average reaction time to decrease throughout the experiment, but each subject might react differently to the sleep deprivation.

We fit a mixed effects model with both random and fixed intercept and slope in time. The model equation for the reaction time of subject i ($i = 1, \ldots, 18$) after t ($t = 0, \ldots, 9$)

days of sleep deprivation is given by

$$\text{Reaction}_{i,t} = \alpha + \boldsymbol{\beta}_1 \cdot t + a_i + \boldsymbol{b}_i \cdot t + \varepsilon_{i,t}. \tag{25}$$

We contaminate the data set in two different ways:

1. Modify the third observation of each individual by multiplying the reaction times by three (*outlying observations*). Outliers like this could, e.g., be caused by an error in the measurement instrument that was used on that day.

2. Modify the first group (individual '308') by multiplying the reaction times by three (*outlying group*). The subject here could happen to be very slow, but this fact should not affect the population average (i.e., fixed effects) too strongly.

Since the predictor in model (25) is exposure time, we do not expect leverage points. Thus, we compare the estimates from our procedure without leverage weights (*rank*, Algorithm A.1) to the standard REML estimator. The original coefficient estimates and their ratios to the estimates from the two contamination settings are displayed in Table 2. For the rank-based model, the estimates of α and β_1 remain stable in all contamination settings, they deviate by at most 2% from the baseline estimates.

Table 2: Coefficients estimated by REML and the rank-based estimator for the unmodified sleep study data and the ratio to coefficient estimates for the two outlier settings; boldface numbers indicate a relative change of more than 10%.

Setting	Method	α	β_1	θ_0	θ_1	σ
baseline	rank	252.10	10.63	31.28	6.56	16.56
	REML	251.41	10.47	25.05	5.99	25.57
1.	rank	1.02	0.99	1.07	**1.23**	**1.26**
	REML	**1.29**	**0.67**	-	-	**7.14**
2.	rank	1.01	1.01	**1.20**	1.00	1.06
	REML	**1.11**	**1.23**	**4.78**	**2.39**	**1.59**

REML returns a singular fit in the *outlying observations* setting, the imposed random effects structure cannot be fitted with **lme4**. As a consequence, the error variance σ is overestimated. The estimate returned by the rank-based method remains more stable. The residuals from REML are strongly skewed (see Figure 4). Due to the outlying observations in each group, the REML regression lines are shifted upwards and a majority of the residuals is less than zero. Two of the outlying values are not recognized as such and fall within the cutoff of 2. For the rank-based method, the residuals are still centered around zero. The outlying values are recognized correctly. A few more observations are marked as conspicuous and could require further attention. However, they already fell outside of the cutoff area without the artificial outliers.

In the *outlying group* setting, we observe problems with the variance structure imposed by the model: REML severely overestimates the variance components due to the outlying random effect. As an indicator whether the imposed variance structure, i.e., the

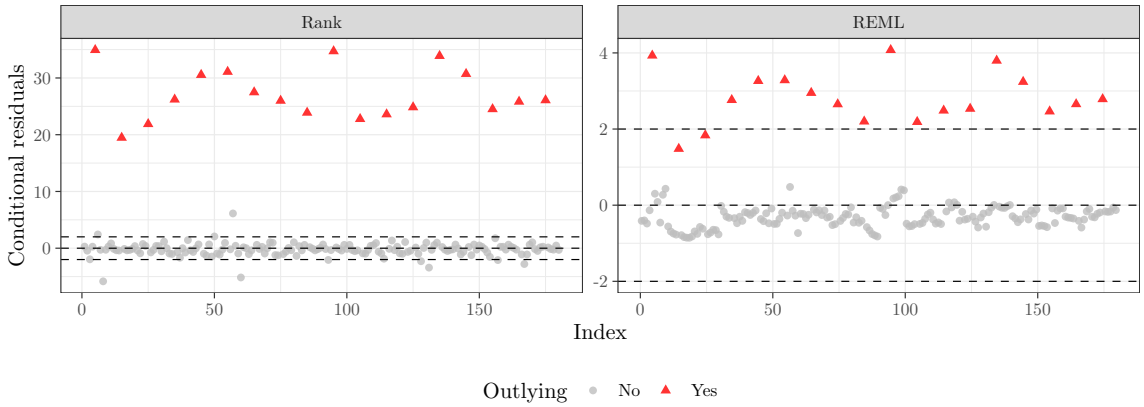

Figure 4: Conditional residuals for the sleep study example in the outlying observations setting; REML and rank-based estimator.

random effects structure, is reasonable, Lesaffre and Verbeke (1998) suggest reporting the following norm

$$d_i = \left|\left| I_{n_i} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{-1/2} \boldsymbol{r}_{i,\text{marg}} \boldsymbol{r}'_{i,\text{marg}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{-1/2} \right|\right|^2 \tag{26}$$

for $i = 1, \ldots, g$. d_i may be interpreted as a residual which measures how well the covariance structure of the observations $\boldsymbol{y}_i$ is captured by the model. Thus, d_i should be small if the variance structure is captured well, and should produce large values if the variance structure in one of the groups does not match that imposed by the model. A plot of the group index against the variance diagnostic (26) yields Figure 5. The outlying group cannot be recognized in the case of REML, the magnitudes of the d_i's are all fairly similar. The outlying individual (308) does not stand out. The d_i's calculated from the rank-based fit clearly capture the outlying group.

The examples of contamination here show that outlying values can be masked by the outcome of the standard REML estimation approach. The standard diagnostic methods fail in this case. We cannot reliably detect the outlying observations or groups. Opposed to this, our robust rank-based estimation method allows to clearly identify the unusual observations and thus enables further analysis.

5.2. Accelerated aging experiments for photovolatic module data

The ADVANCE! project analyzes the aging behavior of photovoltaic (PV) modules (Berger et al. 2021). Different modules were exposed to various climatic conditions in accelerated aging experiments. Such experiments are used as the real-time assessment of material degradation is not possible due to the long time frames in real life. Throughout the experiments, chemical and electric measurements were taken. We are especially interested in modeling the *power at maximum power point* (P_{MPP}) over time, and as a function of other degradation indicators. The goal is to identify degradation pathways and understand the influence of treatment with different climatic settings on changes in material and power loss. The data is available as Knoebl et al. (2024).

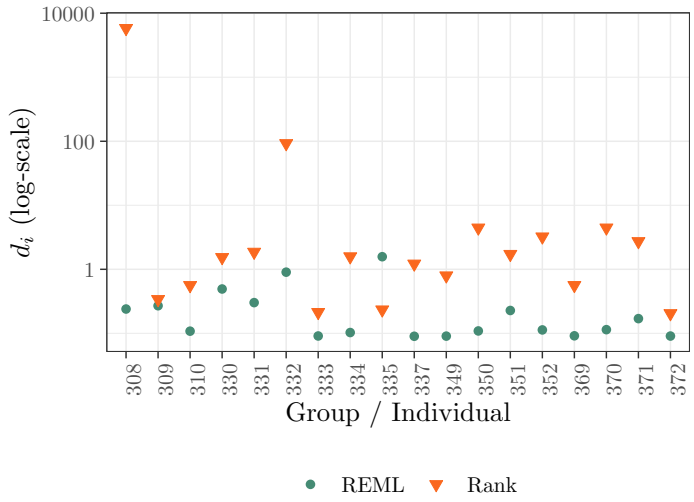

Figure 5: Variance diagnostic (26) for the sleep study example in the outlying group setting; REML and rank-based estimator.

Pairwise comparison of degradation rates

We compare the results from the two climatic settings *Moderate 2* and *Moderate 5*. In both settings three modules were treated with damp-heat (i.e., high temperature and humidity) and irradiance for 1,200 hours. The modules in the Moderate 5-group were additionally exposed to temperature cycles. We fit the following model:

$$(\mathrm{P_{MPP}})_{it} = 1 + \beta_1 \cdot \mathrm{Ramp}(t) + \beta_2 \cdot t^2 + \beta_3 \cdot t^2 \cdot \mathbf{1}\{\text{Moderate 5}\} + b_i \cdot t^2 + \varepsilon_{i,t}. \quad (27)$$

β_1 captures an initial power increase, β_2 the overall aging, and β_3 the additional aging induced by the treatment with temperature cycles. The intercept / initial power is fixed at 1. The random effects b_i account for the module specific degradation rates over time.

As can be seen from Figure 6, one module broke during the aging procedure. Thus, we expect the REML estimate of β_3 to be biased downwards trying to account for the strong power loss of the faulty module. Since we are interested in modeling the continuous material wear, and not the abrupt degradation caused by material failures, the robust estimate gives a more reliable characterization of the degradation rate.

The resulting estimates for $(\beta_1, \beta_2, \beta_3)'$ are $(0.022, -0.014, -0.022)'$ for the rank-based method, and $(0.022, -0.013, -0.031)'$ for REML. Thus, the degradation for *Moderate 5* captured in β_3 is estimated to be about 40% larger by REML than by the rank-based method.

Degradation modeling with spectral measurements

In this analysis, we connect the power loss in $\mathrm{P_{MPP}}$ with the material aging measured using fitted spectra obtained from Fourier-transform infrared (FTIR) spectroscopy. The changes in the spectra over time indicate changes in the chemical composition of the materials. The spectral measurements are observed as functions which are fitted automatically to locate peaks and calculate their area. These fits can go wrong, causing

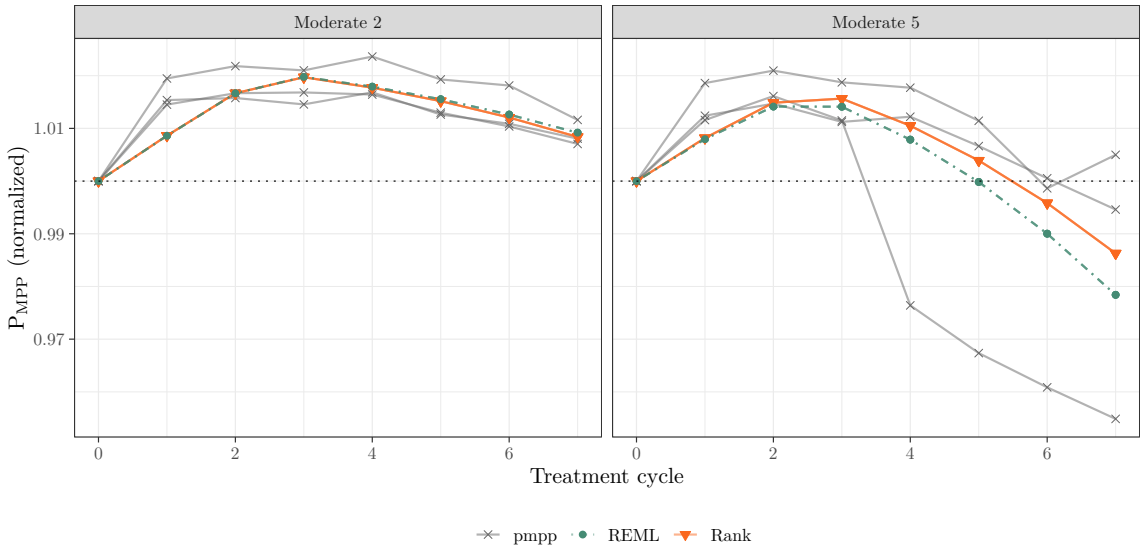

Figure 6: Data and fitted values for pairwise comparison of the *Moderate 2* and *Moderate 5* aging settings.

outlying predictors / leverage points in the resulting data points. We model the degradation for two different climatic settings: *Tropical 1* and *Tropical 2* treated with 85°C / 90°C and 85% / 90% humidity respectively. We aim to explain the power loss over time caused by the treatment with higher temperature and humidity for the Tropical 2-modules through the changes in three peaks areas extracted from the spectral measurements, namely those at wavenumbers (WN) 795, 872 and 3426 cm^{-1}. The standardized peaks and P_{MPP} are shown in Figure 7. As can be seen, the peak areas tend to increase with increasing exposition time. Thus, their influence on the decreasing P_{MPP} should be negative. We fit the model

$$\begin{aligned}(P_{MPP})_{it} &= 1 + \beta_1 \cdot \text{Ramp}(t) + \beta_2 \cdot \text{WN } 795_{it} + \beta_3 \cdot \text{WN } 872_{it} \\ &\quad + \beta_4 \cdot \text{WN } 3426_{it} + b_i \cdot t^2 + \varepsilon_{it}\end{aligned}$$

using REML and our two rank-based methods. The observations that are downweighted by the weighted rank-based estimator are marked as crosses in Figure 7. Hence, the algorithm successfully recognizes the outlying predictors, especially for the peak area at wavenumber 3,426 cm^{-1}.

The fitted coefficients $(\beta_1, \beta_2, \beta_3, \beta_4)'$ are $(0.0428, 0.0006, -0.0051, 0.0007)'$ for REML and $(0.0324, -0.0004, -0.0028, -0.0023)'$ for the weighted rank-based method. The signs of the estimates for β_2 and β_4 differ between the two estimation methods. The sign is positive for REML, but negative for our weighted rank-based method. As noticed before, the shapes of the curves indicate that a negative sign might be more realistic. For β_4, this change in sign is probably caused by the module marked by dashed lines in the WN 3,426 cm^{-1} series. The large leverage points lead the non-robust model to believe the peak has positive influence on the degradation. Accounting for the leverage points with additional weights as we propose for the weighted rank-based procedure is essential. Our rank-based procedure without weights has similar problems as REML.

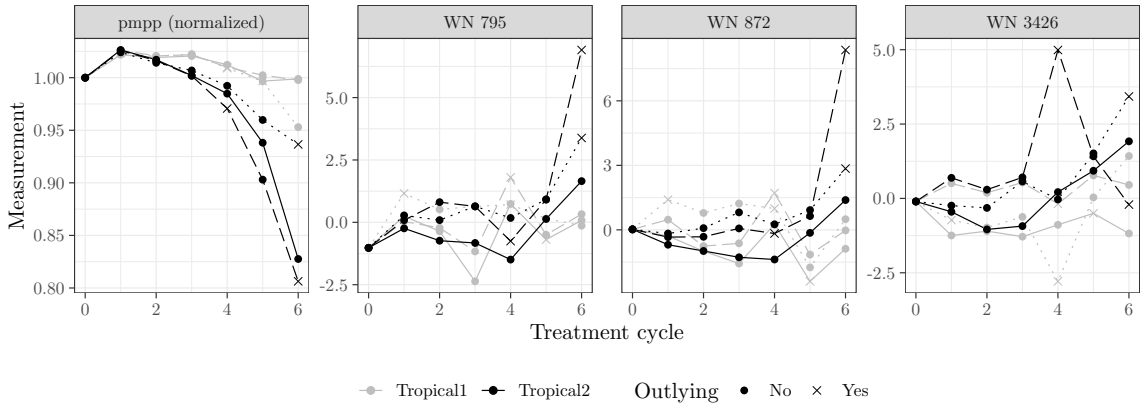

Figure 7: Peaks from FTIR spectra and P_{MPP} for aging settings Tropical 1 and Tropical 2; points with leverage weight (23) < 1 are marked by $\times$.

6. Summary and Outlook

The present work proposes a new approach to the rank-based estimation of mixed effects models. It extends existing methodology for mixed effects models with random intercepts with the possibility to model random slopes. The estimates obtained from the rank-based method are insensitive against outlying response values. In addition, the proposed method yields robustness in the presence of leverage points in the fixed or random effects predictor matrices. Our algorithm converges fast and can be implemented efficiently. An implementation in R is available on GitHub (Brune 2024).

The results of our simulation studies indicate that the bias of the estimates is insignificant: The estimates of the regression coefficients show no evidence of being biased. For small sample sizes, the variance of the random effects, especially of the random intercepts, tends to be slightly overestimated. This is inherent to the iterative nature of the algorithm. A possible mitigation is the derivation of correction factors for the estimation of the random intercepts. We plan to explore this in detail in future research.

Under normality, the estimates obtained from our methodology compare well with those obtained from the classic REML estimation. At the same time, the rank-based approach has advantages in situations where the error- or random effects distributions deviate from normality. Although we fit more parameters than for the standard estimation approaches based on REML, we still gain in efficiency due to our updating procedure. Accounting for the covariance structure reduces the variance of the fixed effects estimator $\hat{\boldsymbol{\beta}}$. The applications show that the resulting estimates enable model diagnostics and help with the identification of unusual observations or groups in the data.

There are different lines of research that can be pursued from this point. One is to explore the consistency and asymptotic distribution of our estimators in theory. This would allow for formal inference on the fixed effects estimates. As the methodology is not yet able to estimate more complex random effects structures such as nested and crossed effects, it would be interesting to extend the model in this direction. An alternative to the rank-based estimators based on Jaeckel's dispersion function applied in this work could be offered by maximum rank correlation estimators, first proposed

by Han (1987). The robustness properties of these estimators have been examined in detail in Alfons et al. (2017).

Acknowledgments

This work was conducted as part of the Austrian Energy Research Program project ADVANCE! (No. 881133), funded by the Austrian Climate and Energy Fund and the Austrian Research Promotion Agency (FFG). We thank the referees for their comments and suggestions that helped us improve the paper.

References

Abebe, A., McKean, J. W., Kloke, J. D., and Bilgic, Y. K. (2016). Iterated reweighted rank-based estimates for GEE models. In Liu, R. Y. and McKean, J. W., editors, *Robust Rank-Based and Nonparametric Methods*, volume 168, pages 61–79. Springer International Publishing, Cham, ISBN: 978-3-319-39063-5 978-3-319-39065-9, DOI: 10.1007/978-3-319-39065-9_4.

Agostinelli, C. and Yohai, V. J. (2016). Composite robust wstimators for linear mixed models. *Journal of the American Statistical Association*, 111(516):1764–1774, DOI: 10.1080/01621459.2015.1115358.

Alfons, A., Croux, C., and Filzmoser, P. (2017). Robust maximum association estimators. *Journal of the American Statistical Association*, 112(517):436–445, DOI: 10.1080/01621459.2016.1148609.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using **lme4**. *Journal of Statistical Software*, 67(1), DOI: 10.18637/jss.v067.i01.

Belenky, G., Wesensten, N. J., Thorne, D. R., Thomas, M. L., Sing, H. C., Redmond, D. P., Russo, M. B., and Balkin, T. J. (2003). Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study. *Journal of Sleep Research*, 12(1):1–12, DOI: 10.1046/j.1365-2869.2003.00337.x.

Berger, K., Knöbl, K., Brune, B., Ortner, I., Grubner, S., Oreski, G., Barretta, C., Kubicek, B., Neumaier, L., Feichtner, M., Voronko, Y., and Eder, G. (2021). Advanced degradation modelling of photovoltaic modules and materials. In *38th European Photovoltaic Solar Energy Conference and Exhibition*, pages 745–748, DOI: 10.4229/EUPVSEC20212021-4AV.1.19.

Bilgic, Y. K. (2012). *Rank-Based Estimation and Prediction for Mixed-Effects Models in Nested Designs*. Dissertation, Western Michigan University, https://www.proquest.com/dissertations-theses/rank-based-estimation-prediction-mixed-effects/docview/1041087025/se-2.

Brune, B. (2024). ***rankLME**: Robust Rank-based Estimation of Linear Mixed Effects Models*, DOI: 10.5281/zenodo.13843795, https://github.com/b-brune/rankLME. R package version 1.0.0.

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96(455):1022–1030, DOI: 10.1198/016214501753209004.

Copt, S. and Victoria-Feser, M.-P. (2006). High-breakdown inference for mixed linear models. *Journal of the American Statistical Association*, 101(473):292–300, DOI: 10.1198/016214505000000772.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model. *Journal of Econometrics*, 35(2-3):303–316, DOI: 10.1016/0304-4076(87)90030-3.

Hettmansperger, T. P. and McKean, J. W. (2011). *Robust Nonparametric Statistical Methods*. Chapman & Hall/Crc Monographs on Statistics & Applied Probability. CRC Press, Boca Raton, FL, 2nd edition, ISBN: 978-1-4398-0908-2.

Hodges, J. L. and Lehmann, E. L. (1963). Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, 34(2):598–611, DOI: 10.1214/aoms/1177704172.

Hubert, M., Debruyne, M., and Rousseeuw, P. J. (2018). Minimum covariance determinant and extensions. *WIREs Computational Statistics*, 10(3), DOI: 10.1002/wics.1421.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *The Annals of Mathematical Statistics*, 43(5):1449–1458, DOI: 10.1214/aoms/1177692377.

Jung, S.-H. and Ying, Z. (2003). Rank-based regression with repeated measurements data. *Biometrika*, 90(3):732–740, DOI: 10.1093/biomet/90.3.732.

Jureckova, J. (1971). Nonparametric estimate of regression coefficients. *The Annals of Mathematical Statistics*, 42(4):1328–1338, DOI: 10.1214/aoms/1177693245.

Kloke, J. D., McKean, J. W., and Rashid, M. M. (2009). Rank-based estimation and associated inferences for linear models with cluster correlated errors. *Journal of the American Statistical Association*, 104(485):384–390, DOI: 10.1198/jasa.2009.0116.

Knoebl, K., Oreski, G., Neumaier, L., and Brune, B. (2024). SQL database with detailed characteristic measurement results of photovoltaic modules that were subjected to accelerated aging sequences. DOI: 10.5281/ZENODO.13354817.

Koller, M. (2013). *Robust Estimation of Linear Mixed Models*. Dissertation, ETH Zurich, https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/65478/eth-6670-02.pdf

Koller, M. (2016). **robustlmm**: An R package for robust estimation of linear mixed-effects models. *Journal of Statistical Software*, 75(6), DOI: 10.18637/jss.v075.i06.

Lesaffre, E. and Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, 54:570–584. DOI: 10.2307/3109764.

McKean, J. W. (2004). Robust analysis of linear models. *Statistical Science*, 19(4):562–570, DOI: 10.1214/088342304000000549.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer, New York, ISBN: 978-0-387-98957-0.

Pinheiro, J. C., Liu, C., and Wu, Y. N. (2001). 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, DOI: 10.1198/10618600152628059.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, DOI: 10.1080/01621459.1993.10476408.

Rousseeuw, P. J. and Van Driessen, K. (2006). Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery*, 12(1):29–45, DOI: 10.1007/s10618-005-0024-4.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA, ISBN: 978-0-470-31648-1 978-0-471-02403-3, DOI: 10.1002/9780470316481.

Wang, Y.-G. and Zhao, Y. (2008). Weighted rank regression for clustered data analysis. *Biometrics*, 64(1):39–45, DOI: 10.1111/j.1541-0420.2007.00842.x.

Wang, Y.-G. and Zhu, M. (2006). Rank-based regression for analysis of repeated measures. *Biometrika*, 93(2):459–464, DOI: 10.1093/biomet/93.2.459.

A. Estimation Algorithms

Algorithm A.1 Rank-based fitting of mixed effects models with random slopes.

Starting with $l = 0$, the l'th iteration of the algorithm is the following:

1. Let

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(l)} = \begin{cases} \boldsymbol{I}_{n_i}, & \text{if } l = 0 \\ \boldsymbol{\Sigma}_{\boldsymbol{y}_i}(\hat{\sigma}_\varepsilon^{(l-1)},\ \hat{\boldsymbol{\theta}}_b^{(l-1)}), & \text{else,} \end{cases}$$

and

$$\boldsymbol{\nu}^{(l)} = \begin{cases} \mathbf{1}_N, & \text{if } l = 0 \\ \left(\nu_{11}^{(l-1)}, \ldots, \nu_{gn_g}^{(l-1)}\right)', & \text{else.} \end{cases}$$

Obtain $\hat{\boldsymbol{\beta}}^{(l)}$ as a solution to the rank-based regression of

$$\boldsymbol{y}_*^{(l)} = (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(l)})^{-1/2} \operatorname{diag}\left(\boldsymbol{\nu}^{(l)}\right) \boldsymbol{y} \quad \text{on} \quad \boldsymbol{X}_*^{(l)} = (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(l)})^{-1/2} \operatorname{diag}\left(\boldsymbol{\nu}^{(l)}\right) \boldsymbol{X}.$$

If the model has an intercept α, estimate

$$\hat{\alpha}^{(l)} = T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(l)}).$$

Otherwise, $\hat{\alpha}^{(l)} = 0$.

2. Calculate the marginal residuals

$$\boldsymbol{r}_{i,\text{marg}}^{(l)} = \boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}^{(l)} - \hat{\alpha}^{(l)}\mathbf{1}_{n_i}.$$

3. Use $\boldsymbol{r}_{i,\text{marg}}^{(l)}$ to obtain predictions $\hat{\boldsymbol{b}}_i^{(l)}$ for the random effects $\boldsymbol{b}_i$ from (13).
If the model has random intercepts a_i, estimate

$$\hat{a}_i^{(l)} = T(\boldsymbol{r}_{i,\text{marg}}^{(l)} - \boldsymbol{Z}\hat{\boldsymbol{b}}_i^{(l)}).$$

Otherwise, $\hat{a}_i^{(l)} = 0$.

4. Obtain conditional residuals

$$\boldsymbol{r}_{i,\text{cond}}^{(l)} = \boldsymbol{y}_i - \hat{\alpha}^{(l)}\mathbf{1}_{n_i} - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}^{(l)} - \hat{a}_i^{(l)}\mathbf{1}_{n_i} - \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i^{(l)}.$$

5. Use the $\boldsymbol{r}_{i,\text{cond}}^{(l)}$ and the realizations $\hat{a}_i^{(l)}$, $\hat{\boldsymbol{b}}_i^{(l)}$ to estimate the variance components $\boldsymbol{\theta}^{(l)}$ and $\sigma_\varepsilon^{(l)}$ with the Q_n estimator (see Rousseeuw and Croux 1993).

6. Determine the outlyingness weights $\nu_{ij}^{(l)}$ from the conditional residuals using (18).

7. If $l > 0$, check for convergence using the criteria

$$\frac{||\hat{\boldsymbol{\beta}}^{(l)} - \hat{\boldsymbol{\beta}}^{(l-1)}||_2}{||\hat{\boldsymbol{\beta}}^{(l-1)}||_2} < \text{tol}, \quad \text{and} \quad \frac{||(\hat{\sigma}_\varepsilon^{(l)}, \hat{\boldsymbol{\theta}}_b^{(l)})' - (\hat{\sigma}_\varepsilon^{(l-1)}, \hat{\boldsymbol{\theta}}_b^{(l-1)})'||_2}{||(\hat{\sigma}_\varepsilon^{(l-1)}, \hat{\boldsymbol{\theta}}_b^{(l-1)})'||_2} < \text{tol}$$

for a prescribed tolerance tol.

If above criteria are not fulfilled or $l = 0$, increase l by one and return to step 1.

Algorithm A.2 Rank-based fitting of mixed effects models with random slopes and leverage weights.

Starting with $l = 0$, the l'th iteration of the algorithm is the following:

1. Calculate the group-specific matrices of robustness weights

$$\tilde{\boldsymbol{W}}_i = \operatorname{diag}(\tilde{w}_{i1}, \ldots, \tilde{w}_{in_i})$$

by applying (23) to the matrices $\boldsymbol{Z}_i$, $i = 1, \ldots, g$. These weights only need to be determined once.

2. Let

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(l)} = \begin{cases} \boldsymbol{I}_{n_i}, & \text{if } l = 0 \\ \boldsymbol{\Sigma}_{\boldsymbol{y}_i}(\hat{\sigma}_\varepsilon^{(l-1)},\ \hat{\boldsymbol{\theta}}_b^{(l-1)}), & \text{else,} \end{cases}$$

and

$$\boldsymbol{\nu}^{(l)} = \begin{cases} \mathbf{1}_N, & \text{if } l = 0 \\ (\nu_{11}^{(l-1)}, \ldots, \nu_{gn_g}^{(l-1)})', & \text{else.} \end{cases}$$

Determine the matrix of weights $\boldsymbol{W}^{(l)}$ as in (23) based on the rescaled matrix of covariates

$$\boldsymbol{X}_*^{(l)} = (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(l)})^{-1/2} \operatorname{diag}\left(\boldsymbol{\nu}^{(l)}\right) \boldsymbol{X}.$$

Obtain $\hat{\boldsymbol{\beta}}_w^{(l)}$ as a solution to the rank-based regression of

$$\boldsymbol{y}_{*,w}^{(l)} = \boldsymbol{W}^{(l)} (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(l)})^{-1/2} \operatorname{diag}\left(\boldsymbol{\nu}^{(l)}\right) \boldsymbol{y}$$

on

$$\boldsymbol{X}_{*,w}^{(l)} = \boldsymbol{W}^{(l)} (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(l)})^{-1/2} \operatorname{diag}\left(\boldsymbol{\nu}^{(l)}\right) \boldsymbol{X}.$$

If the model has an intercept α, estimate

$$\hat{\alpha}^{(l)} = T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_w^{(l)}).$$

otherwise, $\hat{\alpha}^{(l)} = 0$.

3. Calculate the marginal residuals

$$\boldsymbol{r}_{i,\text{marg}}^{(l)} = \boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}_w^{(l)} - \hat{\alpha}^{(l)} \mathbf{1}_{n_i}.$$

4. Use $\boldsymbol{r}_{i,\text{marg}}^{(l)}$ to obtain predictions $\hat{\boldsymbol{b}}_{i,w}^{(l)}$ for the random effects $\boldsymbol{b}_i$ from (22) using the weight matrices $\tilde{\boldsymbol{W}}_i$.

If the model has intercepts a_i, estimate

$$\hat{a}_i^{(l)} = T(\boldsymbol{r}_{i,\text{marg}}^{(l)} - \boldsymbol{Z}\hat{\boldsymbol{b}}_{i,w}^{(l)}).$$

otherwise, $\hat{a}_i^{(l)} = 0$

5. Obtain the final (conditional) residuals

$$\boldsymbol{r}_{i,\text{cond}}^{(l)} = \boldsymbol{y}_i - \hat{\alpha}^{(l)} \mathbf{1}_{n_i} - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}_w^{(l)} - \hat{a}_i^{(l)} \mathbf{1}_{n_i} - \boldsymbol{Z}_i \hat{\boldsymbol{b}}_{i,w}^{(l)}.$$

6. Use the $\boldsymbol{r}_{i,\text{cond}}^{(l)}$ and the realizations $\hat{\boldsymbol{b}}_i^{(l)}$ to estimate the variance components $\boldsymbol{\theta}^{(l)}$ and $\sigma_\varepsilon^{(l)}$ with the Q_n estimator (see Rousseeuw and Croux 1993).

7. Determine the outlyingness weights $\nu_{ij}^{(l)}$ from the conditional residuals using (18).

8. If $l > 0$, check for convergence using the criteria

$$\frac{||\hat{\boldsymbol{\beta}}_w^{(l)} - \hat{\boldsymbol{\beta}}_w^{(l-1)}||_2}{||\hat{\boldsymbol{\beta}}_w^{(l-1)}||_2} < \text{tol}, \quad \text{and} \quad \frac{||(\hat{\sigma}_\varepsilon^{(l)}, \hat{\boldsymbol{\theta}}_b^{(l)})' - (\hat{\sigma}_\varepsilon^{(l-1)}, \hat{\boldsymbol{\theta}}_b^{(l-1)})'||_2}{||(\hat{\sigma}_\varepsilon^{(l-1)}, \hat{\boldsymbol{\theta}}_b^{(l-1)})'||_2} < \text{tol}$$

   for a prescribed tolerance tol.

   If above criteria are not fulfilled or $l = 0$, increase l by one and return to step 2.

9. Given the final estimates $\hat{\boldsymbol{\beta}}_w$ and $\hat{\boldsymbol{b}}_{i,w}$, $i = 1, \ldots, g$, and if $\boldsymbol{Z}_i$ is a subset of $\boldsymbol{X}_i$, perform the adjustment step.

Affiliation:

Barbara Brune
Institute of Statistics and Mathematical Methods in Economics
TU Wien
Wiedner Hauptstraße 8-10, 1040 Wien, Austria
TÜV AUSTRIA Data Intelligence GmbH
Ferdinandstraße 4/1.OG, 1020 Wien, Austria
E-mail: barbara.brune@ymail.com

Irene Ortner
Digitas GmbH
Vienna, Austria
E-mail: irene.ortner@digitas.at

Peter Filzmoser
Institute of Statistics and Mathematical Methods in Economics
TU Wien
Wiedner Hauptstraße 8-10, 1040 Wien, Austria
E-mail: peter.filzmoser@tuwien.ac.at

Journal of Data Science, Statistics, and Visualisation https://jdssv.org/
published by the International Association for Statistical Computing
http://iasc-isi.org/

ISSN 2773-0689
November 2024, Volume IV, Issue VII *Submitted:* 2024-07-23
doi:10.52933/jdssv.v4i7.112 *Accepted:* 2024-10-21