METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

Parametric and Semi-Parametric Bootstrap-Based Confidence Intervals for Robust Linear Mixed Models

Fabio Mason [1], Eva Cantoni [2], Paolo Ghisletta [1,3,4]

[1] *Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland.* **[2]** *Research Center for Statistics and Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland.* **[3]** *Swiss National Centre of Competence in Research LIVES, University of Geneva, Geneva, Switzerland.* **[4]** *Faculty of Psychology, UniDistance Suisse, Brig, Switzerland.*

Methodology, 2021, Vol. 17(4), 271–295, https://doi.org/10.5964/meth.6607

Received: 2021-04-20 • **Accepted:** 2021-08-31 • **Published (VoR):** 2021-12-17

Corresponding Author: Fabio Mason, Université de Genève, boulevard du Pont d'Arve 40, CH-1211 Genève 4, Switzerland. E-mail: fabio.mason@unige.ch

Supplementary Materials: Materials [see Index of Supplementary Materials]

Abstract

The linear mixed model (LMM) is a popular statistical model for the analysis of longitudinal data. However, the robust estimation of and inferential conclusions for the LMM in the presence of outliers (i.e., observations with very low probability of occurrence under Normality) is not part of mainstream longitudinal data analysis. In this work, we compared the coverage rates of confidence intervals (CIs) based on two bootstrap methods, applied to three robust estimation methods. We carried out a simulation experiment to compare CIs under three different conditions: data 1) without contamination, 2) contaminated by within-, or 3) between-participant outliers. Results showed that the semi-parametric bootstrap associated to the composite tau-estimator leads to valid inferential decisions with both uncontaminated and contaminated data. This being the most comprehensive study of CIs applied to robust estimators of the LMM, we provide fully commented R code for all methods applied to a popular example.

Keywords

robustness, linear mixed models, bootstrap, confidence intervals, longitudinal data

The linear mixed model (LMM) has become the preferred choice of analysis in many longitudinal research settings, because it offers many advantages over other traditional methods: 1) it corrects estimation bias and consistency for the statistical dependencies due to multiple assessments on the same participants (which ordinary linear regression

does not); 2) it allows for flexible correlation structures among the repeated measurements, without needing to satisfy stringent assumptions, such as sphericity (which is difficult to achieve but required in repeated measures analysis of variance); and 3) it easily accommodates missing and unbalanced data, under the missing-completely-at-random or missing-at-random assumptions (Laird & Ware, 1982; Snijders & Bosker, 1999; Verbeke & Molenberghs, 2000).

Although the use of the LMM in longitudinal research is well established and frequent (e.g., Singer & Willett, 2003; Verbeke, 1997), the estimation of this class of models in the presence of outliers, that we define as observations associated to a very low probability of occurrence if they are assumed from a normal distribution, is not yet part of mainstream longitudinal data analysis (Koller, 2016). Outliers may be particularly vexing in longitudinal settings, where they can take two forms: 1) within-participant, where a participant with generally ordinary data points has one or a few outlying values, and 2) between-participant, where all data points of a given participant are outliers when compared to those of other participants. With recent technological developments, computational power has increased dramatically, allowing for the application of computer intensive methods also on personal computers (Efron & Hastie, 2016). These developments have facilitated and propagated the use of modern statistical methods to handle outliers, called *robust* methods (Andersen, 2008). We use the term robust in the sense of Huber and Ronchetti (2009), for whom: "robustness signifies insensitivity to small deviations from the assumptions" (p. 1). In this work, we contrast classical (non-robust) to robust estimation methods for the LMM. More precisely, we consider the most recent and comprehensive, to the best of our knowledge, comparison of robust LMM estimators (Agostinelli & Yohai, 2016).

At present, robust estimation of the LMM does not always allow for inferential conclusions, because standard statistical tests for classical estimation (e.g., t, F, or likelihood ratio) are not available with robust estimation (Kuznetsova, Brockhoff, & Christensen, 2017). Consequently, from an applied perspective it becomes difficult, if not impossible, to analyze correlated and contaminated data and draw sound inferential conclusions. Instead of seeking to derive p-values associated with null-hypothesis testing, however, we can rely on confidence intervals (CIs), which offer interpretative advantages over p-values (Cumming, 2014; Greenland et al., 2016; Wasserstein & Lazar, 2016; Wilkinson, 1999). To do so, we apply the bootstrap methodology (Efron, 1979; Goldstein, 2011) and implement it to the LMM estimation methods discussed in Agostinelli and Yohai (2016), and study their coverage (i.e., inclusion of the parameter value) and length.

The Linear Mixed Model

For $i = 1, \ldots, n$, let the model equation for the LMM be:

$$\boldsymbol{y}_i = X_i\boldsymbol{\gamma} + Z_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \text{ with } \boldsymbol{b}_i \perp \boldsymbol{\varepsilon}_i \quad (1)$$

where $\boldsymbol{y}_i$ is a vector of length J_i containing the responses of participant i, $\boldsymbol{\gamma}$ is a vector of coefficients for the p fixed effects, X_i is the $(J_i \times p)$ design matrix for fixed effects, $\boldsymbol{b}_i$ is the vector of random effects of length q, independent of the errors $\boldsymbol{\varepsilon}_i$ and Z_i is the $(J_i \times q)$ design matrix for the random effects. In the LMM language, fixed effects refer to regression coefficients that are constant between participants and thus do not need the subscript i (hence $\boldsymbol{\gamma}$ rather than $\boldsymbol{\gamma}_i$ in Equation 1). In contrast, random effects refer to quantities that can vary between participants and thus are designated with the subscript i (e.g., $\boldsymbol{b}_i$ rather than $\boldsymbol{b}$). The vector $\boldsymbol{\varepsilon}_i = \left(\varepsilon_{i1}, \ldots, \varepsilon_{iJ_i}\right)^T$ contains the J_i individual error terms for participant i. Generally, it is assumed that random effects and error terms are normally distributed around zero (Laird & Ware, 1982):

$$\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \Sigma) \quad \text{and} \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_\varepsilon^2 I\right), \quad (2)$$

where $\Sigma = \Sigma(\boldsymbol{\theta})$ is the $q \times q$ (parametrized) covariance matrix of the random effects and $\sigma_\varepsilon^2 I$ is the $\left(J_i \times J_i\right)$ (diagonal) covariance matrix of the error term. That is, the model assumes that the random effects $\boldsymbol{b}_i$ and errors $\boldsymbol{\varepsilon}_i$ stem from centered normal distributions and therefore are gathered around zero.

Estimation of LMM

The classical estimation method for the LMM is maximum likelihood based on the normal distribution. Conceptually, maximum likelihood (ML) estimates parameters of a model based on the likelihood of observing the data, assuming the model to be correct in the population. To do so, the (log-)likelihood function, which expresses the likelihood of the data as a function of the model parameters, is maximized over all possible parameter values. In the LMM, the parameters to be estimated are the fixed effects $\boldsymbol{\gamma}$ and the variance components $\left(\sigma_\varepsilon^2, \boldsymbol{\theta}\right)$. When the normality assumption on the variance components is fulfilled, we expect high coverage for CIs associated to ML. However, in the presence of either or both between- and within-participant outliers, we expect ML to produce CIs with reduced accuracy (e.g., Copt & Victoria-Feser, 2006; Richardson & Welsh, 1995; Welsh & Richardson, 1997). ML is either carried out in its full version, or in an alternative version, as restricted ML. In analyses not shown here, we obtained virtually identical results from the two versions, so that we only discuss the former. ML is implemented in the function `lmer` of `R` (package `lme4`, Version [1.1-12]; Bates, Mächler, Bolker, & Walker, 2015).

Robust Estimation

Alternatively, robust methods follow the frequently called *central model* (see e.g., Koller, 2013) and consider that the data generating process is:

$$(1 - \delta)F + \delta H, \tag{3}$$

where F is the central model assumed for the majority of the data (e.g. the normal distribution), H is an arbitrary (noncentral) unknown model, different from F, and $\delta \in [0\ ;\ 1]$ is the amount of contamination, that is, the proportion of data not from F. In this approach, the goal of the analysis is to reach inferential conclusions that are valid for the central model F only and are not influenced by the outliers from H. Commonly, this approach involves estimation weights, which are computed so that they are near 1 for the data from F and smaller (with a lower bound at 0) for the data from H. There are several robust methods for the estimation and inference in the LMM based on the "central model". Here, we consider the three methods compared in Agostinelli and Yohai (2016): the S-estimator (Copt & Victoria-Feser, 2006), the composite τ estimator (cTAU, Agostinelli & Yohai, 2016) and Koller's (2016) DAStau method. These methods presume that the model defined in Equation 1 and Equation 2, which assume normality for $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_i$, is true for a majority of the population (thus is the central model), but does not hold for outliers.

Conceptually, we can consider the robust methods as weighted versions of their classical counterparts, which constitute a special case where all observations' weights are equal to 1. Broadly speaking, the weights (between 0 and 1) produced by the robust methods express how likely an observation is to belong to the bulk of the data (and, conversely, how unlikely it is to be an outlier), and the degree to which it is taken into account in the parameter estimation of the central model. The weights can be attributed at the single observation level only (as for cTAU), at the participant level only (S), or at both levels (DAStau).

Robust estimators may differ with respect to their priorities. Bounded influence estimators, such as DAStau, aim at limiting the bias in their parameter estimates, whereas high-breakdown estimators, such as S and cTAU, strive to tolerate high proportions of outliers. However, this latter class typically produces estimators that are less efficient than those from the former category (Huber, 1964).

Below we provide a brief summary of the robust estimators for the LMM (see also section 'Technicalities of the Estimation Methods' in the Supplementary Materials). For a detailed treatment of robustness, we refer interested readers to Maronna, Martin, and Yohai (2006).

S-Estimator

The S-estimator of Copt and Victoria-Feser (2006) gives a set of weights at the participant level (cf. Equation S12 in the Supplementary Materials). These estimators can be tuned through their tuning constant to achieve high-breakdown. We expect CIs associated to S-estimator to perform particularly well in terms of coverage even in the presence of a large number of outliers. The downweighting method at the participant level, however, could lead to a loss of efficiency leading to lower coverage than CIs associated to the other estimators without contamination. The function `varComprob` (package `robustvarComp`, Version [0.1-2]; Agostinelli & Yohai, 2016) implements the original methods choosing the Rocke robustness function.

Composite τ-Estimator (cTAU)

The estimation procedure of the cTAU of Agostinelli and Yohai (2016) returns as many weights as there are pairs of observations per participant (namely $J_i(J_i-1)/2$), see Equation S17 in the Supplementary Materials. This high-breakdown estimator improves efficiency compared to the S-estimator. We thus expect CIs associated to cTAU to have a better coverage without contamination, but equally or slightly more performing than S under contamination, especially with within-participant outliers. This method is also implemented in `R` within the function `varComprob` and we use the default settings.

DAStau Estimator

The DAStau procedure consists of a chain of estimators, including M-estimator and Design Adaptative Scale (DAS) estimator (Koller, 2013, 2016). The software also proposes a quicker version called "DASvar." However, because this yields approximate results, we did not include it in this study. The procedure returns two sets of weights: one for the participants and one for each observation of each participant. DAStau handles both within- and between-participant outliers. Because the DAStau is a bounded influence, but not a high-breakdown, estimator, it is not designed to handle as much contamination as S and cTAU, as discussed in Agostinelli and Yohai (2016). We expect CIs associated to this estimator to better behave than when associated to S and cTAU without contamination but slightly worse under contamination. The function `rlmer` (package `robustlmm`, Version [2.1-2]; Koller, 2016) implements the DAStau method by default (or with the option `method = "DAStau"`). Based on the recommendations in Koller (2013), we chose the smoothed Huber function for all four sets of estimating equations (with default tuning constant for estimating the fixed effects parameters and for the prediction of the random effect, with `k=2.28` for the estimation of σ_ϵ and `k=5.11`, `s=10` for the estimation of the other variance components parameters).

Percentile CIs Based on Bootstrap

A standard procedure to obtain CIs for fixed effects and variance components for LMM parameters with both classical and robust estimators is the bootstrap (e.g., Koller, 2016; Modugno & Giannerini, 2013). Percentile confidence intervals are obtained by constructing B bootstrap samples, typically of the same size as the original sample, then estimating the parameter of interest on each bootstrap sample, and finally obtaining the lower and upper bounds by taking the $\frac{\alpha}{2} \times 100$ and $\left(1 - \frac{\alpha}{2}\right) \times 100$ empirical percentile of the bootstrap distribution of each parameter (which may thus produce asymmetrical CIs). If the bootstrap distribution is skewed and/or biased, Tibshirani and Efron (1993) proposed a refined version, called bias-corrected accelerated (BCa). This CI method performs well in the LMM context (Deen & de Rooij, 2019), but, to the best of our knowledge, has never been associated to robust LMM estimators. We tested this method in one of our simulation conditions, and found that, whereas in classical LMM estimation BCa was slightly superior to the percentile method, with robust LMM estimation results were highly inconsistent, and often worse than for other CI methods. In the end, because of its unpromising results with robust LMM estimation, we did not consider it for our simulation study.

There are several methods to generate bootstrap samples in the LMM setting. Modugno and Giannerini (2013) compared the classical *parametric* bootstrap to their proposed semi-parametric *wild* bootstrap in the LMM estimated by ML only. The former method has stronger assumptions than the latter (i.e., fixed covariates, correct specification of both fixed effects and variance components, and homoscedasticity and normality for random effects). The authors found that, although generally the two methods performed very similarly, the wild method was superior in large samples and with heteroscedastic random effects. We thus consider these two CI methods, but extend the study by Modugno and Giannerini (2013) by applying them to robust LMM estimators.

Parametric Bootstrap

First, the procedure estimates fixed effects ($\boldsymbol{\gamma}$) and variance components $\left(\sigma_\varepsilon^2, \boldsymbol{\theta}\right)$ parameters from model (1) on the original data $\left(\boldsymbol{y}_i, X_i, Z_i; i = 1, \ldots, n\right)$ to produce $\widehat{\boldsymbol{\gamma}}$, $\widehat{\sigma}_\varepsilon^2$, and $\Sigma\left(\widehat{\boldsymbol{\theta}}\right)$. Then, for each of the B bootstrap samples:

1. For $i = 1, \ldots, n$, generate $\boldsymbol{\varepsilon}_i^* \sim \mathcal{N}\left(\mathbf{0}, \widehat{\sigma}_\varepsilon^2\right)$ and $\boldsymbol{b}_i^* \sim \mathcal{N}\left(\mathbf{0}, \Sigma\left(\widehat{\boldsymbol{\theta}}\right)\right)$ to build
$$\boldsymbol{y}_i^* = X_i\widehat{\boldsymbol{\gamma}} + Z_i\boldsymbol{b}_i^* + \boldsymbol{\varepsilon}_i^*$$
2. Estimate $\boldsymbol{\gamma}$, σ_ε^2, and $\boldsymbol{\theta}$ from model (1) on the bootstrap sample $\left(\boldsymbol{y}_i^*, X_i, Z_i; i = 1, \ldots, n\right)$ to produce $\widehat{\boldsymbol{\gamma}}^*$, $\left(\widehat{\sigma}_\varepsilon^2\right)^*$, and $\Sigma\left(\widehat{\boldsymbol{\theta}}^*\right)$.

The estimator used at step 2 can either be the same as, or differ from, that used on the original sample. Because we generated $B = 5000$, and robust LMM estimation can

be computationally extremely intensive[1], we choose to apply ML at step 2 also when a robust estimator was used at step 1. We believe this procedure legitimate because the bootstrap samples are unlikely to contain outliers.

Wild Bootstrap

The wild bootstrap method replaces step 1 of the parametric bootstrap computing "residual" $\tilde{\boldsymbol{v}}_i$ as follows (Modugno & Giannerini, 2013):

$$\tilde{\boldsymbol{v}}_i = \operatorname{diag}\left(I - H_i\right)^{-1/2} \circ \left(\boldsymbol{y}_i - X_i\widehat{\boldsymbol{\gamma}}\right), \tag{4}$$

with $H_i = X_i\left(X^TX\right)^{-1}X_i^T$, where the operator "$\circ$" denotes the (element-wise) Hadamard product.

For $i = 1, \ldots, n$, sample independently w_i^* from the following standardized random variable (Mammen, 1993):

$$w_i^* = \begin{cases} -\dfrac{\sqrt{5}-1}{2} & \text{with probability } p = \left(\sqrt{5}+1\right)/\left(2\sqrt{5}\right) \\ \dfrac{\sqrt{5}+1}{2} & \text{with probability } q = 1 - p \end{cases} \tag{5}$$

According to Modugno and Giannerini (2013), these weights produce better results than the simple weighting scheme usually used with classical linear models, consisting in sampling either 1 or -1 with equal probability of .50. Then construct individual responses

$$\boldsymbol{y}_i^* = X_i\widehat{\boldsymbol{\gamma}} + \tilde{\boldsymbol{v}}_i w_i^* . \tag{6}$$

Monte Carlo Simulation

Inspired by the dataset `sleepstudy` used very often for didactic purposes in LMM packages, such as in Bates et al. (2015) and Koller (2016), we conducted a simulation experiment, where we studied the consequences of various types of data contamination (C) due to outliers on CIs properties. We manipulated (a) the proportion of outliers ($\delta = 0.05$ vs. 0.1), expecting decreased CI coverage with higher δ; and (b) the number of waves of assessment ($J = 4, 8$; cf. Ghisletta et al., 2020), expecting an impact mainly on change-related parameters, that is, γ_1 and σ_1^2 estimates, with higher coverage when $J = 8$.

1) Over 500 large simulated samples (40 participants measured 80 times following the design described in Modugno & Giannerini, 2013) the mean computing time for a CI is 3.034 days with the first method and 0.9 days with the second method on the most powerful computing facility at the University of Geneva.

We initially intended to manipulate also sample size, by contrasting $n = 200$ to $n = 500$. Due to the excessively heavy computational time involved, we obtained results in the uncontaminated and within-participant outlier conditions only, but not in the remaining between-participant contamination conditions. Nevertheless, with greater sample size we observed very similar amounts of parameter bias, but lower CI length and coverage rates (cf., Figures S1-S4 in the Supplementary Materials). We therefore found limited interest in pursuing in this direction.

We based the simulation population values and magnitude of outliers on the rounded ML estimates from the `sleepstudy` data.

Uncontaminated Data (C_0)

We first generated data without contamination from the following model:

$$\boldsymbol{y}_i = 250\ \mathbf{1} + 10\ \boldsymbol{x}_i + b_{0i}\mathbf{1} + b_{1i}\boldsymbol{x}_i + \boldsymbol{\varepsilon}_i = X_i\boldsymbol{\gamma} + Z_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \tag{7}$$

with $\mathbf{1}$ a vector of 1's of dimension J_i, $\boldsymbol{\gamma} = (250, 10)^T$ and with the elements of $\boldsymbol{x}_i$ representing the day of measurement (hence, from 0 to J-1, with $J_i = J$). We additionally assumed

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 790 & -8.5 \\ -8.5 & 40 \end{pmatrix}\right), \tag{8}$$

and

$$\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, 400\ I)\,. \tag{9}$$

We then generated contaminated data under two different conditions: with outlying errors $\boldsymbol{\varepsilon}_i$ ($C_{\boldsymbol{\varepsilon}_i}$) and with outlying random slope effects b_{1i} ($C_{b_{1i}}$), as described below. In the Supplementary Materials (see Figures S7 and S10), we also provide results with outlying random intercept effects b_{0i} ($C_{b_{0i}}$), but do not discuss them here, because they were analogous to those for $C_{b_{1i}}$.

Contamination of the Errors ($C_{\boldsymbol{\varepsilon}_i}$)

This condition simulated within-participant outliers, with outlying single observations, randomly distributed across participants. While we assumed that a proportion (1-δ) of the generated data points follow Equation 9, we also assumed that the remaining proportion δ had outlying individual error terms, so that the full data generating mechanism was defined by Equation 7 and Equation 8 with $\boldsymbol{\varepsilon}_i$, such that

$$\varepsilon_{ij} \sim (1-\delta)\mathcal{N}(0, 400) + \delta\mathcal{N}(-80, 0.25)\,. \tag{10}$$

We centered the contaminated distribution at −80, thus four *SD*s (= $4\sqrt{400}$) below zero, because in the `sleepstudy` application the ML estimate of σ_ε was 25.59 and the smallest residual was −101.18 (hence roughly 4 *SD*s below 0). We arbitrarily chose the variance of the contaminated errors at 0.25 to limit their variability. We expect degraded performance of CIs associated with ML for γ_0 and σ^2_ε. For all methods, effects of within-participant outliers should be more deleterious when $J = 4$ than $J = 8$. This should especially be the case for σ^2_1, because of the greater influence that a single point out of four, compared to eight, can exert on the estimation of this parameter. In general, we expect these effects to be greater in classical than in robust estimators. Finally, because S downweights data only at the participant-level, we expect CIs to have lower coverage than when associated to cTAU, which is also a high-breakdown estimator but weights data at the observation level and thus should be more efficient.

Contamination of Random Slope Effects ($C_{b_{1i}}$)

This condition implemented between-participant outliers, where all values of a participant were particularly deviant. To produce a δ-proportion of outliers at the random slope level, we modified the mean of their distribution by shifting it four *SD*s to the left, generating data from Equation 7 and Equation 9, using the following distributional assumptions for the random effects:

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim (1-\delta)\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 790 & -8.5 \\ -8.5 & 40 \end{pmatrix}\right) + \delta\mathcal{N}\left(\begin{pmatrix} 0 \\ -25 \end{pmatrix}, \begin{pmatrix} 7.9 & -0.085 \\ -0.085 & .4 \end{pmatrix}\right) \tag{11}$$

We expect contamination effects only on the slope parameters. In particular, γ_1 should be underestimated, because of the negative contamination, whereas σ^2_1 should be overestimated, because of an overestimation of the between-individual variability on the slope reducing the coverage rates. Again, across all methods, the estimation of the individual slopes should improve with $J = 8$ compared to $J = 4$. Indeed, with only four individual observations there is less signal to determine the atypical nature of a trajectory.

Study Design

Our contamination method is inspired by Richardson and Welsh (1995), and reproduces the generating process of the central model: most of the data follow a normal distribution, whereas outliers are generated from an alternative normal distribution. Ultimately, our design contained 2 (δ = 0.05, 0.1) × 2 (J = 4, 8) × 3 (C_0, C_{ε_i}, $C_{b_{1i}}$) conditions, for a total of 12 simulation cell conditions.

For each condition we generated 250 data sets (which, according to Equation 1 of Morris, White, and Crowther (2019), is generally enough to obtain a sufficient degree of precision), for a total of 3000 data sets, each to which we applied the possible combinations of the four estimators and the 2 CI estimation methods, producing 8 types of CIs.

For the percentile CIs based on bootstrap, we generated $B = 5000$ bootstrap replications (cf. Deen & de Rooij, 2019) and all CIs were calculated for a .95 confidence level.

Simulation Analysis

For each of the six parameters (γ_0, γ_1, σ^2_ε, σ^2_0, σ^2_1, σ_{10}), we comparatively assessed the coverage of each type of CI. For each parameter, in each cell condition, we defined the actual CI *coverage* as the proportion of the 250 CIs including the true population value. It is important that the observed coverage approaches the nominal coverage (e.g., .95) to allow for sound inferential conclusions.

Results

All barplots figures (see Figure 1, Figure 3, and Figure 5) present coverage from 0 to 1. The horizontal dashed line represents the nominal (.95) coverage level. The barplots are color-coded according to the type of CI method (parametric in black and wild in grey) and each estimation method is indicated on the abscissa. Parameters are indicated on the left top of each panel. Each major panel is split vertically in two subpanels, with $J = 4$ on the left side and $J = 8$ on the right. For the contaminated scenarios (C_{ε_i} and $C_{b_{1i}}$), the two subpanels are further split in two horizontally, with $\delta = 0.05$ on the top and $\delta = 0.1$ on the bottom row (cf. Figure 3 and Figure 5).

For the 8 types of CIs, we also produced line range plots (see Figure 2, Figure 4, and Figure 6) to visualize the 250 CIs produced for parameters of chief interest in each setting to better understand the CIs coverage rates. Parameter values are on the abscissa, and their expected values are represented by the vertical black line. If the expected values is included in the CI, the corresponding line range is dark, else it is light. Each major panels is split vertically in 16 individual panels, first by J ($J = 4$, the first 8 columns and, $J = 8$ the last 8 ones), then by estimators and bootstrap methods. For the contaminated scenarios (C_{ε_i}, see Figure 4 and $C_{b_{1i}}$, see Figure 6), each panel is further split in two horizontally, with $\delta = 0.05$, on the first row and, $\delta = 0.1$ on the second. Complete line range plots for all parameters in all contamination settings are available in the Supplementary Materials Figures S8, S9 and S10.

Uncontaminated Data (C_0)

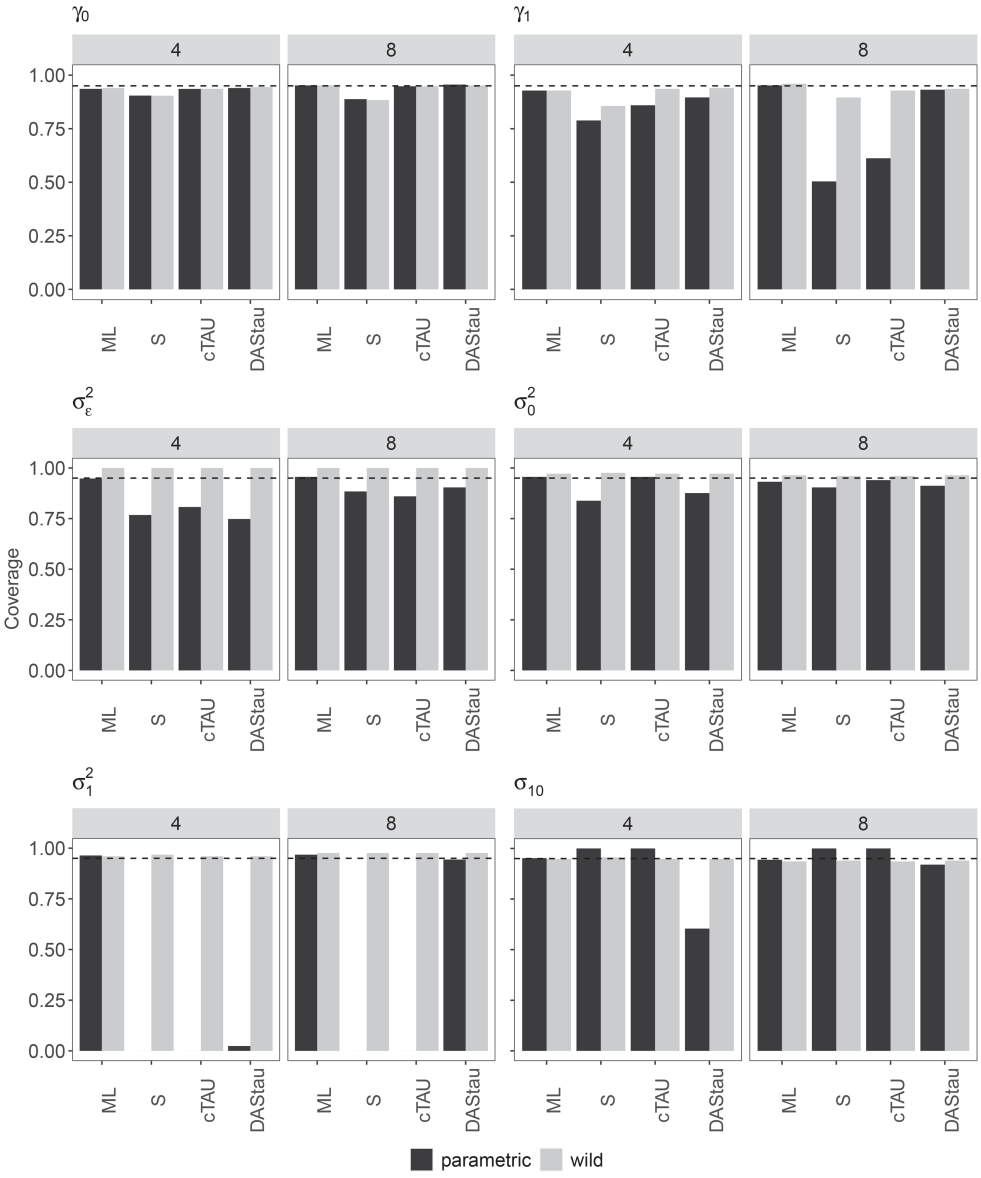
For the fixed effects, when associated to ML, no major differences in coverage emerged between the parametric and the wild bootstrap as we can see on Figure 1. However, when associated to robust estimators (mostly S and cTAU), the wild bootstrap obtained better coverage than the parametric bootstrap for γ_1, whereas coverage was similar for γ_0. Indeed, as we can see on the upper major panel of Figure 2, for γ_1, CIs were much larger with the wild than with the parametric bootstrap, resulting in a lower coverage for

the latter when associated to S and cTAU. Also, ML (with parametric and wild bootstrap), cTAU and DAStau (both with wild bootstrap only) had coverage close to the nominal threshold for both fixed-effect parameters in the uncontaminated condition.

For the variance components (see Figure 1), again, the two types of bootstrap when associated to ML obtained coverage close to the nominal threshold, with a slight difference for σ^2_ε, where the parametric bootstrap was closest to .95. The wild bootstrap obtained better coverage when associated to the robust estimators, particularly for σ^2_1. The lower panel of Figure 2 reveals that the bias estimation of σ^2_1 with S, cTAU, and also, when $J = 4$, with DAStau, affected the parametric bootstrap method and explained the coverage differences noted above.

Figure 1

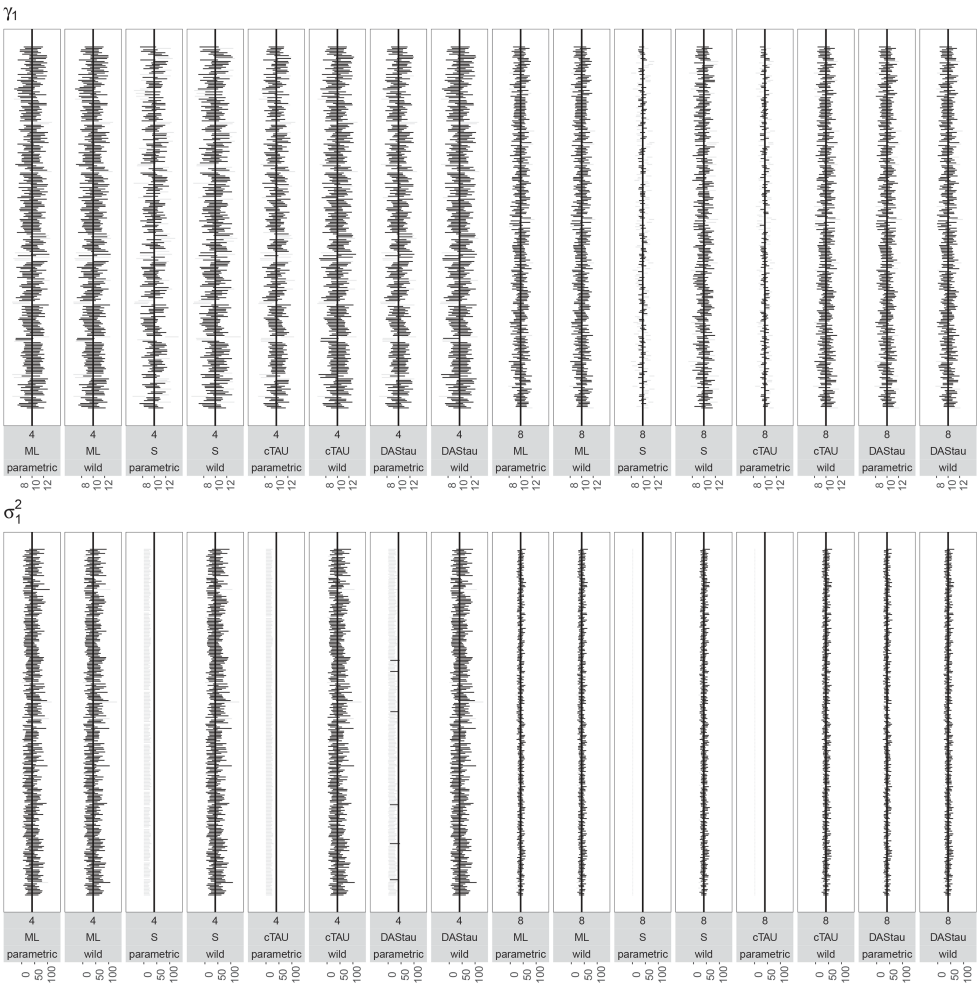
Coverage of CI for the Six Parameters as a Function of J, and the Estimators in the Uncontaminated Condition (C_0)



Note. ML = maximum likelihood; S = Copt and Victoria-Feser's S-Estimator; cTAU = Agostinelli and Yohai's composite τ; DAStau = Koller's DAStau. Black bars indicate parametric bootstrap CIs and grey bars indicate wild bootstrap CIs. Each of the major panels is split in two: $J = 4$ on the left and $J = 8$ on the right. The horizontal dashed lines indicate the .95 nominal confidence level.

Figure 2

CIs for γ_1 and σ_1^2 (Respectively on the Upper and Lower Major Panel) for $J = 4$ and $J = 8$, for Each Estimation and Each Bootstrap Method, in the Uncontaminated Condition (C_0)



Note. ML = maximum likelihood; S = Copt and Victoria-Feser's S-Estimator; cTAU = Agostinelli and Yohai's composite τ; DAStau = Koller's DAStau. CIs for samples 1 to 250 are vertically displayed. The bootstrap type, the estimators and J are indicated in the lower grey boxes. Each of the major panels is split in 16: $J = 4$ on the first eight facets (one by type of CIs) and $J = 8$ on the last eight ones. The black lines at 10 and 40 represent the true underlying value of the parameters.

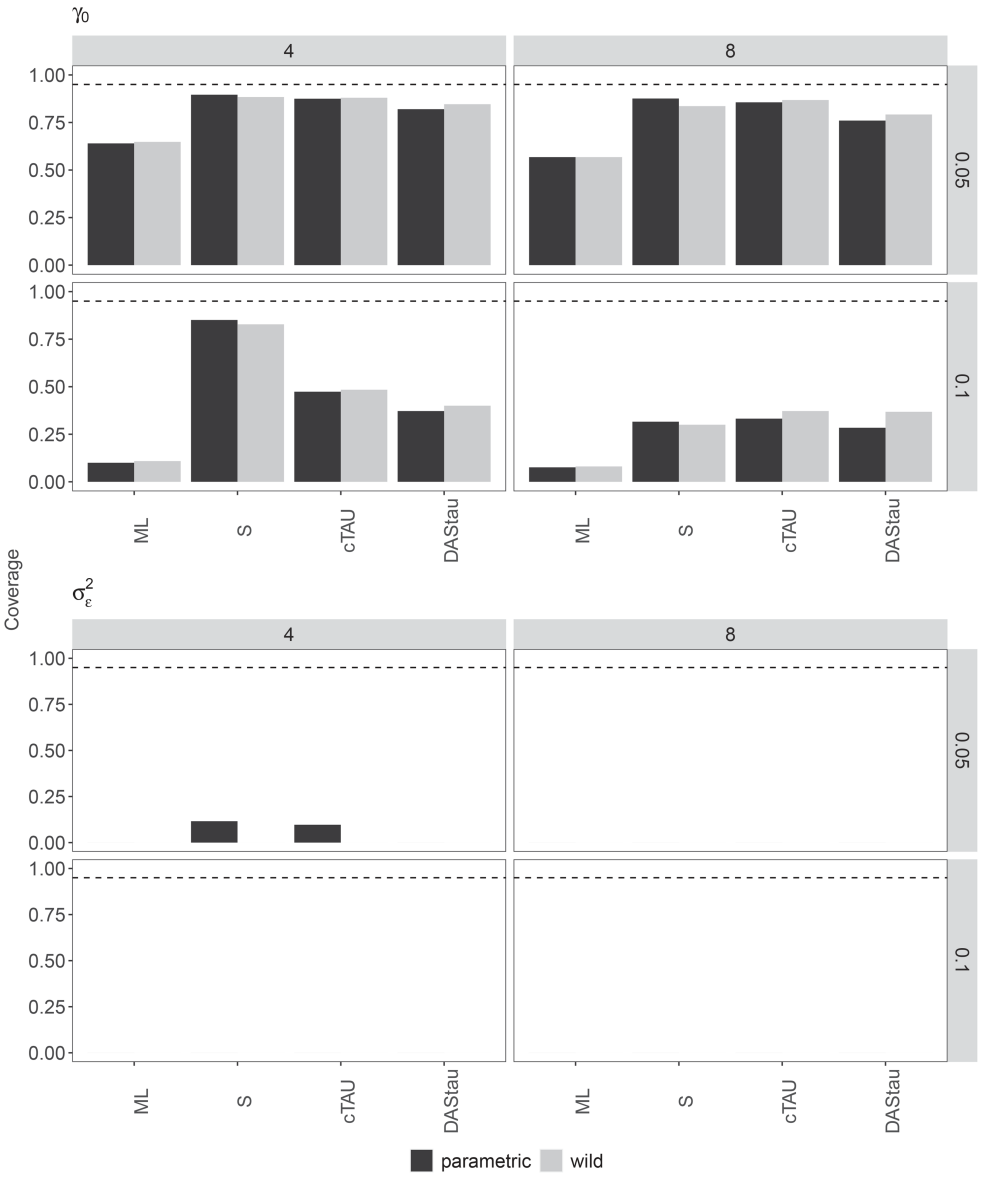
Contamination of ε_i (C_{ε_i})

This contamination condition affected mainly γ_0 and σ^2_ε, whereas results for the remaining four parameters were similar to those observed in C_0 (mainly for γ_1, ML, cTAU and DAStau obtained closer to nominal coverage than S, especially when associated to the wild bootstrap; for the remaining variance components, all estimators performed well when associated to the wild bootstrap; cf. Figures S5 and S8). Figure 3 thus displays coverage CIs for γ_0 and σ^2_ε only. ML obtained the worst coverage for γ_0 in all conditions (cf. upper panel of Figure 3). When $\delta = 0.05$, the two bootstrap methods obtained the highest coverage for γ_0 with S and cTAU, with slightly better coverage for S with the parametric bootstrap, followed by cTAU with the wild bootstrap. When $\delta = 0.1$, all coverages for γ_0 fell drastically, except for S when $J = 4$. The upper panel of Figure 4 reveals that differences in coverage for γ_0 are principally explained by differences in estimation, with more bias for ML than for other estimators, and when $\delta = 0.1$ compared to $\delta = 0.05$. As in C_0, the two bootstrap methods behaved similarly for γ_0.

Coverage rates for σ^2_ε were always exactly null except for S and cTAU with parametric bootstrap when $\delta = 0.05$ and $J = 4$ (but rates were close to .10 only; cf. lower panel of Figure 3). We can see on the lower panel of Figure 4 that all estimates were biased. Overall, bias was larger with $\delta = 0.1$ than with $\delta = 0.05$, parametric bootstrap CIs were shorter than with wild bootstrap, and for all estimation methods, variability in estimation was higher when $J = 4$ than when $J = 8$.

Figure 3

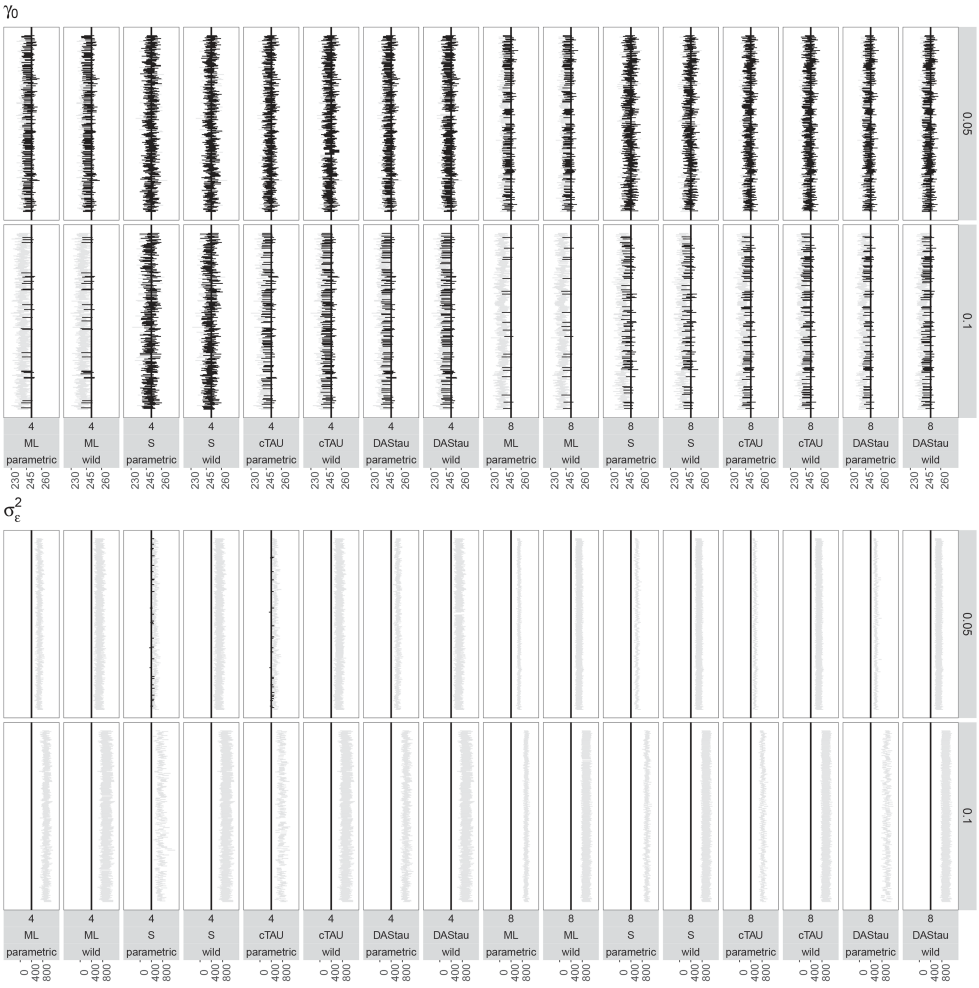
Coverage of CIs for γ_0 and σ^2_ε (Respectively on the Upper and Lower Major Panel) as a Function of J, and the Estimators in the ε_i Contaminated Condition (C_{ε_i})



Note. ML = maximum likelihood; S = Copt and Victoria-Feser's S-Estimator; cTAU = Agostinelli and Yohai's composite τ; DAStau = Koller's DAStau. Black bars indicate parametric bootstrap CIs and grey bars indicate wild bootstrap CIs. Each of the major panels is split in four: $J = 4$ on the left and $J = 8$ on the right, $\delta = 0.05$ on the top and $\delta = 0.1$ on the bottom. The horizontal dashed lines indicate the .95 nominal confidence level.

Figure 4

CIs for γ_0 and σ^2_ε (Respectively on the Upper and Lower Major Panel) for $J = 4$ and $J = 8$, for Each Estimation and Each Bootstrap Method, in the ε_i Contaminated Condition (C_{ε_i})



Note. ML = maximum likelihood; S = Copt and Victoria-Feser's S-Estimator; cTAU = Agostinelli and Yohai's composite τ; DAStau = Koller's DAStau. CIs for samples 1 to 250 are vertically displayed. The bootstrap, the estimators and J are indicated in the lower grey boxes. Each of the major panels is split in 32: $J = 4$ on the first eight facets (one by type of CIs) and $J = 8$ on the last eight ones, $\delta = 0.05$ on the top and $\delta = 0.1$ on the bottom. The black lines at 250 and 400 represent the true underlying value of the parameters.

Contamination of b_{1i} ($C_{b_{1i}}$)

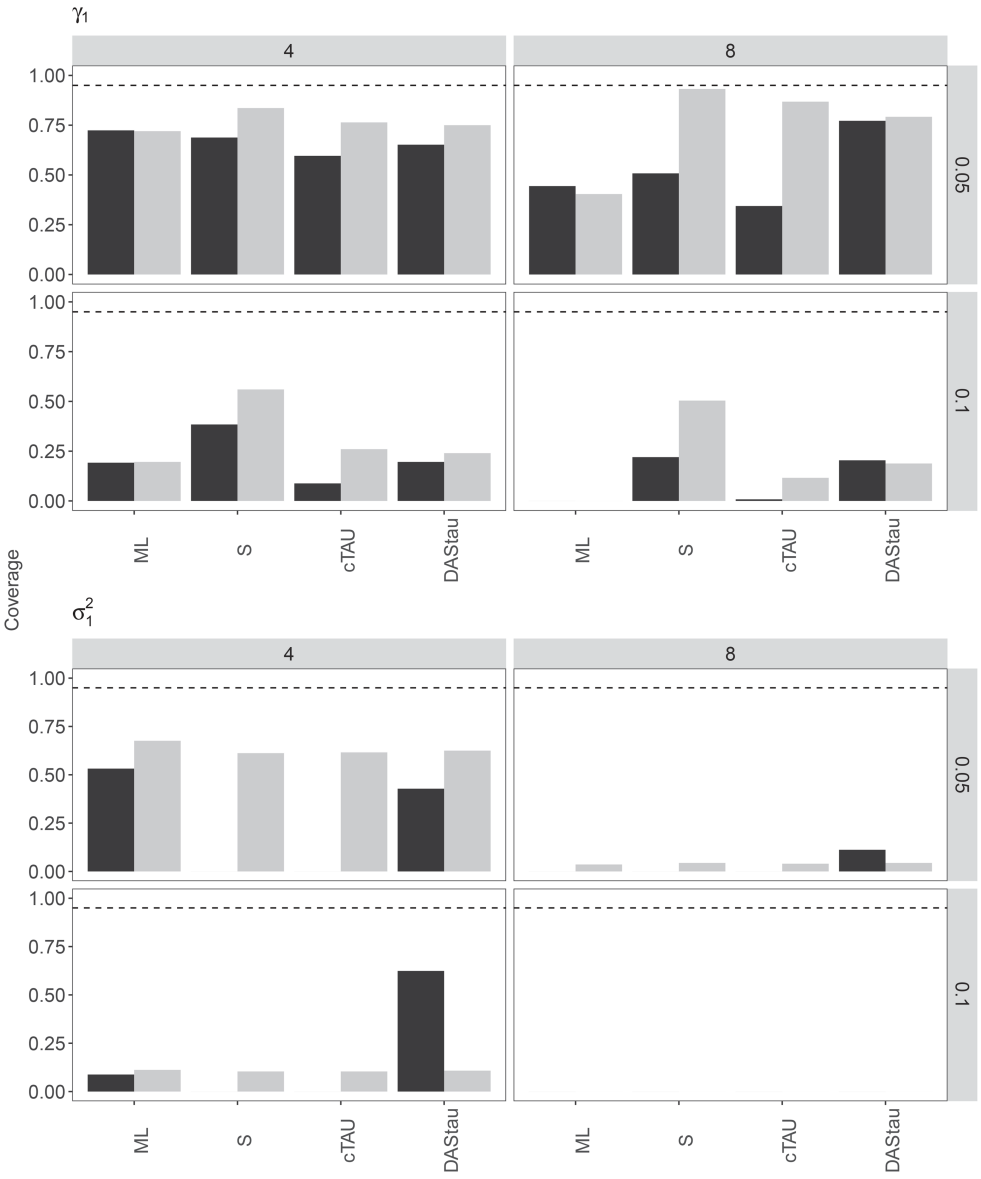
This contamination resulted mainly in effects on CIs for γ_1 and σ_1^2 (see Figure 5 and Figure 6). Again, for the other parameters, results were similar to those obtained in the other simulation conditions and are available in the Supplementary Materials (see Figures S6 and S9), which also contains results when contamination occurred on b_{0i}, affecting coverage rates on γ_0 and σ_0^2 (see Figures S7 and S10 for $C_{b_{0i}}$).

Generally, the wild bootstrap outperformed the parametric bootstrap with robust estimators and, to a lesser extent, also with ML. Again, ML obtained the lowest coverage for γ_1, especially when $J = 8$. When $\delta = 0.05$, S obtained coverage close to the nominal threshold, but only when $J = 8$. cTAU and then DAStau followed next in coverage. Again, coverage dropped drastically when $\delta = 0.1$, but S retained the highest rates. Interestingly, only robust estimators' CIs had better coverage when $J = 8$ than $J = 4$. The upper panel of Figure 6 shows that all CIs were shorter when $J = 8$ than $J = 4$, especially with the parametric bootstrap, which may explain the reduced coverage for ML when $J = 8$. On the contrary, with robust estimators associated to the wild bootstrap, especially when $\delta = 0.05$, CIs appear slightly shifted to the right, towards the true value, thereby resulting in higher coverage.

For σ_1^2, all coverage rates were extremely low (see lower panel of Figure 5). The lower panel of Figure 6 shows that parametric CIs with S and cTAU were very close to zero and extremely small, resulting in nearly null coverage. Again, CIs were larger when $J = 4$ than when $J = 8$ and closer to the expected value when $\delta = 0.05$ than $\delta = 0.1$, explaining the best coverage rate with the wild bootstrap when $\delta = 0.05$ and $J = 4$.

Figure 5

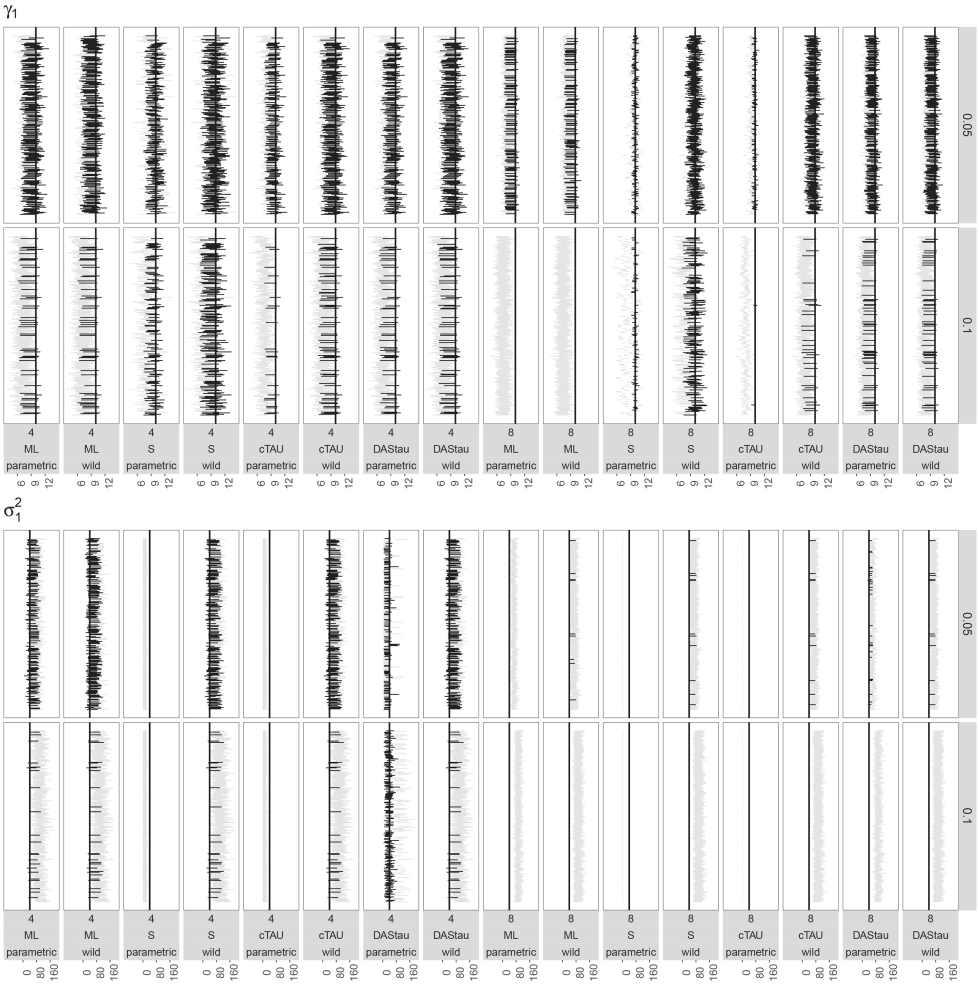
Coverage of CIs for γ_1 and σ_1^2 (Respectively on the Upper and Lower Major Panel) as a Function of J, and the Estimators in the b_{1i} Contaminated Condition ($C_{b_{1i}}$)



Note. ML = maximum likelihood; S = Copt and Victoria-Feser's S-Estimator; cTAU = Agostinelli and Yohai's composite τ; DAStau = Koller's DAStau. Black bars indicate parametric bootstrap CIs and grey bars indicate wild bootstrap CIs. Each of the major panels is split in four: $J = 4$ on the left and $J = 8$ on the right, $\delta = 0.05$ on the top and $\delta = 0.1$ on the bottom. The horizontal dashed lines indicate the .95 nominal confidence level.

Figure 6

CIs for γ_1 and σ_1^2 (Respectively on the Upper and Lower Major Panel) for $J = 4$ and $J = 8$, for Each Estimation and Each Bootstrap Method, in the b_{1i} Contaminated Condition ($C_{b_{1i}}$)



Note. ML = maximum likelihood; S = Copt and Victoria-Feser's S-Estimator; cTAU = Agostinelli and Yohai's composite τ; DAStau = Koller's DAStau. CIs for samples 1 to 250 are vertically displayed. The bootstrap, the estimators and J are indicated in the lower grey boxes. Each of the major panels is split in 32: $J = 4$ on the first eight facets (one by type of CIs) and $J = 8$ on the last eight ones, $\delta = 0.05$ on the top and $\delta = 0.1$ on the bottom. The black lines at 10 and 40 represent the true underlying value of the parameters.

Discussion

In this work we compared the parametric and wild bootstrap methods to compute CIs for both fixed effects and variance components of LMM parameters estimated with one classical and three robust estimators. We compared the eight resulting CI types in terms of coverage under three different conditions: no contamination (C_0), within-participant outliers (C_{ε_i}, based on time-varying errors), and between-participant outliers ($C_{b_{1i}}$, based on time-invariant slopes). We also manipulated number of repeated measurements (J = 4 or 8) and proportion of contaminated data (δ = 0.05 or 0.1). In the end, we extend the work of Agostinelli and Yohai (2016) and of Modugno and Giannerini (2013), by combining the estimation methods studied in the former article to the CI methods investigated in the latter article, thereby providing the thoroughest discussion of CIs from robust estimation in LMM.

Major Statistical Considerations

First, we found that the contamination effects were different across within- and between-participant outliers, but in both cases CI coverage was mainly reduced for specific estimates. For the former, CIs of γ_0 and σ^2_ε were strongly affected, whereas for the latter, CIs of γ_1 and σ^2_1 were impacted under $C_{b_{1i}}$ (and analogously, for γ_0 and σ^2_0 under $C_{b_{0i}}$). But across all contamination conditions, coverage of CIs of the remaining parameters were similar to those obtained without outliers. And, unsurprisingly, the more the outliers, the stronger the effects of the contamination on CI coverage rates.

Second, with ML estimation and uncontaminated data, both parametric and wild CIs obtained excellent coverage across all parameters. Moreover, wild CIs were larger for variance components, especially for σ^2_ε, with rates close to 1, as in Modugno and Giannerini (2013). On the contrary, with robust estimators, the wild bootstrap nearly always outperformed the parametric method in terms of coverage, particularly for γ_1 and the variance components.

Third, without contamination, CI coverage rates with cTAU and DAStau were similar to those of ML for fixed effects and variance components parameters. With S, coverage was lower for the fixed effects parameters, but similar for the variance components. Thus, with uncontaminated data, ML with both bootstrap methods, and cTAU and DAStau with the wild bootstrap produced the best CIs.

Fourth, our results confirm that in the presence of relatively many (i.e., 5% or 10%) outliers, CIs for ML estimates can obtain very low coverage rates. As expected, with 5% of outlying data, robust estimators obtain higher CI coverage than ML. Consistently with Agostinelli and Yohai (2016), who observed that DAStau fixed effect estimates were less accurate than those of S and cTAU under within- and between-cluster outliers, we found that the wild bootstrap obtained slightly lower coverage with DAStau than with S and cTAU for the parameters affected by the contamination condition. With 10% of outliers

in the data, only S maintained a high coverage for γ_0 under C_{ε_i} and $C_{b_{0i}}$ when $J = 4$. But under $C_{b_{1i}}$, the coverage for γ_1 dropped across all estimators, even if it still obtained the best performance. For the variance components affected by contamination, coverage were very low for all methods, independently of the proportion of outliers.

In conclusion, in the realm of our simulation, when there are no or a few outliers, the wild bootstrap associated to the cTAU estimator provides excellent coverage generally for all parameters, except for the variance component parameters particularly affected by the contamination condition. However, when the proportion of outliers becomes larger, CIs obtained with the S estimator and the wild bootstrap method appear superior.

Limitations, Extensions, and Further Applications

Motivated by the empirical `sleepstudy` example that is often used as a didactic example for `R` (Bates et al., 2015; Koller, 2016), we based the population values of our simulation on LMM results from this data set. A characteristic of the empirical LMM results was the nearly null intercept-slope covariance estimate ($\sigma_{10} = -8.5$, implying $r_{10} = -.05$). Of the results we obtained, we are most wary of those about this parameter. As usual, the extension of the conclusions beyond the specific simulation conditions has to be made cautiously.

Also, our example enjoyed balanced data, in that each participant was assessed at the same occasions and had complete data. Two of the robust estimation methods we examined cannot be applied to unbalanced data (namely S and cTAU), so that analysts with such data cannot apply all robust estimators to the LMM. Extending S and cTAU to handle unbalanced data would greatly increase their applicability. Nevertheless, based on the current results, it appears that the use of DAStau with the wild bootstrap when analyzing unbalanced data that possibly contain outliers might be useful. Indeed, this method obtains CIs for most parameters with larger coverage than those based on classical estimators, with and without outliers.

The example did not include explanatory covariates, other than time. Typically, researchers are interested not only in testing for sample heterogeneity in intercept and slope values, but also in trying to explain such interindividual differences by means of person-specific covariates. For instance, in LMM applications to experimental data a main covariate is group membership, which distinguishes participants in the treatment from those in the control group. It seems plausible that the presence of outliers in the response variable may differ between the groups, to the point of possibly clouding the treatment effect. Knowledge about robust estimation of the LMM in such situations should further increase the attractiveness of applying the LMM to experimental data.

Conclusions

Nowadays, several methods for the robust estimation of the LMM are available. In this work, which we believe to be the most comprehensive on CIs with robust estimation in the LMM, we have shown that cTAU with the wild bootstrap can produce sound inferential conclusions, both with and without data contamination.

We are not claiming that all repeated-measure (or, more generally, nested or crossed) data sets necessarily contain outliers that bias results when classical estimation methods are used. Nevertheless, we hope to have raised awareness about the possible detrimental effects of outliers in the context of the LMM and to have provided useful suggestions about alternative estimation methods for such data. We provide `R` code (see Supplementary Materials) to implement these methods, accompanied by detailed comments, so that most users familiar with the `R` language and environment ought to be able to use our code to estimate their LMMs when suspicion arises about the presence of outliers. Our wish is that research in this domain continues to provide tools aimed at reducing the potentially detrimental effects of outliers, and that researchers may find the applications of such tools meaningful to their work.

Funding: The authors have no funding to report.

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have declared that no competing interests exist.

Supplementary Materials

For this article the following Supplementary Materials are available (for access see Index of Supplementary Materials below):
Via the PsychArchives repository:

- Technicalities of estimation methods: Equations S1-S23.
- Additional simulation results: Figures S1-S10.

Via the GitHub repository:

- Code.

Index of Supplementary Materials

Mason, F., Cantoni, E., & Ghisletta, P. (2021). *Supplementary materials to: Parametric and semi-parametric bootstrap-based confidence intervals for robust linear mixed models* [Equations and Figures]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.5302

Mason, F., Cantoni, E., & Ghisletta, P. (2021). *Supplementary materials to: Parametric and semi-parametric bootstrap-based confidence intervals for robust linear mixed models* [Code]. https://github.com/masonFG/CIrobustLMM

References

Agostinelli, C., & Yohai, V. J. (2016). Composite robust estimators for linear mixed models. *Journal of the American Statistical Association, 111*(516), 1764-1774. https://doi.org/10.1080/01621459.2015.1115358

Andersen, R. (2008). *Quantitative applications in the social sciences: Modern methods for robust regression.* Sage Publications. https://doi.org/10.4135/9781412985109

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

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

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7-29. https://doi.org/10.1177/0956797613504966

Deen, M., & de Rooij, M. (2019). ClusterBootstrap: An R package for the analysis of hierarchical data using generalized linear models with the cluster bootstrap. *Behavior Research Methods, 52*(2), 527-590.

Efron, B. (1979). Bootstrap methods: Another look at the jacknife. *The Annals of Statistics, 7*(1), 1-26. https://doi.org/10.1214/aos/1176344552

Efron, B., & Hastie, T. (2016). *Computer age statistical inference: Algorithms, evidence, and data science.* Cambridge University Press.

Ghisletta, P., Mason, F., Oertzen, T. V., Hertzog, C., Nilsson, L.-G., & Lindenberger, U. (2020). On the use of growth models to study normal cognitive aging. *International Journal of Behavioral Development: Methods and Measures, 44*(1), 88-96. https://doi.org/10.1177/0165025419851576

Goldstein, H. (2011). *Multilevel statistical models.* John Wiley and Sons.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology, 31*(4), 337-350. https://doi.org/10.1007/s10654-016-0149-3

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association, 69*, 383-393.

Henderson, C. R., Kempthorne, O., Searle, S. R., & Von Krosigk, C. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics, 15*(2), 192-218. https://doi.org/10.2307/2527669

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics, 35*, 73-101.

Huber, P. J., & Ronchetti, E. M. (2009). Robust statistics 2009. John Wiley & Sons. https://doi.org/10.1002/9780470434697

Koller, M. (2013). Robust estimation of linear mixed models. Doctoral dissertation Eidgenössische Technische Hochschule Zurich. https://doi.org/10.3929/ethz-a-007632241

Koller, M. (2016). robustlmm: An R package for robust estimation of linear mixed-effects models. *Journal of Statistical Software, 75*(6), 1-24. https://doi.org/10.18637/jss.v075.i06

Koller, M., & Stahel, W. A. (2011). Sharpening wald-type inference in robust regression for small samples. *Computational Statistics & Data Analysis, 55*(8), 2504-2515. https://doi.org/10.1016/j.csda.2011.02.014

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1-26. https://doi.org/10.18637/jss.v082.i13

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*(4), 963-974. https://doi.org/10.2307/2529876

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics, 21*(1), 255-285. https://doi.org/10.1214/aos/1176349025

Maronna, R., Martin, R., & Yohai, V. (2006). *Robust statistics*. Wiley.

Modugno, L., & Giannerini, S. (2013). The wild bootstrap for multilevel models. *Communications in Statistics – Theory and Methods, 44*(22), 4812-4825. https://doi.org/10.1080/03610926.2013.802807

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine, 38*(11), 2074-2102.

Muler, N., & Yohai, V. J. (2002). Robust estimates for ARCH processes. *Journal of Time Series Analysis, 23*(3), 341-375. https://doi.org/10.1111/1467-9892.00268

Richardson, A. M., & Welsh, A. H. (1995). Robust restricted maximum likelihood in mixed linear models. *Biometrics, 51*, 1429-1439. https://doi.org/10.2307/2533273

Rocke, D. M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension. *The Annals of Statistics, 24*, 1327-1345. https://doi.org/10.1214/aos/1032526972

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*(388), 871-880.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage Publications.

Tibshirani, R. J., & Efron, B. (1993). An introduction to the bootstrap. *Monographs on Statistics and Applied Probability, 57*, 1-436.

Verbeke, G. (1997). *Linear mixed models for longitudinal data*. In G. Verbeke & G. Molenberghs (Eds.), *Linear mixed models in practice* (pp. 63-153). Springer. https://doi.org/10.1007/978-1-4612-2294-1_3

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer. https://doi.org/10.1007/b98969

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129-133. https://doi.org/10.1080/00031305.2016.1154108

Welsh, A., & Richardson, A. (1997). 13 Approaches to the robust estimation of mixed models. *Handbook of Statistics, 15*, 343-384.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594-604. https://doi.org/10.1037/0003-066X.54.8.594

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics. 15*, 642-656.