S-Estimation in linear models with structured covariance matrices

Lopuhaä, Hendrik Paul; Gares, Valerie; Ruiz-Gazen, Anne

DOI
10.1214/23-AOS2334

Publication date
2023

Document Version
Final published version

Published in
Annals of Statistics

Citation (APA)
Lopuhaä, H. P., Gares, V., & Ruiz-Gazen, A. (2023). S-Estimation in linear models with structured covariance matrices. *Annals of Statistics*, *51*(6), 2415-2439. https://doi.org/10.1214/23-AOS2334

The Annals of Statistics
2023, Vol. 51, No. 6, 2415–2439
https://doi.org/10.1214/23-AOS2334
This research was funded, in whole or in part, by [Agence National de la Recherche, ANR-17-EURE-0010]. A CC BY 4.0 license is applied to this article arising from this submission, in accordance with the grant's open access conditions

S-ESTIMATION IN LINEAR MODELS WITH STRUCTURED COVARIANCE MATRICES

BY HENDRIK PAUL LOPUHAÄ[1,a], VALERIE GARES[2,b] AND ANNE RUIZ-GAZEN[3,c]

[1]*Delft University of Technology,* [a]*h.p.lopuhaa@tudelft.nl*
[2]*Institut National des Sciences Appliqués de Rennes,* [b]*Valerie.Gares@insa-rennes.fr*
[3]*Toulouse School of Economics,* [c]*anne.ruiz-gazen@tse-fr.eu*

We provide a unified approach to S-estimation in balanced linear models with structured covariance matrices. Of main interest are S-estimators for linear mixed effects models, but our approach also includes S-estimators in several other standard multivariate models, such as multiple regression, multivariate regression and multivariate location and scatter. We provide sufficient conditions for the existence of S-functionals and S-estimators, establish asymptotic properties such as consistency and asymptotic normality, and derive their robustness properties in terms of breakdown point and influence function. All the results are obtained for general identifiable covariance structures and are established under mild conditions on the distribution of the observations, which goes far beyond models with elliptically contoured densities. Some of our results are new and others are more general than existing ones in the literature. In this way, this manuscript completes and improves results on S-estimation in a wide variety of multivariate models. We illustrate our results by means of a simulation study and an application to data from a trial on the treatment of lead-exposed children.

1. Introduction. Linear models 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, $i = 1, \dots, n$ 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. One possibility is the linear mixed effects model, in which the random effects together with the measurement error yields a specific covariance structure depending on a vector $\boldsymbol{\theta}$ consisting of some unknown covariance parameters. Other covariance structures may arise, for example, if the $\mathbf{u}_i$ are the outcome of a time series; see, for example, [16] or [12] for different possible covariance structures.

Maximum likelihood estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ has been studied, for example, in [14, 19, 32]; see also [9, 12]. To be resistant against outliers, robust methods have been investigated for linear mixed effects models, for example, in [1, 4–6, 15, 29]. This mostly concerns S-estimators, originally introduced in the multiple regression context by Rousseeuw and Yohai [34] and extended to multivariate location and scatter in [8, 20], to multivariate regression in [39], and to linear mixed effects models in [4, 6, 15]. S-estimators are well-known smooth versions of the minimum volume ellipsoid estimator [33] that are highly resistant against outliers. As such, S-estimators have gained popularity as robust estimators, but they may also serve as initial estimators to further improve the efficiency. However, the theory about these estimators is far from complete, even in balanced models where the number of observed responses is the same for all subjects.

Received June 2022; revised October 2023.
MSC2020 subject classifications. 62F35, 62F12, 62H12, 62J05.
Key words and phrases. S-estimators, S-functionals, breakdown point, influence function, asymptotic distribution, structured covariance, repeated measurements.

In view of this, we provide a unified approach to S-estimation in balanced linear models with structured covariance matrices, and postpone a unified approach for unbalanced models to a future paper. The balanced setup is already quite flexible and includes several specific multivariate statistical models. Of main interest are S-estimators for linear mixed effects models, but our approach also includes S-estimators in several other standard multivariate models, such as multiple regression, multivariate regression and multivariate location and scatter. We provide sufficient conditions for the existence of S-functionals and S-estimators, establish their asymptotic properties, such as consistency and asymptotic normality, and derive their robustness properties in terms of breakdown point and influence function. All results are obtained for a large class of identifiable covariance structures, and are established under very mild conditions on the distribution of the observations, which goes far beyond models with elliptically contoured densities. In this way, some of our results are new and others are more general than existing ones in the literature.

Existence of S-estimators and S-functionals is established under mild conditions. Although existence of the estimators seems a basic requirement, such results are missing for instance for multivariate regression in [39] and for linear mixed effects models in [4, 6]. We obtain robustness properties for S-estimators, such as breakdown point and influence function, under mild conditions on collections of observations and under mild conditions on the distribution of the observations. High breakdown and a bounded influence function seem basic requirements for a robust method, but both properties are not available for linear mixed effects models [4, 6]. For multivariate regression [39], the influence function is only determined at distributions with an elliptical contoured density. Finally, we establish consistency and asymptotic normality for S-estimators under mild conditions on the distribution of the observations. A rigorous derivation is missing for multivariate regression [39], or is only available for observations from a normal distribution [4, 34].

We apply our asymptotic results, such as influence function and asymptotic normality, to the special case for which the distribution of the observations corresponds to an elliptically contoured density. In this way, we retrieve earlier results found in [20, 34, 39]. Somewhat surprisingly, the asymptotic variances of our S-estimators for linear mixed effects models in which the response has an elliptically contoured density, differ from the ones found in [6]. We investigate this difference by means of a simulation study.

The paper is organized as follows. In Section 2, we explain the model in detail and provide some examples of standard multivariate models that are included in our setup. In Section 3, we define the S-estimator and S-functional and in Section 4 we give conditions under which they exist. In Section 5, we establish continuity of the S-functional, which is then used to obtain consistency of the S-estimator. Section 6 deals with the breakdown point. Section 7 provides the preparation for Sections 8 and 9 in which we obtain the influence function and establish asymptotic normality. Finally, in Section 10, we illustrate our results by means of a simulation and investigate the performance of our estimators by means of an application to data from a trial on the treatment of lead-exposed children. All proofs are available as Supplementary Material [25].

2. Balanced models with structured covariances. We consider independent observations $(\mathbf{y}_1, \mathbf{X}_1), \dots, (\mathbf{y}_n, \mathbf{X}_n)$, for which we assume the following model:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{u}_i, \quad i = 1, \dots, n, \tag{1}$$

where $\mathbf{y}_i \in \mathbb{R}^k$ contains repeated measurements for the ith subject, $\boldsymbol{\beta} \in \mathbb{R}^q$ is an unknown parameter vector, $\mathbf{X}_i \in \mathbb{R}^{k \times q}$ is a known design matrix and $\mathbf{u}_i \in \mathbb{R}^k$ are unobservable independent mean zero random vectors with covariance matrix $\mathbf{V} \in \mathrm{PDS}(k)$, the class of positive definite symmetric $k \times k$ matrices. The model is balanced in the sense that all $\mathbf{y}_i$ have the same dimension. Furthermore, 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 $\boldsymbol{\theta} \in \mathbb{R}^l$. We first discuss some examples that are covered by this setup.

EXAMPLE 1. An important case of interest is the (balanced) linear mixed effects model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}\gamma_i + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, n.$$

This model arises from $\mathbf{u}_i = \mathbf{Z}\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i$, for $i = 1, \dots, n$, where $\mathbf{Z} \in \mathbb{R}^{k\times g}$ is known and $\boldsymbol{\gamma}_i \in \mathbb{R}^g$ and $\boldsymbol{\epsilon}_i \in \mathbb{R}^k$ are independent mean zero random variables, with unknown covariance matrices $\mathbf{G}$ and $\mathbf{R}$, respectively. In this case, $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$ and $\boldsymbol{\theta} = (\mathrm{vech}(\mathbf{G})^T, \mathrm{vech}(\mathbf{R})^T)^T$, where

$$\mathrm{vech}(\mathbf{A}) = (a_{11}, \dots, a_{k1}, a_{22}, \dots, a_{kk}) \tag{2}$$

is the unique $k(k+1)/2$-vector that stacks the columns of the lower triangle elements of a symmetric matrix $\mathbf{A}$. In full generality, the model is usually overparametrized and one may run into identifiability problems. A more feasible example is obtained by taking $\mathbf{R} = \sigma_0^2\mathbf{I}_k$, $\mathbf{Z} = [\mathbf{Z}_1 \cdots \mathbf{Z}_r]$ and $\boldsymbol{\gamma}_i = (\gamma_{i1} \dots, \gamma_{ir})^T$, where the $\mathbf{Z}_j$'s are known $k \times g_j$ design matrices and the $\gamma_{ij} \in \mathbb{R}^{g_j}$ are independent mean zero random variables with covariance matrix $\sigma_j^2\mathbf{I}_{g_j}$, for $j = 1, \dots, r$. This leads to

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \sum_{j=1}^{r}\mathbf{Z}_j\gamma_{ij} + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, n, \tag{3}$$

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

EXAMPLE 2. An example with an unstructured covariance is the multivariate linear regression model

$$\mathbf{y}_i = \mathbf{B}^T\mathbf{x}_i + \mathbf{u}_i, \quad i = 1, \dots, n, \tag{4}$$

where $\mathbf{B} \in \mathbb{R}^{q\times k}$ is a matrix of unknown parameters, $\mathbf{x}_i \in \mathbb{R}^q$ is known and $\mathbf{u}_i$, for $i = 1, \dots, n$, are independent mean zero random variables with covariance matrix $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{C} \in \mathrm{PDS}(k)$. In this case, the vector of unknown covariance parameters is given by

$$\boldsymbol{\theta} = \mathrm{vech}(\mathbf{C}) = (c_{11}, \dots, c_{1k}, c_{22}, \dots, c_{kk})^T \in \mathbb{R}^{\frac{1}{2}k(k+1)}. \tag{5}$$

The model can be obtained as a special case of (1), by taking $\mathbf{X}_i = \mathbf{x}_i^T \otimes \mathbf{I}_k$ and $\boldsymbol{\beta} = \mathrm{vec}(\mathbf{B}^T)$, where $\otimes$ denotes the Kronecker product and $\mathrm{vec}(\cdot)$ is the k^2-vector that stacks the columns of a matrix.

Clearly, the linear multiple regression model is a special case with $k = 1$. Of particular interest may be the SUR model [42]. This concerns seemingly unrelated multiple linear regression models, $\mathbf{y}_s = \mathbf{X}_s\boldsymbol{\beta}_s + \boldsymbol{\epsilon}_s$, for $s = 1, \dots, p$, with $\mathbf{y}_s \in \mathbb{R}^n$, that can be reformulated as a combined multiple linear regression model with np observations and a covariance matrix $\mathbf{V}(\boldsymbol{\theta}) = \boldsymbol{\Sigma} \otimes \mathbf{I}_n$, where $\boldsymbol{\theta} = \mathrm{vech}(\boldsymbol{\Sigma})$ and $\mathrm{cov}(\boldsymbol{\epsilon}_s, \boldsymbol{\epsilon}_t) = \sigma_{st}\mathbf{I}_n$, for $s, t = 1, \dots, p$.

EXAMPLE 3. Model (1) also includes examples, for which $\mathbf{u}_1, \dots, \mathbf{u}_n$ are generated from a time series. One example is the case where $\mathbf{u}_i$ has a covariance matrix with elements

$$v_{st} = \sigma^2\rho^{|s-t|}, \quad s, t = 1, \dots, k.$$

This arises when the $\mathbf{u}_i$'s are generated by an autoregressive process of order one. The vector of unknown covariance parameters is $\boldsymbol{\theta} = (\sigma^2, \rho) \in (0, \infty) \times [-1, 1]$. A general stationary process leads to

$$v_{st} = \theta_{|s-t|+1}, \quad s, t = 1, \dots, k, \tag{6}$$

in which case $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^T \in \mathbb{R}^k$, where $\theta_{|s-t|+1}$ represents the autocovariance over lag $|s-t|$.

EXAMPLE 4. Also the multivariate location-scale model can be obtained as a special case of (1), by taking $\mathbf{X}_i = \mathbf{I}_k$, the $k \times k$ identity matrix. In this case, $\boldsymbol{\beta} \in \mathbb{R}^k$ is the unknown location parameter and $\mathbf{V}(\boldsymbol{\theta})$ is the unstructured covariance matrix as in Example 2, with $\boldsymbol{\theta}$ as in (5).

Throughout the manuscript, we will assume that the parameter $\boldsymbol{\theta}$ is identifiable in the sense that

$$\mathbf{V}(\boldsymbol{\theta}_1) = \mathbf{V}(\boldsymbol{\theta}_2) \quad \Rightarrow \quad \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2. \tag{7}$$

This is true for all models in Examples 2, 3 and 4. This may not be true in general for the linear mixed effects model in Example 1 with unknown vech($\mathbf{G}$) and vech($\mathbf{R}$). For linear mixed effects models in (3), identifiability of $\boldsymbol{\theta} = (\sigma_0^2, \sigma_1^2, \dots, \sigma_r^2)$ holds for particular choices of the design matrices $\mathbf{Z}_1, \dots, \mathbf{Z}_r$.

3. Definitions. We start by representing our observations as points in $\mathbb{R}^k \times \mathbb{R}^{kq}$ in the following way. For $r = 1, \dots, k$, let $\mathbf{x}_r^T$ denote the rth row of the $k \times q$ matrix $\mathbf{X}$, so that $\mathbf{x}_r \in \mathbb{R}^q$. We represent the pair $\mathbf{s} = (\mathbf{y}, \mathbf{X})$ as an element in $\mathbb{R}^k \times \mathbb{R}^{kq}$ defined by $\mathbf{s}^T = (\mathbf{y}^T, \mathbf{x}_1^T, \dots, \mathbf{x}_k^T)$. In this way, our observations can be represented as $\mathbf{s}_1, \dots, \mathbf{s}_n$, with $\mathbf{s}_i = (\mathbf{y}_i, \mathbf{X}_i) \in \mathbb{R}^k \times \mathbb{R}^{kq}$.

3.1. *S-estimator.* S-estimators are defined by means of a function $\rho : \mathbb{R} \to [0, \infty)$ that satisfies the following properties:

(R1) ρ is symmetric around zero with $\rho(0) = 0$ and ρ is continuous at zero;

(R2) There exists a finite constant $c_0 > 0$, such that ρ is nondecreasing on $[0, c_0]$ and constant on $[c_0, \infty)$; put $a_0 = \sup \rho$.

The S-estimator $\boldsymbol{\xi}_n = (\boldsymbol{\beta}_n, \boldsymbol{\theta}_n)$ is defined as the solution to the following minimization problem:

$$\begin{aligned} &\min_{\boldsymbol{\beta},\boldsymbol{\theta}} \det\bigl(\mathbf{V}(\boldsymbol{\theta})\bigr) \\ &\text{subject to} \\ &\frac{1}{n}\sum_{i=1}^n \rho\Bigl(\sqrt{(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}(\boldsymbol{\theta})^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})}\Bigr) \le b_0, \end{aligned} \tag{8}$$

where the minimum is taken over all $\boldsymbol{\beta} \in \mathbb{R}^q$ and $\boldsymbol{\theta} \in \mathbb{R}^l$, such that $\mathbf{V}(\boldsymbol{\theta}) \in \text{PDS}(k)$, with ρ satisfying (R1)–(R2).

The S-estimator defined by (8) for the setup in (1) includes several specific cases that have been considered in the literature. The original regression S-estimator introduced by Rousseeuw and Yohai [34] is obtained as a special case by taking $\mathbf{X}_i = \mathbf{x}_i^T$ a $1 \times q$ vector and $\mathbf{V}(\boldsymbol{\theta}) = \sigma^2 > 0$. S-estimators for multivariate location and scale, as considered in Davies [8] and Lopuhaä [20] can be obtained by taking $\mathbf{X}_i$ and $\mathbf{V}(\boldsymbol{\theta})$ as in Example 4. For the multivariate regression model in Example 2, S-estimators have been considered by Van Aelst and Willems [39]. Bilodeau and Duchesne [3] and Peremans and Van Aelst [28] investigated S-estimators for SUR models. Copt and Victoria-Feser [6] and Chervoneva and Vishnyakov [4] consider S-estimators for the parameters in the linear mixed effects model (3).

The constant $0 < b_0 < a_0$ in (8) can be chosen in agreement with an assumed underlying distribution. For the multivariate regression model in [39], it is assumed that $\mathbf{y}_i \mid \mathbf{X}_i$ has an elliptically contoured density of the form

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{y}) = \det(\boldsymbol{\Sigma})^{-1/2} h\big((\mathbf{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\big), \tag{9}$$

with $\boldsymbol{\mu} = \mathbf{X}_i\boldsymbol{\beta}$ and $\boldsymbol{\Sigma} = \mathbf{V}(\boldsymbol{\theta})$ and $h : [0,\infty) \to [0,\infty)$. For the linear mixed effects model in [6], it is assumed that $\mathbf{y}_i \mid \mathbf{X}_i$ has a multivariate normal distribution, which is a special case of (9) with $h(t) = (2\pi)^{-k/2}\exp(-t/2)$. When the underlying distribution corresponds to a density of the form (9), then a natural choice is $b_0 = \mathbb{E}_{\mathbf{0},\mathbf{I}_k}\rho(\|\mathbf{z}\|)$, where $\mathbf{z}$ has density (9) with $(\boldsymbol{\mu},\boldsymbol{\Sigma}) = (\mathbf{0},\mathbf{I}_k)$. Finally, it should be emphasized that the ratio b_0/a_0 determines the breakdown point of the S-estimator (see Theorem 6.1), as well as its limiting variance (see Corollary 9.2). By choosing the constant c_0 in (R2), one then has to make a trade-off between robustness and efficiency.

Note that at this point we do not assume smoothness of ρ or strict monotonicity on $[0, c_0]$. This means that (R1)–(R2) allow the function $\rho(d) = 1 - \mathbb{1}_{[-c_0,c_0]}(d)$, which corresponds to the minimum volume ellipsoid (MVE) estimator in location-scale models (see [33]) and to the least median of squares estimator in linear regression models (see [35]). Indeed, with $\rho(d) = 1 - \mathbb{1}_{[-c_0,c_0]}(d)$, the S-estimator $(\boldsymbol{\beta}_n, \boldsymbol{\theta}_n)$ corresponds to the smallest cylinder

$$\mathcal{C}(\boldsymbol{\beta},\boldsymbol{\theta},c_0) = \big\{(\mathbf{y},\mathbf{X}) \in \mathbb{R}^k \times \mathbb{R}^{kq} : (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}(\boldsymbol{\theta})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \le c_0^2\big\} \tag{10}$$

that contains at least $n - nb_0$ points.

REMARK 1. Clearly, the definition of the S-estimator in (8) has great similarities with the S-estimator for multivariate location and covariance (see [8] and [20]), defined as the solution $(\mathbf{t}_n, \mathbf{C}_n)$ to the minimization problem

$$\begin{aligned} &\min_{\mathbf{t},\mathbf{C}} \det(\mathbf{C}) \\ &\text{subject to} \\ &\frac{1}{n}\sum_{i=1}^n \rho\Big(\sqrt{(\mathbf{y}_i-\mathbf{t})^T\mathbf{C}^{-1}(\mathbf{y}_i-\mathbf{t})}\Big) \le b_0, \end{aligned} \tag{11}$$

where the minimum is taken over all $\mathbf{t} \in \mathbb{R}^k$ and $\mathbf{C} \in \mathrm{PDS}(k)$. Even more so, if all $\mathbf{X}_i$ are assumed to be equal to the same design matrix $\mathbf{X}$ of full rank, as was done in [5, 6]. However, there is a subtle, but important difference between minimization problems (11) and (8). The important difference is that in (11) we minimize over *all* positive definite symmetric $k \times k$ matrices $\mathbf{C}$, whereas in (8), we *only* minimize over positive definite symmetric $k \times k$ matrices $\mathbf{V}(\boldsymbol{\theta})$, which can arise as the image of the mapping $\boldsymbol{\theta} \mapsto \mathbf{V}(\boldsymbol{\theta})$. The latter collection is a subset of the other:

$$\big\{\mathbf{V}(\boldsymbol{\theta}) \in \mathrm{PDS}(k) : \boldsymbol{\theta} \in \mathbb{R}^l\big\} \subset \mathrm{PDS}(k),$$

and will typically be a strictly smaller subset. This means that the properties of $\mathbf{V}(\boldsymbol{\theta}_n)$ and $\mathbf{C}_n$ are related, but the properties of $\mathbf{V}(\boldsymbol{\theta}_n)$ *cannot* simply be derived from properties of $\mathbf{C}_n$, not even in the case where all $\mathbf{X}_i$ are equal to the same $\mathbf{X}$. In fact, this will lead to limiting covariances that differ from the ones found in [6]; see Corollary 9.2.

3.2. *S-functional.* The concept of S-functional is needed to investigate local robustness properties of the corresponding S-estimator, such as the influence function (see Section 8). Let $\mathbf{s} = (\mathbf{y}, \mathbf{X})$ have a probability distribution P on $\mathbb{R}^k \times \mathbb{R}^{kq}$. The S-functional $\boldsymbol{\xi}(P) = (\boldsymbol{\beta}(P), \boldsymbol{\theta}(P))$ is defined as the solution to the following minimization problem:

$$\min_{\boldsymbol{\beta},\boldsymbol{\theta}} \det\bigl(\mathbf{V}(\boldsymbol{\theta})\bigr) \quad \text{subject to} \quad \int \rho\Bigl(\sqrt{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}(\boldsymbol{\theta})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}\Bigr)\,\mathrm{d}P(\mathbf{y},\mathbf{X}) \le b_0, \tag{12}$$

where the minimum is taken over all $\boldsymbol{\beta} \in \mathbb{R}^q$ and $\boldsymbol{\theta} \in \mathbb{R}^l$, such that $\mathbf{V}(\boldsymbol{\theta}) \in \mathrm{PDS}(k)$, with ρ satisfying (R1)–(R2).

As a special case, we obtain the S-estimator $\boldsymbol{\xi}_n = (\boldsymbol{\beta}_n, \boldsymbol{\theta}_n)$ by taking $P = \mathbb{P}_n$, the empirical measure corresponding to the observations $(\mathbf{y}_1, \mathbf{X}_1), \ldots, (\mathbf{y}_n, \mathbf{X}_n)$. In view of this connection, existence and consistency of solutions to (8) will follow from general results on the existence and the continuity of solutions to (12).

The definition of the S-functionals for the multivariate location-scale model given in Lopuhaä [20] and for the multivariate regression model given by Van Aelst and Willems [39] can be obtained as special cases of (12), by choosing $\mathbf{X}$, $\boldsymbol{\beta}$ and $\mathbf{V}(\boldsymbol{\theta})$ as in Examples 4 and 2, respectively. Copt and Victoria-Feser [6] do not pay attention to S-functionals or the influence function in the linear mixed effects model (3). However, S-functionals for linear mixed effects models can be also be obtained as a special case of (12), by choosing $\mathbf{X}$, $\boldsymbol{\beta}$ and $\mathbf{V}(\boldsymbol{\theta})$ as in Example 1.

4. Existence.

We will first establish existence of the S-functional $\boldsymbol{\xi}(P)$ defined by (12), under particular conditions on the probability measure P. As a consequence, this will also yield the existence of the S-estimator, defined by (8). Recall that $(\mathbf{y}_1, \mathbf{X}_1), \ldots, (\mathbf{y}_n, \mathbf{X}_n)$ are represented as points in $\mathbb{R}^k \times \mathbb{R}^{kq}$. Note, however, that for linear models with intercept, the first column of each $\mathbf{X}_i$ consists of 1's. This means that the points $(\mathbf{y}_i, \mathbf{X}_i)$ are concentrated in a lower-dimensional subset of $\mathbb{R}^k \times \mathbb{R}^{kq}$. A similar situation occurs when all $\mathbf{X}_i$ are equal to the same design matrix, such as in [6]. In view of this, define $\mathcal{X} \subset \mathbb{R}^{kq}$ as the subset with the lowest dimension $p = \dim(\mathcal{X}) \le kq$ satisfying

$$P(\mathbf{X} \in \mathcal{X}) = 1. \tag{13}$$

Hence, P is then concentrated on the subset $\mathbb{R}^k \times \mathcal{X}$ of $\mathbb{R}^k \times \mathbb{R}^{kq}$, which is of dimension $k + p$, which may be smaller than $k + kq$.

The first condition that we require expresses the fact that P cannot have too much mass at infinity in relation to the ratio $r = b_0/a_0$.

(C1$_\epsilon$) There exists a compact set $K_\epsilon \subset \mathbb{R}^k \times \mathcal{X}$, such that $P(K_\epsilon) \ge r + \epsilon$.

The second condition requires that P cannot have too much mass at arbitrarily thin strips in $\mathbb{R}^k \times \mathcal{X}$. For $\boldsymbol{\alpha} \in \mathbb{R}^{k+kq}$, such that $\|\boldsymbol{\alpha}\| = 1$, $\ell \in \mathbb{R}$ and $\delta \ge 0$, we define a strip $H(\boldsymbol{\alpha}, \ell, \delta)$ as follows:

$$H(\boldsymbol{\alpha}, \ell, \delta) = \bigl\{\mathbf{s} \in \mathbb{R}^k \times \mathbb{R}^{kq} : \ell - \delta/2 \le \boldsymbol{\alpha}^T\mathbf{s} \le \ell + \delta/2\bigr\}. \tag{14}$$

Defined in this way, a strip is the area between two parallel hyperplanes, which are symmetric around the hyperplane $H(\boldsymbol{\alpha}, \ell, 0) = \{\mathbf{s} \in \mathbb{R}^k \times \mathbb{R}^{kq} : \boldsymbol{\alpha}^T\mathbf{s} = \ell\}$. Since the distance between two parallel hyperplanes $\boldsymbol{\alpha}^T\mathbf{s} = \ell_1$ and $\boldsymbol{\alpha}^T\mathbf{s} = \ell_2$ is $|\ell_1 - \ell_2|$, the strip $H(\boldsymbol{\alpha}, \ell, \delta)$ defined as in (14) has width δ. We require the following condition.

(C2$_\epsilon$) The value

$$\delta_\epsilon = \inf\{\delta : P\big(H(\boldsymbol{\alpha}, \ell, \delta)\big) \geq \epsilon, \boldsymbol{\alpha} \in \mathbb{R}^{k+kq}, \|\boldsymbol{\alpha}\| = 1, \ell \in \mathbb{R}, \delta \geq 0\}$$

is strictly positive.

According to (13), in (C2$_\epsilon$) one only needs to consider strips in $\mathbb{R}^k \times \mathcal{X}$.

Both conditions are satisfied for any $0 < \epsilon \leq 1 - r$ by any probability measure P that is absolutely continuous. Clearly, condition (C1$_\epsilon$) holds for any $0 \leq \epsilon \leq 1 - r$ for the empirical measure $\mathbb{P}_n$ corresponding to a collection of n points $\mathcal{S}_n = \{\mathbf{s}_1, \dots, \mathbf{s}_n\} \subset \mathbb{R}^k \times \mathcal{X}$. Condition (C2$_\epsilon$) for $\epsilon = (k + p + 1)/n$ is also satisfied by the empirical measure $\mathbb{P}_n$, when the collection $\mathcal{S}_n$ is in *general position*, that is, no subset $J \subset \mathcal{S}_n$ of $k + p + 1$ points is contained in the same hyperplane in $\mathbb{R}^k \times \mathcal{X}$. Conditions (C1$_\epsilon$) and (C2$_\epsilon$) together are similar to condition (C$_\epsilon$) in [20]. The reason that (C1$_\epsilon$) slightly deviates from [20] is to handle the presence of $\mathbf{X}$ in minimization problem (12).

REMARK 2. Note that condition (C2$_\epsilon$) is equivalent with

$$\omega_\epsilon = \inf_{P(J)\geq\epsilon} \inf_{\|\boldsymbol{\alpha}\|=1} \inf_{\ell\in\mathbb{R}} \sup_{\mathbf{s}\in J} \big|\boldsymbol{\alpha}^T\mathbf{s} - \ell\big| > 0, \tag{15}$$

where the infima are taken over all subsets $J \subset \mathbb{R}^k \times \mathcal{X}$ with $P(J) \geq \epsilon$, all vectors $\boldsymbol{\alpha} \in \mathbb{R}^{k+kq}$, with $\|\boldsymbol{\alpha}\| = 1$ and levels $\ell \in \mathbb{R}$. Details can be found in [25].

To establish existence of the S-functional, we follow the reasoning in [20]. The idea is to argue that one can restrict oneself to a compact set for finding solutions to (12). When the object function in (12) is continuous, this immediately yields existence of a solution of (12). To this end, we assume the following condition.

(V1) The mapping $\boldsymbol{\theta} \mapsto \mathbf{V}(\boldsymbol{\theta})$ is continuous.

The lemma below is fundamental for the existence of the S-functional. It requires that the identity is in $\mathcal{V} = \{\mathbf{V}(\boldsymbol{\theta}) \in \text{PDS}(k) : \boldsymbol{\theta} \in \mathbb{R}^l\}$ and that $\mathcal{V}$ is closed under multiplication with a positive scalar.

(V2) There exists a $\boldsymbol{\theta} \in \mathbb{R}^l$, such that $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{I}_k$. For any $\mathbf{V}(\boldsymbol{\theta}) \in \mathcal{V}$ and any $\alpha > 0$, it holds that $\alpha\mathbf{V}(\boldsymbol{\theta}) = \mathbf{V}(\boldsymbol{\theta}')$, for some $\boldsymbol{\theta}' \in \mathbb{R}^l$.

Conditions (V1)–(V2) are not very restrictive. For example, all models in Examples 1 to 4 satisfy these conditions.

For any $k \times k$ matrix $\mathbf{A}$, let $\lambda_k(\mathbf{A}) \leq \cdots \leq \lambda_1(\mathbf{A})$ denote the eigenvalues of $\mathbf{A}$. We then have the following key lemma for the existence of S-functionals. The lemma is similar to Lemma 1 in [20] and its proof can be found in [25].

LEMMA 4.1. *Let* $(\boldsymbol{\beta}, \boldsymbol{\theta}) \in \mathbb{R}^q \times \mathbb{R}^l$, $0 < m_0 < \infty$, $0 < c < \infty$, *and* $0 < \epsilon < 1$, *and suppose that the mapping* $\boldsymbol{\theta} \mapsto \mathbf{V}(\boldsymbol{\theta})$ *satisfies* (V2). *Then the following properties hold*:

(i) *If P satisfies* (C2$_\epsilon$) *and* $P(\mathcal{C}(\boldsymbol{\beta}, \boldsymbol{\theta}, c))) \geq \epsilon$, *then* $\lambda_k(\mathbf{V}(\boldsymbol{\theta})) \geq a_1 > 0$, *where a_1 only depends on c and the width δ_ϵ from condition* (C2$_\epsilon$).

(ii) *Suppose* $\int \rho(\|\mathbf{y}\|/m_0)\,\mathrm{d}P(\mathbf{s}) \leq b_0$. *Then for any solution* $(\boldsymbol{\beta}, \boldsymbol{\theta})$ *of* (12), *which is such that* $\lambda_k(\mathbf{V}(\boldsymbol{\theta})) \geq a_1 > 0$, *it holds that* $\lambda_1(\mathbf{V}(\boldsymbol{\theta})) \leq a_2 < \infty$, *where a_2 only depends on a_1 and m_0*.

(iii) *Let P satisfy* (C2$_\epsilon$) *and suppose that* $P(\mathcal{C}(\boldsymbol{\beta}, \boldsymbol{\theta}, c)) \geq a > 0$. *Suppose there exists a compact set* $K \in \mathbb{R}^k \times \mathcal{X}$, *such that* $P(K) \geq 1 - a + \epsilon$. *If* $\lambda_1(\mathbf{V}(\boldsymbol{\theta})) \leq a_2 < \infty$, *then* $\|\boldsymbol{\beta}\| \leq M < \infty$, *where M only depends on c, a_2, the set K and a constant $\gamma_\epsilon > 0$ that can be deduced from condition* (C2$_\epsilon$).

Lemma 4.1 will ensure that there exists a compact set that contains all pairs $(\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\theta}))$ that correspond to possible solutions $(\boldsymbol{\beta}, \boldsymbol{\theta})$ of (12). To establish that possible solutions $(\boldsymbol{\beta}, \boldsymbol{\theta})$ of (12) are in a compact set, we need that the preimage $\{\boldsymbol{\theta} \in \mathbb{R}^l : \mathbf{V}(\boldsymbol{\theta}) \in K\}$ of a compact set $K \subset \mathbb{R}^{k\times k}$ is again compact. Recall that subsets of $\mathbb{R}^l$ are compact if and only if they are closed and bounded, and note that the preimage of a continuous mapping of a closed set is closed. Hence, in view of condition (V1), it suffices to require the following condition.

(V3) The mapping $\boldsymbol{\theta} \mapsto \mathbf{V}(\boldsymbol{\theta})$ is such that the preimage of a bounded set is bounded.

Condition (V3) is satisfied by all models in Examples 1 to 4, including the linear mixed effects model of Example 1, as long as the matrix $\mathbf{Z}$ is of full rank. We then have the following theorem.

THEOREM 4.2. *Consider minimization problem* (12) *with* ρ *satisfying* (R1)–(R2). *Suppose that* P *satisfies* ($\mathrm{C1}_\epsilon$) *and* ($\mathrm{C2}_\epsilon$), *for some* $0 < \epsilon \leq 1 - r$, *where* $r = b_0/a_0$, *and suppose that* $\mathbf{V}$ *satisfies* (V1)–(V3). *Then there exists at least one solution to* (12).

The theorem has a direct corollary for the existence of the S-estimator, when dealing with a collections of points. Let $\mathcal{S}_n = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, with $\mathbf{s}_i = (\mathbf{y}_i, \mathbf{X}_i)$ be a collection of n points in $\mathbb{R}^k \times \mathcal{X}$. Define

(16) $\kappa(\mathcal{S}_n) =$ maximal number of points of $\mathcal{S}_n$ lying on the same hyperplane in $\mathbb{R}^k \times \mathcal{X}$.

For example, if the distribution P is absolutely continuous, then $\kappa(\mathcal{S}_n) \leq k + p$ with probability one. We then have the following corollary.

COROLLARY 4.3. *Consider minimization problem* (8) *with* ρ *satisfying* (R1)–(R2), *for a collection* $\mathcal{S}_n = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subset \mathbb{R}^k \times \mathcal{X}$, *with* $\mathbf{s}_i = (\mathbf{y}_i, \mathbf{X}_i)$, *for* $i = 1, \ldots, n$. *Suppose that* $\mathbf{V}$ *satisfies* (V1)–(V3). *If* $\kappa(\mathcal{S}_n) + 1 \leq n(1 - r)$, *where* $r = b_0/a_0$, *then there exists at least one solution* $\boldsymbol{\xi}_n = (\boldsymbol{\beta}_n, \boldsymbol{\theta}_n)$ *to minimization problem* (8).

Copt and Victoria-Feser [6] consider S-estimators for the linear mixed effects model (3). Despite their Proposition 1 about the asymptotic behavior of solutions to their S-minimization problem [6], equation (7), the actual existence of such a solution is not established. However, this now follows from our Corollary 4.3. In their case, $\mathbf{V}(\boldsymbol{\theta})$ satisfies conditions (V1) and (V2). It can be seen, that if all matrices $\mathbf{Z}_j$, for $j = 1, \ldots, r$, are of full rank, then $\mathbf{V}(\boldsymbol{\theta})$ also satisfies (V3). The translated biweight ρ-function proposed in [6] satisfies (R1)–(R2). Finally, under their assumption that $\mathbf{X}_i = \mathbf{X}$ is the same and $\mathbf{y}_i \mid \mathbf{X} \sim N_k(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\theta}))$, it follows that $\kappa(\mathcal{S}_n) \leq k$. It then follows from Corollary 4.3 that with $b_0 \leq a_0(n - k - 1)/n$, at least one solution to their S-minimization problem exists.

For the multivariate regression model from Example 2, Van Aelst and Willems [39] do not explicitly prove existence of the S-estimator. Since in their case, $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{C} \in \mathrm{PDS}(k)$ satisfies (V1)–(V3) and the conditions imposed in [39] on the ρ-function satisfy (R1)–(R2), the existence of their S-estimator now also follows from Corollary 4.3, when b_0 is chosen suitably.

Existence of S-estimators is obtained from existence of S-functionals at the empirical measure $\mathbb{P}_n$. The following corollary shows that existence can be established in general, for probability measures that are close to P. It requires the following condition on P.

(C3) Let $\mathfrak{C}$ be the class of all measurable convex subsets of $\mathbb{R}^k \times \mathbb{R}^{kq}$. Every $C \in \mathfrak{C}$ is a P-continuity set, that is, $P(\partial C) = 0$, where ∂C denotes the boundary of C.

COROLLARY 4.4. *Suppose that ρ satisfies* (R1)–(R2) *and* $\mathbf{V}$ *satisfies* (V1)–(V3). *Let P_t, $t \geq 0$ be a sequence of probability measures on $\mathbb{R}^k \times \mathbb{R}^{kq}$ that converges weakly to P, as $t \to \infty$. Suppose that P satisfies* (C3), *as well as* ($C1_{\epsilon'}$) *and* ($C2_{\epsilon}$), *for some $0 < \epsilon < \epsilon' \leq 1 - r = b_0/a_0$. Then, for t sufficiently large, the minimization problem* (12) *with probability measure P_t has at least one solution $\boldsymbol{\xi}(P_t)$.*

Condition (C3) is needed to apply Theorem 4.2 in [31]. Clearly, this condition is satisfied if P is absolutely continuous.

5. Continuity and consistency. Consider a sequence P_t, $t \geq 0$, of probability measures on $\mathbb{R}^k \times \mathbb{R}^{kq}$ that converges weakly to P, as $t \to \infty$. By continuity of the S-functional $\boldsymbol{\xi}(P)$, we mean that $\boldsymbol{\xi}(P_t) \to \boldsymbol{\xi}(P)$, as $t \to \infty$. An example of such a sequence is the sequence of empirical measures $\mathbb{P}_n$, $n = 1, 2, \ldots$, that converges weakly to P, almost surely. Continuity of the S-functional for this sequence would then mean that the S-estimator $\boldsymbol{\xi}_n$ is consistent, that is, $\boldsymbol{\xi}_n = \boldsymbol{\xi}(\mathbb{P}_n) \to \boldsymbol{\xi}(P)$, almost surely.

We require an additional condition for the function ρ.

(R3) ρ is continuous and strictly increasing on $[0, c_0]$.

For $\mathbf{s} = (\mathbf{y}, \mathbf{X})$ and $\boldsymbol{\xi} = (\boldsymbol{\beta}, \boldsymbol{\theta})$, define the Mahalanobis distances by

$$d^2(\mathbf{s}, \boldsymbol{\xi}) = d^2(\mathbf{s}, \boldsymbol{\beta}, \boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}(\boldsymbol{\theta})^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{17}$$

We then have the following theorem for the S-functional $\boldsymbol{\xi}(P) = (\boldsymbol{\beta}(P), \boldsymbol{\theta}(P))$.

THEOREM 5.1. *Let P_t, $t \geq 0$ be a sequence of probability measures on $\mathbb{R}^k \times \mathbb{R}^{kq}$ that converges weakly to P, as $t \to \infty$, and let $\boldsymbol{\xi}(P_t)$ be a solution to minimization problem* (12) *with probability measure P_t. Suppose that ρ satisfies* (R1)–(R3) *and* $\mathbf{V}$ *satisfies* (V1)–(V3). *Suppose that P satisfies* (C3), *as well as* ($C1_{\epsilon'}$) *and* ($C2_{\epsilon}$), *for some $0 < \epsilon < \epsilon' \leq 1 - r = b_0/a_0$. If the solution $\boldsymbol{\xi}(P)$ of* (12) *is unique, then for any sequence of solutions $\boldsymbol{\xi}(P_t)$, $t \geq 0$, it holds* $\lim_{t \to \infty} \boldsymbol{\xi}(P_t) = \boldsymbol{\xi}(P)$.

Theorem 5.1 is an extension of Theorem 3.1 in [20] on the continuity of S-functionals for multivariate location and scale. Continuity of S-functionals for multiple regression has been investigated in [11].

Continuity of the S-functional will be used to derive the influence function of the S-estimator in Section 8. Another nice consequence of the continuity of the S-functional is that one can directly obtain consistency of the S-estimator. Consider the S-estimator $\boldsymbol{\xi}_n$ defined by minimization problem (8). Recall that $\boldsymbol{\xi}_n = \boldsymbol{\xi}(\mathbb{P}_n)$, so that we can use Theorem 5.1 to establish consistency of the S-estimator.

COROLLARY 5.2. *Let $\boldsymbol{\xi}_n$ be a solution to minimization problem* (8). *Suppose ρ satisfies* (R1)–(R3) *and* $\mathbf{V}$ *satisfies* (V1)–(V3). *Suppose that P satisfies* (C3) *as well as* ($C1_{\epsilon'}$) *and* ($C2_{\epsilon}$), *for some $0 < \epsilon < \epsilon' \leq 1 - r = b_0/a_0$. If the solution $\boldsymbol{\xi}(P)$ of* (12) *is unique, then* $\lim_{n \to \infty} \boldsymbol{\xi}_n = \boldsymbol{\xi}(P)$, *with probability one.*

Theorem 5.1 and Corollary 5.2 require that $\boldsymbol{\xi}(P)$ is the unique solution to minimization problem (12). An example of a distribution P for which $\boldsymbol{\xi}(P)$ is unique, is when P is such that $\mathbf{y} \mid \mathbf{X}$ has an elliptically contoured density (9). This situation is very similar to that of multivariate location-scale S-estimators, for which Davies [8], Theorem 1, shows that the corresponding S-minimization problem (12) has a unique solution. This has been extended by Tatsuoka and Tyler [37], Theorem 4.2 together with Theorem 2.1, to a broader class, consisting of affine transformations of distributions on $\mathbb{R}^k$, which are invariant under permutations and sign changes of its components and which have densities g such that $g \circ \exp$ is Schur-concave (see [37] for details), that is,

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{y}) = \det(\boldsymbol{\Sigma})^{-1/2} g\big(\boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})\big). \tag{18}$$

The next theorem is a direct consequence of that result. Its proof can be found in [25]. Note that elliptically contoured densities in (9) are special cases of (18) by taking $g(\mathbf{z}) = h(\mathbf{z}^T\mathbf{z})$.

THEOREM 5.3. *Suppose that $\rho : \mathbb{R} \to [0, \infty)$ satisfies* (R1)–(R2) *and that there exists at least one solution $(\boldsymbol{\beta}(P), \boldsymbol{\theta}(P))$ to* (12). *Suppose that P is such $\mathbf{y} \mid \mathbf{X}$ has density $f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ from* (18), *with $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}_0$ and $\boldsymbol{\Sigma} = \mathbf{V}(\boldsymbol{\theta}_0)$. Suppose that either ρ is strictly increasing on* $[0, c_0]$ *or g in* (18) *is strictly M-concave* (*see* [37], *Definition* 4.4). *If $\mathbf{V}$ satisfies* (V1)–(V3) *and $\mathbf{X}^T\mathbf{X}$ is nonsingular with probability one, then $(\boldsymbol{\beta}(P), \boldsymbol{\theta}(P)) = (\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$ with probability one.*

Uniqueness for the S-functional in general cannot be expected. Some kind of symmetry and unimodality seems to be needed to assure uniqueness. For a more detailed discussion, we refer to Tatsuoka and Tyler [37], who provide some counterexamples. An elliptically contoured density for $\mathbf{y}_i \mid \mathbf{X}_i$ in the context of S-estimators for specific cases of the model (1) has been assumed in [8] for the multivariate location-scale model of Example 4, in [39] for the multivariate regression model of Example 2 and in [6] for the linear mixed effects model (3). More precisely, in [6] it is assumed that $\mathbf{X}_i = \mathbf{X}$ and that $\mathbf{y}_i \mid \mathbf{X}$ has a multivariate normal distribution. In that case, the function g in (18) satisfies all the conditions of Theorem 5.3.

Uniqueness of the S-estimator seems difficult to establish. Davies [7] proves uniqueness of the MVE, which corresponds to a location-scale S-estimator with $\rho(d) = 1 - \mathbb{1}_{[-c_0,c_0]}(d)$, at samples from an elliptically contoured density. However, it does not seem straightforward to extend his result to general S-estimators. Tatsuoka and Tyler [37] conjecture that location-scale S-estimators are unique with probability 1 when sampling from an absolutely continuous distribution, even one for which the corresponding S-functional is not unique. Finally, for large enough random samples from a distribution for which the S-functional is uniquely defined, strong consistency of the S-estimators, as established in Corollary 5.2, ensures that all possible solutions to (3.1) are within a shrinking compact set as the sample size goes to infinity.

6. Global robustness: The breakdown point. Consider a collection of points $\mathcal{S}_n = \{\mathbf{s}_i = (\mathbf{y}_i, \mathbf{X}_i), i = 1, \ldots, n\} \subset \mathbb{R}^k \times \mathcal{X}$. To emphasize the dependence on the collection $\mathcal{S}_n$, denote by $\boldsymbol{\xi}_n(\mathcal{S}_n) = (\boldsymbol{\beta}_n(\mathcal{S}_n), \boldsymbol{\theta}_n(\mathcal{S}_n))$, the S-estimator, as defined in (8). To investigate the global robustness of S-estimators, we compute that finite-sample (replacement) breakdown point. For a given collection $\mathcal{S}_n$ the finite-sample breakdown point (see Donoho and Huber [10]) of a regression S-estimator $\boldsymbol{\beta}_n$ is defined as the smallest proportion of points from $\mathcal{S}_n$ that one needs to replace in order to carry the estimator over all bounds. More precisely,

$$\epsilon_n^*(\boldsymbol{\beta}_n, \mathcal{S}_n) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{\mathcal{S}'_m} \|\boldsymbol{\beta}_n(\mathcal{S}_n) - \boldsymbol{\beta}_n(\mathcal{S}'_m)\| = \infty \right\}, \tag{19}$$

where the minimum runs over all possible collections $\mathcal{S}'_m$ that can be obtained from $\mathcal{S}_n$ by replacing m points of $\mathcal{S}_n$ by arbitrary points in $\mathbb{R}^k \times \mathcal{X}$.

The estimator $\boldsymbol{\theta}_n$ determines the covariance estimator $\mathbf{V}_n = \mathbf{V}(\boldsymbol{\theta}_n)$. For this reason, it seems natural to let the breakdown point of $\boldsymbol{\theta}_n$ correspond to the breakdown of a covariance estimator. We define the finite sample (replacement) breakdown point of the S-estimator

$\boldsymbol{\theta}_n$ at a collection $\mathcal{S}_n$ as

$$\epsilon_n^*(\boldsymbol{\theta}_n, \mathcal{S}_n) = \min_{1\leq m\leq n} \left\{ \frac{m}{n} : \sup_{\mathcal{S}'_m} \operatorname{dist}(\mathbf{V}(\boldsymbol{\theta}_n(\mathcal{S}_n))), \mathbf{V}(\boldsymbol{\theta}_n(\mathcal{S}'_m)) = \infty \right\}, \tag{20}$$

with $\operatorname{dist}(\cdot,\cdot)$ defined as $\operatorname{dist}(\mathbf{A},\mathbf{B}) = \max\{|\lambda_1(\mathbf{A}) - \lambda_1(\mathbf{B})|, |\lambda_k(\mathbf{A})^{-1} - \lambda_k(\mathbf{B})^{-1}|\}$, where the minimum runs over all possible collections $\mathcal{S}'_m$ that can be obtained from $\mathcal{S}_n$ by replacing m points of $\mathcal{S}_n$ by arbitrary points in $\mathbb{R}^k \times \mathcal{X}$. So, the breakdown point of $\boldsymbol{\theta}_n$ is the smallest proportion of points from $\mathcal{S}_n$ that one needs to replace in order to make the largest eigenvalue of $\mathbf{V}(\boldsymbol{\theta}(\mathcal{S}'_m))$ arbitrarily large (explosion), or to make the smallest eigenvalue of $\mathbf{V}(\boldsymbol{\theta}(\mathcal{S}'_m))$ arbitrarily small (implosion).

Good global robustness is illustrated by a high breakdown point. The breakdown point of the S-estimators is given in the theorem below. It extends the results for S-estimators of multivariate location and scale; see [8] and [26], and S-estimators for multivariate regression; see [39]. For S-estimators in the linear mixed effects model considered in [6], the breakdown point has not been established. This will now follow as a special case from the next theorem. Its proof can be found in [25].

THEOREM 6.1. *Consider minimization problem* (8) *with* ρ *satisfying* (R1)–(R2). *Suppose that* $\mathbf{V}$ *satisfies* (V1)–(V3). *Let* $\mathcal{S}_n \subset \mathbb{R}^k \times \mathcal{X}$ *be a collection of* n *points* $\mathbf{s}_i = (\mathbf{y}_i, \mathbf{X}_i)$, $i = 1, \ldots, n$. *Let* $r = b_0/a_0$ *and suppose that* $0 < r \leq (n - \kappa(\mathcal{S}_n))/(2n)$, *where* $\kappa(\mathcal{S}_n)$ *is defined by* (16). *Then for any solution* $(\boldsymbol{\beta}_n, \boldsymbol{\theta}_n)$ *of minimization problem* (12),

$$\frac{\lfloor (n+1)/2 \rfloor}{n} \geq \epsilon_n^*(\boldsymbol{\beta}_n, \mathcal{S}_n) \geq \frac{\lceil nr \rceil}{n},$$

$$\epsilon_n^*(\boldsymbol{\theta}_n, \mathcal{S}_n) = \frac{\lceil nr \rceil}{n}.$$

The largest possible value of the breakdown point occurs when $r = (n - \kappa(\mathcal{S}_n))/(2n)$, in which case $\lceil nr \rceil / n = \lceil (n - \kappa(\mathcal{S}_n))/2 \rceil / n = \lfloor (n - \kappa(\mathcal{S}_n) + 1)/2 \rfloor / n$. When the collection $\mathcal{S}_n$ is in general position, then $\kappa(\mathcal{S}_n) = k + p$. In that case, the breakdown point of both estimators is at least equal to $\lfloor (n - k - p + 1)/2 \rfloor / n$. When all $\mathbf{X}_i$ are equal to the same $\mathbf{X}$, in [5, 6], one has $p = 0$ and $\kappa(\mathcal{S}_n) = k$. In that case, the breakdown point of $\boldsymbol{\theta}_n$ is equal to $\lfloor (n - k + 1)/2 \rfloor / n$. This coincides with the maximal breakdown point for affine equivariant estimators for $k \times k$ covariance matrices (see [8], Theorem 6).

REMARK 3. Van Aelst and Willems [39] also take into account $r > (n - \kappa(\mathcal{S}_n))/(2n)$. For this case, they show that the breakdown point is $(\lceil n - nr \rceil - \kappa(\mathcal{S}_n))/n$. We could not extend this to our general setup, but we are able to show that solutions to (8) do not break down, when replacing at most $\lceil n - nr \rceil - \kappa(\mathcal{S}_n) - 1$ points.

7. Score equations. Up to this point, properties of S-functionals and S-estimators have been derived from the minimization problems (8) and (12). To obtain the influence function and to establish the limiting distribution of S-estimators, we use the score equations that can be found by differentiation of the Lagrangian corresponding to the constrained minimization problems. To this end, we require the following additional condition on the function ρ:

(R4) ρ is continuously differentiable and $u(s) = \rho'(s)/s$ is continuous,

and the following condition on the mapping $\boldsymbol{\theta} \mapsto \mathbf{V}(\boldsymbol{\theta})$,

(V4) $\mathbf{V}(\boldsymbol{\theta})$ is continuously differentiable.

Obviously, condition (V4) implies the former condition (V1).

7.1. *General covariance structures.* Let $\boldsymbol{\xi}_P = (\boldsymbol{\beta}_P, \boldsymbol{\theta}_P)$ be a solution to minimization problem (12). If we denote the corresponding Lagrange multiplier by λ_P, then the pair $(\boldsymbol{\xi}_P, \lambda_P)$ is a zero of all partial derivatives $\partial L_P/\partial\boldsymbol{\beta}$, $\partial L_P/\partial\boldsymbol{\theta}$ and $\partial L_P/\partial\lambda$, where L_P is the Lagrangian given by

$$L_P(\boldsymbol{\xi}, \lambda) = \log\det(\mathbf{V}(\boldsymbol{\theta})) - \lambda\left\{\int \rho\left(\sqrt{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}(\boldsymbol{\theta})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}\right) \mathrm{d}P(\mathbf{y},\mathbf{X}) - b_0\right\}.$$

If $\mathbb{E}_P\|\mathbf{X}\| < \infty$, then under conditions (R4) and (V4), one may interchange the order of integration and differentiation in $\partial L_P/\partial\boldsymbol{\beta}$ and $\partial L_P/\partial\boldsymbol{\theta}$, on a neighborhood of $\boldsymbol{\xi}_P$. It follows that besides the constraint in (12), the pair $(\boldsymbol{\xi}_P, \lambda_P)$ satisfies

$$\begin{aligned} &\int u(d)\mathbf{X}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,\mathrm{d}P(\mathbf{s}) = \mathbf{0}, \\ &\operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\right) + \frac{\lambda}{2}\int u(d)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,\mathrm{d}P(\mathbf{s}) = 0, \end{aligned} \tag{21}$$

for $j = 1,\ldots,l$, where $u(s) = \rho'(s)/s$ and $d = d(\mathbf{s},\boldsymbol{\xi})$ is defined by (17), and where we abbreviate $\mathbf{V}(\boldsymbol{\theta})$ by $\mathbf{V}$. To solve λ_P from the second set of equations, we multiply the jth equation by θ_j and then sum over $j = 1,\ldots,l$. This leads to

$$\operatorname{tr}\left(\mathbf{V}^{-1}\sum_{j=1}^{l}\theta_j\frac{\partial\mathbf{V}}{\partial\theta_j}\right) + \frac{\lambda}{2}\int u(d)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\left(\sum_{j=1}^{l}\theta_j\frac{\partial\mathbf{V}}{\partial\theta_j}\right)\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,\mathrm{d}P(\mathbf{s}) = 0,$$

which is solved by

$$\lambda_P = \frac{-2\operatorname{tr}(\mathbf{V}^{-1}\sum_{j=1}^{l}\theta_j(\partial\mathbf{V}/\partial\theta_j))}{\int u(d)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\sum_{j=1}^{l}\theta_j(\partial\mathbf{V}/\partial\theta_j))\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,\mathrm{d}P(\mathbf{s})}.$$

When we insert this back into the second equation in (21), we find

$$\begin{aligned} &\operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\right)\int u(d)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\left(\sum_{t=1}^{l}\theta_t\frac{\partial\mathbf{V}}{\partial\theta_t}\right)\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,\mathrm{d}P(\mathbf{s}) \\ &\quad - \operatorname{tr}\left(\mathbf{V}^{-1}\sum_{t=1}^{l}\theta_t\frac{\partial\mathbf{V}}{\partial\theta_t}\right)\int u(d)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,\mathrm{d}P(\mathbf{s}) = 0, \end{aligned}$$

or briefly

$$\int u(d)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\mathbf{H}_j\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,\mathrm{d}P(\mathbf{s}) = 0, \quad j = 1,\ldots,l, \tag{22}$$

where

$$\mathbf{H}_j = \operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\right)\left(\sum_{t=1}^{l}\theta_t\frac{\partial\mathbf{V}}{\partial\theta_t}\right) - \operatorname{tr}\left(\mathbf{V}^{-1}\sum_{t=1}^{l}\theta_t\frac{\partial\mathbf{V}}{\partial\theta_t}\right)\frac{\partial\mathbf{V}}{\partial\theta_j}. \tag{23}$$

Because $\sum_{j=1}^{l}\theta_j\mathbf{H}_j = \mathbf{0}$, the system of equations (22) is linearly dependent. Similar to [20] we subtract the S-constraint from each equation. For each $j = 1,\ldots,l$, we subtract the term

$$\operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\right)(\rho(d) - b_0)$$

from the left-hand side of equation (22). We then find that any solution $\boldsymbol{\xi}_P$ of (12) satisfies the following equation:

$$\int \Psi(\mathbf{s},\boldsymbol{\xi})\,\mathrm{d}P(\mathbf{s}) = \mathbf{0}, \tag{24}$$

where $\Psi = (\Psi_{\boldsymbol{\beta}}, \Psi_{\boldsymbol{\theta}})$, with $\Psi_{\boldsymbol{\theta}} = (\Psi_{\boldsymbol{\theta},1}, \dots, \Psi_{\boldsymbol{\theta},l})$, where

$$\Psi_{\boldsymbol{\beta}}(\mathbf{s}, \boldsymbol{\xi}) = u(d)\mathbf{X}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

$$\Psi_{\boldsymbol{\theta},j}(\mathbf{s}, \boldsymbol{\xi}) = u(d)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\mathbf{H}_j\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\right)(\rho(d) - b_0), \tag{25}$$

for $j = 1, \dots, l$, where $\mathbf{H}_j$ and $d = d(\mathbf{s}, \boldsymbol{\xi})$ are defined in (23) and (17), respectively, and where we abbreviate $\mathbf{V}(\boldsymbol{\theta})$ by $\mathbf{V}$.

The regression score equation for $\Psi_{\boldsymbol{\beta}}$ with the empirical measure $\mathbb{P}_n$ for P in (24) coincides with the one for the regression S-estimator in the linear mixed effects model (3) considered in [6] (see their equation (10)). The empirical regression score equation also coincides with the one for the regression S-estimator in the multivariate regression model of Example 2 considered in [39] (see equation (2.2) in [38]). Similarly, the empirical score equation for $\Psi_{\boldsymbol{\beta}}$ coincides with the one for the location S-estimator of Example 4 considered in [20].

For general covariance structures, the empirical covariance score equation for $\Psi_{\boldsymbol{\theta}}$ does not compare directly to existing equations in the literature. However, as we will see in the next subsection, similar comparisons are available for models with a linear covariance structure.

7.2. *Linear covariance structures.* In the previous section, we solved λ from (21) and subtracted the S-constraint, leading to score equation (24) with Ψ given in (25). The fact that this was done in a specific way has the following reason. In cases where $\mathbf{V}(\boldsymbol{\theta})$ is linear, say

$$\mathbf{V}(\boldsymbol{\theta}) = \sum_{j=1}^{l} \theta_j \mathbf{L}_j, \tag{26}$$

the function $\Psi_{\boldsymbol{\theta}}$ simplifies a lot and can also be related to the covariance psi-function in [20]. Typical models of interest that have a covariance matrix of this type are the mixed linear effects model from Example 1 and the multivariate regression model and SUR model from Example 2. But also the multivariate location-scale model from Example 4 and the time-series model (6) from Example 3 have linear covariance structures.

When $\mathbf{V}$ is of the form (26), then $\partial\mathbf{V}/\partial\theta_j = \mathbf{L}_j$ and $\sum_{j=1}^{l}\theta_j(\partial\mathbf{V}/\partial\theta_j) = \mathbf{V}$. In this case, (23) simplifies to $\mathbf{H}_j = \mathrm{tr}(\mathbf{V}^{-1}\mathbf{L}_j)\mathbf{V} - k\mathbf{L}_j$, and $\Psi_{\boldsymbol{\theta},j}$ in (25) becomes

$$\Psi_{\boldsymbol{\theta},j}(\mathbf{s}, \boldsymbol{\xi}) = \mathrm{tr}(\mathbf{V}^{-1}\mathbf{L}_j)v(d) - ku(d)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\mathbf{L}_j\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

where $u(s)$ is defined in (R4) and

$$v(s) = u(s)s^2 - \rho(s) + b_0. \tag{27}$$

Using that $\mathrm{tr}(\mathbf{A}^T\mathbf{B}) = \mathrm{vec}(\mathbf{A})^T \mathrm{vec}(\mathbf{B})$, this can be written as

$$\Psi_{\boldsymbol{\theta},j}(\mathbf{s}, \boldsymbol{\xi}) = -\,\mathrm{vec}\big(ku(d)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T - v(d)\mathbf{V}\big)^T \mathrm{vec}\big(\mathbf{V}^{-1}\mathbf{L}_j\mathbf{V}^{-1}\big).$$

On the right-hand side, we recognize $ku(d)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T - v(d)\mathbf{V}$, being the covariance psi-function that also appears in (2.8) in [20]. For our purposes, we define

$$\Psi_{\mathbf{V}}(\mathbf{s}, \boldsymbol{\xi}) = ku\big(d(\mathbf{s}, \boldsymbol{\xi})\big)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T - v\big(d(\mathbf{s}, \boldsymbol{\xi})\big)\mathbf{V}. \tag{28}$$

The functions $\Psi_{\boldsymbol{\theta},j}$, for $j = 1, \dots, l$, can be combined in one expression for the vector valued function $\Psi_{\boldsymbol{\theta}}$ as follows. First, note that

$$\mathrm{vec}\big(\mathbf{V}^{-1}\mathbf{L}_j\mathbf{V}^{-1}\big) = \big(\mathbf{V}^{-1} \otimes \mathbf{V}^{-1}\big)\,\mathrm{vec}(\mathbf{L}_j)$$

for $j = 1, \dots, l$. Define the $k^2 \times l$ matrix

$$\mathbf{L} = \big[\mathrm{vec}(\mathbf{L}_1) \cdots \mathrm{vec}(\mathbf{L}_l)\big]. \tag{29}$$

Then the column vector $\Psi_{\boldsymbol{\theta}} = (\Psi_{\boldsymbol{\theta},1}, \dots, \Psi_{\boldsymbol{\theta},l})$ can be written as

$$\Psi_{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\xi}) = -\mathbf{L}^T \left(\mathbf{V}^{-1} \otimes \mathbf{V}^{-1}\right) \operatorname{vec}\left(\Psi_{\mathbf{V}}(\mathbf{s}, \boldsymbol{\xi})\right),$$

where $\Psi_{\mathbf{V}}$ is defined in (28) and $\mathbf{L}$ in (29). Note that the dependence on $\mathbf{s} = (\mathbf{y}, \mathbf{X})$ in $\Psi_{\boldsymbol{\theta}}$ is only through the function $\Psi_{\mathbf{V}}$. We conclude that in the case of a linear covariance structure, any solution $\boldsymbol{\xi}_P$ of (12) satisfies (24), where $\Psi = (\Psi_{\boldsymbol{\beta}}, \Psi_{\boldsymbol{\theta}})$, with

$$\begin{aligned} \Psi_{\boldsymbol{\beta}}(\mathbf{s}, \boldsymbol{\xi}) &= u(d)\mathbf{X}^T \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \\ \Psi_{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\xi}) &= -\mathbf{L}^T \left(\mathbf{V}^{-1} \otimes \mathbf{V}^{-1}\right) \operatorname{vec}\left(\Psi_{\mathbf{V}}(\mathbf{s}, \boldsymbol{\xi})\right), \end{aligned} \tag{30}$$

where $d = d(\mathbf{s}, \boldsymbol{\xi})$ is defined in (17), and where we abbreviate $\mathbf{V}(\boldsymbol{\theta})$ by $\mathbf{V}$.

For the multivariate regression model in Example 2, one has $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{C}$, where $\boldsymbol{\theta} = \operatorname{vech}(\mathbf{C})$. The matrix $\mathbf{L} = \partial \operatorname{vec}(\mathbf{V})/\partial \boldsymbol{\theta}^T$ is then equal to the so-called duplication matrix $\mathcal{D}_k$, which is the unique $k^2 \times k(k+1)/2$ matrix, with the properties $\mathcal{D}_k \operatorname{vech}(\mathbf{C}) = \operatorname{vec}(\mathbf{C})$ and $(\mathcal{D}_k^T \mathcal{D}_k)^{-1} \mathcal{D}_k^T \operatorname{vec}(\mathbf{C}) = \operatorname{vech}(\mathbf{C})$ (e.g., see [27], Chapter 3, Section 8). Because $\mathbf{V}$ has full rank, it follows that equation (24) holds for $\Psi = (\Psi_{\boldsymbol{\beta}}, \Psi_{\mathbf{V}})$. The resulting score equations for the empirical measure $\mathbb{P}_n$ corresponding to observations $(\mathbf{y}_i, \mathbf{X}_i)$, for $i = 1, \dots, n$, are then equivalent with the ones found in [39].

For the linear mixed effects model (3), the covariance matrix $\mathbf{V}(\boldsymbol{\theta})$ has a linear structure with the vector $\boldsymbol{\theta} = (\sigma_0^2, \dots, \sigma_r^2)$ of unknown covariance parameters. The matrix $\mathbf{L}$ is then a $k^2 \times (r+1)$ matrix and will typically be of rank $r + 1 < k^2$. As a consequence, in this case one cannot further simplify equation (24), by removing the factor $\mathbf{L}^T(\mathbf{V}^{-1} \otimes \mathbf{V}^{-1})$ from the function $\Psi_{\boldsymbol{\theta}}$. The score equation for $\Psi_{\boldsymbol{\beta}}$ resulting from the empirical measure $\mathbb{P}_n$ corresponding to observations $(\mathbf{y}_i, \mathbf{X}_i)$, for $i = 1, \dots, n$, is the same as the one obtained in [6]. The corresponding score equation for $\Psi_{\boldsymbol{\theta}}$ differs slightly from the one in [6], because the authors do not subtract a term with $\rho(d) - b_0$ to remove the linear dependency of the equations (22).

8. Local robustness: The influence function. For $0 < h < 1$ and $\mathbf{s} = (\mathbf{y}, \mathbf{X}) \in \mathbb{R}^k \times \mathbb{R}^{kq}$ fixed, define the perturbed probability measure $P_{h,\mathbf{s}} = (1-h)P + h\delta_{\mathbf{s}}$, where $\delta_{\mathbf{s}}$ denotes the Dirac measure at $\mathbf{s} \in \mathbb{R}^k \times \mathbb{R}^{kq}$. The *influence function* of the functional $\boldsymbol{\xi}(\cdot)$ at probability measure P is defined as

$$\operatorname{IF}(\mathbf{s}; \boldsymbol{\xi}, P) = \lim_{h \downarrow 0} \frac{\boldsymbol{\xi}((1-h)P + h\delta_{\mathbf{s}}) - \boldsymbol{\xi}(P)}{h}, \tag{31}$$

if this limit exists. In contrast to the global robustness measured by the breakdown point, the influence function measures the local robustness. It describes the effect of an infinitesimal contamination at a single point $\mathbf{s}$ on the functional (see Hampel [13]). Good local robustness is therefore illustrated by a bounded influence function.

8.1. *The general case.* The theorem below gives the influence function for the S-functional $\boldsymbol{\xi}$. It extends the result for S-functionals of multivariate location and scale [20]. Under the assumption that the limit in (31) exists and P has an elliptical contoured density (9), Van Aelst and Willems [39] relate the influence function for S-functionals of multivariate regression to that of S-functionals of multivariate location and scale. For the linear mixed effects model considered in [6], the influence function has not been established. The influence function for these functionals now follows as a special case from the theorem below.

We will show that the limit in (31) exists and derive its expression at general P. Since the value of $\boldsymbol{\theta}$ determines the covariance matrix $\mathbf{V}(\boldsymbol{\theta})$, we also include the influence function of the covariance functional. Consider the S-functional at $P_{h,\mathbf{s}_0}$. From the Portmanteau theorem [2], Theorem 2.1, it can easily be seen that $P_{h,\mathbf{s}_0} \to P$, weakly, as $h \downarrow 0$. Therefore, under the conditions of Corollary 4.4 and Theorem 5.1, it follows that there exist solutions $\boldsymbol{\xi}(P_{h,\mathbf{s}_0})$ and $\boldsymbol{\xi}(P)$ to minimization problem (12) at $P_{h,\mathbf{s}_0}$ and P, respectively, and that $\boldsymbol{\xi}(P_{h,\mathbf{s}_0}) \to \boldsymbol{\xi}(P)$, as $h \downarrow 0$.

THEOREM 8.1. *Let $\boldsymbol{\xi}(P_{h,\mathbf{s}_0})$ and $\boldsymbol{\xi}(P)$ be solutions to minimization problems* (12) *at $P_{h,\mathbf{s}_0}$ and P, respectively, and suppose that $\boldsymbol{\xi}(P_{h,\mathbf{s}_0}) \to \boldsymbol{\xi}(P)$, as $h \downarrow 0$. Suppose that ρ satisfies* (R4) *and* $\mathbf{V}$ *satisfies* (V4). *Let Ψ be defined in* (25) *and suppose that*

$$\Lambda(\boldsymbol{\xi}) = \int \Psi(\mathbf{s}, \boldsymbol{\xi})\, \mathrm{d}P(\mathbf{s}), \tag{32}$$

is continuously differentiable with a nonsingular derivative $\mathbf{D}(P)$ at $\boldsymbol{\xi}(P)$. Then for $\mathbf{s}_0 \in \mathbb{R}^k \times \mathbb{R}^{kq}$,

$$\mathrm{IF}(\mathbf{s}_0; \boldsymbol{\xi}, P) = -\mathbf{D}(P)^{-1}\Psi\big(\mathbf{s}_0, \boldsymbol{\xi}(P)\big).$$

For the covariance functional $\mathbf{C}(P) = \mathbf{V}(\boldsymbol{\theta}(P))$, it holds that

$$\mathrm{IF}\big(\mathbf{s}_0; \mathrm{vec}(\mathbf{C}), P\big) = \left(\frac{\partial\, \mathrm{vec}(\mathbf{V}(\boldsymbol{\theta}(P)))}{\partial \boldsymbol{\theta}^T}\right) \mathrm{IF}(\mathbf{s}_0; \boldsymbol{\theta}, P).$$

To investigate the local robustness of S-estimators, we derive the following bound on the influence function for $\boldsymbol{\xi}(P)$.

COROLLARY 8.2. *Suppose that ρ satisfies* (R2) *and* (R4), *and* $\mathbf{V}$ *satisfies* (V4). *Then there exist $0 < C_1 < \infty$ and $0 < C_2 < \infty$, only depending on P, such that for $\mathbf{s} = (\mathbf{y}, \mathbf{X})$ it holds that $\|\mathrm{IF}(\mathbf{s}, \boldsymbol{\xi}(P))\| \leq C_1 + C_2\|\mathbf{X}\|$.*

Its proof can be found in [25].

8.2. *Elliptically contoured densities.* When P is such that $\mathbf{y} \mid \mathbf{X}$ has an elliptically contoured density (9) and $\mathbf{V}(\boldsymbol{\theta})$ is linear, we can obtain a more detailed expression for the influence function. This requires the following condition on the function ρ,

(R5) ρ is twice continuously differentiable,

and the following condition on the mapping $\boldsymbol{\theta} \mapsto \mathbf{V}(\boldsymbol{\theta})$,

(V5) $\mathbf{V}(\boldsymbol{\theta})$ is twice continuously differentiable.

Conditions (R5) and (V5) are needed to establish that Λ, as defined in (32), is continuously differentiable. Clearly, condition (V5) implies former conditions (V4) and (V1).

Suppose that P is such that $\mathbf{y} \mid \mathbf{X}$ has elliptically contoured density $f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ from (9), with $\boldsymbol{\mu} \in \mathbb{R}^k$ and $\boldsymbol{\Sigma} \in \mathrm{PDS}(k)$. When the S-functional is affine equivariant, it suffices to determine the influence function for the case $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{0}, \mathbf{I}_k)$. However, this does not hold in general for the S-functionals in our setting. The reason is that, for a $k \times k$ nonsingular matrix $\mathbf{A}$ and $\boldsymbol{\theta} \in \mathbb{R}^l$, the matrix $\mathbf{A}\mathbf{V}(\boldsymbol{\theta})\mathbf{A}^T$ may not be of the form $\mathbf{V}(\boldsymbol{\theta}')$, for some $\boldsymbol{\theta}' \in \mathbb{R}^l$. Examples are the (linear) covariance structure that corresponds to the linear mixed effects model (3) considered in [6] or the models discussed in Example 3.

Nevertheless, note that for the general case with $\boldsymbol{\mu} \in \mathbb{R}^k$ and $\boldsymbol{\Sigma} \in \mathrm{PDS}(k)$, we can still use the fact that, conditionally on $\mathbf{X}$, the distribution of $\mathbf{y}$ is the same as that of $\boldsymbol{\Sigma}^{1/2}\mathbf{z} + \boldsymbol{\mu}$, where $\mathbf{z}$ has a spherical density $f_{\mathbf{0},\mathbf{I}_k}$. As a consequence, we can still obtain the following result, which enables one to determine the influence functions of the functionals $\boldsymbol{\beta}(P)$ and $\boldsymbol{\theta}(P)$ separately.

If P itself is also absolutely continuous, then it satisfies (C3), as well as ($C1_{\epsilon'}$) and ($C2_{\epsilon}$), for any $0 < \epsilon < \epsilon' \leq 1 - r$. When ρ and $\mathbf{V}$ satisfy (R1)–(R3) and (V1)–(V3), it follows from Theorem 4.2 and Corollary 4.4 that $\boldsymbol{\xi}(P)$ and $\boldsymbol{\xi}(P_{h,\mathbf{s}})$ exist, for h sufficiently small. If h in (9) is nonincreasing and not constant on $[0, c_0^2]$, then $\boldsymbol{\xi}(P)$ is unique, according to Theorem 5.3, so that $\boldsymbol{\xi}(P_{h,\mathbf{s}}) \to \boldsymbol{\xi}(P)$, as $h \downarrow 0$. Hence, in order to apply Theorem 8.1, it remains to show that Λ in (32) is continuously differentiable with a nonsingular derivative at $\boldsymbol{\xi}(P)$. As a first step, we obtain that the derivative of Λ is a block matrix.

LEMMA 8.3. *Suppose that P is such that $\mathbf{y} \mid \mathbf{X}$ has an elliptically contoured density $f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ from* (9) *and* $\mathbb{E}\|\mathbf{X}\|^2 < \infty$. *Suppose that $\boldsymbol{\xi}(P)$ is a solution to the corresponding minimization problem* (12), *such that $(\mathbf{X}\boldsymbol{\beta}(P), \mathbf{V}(\boldsymbol{\theta}(P))) = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Suppose that ρ satisfies* (R2), (R4)–(R5) *and that* $\mathbf{V}$ *satisfies* (V5) *and has a linear structure* (26). *Let Λ be defined in* (32) *with Ψ defined in* (30). *Then*

$$\frac{\partial \Lambda(\boldsymbol{\xi}(P))}{\partial \boldsymbol{\xi}} = \begin{pmatrix} \dfrac{\partial \Lambda_{\boldsymbol{\beta}}(\boldsymbol{\xi}(P))}{\partial \boldsymbol{\beta}} & \mathbf{0} \\ \mathbf{0} & \dfrac{\partial \Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}(P))}{\partial \boldsymbol{\theta}} \end{pmatrix},$$

where

$$\frac{\partial \Lambda_{\boldsymbol{\beta}}(\boldsymbol{\xi}(P))}{\partial \boldsymbol{\beta}} = -\alpha \mathbb{E}[\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X}], \tag{33}$$

with

$$\alpha = \mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\left(1 - \frac{1}{k}\right)\frac{\rho'(\|\mathbf{z}\|)}{\|\mathbf{z}\|} + \frac{1}{k}\rho''(\|\mathbf{z}\|)\right], \tag{34}$$

and

$$\frac{\partial \Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}(P))}{\partial \boldsymbol{\theta}} = \gamma_1 \mathbf{L}^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{L} - \gamma_2 \mathbf{L}^T \operatorname{vec}(\boldsymbol{\Sigma}^{-1}) \operatorname{vec}(\boldsymbol{\Sigma}^{-1})^T \mathbf{L},$$

where $\mathbf{L} = \partial \operatorname{vec}(\mathbf{V}(\boldsymbol{\theta}(P)))/\partial \boldsymbol{\theta}^T$ is the $k^2 \times l$ matrix given in (29) *and*

$$\begin{aligned} \gamma_1 &= \frac{\mathbb{E}_{\mathbf{0},\mathbf{I}_k}[\rho''(\|\mathbf{z}\|)\|\mathbf{z}\|^2 + (k+1)\rho'(\|\mathbf{z}\|)\|\mathbf{z}\|]}{k+2}, \\ \gamma_2 &= \frac{\mathbb{E}_{\mathbf{0},\mathbf{I}_k}[2\rho''(\|\mathbf{z}\|)\|\mathbf{z}\|^2 + k\rho'(\|\mathbf{z}\|)\|\mathbf{z}\|]}{2k(k+2)}, \end{aligned} \tag{35}$$

The proof is tedious, but straightforward, and can be found in [25].

REMARK 4. The proof of Lemma 8.3 uses the fact that

$$\frac{\partial \Lambda(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = \int \frac{\partial \Psi(\mathbf{s}, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}} \, \mathrm{d}P(\mathbf{s}),$$

for all $\boldsymbol{\xi}$ in a neighborhood of $\boldsymbol{\xi}(P)$. This holds for general P and any covariance structure $\mathbf{V}(\boldsymbol{\theta})$ that satisfies (V2)–(V3) and (V5); see Lemma 11.3 in [25]. Furthermore, Lemma 8.3 is obtained for a linear covariance structure. However, with some additional technicalities, this result can also be shown to hold for Ψ defined in (25) corresponding to general covariance structures. For general covariance structures, one still obtains (33), and that

$$\begin{aligned} \frac{\partial \Lambda_{\boldsymbol{\theta},j}(\boldsymbol{\xi}(P))}{\partial \theta_s} = &-\alpha_1 \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1} \frac{\partial \mathbf{V}(\boldsymbol{\theta}(P))}{\partial \theta_s} \boldsymbol{\Sigma}^{-1} \mathbf{H}_j\right) \\ &+ \alpha_2 \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1} \frac{\partial \mathbf{V}(\boldsymbol{\theta}(P))}{\partial \theta_s}\right) \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1} \frac{\partial \mathbf{V}(\boldsymbol{\theta}(P))}{\partial \theta_j}\right), \end{aligned}$$

for $j, s = 1, \ldots, l$, and where $\alpha_1 = \gamma_1/k$ and $\alpha_2 = \gamma_1/k - \gamma_2$, with γ_1, γ_2 from (35), and where $\mathbf{H}_j$ is defined in (23).

The next corollary gives expressions for the influence functions of the functionals $\boldsymbol{\beta}(P)$ and $\boldsymbol{\theta}(P)$ separately, at a distribution P that is such that $\mathbf{y} \mid \mathbf{X}$ has an elliptically contoured density. The proof is tedious, but straightforward, and can be found in [25].

COROLLARY 8.4. *Suppose that P is such that $\mathbf{y} \mid \mathbf{X}$ has an elliptically contoured density $f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ from* (9)*, such that $(\mathbf{X}\boldsymbol{\beta}(P), \mathbf{V}(\boldsymbol{\theta}(P))) = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\boldsymbol{\xi}(P_{h,\mathbf{s}_0})$ and $\boldsymbol{\xi}(P)$ be a solution to minimization problem* (12) *at $P_{h,\mathbf{s}_0}$ and P, respectively, and suppose that $\boldsymbol{\xi}(P_{h,\mathbf{s}_0}) \to \boldsymbol{\xi}(P)$, as $h \downarrow 0$. Suppose that $\mathbb{E}\|\mathbf{X}\|^2 < \infty$ and suppose that ρ satisfies* (R2)–(R5) *and that $\mathbf{V}$ satisfies* (V5)*, and has a linear structure* (26)*. Let α, γ_1, and γ_2 be defined in* (34) *and* (35)*, and suppose that $\mathbb{E}_{\mathbf{0},\mathbf{I}_k}[\rho''(\|\mathbf{z}\|)] > 0$. If $\mathbf{X}$ has full rank with probability one, then*

$$\mathrm{IF}(\mathbf{s}_0, \boldsymbol{\beta}, P) = \frac{u(d_0)}{\alpha}(\mathbb{E}[\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}])^{-1}\mathbf{X}_0^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta}(P)),$$

where $d_0^2 = (\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta}(P))^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta}(P))$ and $u(s) = \rho'(s)/s$. If $\gamma_1 > 0$ and the $k^2 \times l$ matrix $\mathbf{L}$, as defined in (29)*, has full rank, then* $\mathrm{IF}(\mathbf{s}_0, \boldsymbol{\theta}, P)$ *is given by*

$$\begin{aligned}&\frac{ku(d_0)}{\gamma_1}(\mathbf{L}^T(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\mathbf{L})^{-1}\mathbf{L}^T\operatorname{vec}\big(\boldsymbol{\Sigma}^{-1}(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta}(P))(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta}(P))^T\boldsymbol{\Sigma}^{-1}\big)\\&\quad + \left(-\frac{u(d_0)d_0^2}{\gamma_1} + \frac{\rho(d_0) - b_0}{\gamma_1 - k\gamma_2}\right)\boldsymbol{\theta}(P).\end{aligned}$$

Note that since $\mathbf{L}\boldsymbol{\theta}(P) = \operatorname{vec}(\mathbf{V}(\boldsymbol{\theta}(P))) = \operatorname{vec}(\boldsymbol{\Sigma})$, we can immediately obtain the influence function for the covariance functional $\mathbf{C}(P) = \mathbf{V}(\boldsymbol{\theta}(P))$. From Theorem 8.1, it immediately follows that $\mathrm{IF}(\mathbf{s}_0, \operatorname{vec}(\mathbf{C}), P)$ is given by

$$\begin{aligned}&\frac{ku(d_0)}{\gamma_1}\mathbf{L}(\mathbf{L}^T(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\mathbf{L})^{-1}\mathbf{L}^T\operatorname{vec}\big(\boldsymbol{\Sigma}^{-1}(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta}(P))(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta}(P))^T\boldsymbol{\Sigma}^{-1}\big)\\&\quad + \left(-\frac{u(d_0)d_0^2}{\gamma_1} + \frac{\rho(d_0) - b_0}{\gamma_1 - k\gamma_2}\right)\operatorname{vec}(\boldsymbol{\Sigma}).\end{aligned}$$

Since the functions $u(s)s = \rho'(s)$, $u(s)s^2 = \rho'(s)s$, and $\rho(s)$ are bounded, it follows that $\mathrm{IF}(\mathbf{s}, \boldsymbol{\theta}, P)$ and $\mathrm{IF}(\mathbf{s}, \operatorname{vec}(\mathbf{C}), P)$ are bounded uniformly in both $\mathbf{y}$ and $\mathbf{X}$, whereas $\mathrm{IF}(\mathbf{s}, \boldsymbol{\beta}, P)$ is bounded uniformly in $\mathbf{y}$, but not in $\mathbf{X}$. This illustrates the phenomenon in linear regression that leverage points can have a high effect on the regression S-estimator.

For the S-estimators in the linear mixed effects model (3) with normal errors considered in [6], the influence function is not available. The expression can now be obtained from Corollary 8.4. The expression for $\mathrm{IF}(\mathbf{s}, \boldsymbol{\beta}, P)$ in Corollary 8.4 coincides with the one found for the multivariate regression S-functional in [39], where $\alpha > 0$ is the same constant as the one in the expression of the influence function for the location S-functional in [20]. Furthermore, for the multivariate regression model, one has $\boldsymbol{\theta} = \operatorname{vech}(\mathbf{C})$ and the matrix $\mathbf{L}$ is equal to the duplication matrix $\mathcal{D}_k$. From the properties of $\mathcal{D}_k$, the expressions for the influence functions simplify. One finds in this case that

$$\mathrm{IF}(\mathbf{s}, \boldsymbol{\theta}, P) = \frac{ku(d)}{\gamma_1}\operatorname{vech}\big((\mathbf{y} - \mathbf{X}\boldsymbol{\beta}(P))(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}(P))^T\big) + \left(-\frac{u(d)d^2}{\gamma_1} + \frac{\rho(d) - b_0}{\gamma_1 - k\gamma_2}\right)\boldsymbol{\theta}(P)$$

and the influence function of the covariance functional $\mathbf{C}(P) = \mathbf{V}(\boldsymbol{\theta}(P))$ itself is given by

$$\mathrm{IF}(\mathbf{s}, \mathbf{C}, P) = \frac{ku(d)}{\gamma_1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}(P))(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}(P))^T + \left(-\frac{u(d)d^2}{\gamma_1} + \frac{\rho(d) - b_0}{\gamma_1 - k\gamma_2}\right)\boldsymbol{\Sigma}.$$

This coincides with the expressions found for the covariance S-functionals in [39] and in [20].

9. Asymptotic normality. To establish asymptotic normality of the S-estimators, we use the score equations obtained from differentiation of the Lagrangian corresponding to minimization problem (8). In the same way as before, we obtain score equation (24), with P equal to the empirical measure $\mathbb{P}_n$ corresponding to observations $\mathbf{s}_1, \dots, \mathbf{s}_n$, with $\mathbf{s}_i = (\mathbf{y}_i, \mathbf{X}_i) \in \mathbb{R}^k \times \mathbb{R}^{kq}$. From (24), we see that any solution $\boldsymbol{\xi}_n = \boldsymbol{\xi}(\mathbb{P}_n)$ to the S-minimization problem (8) must satisfy

$$\int \Psi(\mathbf{s}, \boldsymbol{\xi}_n)\, \mathrm{d}\mathbb{P}_n(\mathbf{s}) = \mathbf{0}, \tag{36}$$

where $\Psi = (\Psi_{\boldsymbol{\beta}}, \Psi_{\boldsymbol{\theta}})$ is defined in (25).

9.1. *General case.* Writing $\boldsymbol{\xi}_P = \boldsymbol{\xi}(P)$, we decompose (36) as follows:

$$\begin{aligned} 0 = &\int \Psi(\mathbf{s}, \boldsymbol{\xi}_n)\, \mathrm{d}P(\mathbf{s}) + \int \Psi(\mathbf{s}, \boldsymbol{\xi}_P)\, \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) \\ &+ \int \bigl(\Psi(\mathbf{s}, \boldsymbol{\xi}_n) - \Psi(\mathbf{s}, \boldsymbol{\xi}_P)\bigr)\, \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}). \end{aligned} \tag{37}$$

The essential step in establishing asymptotic normality of $\boldsymbol{\xi}_n$, is to show that the third term on the right-hand side of (37) is of the order $o_P(n^{-1/2})$. To this end, we will apply results from empirical process theory as developed in Pollard [30]. This leads to the following theorem.

THEOREM 9.1. *Suppose that ρ satisfies* (R1)–(R2) *and* (R4), *such that $u(s)$ is of bounded variation, and suppose that* $\mathbf{V}$ *satisfies* (V4). *Let $\boldsymbol{\xi}_n$ and $\boldsymbol{\xi}(P)$ be solutions to minimization problems* (8) *and* (12), *and suppose that $\boldsymbol{\xi}_n \to \boldsymbol{\xi}(P)$ in probability. Suppose that Λ, as defined in* (32) *with Ψ defined in* (25), *is continuously differentiable with a nonsingular derivative $\mathbf{D}(P)$ at $\boldsymbol{\xi}(P)$ and suppose that $\mathbb{E}\|\mathbf{X}\|^2 < \infty$. Then $\sqrt{n}(\boldsymbol{\xi}_n - \boldsymbol{\xi}(P))$ is asymptotically normal with mean zero and covariance matrix* $\mathbf{D}(P)^{-1}\mathbb{E}[\Psi(\mathbf{s}, \boldsymbol{\xi}(P))\Psi(\mathbf{s}, \boldsymbol{\xi}(P))^T]\mathbf{D}(P)^{-1}$.

Theorem 9.1 is similar to Theorem 4.1 in [20]. Note that Theorem 9.1 confirms the well-known heuristic that relates the limiting covariance of $\sqrt{n}(\boldsymbol{\xi}_n - \boldsymbol{\xi}(P))$ to the influence function of the functional $\boldsymbol{\xi}(\cdot)$ given in Theorem 8.1,

$$\mathbf{D}(P)^{-1}\mathbb{E}\bigl[\Psi\bigl(\mathbf{s}, \boldsymbol{\xi}(P)\bigr)\Psi\bigl(\mathbf{s}, \boldsymbol{\xi}(P)\bigr)^T\bigr]\mathbf{D}(P)^{-1} = \mathbb{E}\bigl[\mathrm{IF}(\mathbf{s}, \boldsymbol{\xi}, P)\,\mathrm{IF}(\mathbf{s}, \boldsymbol{\xi}, P)^T\bigr]. \tag{38}$$

Van Aelst and Willems [39] consider the limiting behavior of S-estimators in the multivariate regression model of Example 2, but only under P for which $\mathbf{y} \mid \mathbf{X}$ has an elliptical contoured density. Copt and Victoria-Feser [6] consider asymptotic normality for S-estimators in the linear mixed effects model (3) with a constant design matrix $\mathbf{X}_i = \mathbf{X}$ and only consider P for which $\mathbf{y} \mid \mathbf{X}$ has an multivariate normal distribution.

9.2. *Elliptically contoured densities.* Consider the special case that P is such that $\mathbf{y} \mid \mathbf{X}$ has an elliptically contoured density $f_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}$ from (9), with $\boldsymbol{\mu} \in \mathbb{R}^k$ and $\boldsymbol{\Sigma} \in \mathrm{PDS}(k)$. As before, in determining the limiting normal distribution of the individual S-estimators, we cannot use affine equivariance and restrict ourselves to the case $(\mathbf{0}, \mathbf{I}_k)$. Instead, we use some of the results obtained in Section 8.2 to establish the limiting normal distributions of the S-estimators $\boldsymbol{\beta}_n = \boldsymbol{\beta}(\mathbb{P}_n)$, $\boldsymbol{\theta}_n = \boldsymbol{\theta}(\mathbb{P}_n)$ and $\mathbf{C}_n = \mathbf{V}(\boldsymbol{\theta}(\mathbb{P}_n))$.

COROLLARY 9.2. *Suppose that P is such that $\mathbf{y} \mid \mathbf{X}$ has an elliptically contoured density $f_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}$ from* (9), *such that $(\mathbf{X}\boldsymbol{\beta}(P), \mathbf{V}(\boldsymbol{\theta}(P))) = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\boldsymbol{\xi}_n$ and $\boldsymbol{\xi}(P)$ be solutions to minimization problems* (8) *and* (12), *and suppose that $\boldsymbol{\xi}_n \to \boldsymbol{\xi}(P)$ in probability. Suppose* *that* $\mathbb{E}\|\mathbf{X}\|^2 < \infty$ *and suppose that* ρ *satisfies* (R2)–(R5), *such that* $u(s)$ *is of bounded variation. Suppose that* $\mathbf{V}$ *satisfies* (V5), *and has a linear structure* (26). *Let* α, γ_1 *and* γ_2 *be defined in* (34) *and* (35), *and suppose that* $\mathbb{E}_{\mathbf{0},\mathbf{I}_k}[\rho''(\|\mathbf{z}\|)] > 0$. *If* $\mathbf{X}$ *has full rank with probability one, then* $\sqrt{n}(\boldsymbol{\beta}_n - \boldsymbol{\beta}(P))$ *is asymptotically normal with mean zero and covariance matrix*

$$\frac{\mathbb{E}_{\mathbf{0},\mathbf{I}_k}[\rho'(\|\mathbf{z}\|)^2]}{k\alpha^2}\left(\mathbb{E}[\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}]\right)^{-1}.$$

If $\gamma_1 > 0$ *and the* $k^2 \times l$ *matrix* $\mathbf{L}$, *as defined in* (29), *has full rank then* $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}(P))$ *is asymptotically normal with mean zero and covariance matrix*

$$2\sigma_1\left(\mathbf{L}^T(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\mathbf{L}\right)^{-1} + \sigma_2\boldsymbol{\theta}(P)\boldsymbol{\theta}(P)^T,$$

where

$$\sigma_1 = \frac{k(k+2)\mathbb{E}_{\mathbf{0},\mathbf{I}_k}[\rho'(\|\mathbf{z}\|)^2\|\mathbf{z}\|^2]}{(\mathbb{E}_{\mathbf{0},\mathbf{I}_k}[\rho''(\|\mathbf{z}\|)\|\mathbf{z}\|^2 + (k+1)\rho'(\|\mathbf{z}\|)\|\mathbf{z}\|])^2},$$

$$\sigma_2 = -\frac{2}{k}\sigma_1 + \frac{4\mathbb{E}_{\mathbf{0},\mathbf{I}_k}[(\rho(\|\mathbf{z}\|) - b_0)^2]}{(\mathbb{E}_{\mathbf{0},\mathbf{I}_k}[\rho'(\|\mathbf{z}\|)\|\mathbf{z}\|])^2}.$$

Due to the linearity of $\mathbf{V}$, we can immediately establish asymptotic normality of the covariance estimator $\mathbf{C}_n = \mathbf{V}(\boldsymbol{\theta}_n)$. From Corollary 9.2, it follows that

$$\sqrt{n}\left(\text{vec}(\mathbf{C}_n) - \text{vec}(\boldsymbol{\Sigma})\right) = \sqrt{n}\left(\mathbf{L}\boldsymbol{\theta}_n - \mathbf{L}\boldsymbol{\theta}(P)\right) = \mathbf{L}\sqrt{n}\left(\boldsymbol{\theta}_n - \boldsymbol{\theta}(P)\right).$$

It follows that the limiting covariance of $\sqrt{n}(\text{vec}(\mathbf{V}(\boldsymbol{\theta}_n)) - \text{vec}(\boldsymbol{\Sigma}))$ is given by

$$2\sigma_1\mathbf{L}\left(\mathbf{L}^T(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\mathbf{L}\right)^{-1}\mathbf{L}^T + \sigma_2\,\text{vec}(\boldsymbol{\Sigma})\,\text{vec}(\boldsymbol{\Sigma})^T.$$

Corollary 9.2 is a direct consequence of Theorem 9.1. Its proof, in particular the derivations of the expressions for the limiting covariances, can be found in the Supplementary Material [25]. Note that the constants $\mathbb{E}_{\mathbf{0},\mathbf{I}_k}[\rho'(\|\mathbf{z}\|)^2]/(k\alpha^2)$, σ_1 and σ_2, are the same as the ones found in [20] for the location and covariance S-estimators, respectively. In fact, Corollary 9.2 is an extension of Corollary 5.1 in [20] for S-estimators in the multivariate location-scale model of Example 4.

Asymptotic normality of S-estimators in the multivariate regression model of Example 2 follows from Corollary 9.2. These estimators have been considered in [39], but asymptotic normality has not been established. Under the assumption that the heuristic (38) holds, asymptotic relative efficiencies are computed on the basis of this heuristic. Indeed, now that Corollary 9.2 has been established, one may check that (38) holds.

Finally, note that the limiting covariances of $\sqrt{n}(\boldsymbol{\beta}_n - \boldsymbol{\beta}(P))$ and $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}(P))$ in Corollary 9.2 differ from the ones found in [6] for the linear mixed effects model (3) with $\mathbf{X}_i = \mathbf{X}$, for $i = 1, \dots, n$. The results in [6] are obtained by reparameterizing $\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu}$ and interpreting the model as a multivariate location-scale model. Then building on the results in [20] for S-estimators of multivariate location-scale, the limiting covariances in [6] are found by application of the delta method. However, in view of Remark 1 this does not seem to be a correct approach.

REMARK 5. Although our expressions for the limiting covariances in Corollary 9.2 differ from the ones found in Proposition 1 in [6], somewhat surprisingly, they yield the same matrices for the example discussed in Section 5.1 in [6]. However, this is a consequence of the specific structure of the design matrices $\mathbf{X}$ and $\mathbf{Z}$ in this example. One can easily find other design matrices for which the limiting covariances in Corollary 9.2 yield different matrices as the ones found in [6]. Moreover, the corresponding confidence regions based on the expressions in Corollary 9.2 can be substantially smaller than the ones based on the expressions found in [6]. See the simulation in Section 10.

The constant $\mathbb{E}_{\mathbf{0},\mathbf{I}_k}[\rho'(\|\mathbf{z}\|)^2]/(k\alpha^2)$ represents the asymptotic efficiency of the S-estimator for $\boldsymbol{\beta}$ relative to the least squares estimator. As a consequence, the cut-off constant c_0 of the ρ-function can be tuned in such a way that the asymptotic efficiency relative to the least squares estimator is high at the normal distribution. However, the constant c_0 also determines the breakdown point of the corresponding S-estimator. Indeed, since a natural choice in (11) is $b_0 = \mathbb{E}_{\mathbf{0},\mathbf{I}_k}[\rho(\|\mathbf{z}\|)]$, the lower bound on the asymptotic breakdown point $b_0/\rho(\infty)$, obtained in Theorem 6.1, is determined by c_0. Unfortunately, this forces a trade-off between efficiency and breakdown point. Typically, large values of c_0 correspond to high efficiency and low breakdown point, and vice versa for moderate values of c_0. Details of varying breakdown points and corresponding efficiencies can be found in the tables given in [20, 34] and [39].

A possible remedy is to use a high breakdown S-estimator followed by a second step that retains the breakdown point and improves the efficiency. Well-known popular examples are MM-estimators introduced by Yohai [40] for multiple linear regression, and extensions thereof to multivariate location and scatter in [22, 36], to multivariate linear regression in [18], and to linear mixed effects models in [6]. A general unified approach that covers these estimators can be found in [24]. Other attempts to combine high breakdown point and high efficieny are τ-estimators introduced by Yohai and Zamar [41] for multiple regression and extended to multivariate location and scatter in [21], or reweighted estimators [23] and CM-estimators [17] for multivariate location and scatter.

10. Simulation and data example. We compare the asymptotic results of the S-estimators with their finite sample behavior by means of a simulation. Moreover, we investigate the differences between the expressions found in Corollary 9.2 and the ones in Copt and Victoria-Feser [6]. To this end, we will study the behavior of the estimators for samples generated from a model that is close to the one in [6]:

$$\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta} + \gamma_i \mathbf{Z} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{39}$$

a linear mixed effects model with $\mathbf{y}_i$ in dimension $k = 4$ and all subjects with the same design matrix $\mathbf{X}$ for the fixed effects $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$. Following the setup in [6], the matrix $\mathbf{X}$ is built as follows. The first column of $\mathbf{X}$ is taken to be a vector $\mathbf{1}$ consisting of ones of length four. The four x-values in the second column are generated from a standard normal, and then $\mathbf{X}$ is rescaled to a new matrix $\mathbf{X} = [\mathbf{1} \quad \mathbf{x}]$, such that $\mathbf{X}^T\mathbf{X} = 4\mathbf{I}_2$. For our simulation, we used

$$\mathbf{X} = \begin{pmatrix} 1 & -0.9504967 \\ 1 & -0.5428346 \\ 1 & 1.6650521 \\ 1 & -0.1717207 \end{pmatrix}.$$

The random effects γ_i are independent $N(0, \sigma_\gamma^2)$ distributed random variables, which are independent from the measurement error $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{R})$. This leads to a structured covariance $\boldsymbol{\Sigma} = \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}^T + \sigma_\epsilon^2 \mathbf{R}$, with covariance parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$, where $\theta_1 = \sigma_\gamma^2$ and $\theta_2 = \sigma_\epsilon^2$. Following the setup in [6], we set $\beta_1 = \beta_2 = 1$ and $\theta_1 = \theta_2 = 1$.

In [6], the authors took $\mathbf{Z} = (1, 1, 1, 1)^T$ and $\mathbf{R} = \mathbf{I}_4$. With these choices, the expression

$$\mathrm{Var}_{\mathrm{CVF}}(\boldsymbol{\beta}_n) = \frac{\mathbb{E}_{\mathbf{0},\mathbf{I}_k}[\rho'(\|\mathbf{z}\|)^2]}{k\alpha^2} (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \tag{40}$$

found in [6] for the limiting covariance matrix of $\sqrt{n}(\boldsymbol{\beta}_n - \boldsymbol{\beta})$ (see (14) in [6]) is equal to our expression

$$\mathrm{Var}_{\mathrm{LGRG}}(\boldsymbol{\beta}_n) = \frac{\mathbb{E}_{\mathbf{0},\mathbf{I}_k}[\rho'(\|\mathbf{z}\|)^2]}{k\alpha^2}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}, \tag{41}$$

found in Corollary 9.2, and similarly for the limiting covariance matrix of $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta})$. However, this is just the consequence of the extreme simple choices for $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{R}$. Already, if we keep $\mathbf{X}$ as it is, and only take a slight variation of either $\mathbf{Z}$ or $\mathbf{R}$, one finds severe differences between (40) and (41), and similarly for the expression of the limiting covariance matrix of $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta})$.

We considered the following two alternatives:

1. take $\mathbf{Z} = (1, 2, 3, 4)^T$ and leave $\mathbf{X}$ and $\mathbf{R} = \mathbf{I}_4$ as they are;
2. take $\mathbf{R} = (1, 4, 9, 16)^T$ and leave $\mathbf{X}$ and $\mathbf{Z} = (1, 1, 1, 1)^T$ as they are.

We generated 10,000 samples of size $n = 100$ according to model (39) and computed the value of S-estimators $\boldsymbol{\beta}_n$ and $\boldsymbol{\theta}_n$ by means of Tukey's biweight

$$\rho_{\mathrm{B}}(s; c) = \begin{cases} s^2/2 - s^4/(2c^2) + s^6/(6c^4), & |s| \le c, \\ c^2/6, & |s| > c, \end{cases} \tag{42}$$

and $b_0 = \mathbb{E}_{\mathbf{0},\mathbf{I}_k}[\rho_B(\|\mathbf{z}\|; c_0)]$, with the cut-off value c_0 chosen such that $b_0/a_0 = 0.5$. According to Theorem 6.1, this corresponds to the (asymptotic) breakdown point 50%. Details of the algorithm used to compute the S-estimators and a brief discussion on its behavior can be found in [25].

Figure 1 displays the limiting marginal and joined distributions of $\sqrt{n}(\boldsymbol{\beta}_n - \boldsymbol{\beta})$ in the first row, where we generated the samples with alternative 1. The histograms and scatterplot correspond to the 10,000 different values of $\sqrt{n}(\boldsymbol{\beta}_n - \boldsymbol{\beta})$. The dashed curves correspond to

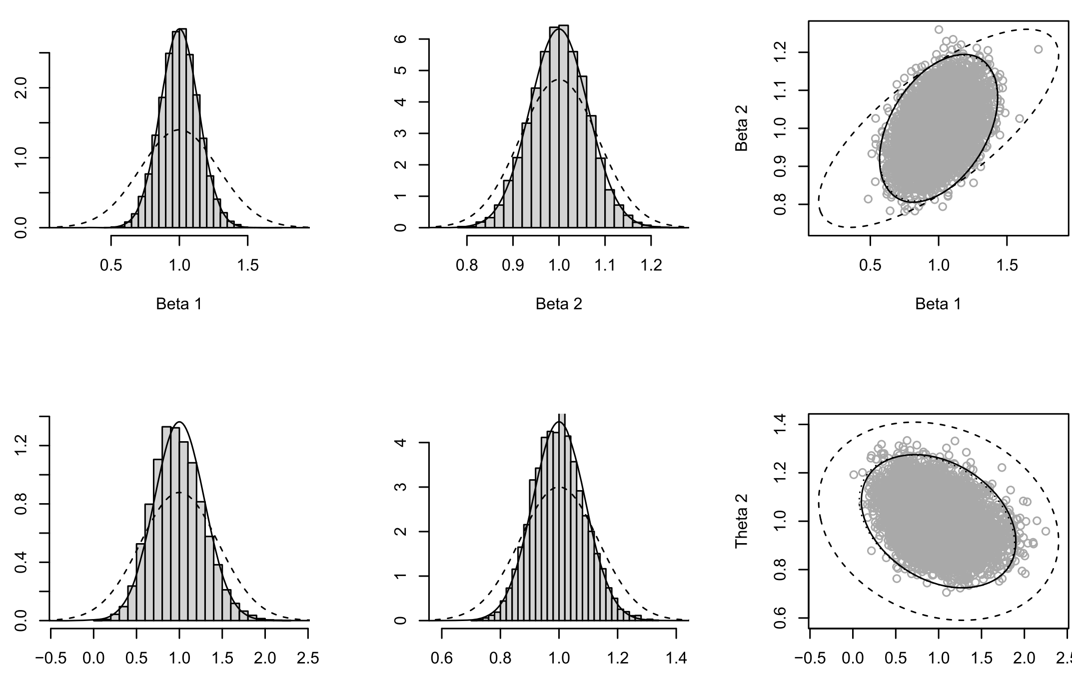

FIG. 1. *Empirical marginal and joined distributions together with limiting marginal and joined distributions of* $\sqrt{n}(\boldsymbol{\beta}_n - \boldsymbol{\beta})$ (*first row*) *and* $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta})$ (*second row*).

the densities and 95% contour lines of the theoretical limiting marginal and joined normal distributions using the covariance matrix in (40). The solid curves correspond to the marginal and joined normal distributions using the covariance matrix in (41). The empirical contour lines based on the sample mean and sample covariance of the 10,000 estimates are plotted in dotted lines, but they are almost indistinguishable from the solid contour lines. We find

$$\mathrm{Var}_{\mathrm{CVF}}(\boldsymbol{\beta}_n) = \begin{pmatrix} 8.13 & 1.78 \\ 1.78 & 0.72 \end{pmatrix} \quad \text{and} \quad \mathrm{Var}_{\mathrm{LGRG}}(\boldsymbol{\beta}_n) = \begin{pmatrix} 1.97 & 0.38 \\ 0.38 & 0.40 \end{pmatrix}.$$

Clearly, the histograms of the repeated estimates for β_1 and β_2 match the graphs of the (marginal) normal densities with the variances given by $\mathrm{Var}_{\mathrm{LGRG}}(\boldsymbol{\beta}_n)$, and the scatterplot matches with the 95% contour line corresponding to $\mathrm{Var}_{\mathrm{LGRG}}(\boldsymbol{\beta}_n)$. Note that the differences with $\mathrm{Var}_{\mathrm{CVF}}(\boldsymbol{\beta}_n)$ are quite severe. For example, this yields that the length of the confidence interval for β_1 based on $\mathrm{Var}_{\mathrm{CVF}}(\boldsymbol{\beta}_n)$ will be two times larger than the one based on $\mathrm{Var}_{\mathrm{LGRG}}(\boldsymbol{\beta}_n)$.

The second row in Figure 1 displays the limiting distributions of $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta})$, where we generated the samples with alternative 2. In [6], the limiting covariance matrix was given by (see (15) in [6]) $(\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T\mathbf{V}_{\boldsymbol{\Sigma}}\mathbf{L}(\mathbf{L}^T\mathbf{L})^{-1}$, where $\mathbf{V}_{\boldsymbol{\Sigma}} = \sigma_1(\mathbf{I}_{k^2} + \mathbf{K}_{k,k})(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) + \sigma_2 \operatorname{vec}(\boldsymbol{\Sigma}) \operatorname{vec}(\boldsymbol{\Sigma})^T$; see Corollary 5.1 in [20]. Because $(\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T(\mathbf{I}_{k^2} + \mathbf{K}_{k,k}) = 2(\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T$ and $(\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T \operatorname{vec}(\boldsymbol{\Sigma}) = \boldsymbol{\theta}(P)$, the expression given in [6] becomes

$$\mathrm{Var}_{\mathrm{CVF}}(\boldsymbol{\theta}_n) = 2\sigma_1(\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})\mathbf{L}(\mathbf{L}^T\mathbf{L})^{-1} + \sigma_2\boldsymbol{\theta}(P)\boldsymbol{\theta}(P)^T.$$

This differs from our Corollary 9.2, which gives

$$\mathrm{Var}_{\mathrm{LGRG}}(\boldsymbol{\theta}_n) = 2\sigma_1(\mathbf{L}^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{L})^{-1} + \sigma_2\boldsymbol{\theta}(P)\boldsymbol{\theta}(P)^T.$$

For the choices of $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{R}$ in [6], both expressions are equal. However, for the alternative choice for $\mathbf{R}$ made in alternative 2, one finds

$$\mathrm{Var}_{\mathrm{CVF}}(\boldsymbol{\theta}_n) = \begin{pmatrix} 20.63 & -1.22 \\ -1.22 & 1.77 \end{pmatrix} \quad \text{and} \quad \mathrm{Var}_{\mathrm{LGRG}}(\boldsymbol{\theta}_n) = \begin{pmatrix} 8.57 & -0.82 \\ -0.82 & 0.80 \end{pmatrix}.$$

Again the differences are quite large. For example, as a consequence the length of the confidence interval for θ_1 based on $\mathrm{Var}_{\mathrm{CVF}}(\boldsymbol{\theta}_n)$ will be 1.5 times larger than the one based on $\mathrm{Var}_{\mathrm{LGRG}}(\boldsymbol{\theta}_n)$.

Finally, we illustrate the performance of S-estimators by an application to data from a trial on the treatment of lead-exposed children. This data set is discussed in [12] and consists of four repeated measurements of blood lead levels obtained at baseline (or week 0), week 1, week 4 and week 6 on 100 children who were randomly assigned to chelation treatment with succimer (a chelation agent) or placebo. On the basis of a graphical display of the mean response over time, it is suggested in [12] that a quadratic trend over time seems suitable. We fitted the following model:

$$y_{ij} = \beta_0 + \beta_1\delta_i + (\beta_3 + \beta_4\delta_i)t_j + (\beta_5 + \beta_6\delta_i)t_j^2 + \gamma_{1i} + \gamma_{2i}t_j + \gamma_{3i}t_j^2 + \epsilon_{ij},$$

for $i = 1, \dots, 100$ and $j = 1, \dots, 4$, where $(t_1, \dots, t_4) = (0, 1, 4, 6)$ refer to the different weeks, y_{ij} is the blood lead level (mcg/dL) of subject i obtained at time t_j and $\delta_i = 0$ if the ith subject is in the placebo group and $\delta_i = 1$, otherwise. The random effects $\boldsymbol{\gamma}_i = (\gamma_{1i}, \gamma_{2i}, \gamma_{3i})$, $i = 1, \dots, 100$ are assumed to be independent mean zero normal random vectors with a diagonal covariance matrix consisting of variances $\sigma_{\gamma_1}^2$, $\sigma_{\gamma_2}^2$ and $\sigma_{\gamma_3}^2$, respectively. The measurement errors $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{i4})$, $i = 1, \dots, 100$ are assumed to be independent mean zero random vectors with covariance matrix $\sigma_\epsilon^2\mathbf{I}_4$, also being independent of the random effects. In this way, we are fitting a balanced linear mixed effects model with unknown parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_6)$ and $\boldsymbol{\theta} = (\sigma_{\gamma_1}^2, \sigma_{\gamma_2}^2, \sigma_{\gamma_3}^2, \sigma_\epsilon^2)$, and a linear covariance structure.

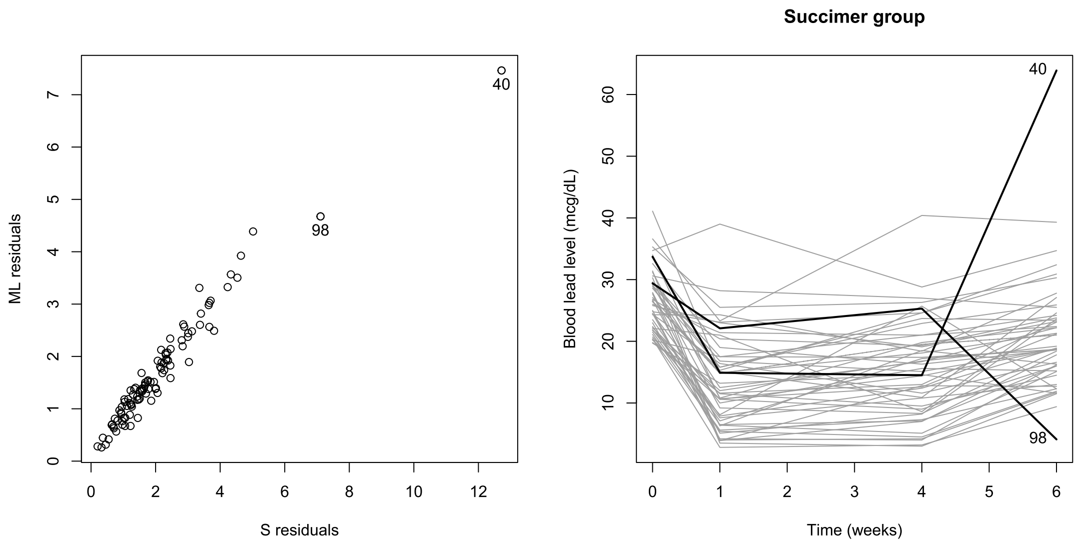

FIG. 2. *Left picture*: *Standardized residuals for the S-estimates* (*horizontal axis*) *and the ML estimates* (*vertical axis*). *Right picture*: *observations for the subjects in the treatment group*.

We estimated $(\boldsymbol{\beta}, \boldsymbol{\theta})$ by means of maximum likelihood and by means of the S-estimator corresponding to Tukey's biweight defined in (42). The tuning constant was chosen to be $c = 4.097$, which corresponds to asymptotic breakdown point 0.5. For each estimate $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$, we determined the estimate $\mathbf{V}(\widehat{\boldsymbol{\theta}})$ for the structured covariance and the standardized residuals for each subject

$$\mathrm{RES}_i = \sqrt{(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})^T \mathbf{V}(\widehat{\boldsymbol{\theta}})^{-1}(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})}.$$

The residuals for both estimation procedures are visible in the left picture of Figure 2, with the residuals determined from the S-estimate on the horizonal axis and the ones determined from the ML estimate on the vertical axis. Both estimates identify subject 40 as an outlier, but only the robust S-estimate also clearly identifies observation 98 as outlier. The extreme large observation in week 6 seems to be the reason that observation 40 is identified as an outlier by both methods; see the right picture in Figure 2. Observation 98 also seems to deviate from the overall quadratic trend by having a suspicious low observation in week 6. The corresponding S-residual clearly sticks out from the other S-residuals, whereas this is much less so for the corresponding ML residual.

Acknowledgments. We thank both referees for pointing out relevant references and we thank the second referee whose comments lead to an improvement of Theorem 5.3.

Funding. The third author was partially supported by by the French *Agence Nationale de la Recherche* through the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EURE-0010.

SUPPLEMENTARY MATERIAL

Supplement to "S-estimation in linear models with structured covariance matrices" (DOI: 10.1214/23-AOS2334SUPP; .pdf). This supplement contains all proofs.

REFERENCES

[1] AGOSTINELLI, C. and YOHAI, V. J. (2016). Composite robust estimators for linear mixed models. *J. Amer. Statist. Assoc.* **111** 1764–1774. MR3601734 https://doi.org/10.1080/01621459.2015.1115358

[2] BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York. MR0233396

[3] BILODEAU, M. and DUCHESNE, P. (2000). Robust estimation of the SUR model. *Canad. J. Statist.* **28** 277–288. MR1921613 https://doi.org/10.2307/3315978

[4] CHERVONEVA, I. and VISHNYAKOV, M. (2014). Generalized S-estimators for linear mixed effects models. *Statist. Sinica* **24** 1257–1276. MR3241287

[5] COPT, S. and HERITIER, S. (2007). Robust alternatives to the F-test in mixed linear models based on MM-estimates. *Biometrics* **63** 1045–1052, 1310. MR2414581 https://doi.org/10.1111/j.1541-0420.2007.00804.x

[6] COPT, S. and VICTORIA-FESER, M.-P. (2006). High-breakdown inference for mixed linear models. *J. Amer. Statist. Assoc.* **101** 292–300. MR2268046 https://doi.org/10.1198/016214505000000772

[7] DAVIES, L. (1992). The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *Ann. Statist.* **20** 1828–1843. MR1193314 https://doi.org/10.1214/aos/1176348891

[8] DAVIES, P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *Ann. Statist.* **15** 1269–1292. MR0902258 https://doi.org/10.1214/aos/1176350505

[9] DEMIDENKO, E. (2013). *Mixed Models: Theory and Applications with R*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3235905

[10] DONOHO, D. and HUBER, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann. Wadsworth Statist./Probab. Ser.* 157–184. Wadsworth, Belmont, CA. MR0689745

[11] FASANO, M. V., MARONNA, R. A., SUED, M. and YOHAI, V. J. (2012). Continuity and differentiability of regression M functionals. *Bernoulli* **18** 1284–1309. MR2995796 https://doi.org/10.3150/11-BEJ368

[12] FITZMAURICE, G. M., LAIRD, N. M. and WARE, J. H. (2011). *Applied Longitudinal Analysis*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2830137

[13] HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393. MR0362657

[14] HARTLEY, H. O. and RAO, J. N. K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika* **54** 93–108. MR0216684 https://doi.org/10.1093/biomet/54.1-2.93

[15] HERITIER, S., CANTONI, E., COPT, S. and VICTORIA-FESER, M.-P. (2009). *Robust Methods in Biostatistics. Wiley Series in Probability and Statistics*. Wiley, Chichester. MR2604994 https://doi.org/10.1002/9780470740538

[16] JENNRICH, R. I. and SCHLUCHTER, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* **42** 805–820. MR0872961 https://doi.org/10.2307/2530695

[17] KENT, J. T. and TYLER, D. E. (1996). Constrained M-estimation for multivariate location and scatter. *Ann. Statist.* **24** 1346–1370. MR1401854 https://doi.org/10.1214/aos/1032526973

[18] KUDRASZOW, N. L. and MARONNA, R. A. (2011). Estimates of MM type for the multivariate linear model. *J. Multivariate Anal.* **102** 1280–1292. MR2811617 https://doi.org/10.1016/j.jmva.2011.04.011

[19] LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** 963–974.

[20] LOPUHAÄ, H. P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *Ann. Statist.* **17** 1662–1683. MR1026304 https://doi.org/10.1214/aos/1176347386

[21] LOPUHAÄ, H. P. (1991). Multivariate τ-estimators for location and scatter. *Canad. J. Statist.* **19** 307–321. MR1144148 https://doi.org/10.2307/3315396

[22] LOPUHAÄ, H. P. (1992). Highly efficient estimators of multivariate location with high breakdown point. *Ann. Statist.* **20** 398–413. MR1150351 https://doi.org/10.1214/aos/1176348529

[23] LOPUHAÄ, H. P. (1999). Asymptotics of reweighted estimators of multivariate location and scatter. *Ann. Statist.* **27** 1638–1665. MR1742503 https://doi.org/10.1214/aos/1017939145

[24] LOPUHAÄ, H. P. (2023). Highly efficient estimators with high breakdown point for linear models with structured covariance matrices. *Econom. Stat.* **242**. https://doi.org/10.1016/j.ecosta.2023.03.003

[25] LOPUHAÄ, H. P., GARES, V. and RUIZ-GAZEN, A. (2023). Supplement to "S-estimation in linear models with structured covariance matrices." https://doi.org/10.1214/23-AOS2334SUPP

[26] LOPUHAÄ, H. P. and ROUSSEEUW, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Statist.* **19** 229–248. MR1091847 https://doi.org/10.1214/aos/1176347978

[27] MAGNUS, J. R. and NEUDECKER, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, Chichester. MR0940471

[28] PEREMANS, K. and VAN AELST, S. (2018). Robust inference for seemingly unrelated regression models. *J. Multivariate Anal.* **167** 212–224. MR3830643 https://doi.org/10.1016/j.jmva.2018.05.002

[29] 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. *J. Comput. Graph. Statist.* **10** 249–276. MR1939700 https://doi.org/10.1198/10618600152628059

[30] POLLARD, D. (1984). *Convergence of Stochastic Processes. Springer Series in Statistics*. Springer, New York. MR0762984 https://doi.org/10.1007/978-1-4612-5254-2

[31] RANGA RAO, R. (1962). Relations between weak and uniform convergence of measures with applications. *Ann. Math. Stat.* **33** 659–680. MR0137809 https://doi.org/10.1214/aoms/1177704588

[32] RAO, C. R. (1972). Estimation of variance and covariance components in linear models. *J. Amer. Statist. Assoc.* **67** 112–115. MR0314185

[33] ROUSSEEUW, P. (1985). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications, Vol. B* (*Bad Tatzmannsdorf*, 1983) 283–297. Reidel, Dordrecht. MR0851060

[34] ROUSSEEUW, P. and YOHAI, V. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis* (*Heidelberg*, 1983). *Lect. Notes Stat.* **26** 256–272. Springer, New York. MR0786313 https://doi.org/10.1007/978-1-4615-7821-5_15

[35] ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880. MR0770281

[36] SALIBIÁN-BARRERA, M., VAN AELST, S. and WILLEMS, G. (2006). Principal components analysis based on multivariate MM estimators with fast and robust bootstrap. *J. Amer. Statist. Assoc.* **101** 1198–1211. MR2328307 https://doi.org/10.1198/016214506000000096

[37] TATSUOKA, K. S. and TYLER, D. E. (2000). On the uniqueness of S-functionals and M-functionals under nonelliptical distributions. *Ann. Statist.* **28** 1219–1243. MR1811326 https://doi.org/10.1214/aos/1015956714

[38] VAN AELST, S. and WILLEMS, G. (2004). Multivariate regression S-estimators for robust estimation and inference, 2004, Preprint received by personal communication.

[39] VAN AELST, S. and WILLEMS, G. (2005). Multivariate regression S-estimators for robust estimation and inference. *Statist. Sinica* **15** 981–1001. MR2234409

[40] YOHAI, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.* **15** 642–656. MR0888431 https://doi.org/10.1214/aos/1176350366

[41] YOHAI, V. J. and ZAMAR, R. H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *J. Amer. Statist. Assoc.* **83** 406–413. MR0971366

[42] ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.* **57** 348–368. MR0139235