Robust estimation in linear mixed effects models

Valérie Garès[1], Rik Lopuhaä[2] & Anne Ruiz Gazen[3]

[1] *Univ. Rennes INSA, CNRS, IRMAR - UMR 6625, Rennes, France,*
valerie.gares@insa-rennes.fr
[2] *Delft University of Technology*
[3] *Toulouse School of Economics*

Résumé. Les modèles linéaires à effets mixtes sont largement utilisés pour étudier des réponses corrélées notamment pour l'analyse de données longitudinales, de données de croissance ou des mesures répétées. Les estimateurs classiques de ces modèles, tels que les estimateurs du maximum de vraisemblance, sont basés sur des hypothèses de normalité et sont sensibles aux valeurs atypiques. Il est donc important d'explorer des estimateurs robustes dans ce contexte. Nous nous concentrons sur les modèles linéaires à effets mixtes équilibrés et proposons un aperçu des méthodes d'estimation robustes qui ont été étudiées et revisitées ces dernières années, tels que les estimateurs S, MM et l'estimateur tau composites. Lors de la présentation, nous rappellerons brièvement leur définition et leurs propriétés théoriques, et les comparerons via une étude par simulations.

Mots-clés. Estimateurs composites, Estimateurs robustes, Modèles linéaires mixtes, MM estimateurs, S estimateurs, tau estimateurs.

Abstract. Linear mixed effects models are widely used to study correlated responses, particularly for the analysis of longitudinal data, growth data or repeated measures. Conventional estimators of these models, such as maximum likelihood estimators, are based on assumptions of normality and are sensitive to outliers. It is therefore important to explore robust estimators in this context. We focus on balanced linear mixed effects models and provide an overview of robust estimation methods that have been studied and revisited in recent years, such as the S, the MM and the composite tau estimators, and their composite counterparts. We will briefly recall their definition and their theoretical properties, and we will compare them via a simulation study.

Keywords. Composite estimators, Linear mixed effects models, MM-estimators, Robust estimators, S-estimators, tau-estimators.

1 Introduction

Linear models with structured covariance matrices are widely used and provide a versatile approach for analyzing correlated responses, such as longitudinal data, growth data or repeated measurements. In such models, each subject i, $i = 1, \ldots, n$, is observed at k_i occasions, and the vector of responses $\mathbf{y}_i$ is assumed to arise from the model

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

where $\mathbf{X}_i$ is the design matrix for the ith subject and $\mathbf{u}_i$ is a vector whose covariance matrix can be used to model the correlation between the responses. We consider a structured covariance matrix, that is, the matrix $\mathbf{V} = \mathbf{V}(\boldsymbol{\theta})$ is a known function of unknown covariance parameters combined in a vector $\boldsymbol{\theta} \in \mathbb{R}^l$.

The balanced linear mixed effects model is a well-known example of a linear model with a structured covariance matrix. We consider independent observations $(\mathbf{y}_1, \mathbf{X}_1), \ldots, (\mathbf{y}_n, \mathbf{X}_n)$ such that

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \sum_{j=1}^{r} \mathbf{Z}_j\gamma_{ij} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n. \tag{1}$$

For each subject $i = 1, \ldots, n$, $\mathbf{y}_i \in \mathbb{R}^k$ is the response vector and $\mathbf{X}_i \in \mathbb{R}^{k\times q}$ the known design matrix for the fixed effects. The vector $\boldsymbol{\beta} \in \mathbb{R}^q$ is the unknown fixed effects parameter. The $\mathbf{Z}_j$'s are known $k \times g_j$ design matrices for the random effects $\gamma_{ij} \in \mathbb{R}^{g_j}$, which are assumed to be independent mean zero random vectors with covariance matrix $\sigma_j^2\mathbf{I}_{g_j}$, for $j = 1, \ldots, r$. The error terms $\boldsymbol{\epsilon}_i$, $i = 1, \ldots, n$, belong to $\mathbb{R}^k$, are independent mean zero random vectors with covariance matrix $\sigma_0^2\mathbf{I}_k$, and are independent from the γ_{ij}'s. This means that we concentrate on models for which

$$\mathbf{V}(\boldsymbol{\theta}) = \sum_{j=1}^{r} \sigma_j^2\mathbf{Z}_j\mathbf{Z}_j^T + \sigma_0^2\mathbf{I}_k \text{ and } \boldsymbol{\theta} = (\sigma_0^2, \sigma_1^2, \ldots, \sigma_r^2).$$

Common estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are maximum likelihood estimators, derived under Gaussian assumptions for the random effects and for the error terms (see, e.g., Hartley and Rao, 1967; Laird and Ware, 1982). It is well-known that these estimators rely heavily on the Gaussian assumptions and are highly sensitive to outlying observations (see, e.g., Welsh and Richardson, 1997). To overcome this problem, several robust estimation methods have been proposed and studied in the last thirty years (see, e.g., Agostinelli and Yohai, 2016; Mason et al., 2021; Lopuhaä, 2023, for some recent overviews and studies).

Among the existing robust approaches, some robust estimators are aimed at resisting high proportions of outliers in different contamination models, and we will focus on such estimators. In the robust statistical literature, there exist two contamination models: the Classical (Tukey-Huber) Contamination Model (CCM) and the Independent Contamination Model (ICM). While CCM (also called "case-wise" contamination) considers that the contamination is at the level of the subjects or cases, ICM (also called "cell-wise" contamination) considers the possibility to contaminate data sets at the level of the cells. Initially, the robust statistics literature focused on the CCM context and proposed robust estimators that were able to cope with a proportion of outliers close to 50% without breaking down. However, such estimators may not be robust in the ICM context, since even a small fraction of contaminated cells may lead to more than 50% of contaminated cases.

The aim of this presentation is to give a short overview of some highly robust estimators in the CCM and ICM contexts, and compare their behavior on a simulation study. In Section 2, we introduce briefly some robust estimators we want to compare. In Section 3, we present the simulation scenarios we will consider. Results will be detailed and discussed during the presentation.

2 Robust estimators for balanced linear mixed effects models

The first robust estimators for linear mixed effects models were based on weighted versions of the likelihood function (see, e.g., Welsh and Richardson, 1997).

Then, the well-known S-estimators were extended to linear mixed effects models in Copt and Victoria-Feser (2006) (see also Heritier et al., 2009). S-estimators are smooth versions of the minimum volume ellipsoid estimator proposed in Rousseeuw (1985). They are called high-breakdown point estimators because they can resist to a large proportion of case-wise outliers without breaking down. The theoretical properties of S-estimators and in particular their asymptotic distribution, were revisited very recently in Lopuhaä et al. (2023) for general linear models with structured covariance.

S-estimators may also serve as initial estimators for MM estimators of the fixed effects parameter $\boldsymbol{\beta}$ proposed by Copt and Heritier (2007), and revisited recently by Lopuhaä (2023). MM-estimators are expected to be more efficient than S-estimators while maintaining a high breakdown point in the CCM context.

S and MM-estimators are not able to cope with case-wise contamination rate larger than 50% that may arise in the ICM context. To overcome this problem, Agostinelli and Yohai (2016) proposed a composite approach for tau-estimators (Yohai and Zamar, 1988) inspired by Lindsay (1988). The proposed composite tau estimator is highly robust not only under CCM but also under ICM.

Our objective is to compare the different estimators for CCM and ICM through a Monte Carlo study detailed below.

3 Simulation setup

We will propose a Monte Carlo study of the behavior of the above estimators for samples generated using models inspired by the simulation setup of Mason et al. (2021). We consider uncontaminated data and several types of contaminated data (i) in the random effects, (ii) in the measurement error terms, and (iii) in the design matrix of the fixed effects, according to CCM and ICM. We go beyond Mason et al. (2021) by considering not only ICM but also CCM, by taking larger proportions of contamination and by looking at two types of contamination in the design matrix of the fixed effects. Let us detail the generated data and the setup.

Uncontaminated data. Let us consider a linear mixed effects model with $\mathbf{y}_i$ in dimension $k = 4$ such that:

$$\mathbf{y}_i = 250\,\mathbf{1} + 10\,\mathbf{x}_i + \gamma_{0i}\mathbf{1} + \gamma_{1i}\mathbf{x}_i + \epsilon_i = \mathbf{X}\boldsymbol{\beta} + \gamma_i\mathbf{Z} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{2}$$

with $\mathbf{1}$ the vector of ones of dimension 4, $\mathbf{x}_i = (0, 1, 2, 3)^T$, the fixed effects $\boldsymbol{\beta} = (250, 10)^T$,

the random effects

$$\gamma_i = \begin{pmatrix} \gamma_{0i} \\ \gamma_{1i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 790 & -8.5 \\ -8.5 & 40 \end{pmatrix} \right), \quad \text{and the error terms} \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, 400\mathbf{I}).$$

For the contaminated data, we assume that a proportion $(1-\delta)$ of subjects for the CCM, and cells for the ICM, is generated following the model defined in equation (2), while the remaining proportion δ of subjects for the CCM, and cells for the ICM, are outlying.

Contamination of the errors. Let us consider a shift $m_\epsilon > 0$.

- For CCM, $\boldsymbol{\epsilon}_i \sim (1-\delta)\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, 400\,\mathbf{I} \right) + \delta\mathcal{N}\left(\begin{pmatrix} m_\epsilon \\ 0 \\ 0 \\ 0 \end{pmatrix}, 400\,\mathbf{I} \right)$.
- For ICM, $\boldsymbol{\epsilon}_{ij} \sim (1-\delta)\mathcal{N}(0, 400) + \delta\mathcal{N}(m_\epsilon, 0.25)$,

Contamination of random slope effects. Let us consider a shift $m_\gamma > 0$.

$$\gamma_i \sim (1-\delta)\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 790 & -8.5 \\ -8.5 & 40 \end{pmatrix} \right) + \delta\mathcal{N}\left(\begin{pmatrix} 0 \\ m_\gamma \end{pmatrix}, \begin{pmatrix} 7.9 & -0.085 \\ -0.085 & 0.4 \end{pmatrix} \right)$$

Contamination of the design matrix of the fixed effects. We consider two different contamination frameworks. The first one is such that data are generated according to model (2) and, for a given $\alpha > 1$:

- For CCM, a proportion δ of $\mathbf{x}_i$ is replaced by $\alpha\mathbf{x}_i$.
- For ICM, a proportion δ of x_{ij} is replaced by αx_{ij}.

For the second framework, the design matrix of the uncontaminated data is such that the $\mathbf{x}_i$'s are generated independently and follow a standard Gaussian distribution in 4 dimensions.

- For CCM,
$$\boldsymbol{x}_i \sim (1-\delta)\mathcal{N}\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \mathbf{I} \right) + \delta\mathcal{N}\left(\begin{pmatrix} m_x \\ \vdots \\ m_x \end{pmatrix}, \mathbf{I} \right).$$
- For ICM, $x_{ij} \sim (1-\delta)\mathcal{N}(0, 1) + \delta\mathcal{N}(m_x, 1)$.

In order to investigate the impact of the different contamination parameters, we will choose $\delta \in \{0, 5, 10, 20, 30\}$, the shifts $m_\epsilon \in \{-40, -80, -160\}$, $m_\gamma \in \{-40, -80, -160\}$, $m_x \in \{0.5, 1, 5, 10\}$, and the parameter $\alpha \in \{2, 5, 10, 50, 100\}$. For these different parameters, we will draw parallel boxplots for the maximum likelihood estimate and for the robust S, MM and composite tau estimates. We will compare the results and give some recommendations.

References

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

Copt, S. and Heritier, S. (2007). Robust alternatives to the f-test in mixed linear models based on MM-estimates. *Biometrics*, 63(4):1045–1052.

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.

Hartley, H. O. and Rao, J. N. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54(1-2):93–108.

Heritier, S., Cantoni, E.and Copt, S., and Victoria-Feser, M.-P. (2009). *Robust methods in biostatistics*. John Wiley & Sons.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

Lindsay, B. G. (1988). Composite likelihood methods. *Comtemporary Mathematics*, 80(1):221–239.

Lopuhaä, H. P. (2023). Highly efficient estimators with high breakdown point for linear models with structured covariance matrices. *Econometrics and Statistics*.

Lopuhaä, H. P., Gares, V., and Ruiz-Gazen, A. (2023). S-estimation in linear models with structured covariance matrices. *Annals of Statistics*, 51(6):2415–2439.

Mason, F., Cantoni, E., and Ghisletta, P. (2021). Parametric and semi-parametric bootstrap-based confidence intervals for robust linear mixed models. *Methodology*, 17(4):271–295.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8(283-297):37.

Welsh, A. and Richardson, A. (1997). 13 approaches to the robust estimation of mixed models. *Handbook of statistics*, 15:343–384.

Yohai, V. J. and Zamar, R. H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American statistical association*, 83(402):406–413.