WORKING PAPERS

N° 1343

June 2022

"S-estimation in Linear Models with Structured Covariance Matrices"

Henrik Paul Lopuhaä, Valérie Gares and Anne Ruiz-Gazen

Toulouse School of Economics

S-estimation in Linear Models with Structured Covariance Matrices *

Hendrik Paul Lopuhaä[1], Valerie Gares[2], and Anne Ruiz-Gazen[3]

[1] *Delft University of Technology*
[2] *Institut National des Sciences Appliqués de Rennes*
[3] *Toulouse School of Economics*

June 20, 2022

Abstract

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, $i = 1, \ldots, n$, is observed at k_i occasions, and the vector of responses $\mathbf{y}_i$ is assumed to arise from the model

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

where $\mathbf{X}_i$ is the design matrix for the ith subject and $\mathbf{u}_i$ is a vector whose covariance matrix can be used to model the correlation between the responses. 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 e.g., [14] or [10], for different possible covariance structures.

Maximum likelihood estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ has been studied, e.g., in [12, 25, 15], see also [10, 7]. To be resistant against outliers, robust methods have been investigated for linear mixed effects models, e.g., in [22, 5, 4, 13, 1, 3]. This mostly concerns S-estimators, originally introduced in the multiple regression context by Rousseeuw and Yohai [26] and extended to multivariate location and scatter in [6, 16], to multivariate regression in [29], and to linear mixed effects models in [5, 13, 3].

*This work has been partly supported by the French *Agence Nationale de la Recherche* through the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EURE-0010.

S-estimators are well known smooth versions of the minimum volume ellipsoid estimator [27] 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.

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 [30] and for linear mixed effects models in [5, 3]. 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 [5, 3]. For multivariate regression [30], 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 [30], or is only available for observations from a normal distribution [26, 3].

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 [26, 16, 30]. 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 [5]. 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 and some technical lemmas are put in an Appendix at the end of the paper. Other long and technical proofs are available as supplemental material [19].

2 Balanced models with structured covariances

We consider independent observations $(\mathbf{y}_1, \mathbf{X}_1), \ldots, (\mathbf{y}_n, \mathbf{X}_n)$, for which we assume the following model

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{u}_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\mathbf{y}_i \in \mathbb{R}^k$ contains repeated measurements for the i-th 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*

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n. \tag{2.2}$$

This model arises from $\mathbf{u}_i = \mathbf{Z}\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i$, *for* $i = 1, \ldots, n$, *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* $\boldsymbol{\theta} = (\text{vech}(\mathbf{G})^T, \text{vech}(\mathbf{R})^T)^T$, *where*

$$\text{vech}(\mathbf{A}) = (a_{11}, \ldots, a_{k1}, a_{22}, \ldots, a_{kk}) \tag{2.3}$$

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* $\boldsymbol{\gamma}_i = (\gamma_{i1} \ldots, \gamma_{ir})^T$, *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* $j = 1, \ldots, r$. *This leads to*

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \sum_{j=1}^{r} \mathbf{Z}_j\gamma_{ij} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{2.4}$$

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* $\boldsymbol{\theta} = (\sigma_0^2, \sigma_1^2, \ldots, \sigma_r^2)$.

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

$$\mathbf{y}_i = \mathbf{B}^T\mathbf{x}_i + \mathbf{u}_i, \qquad i = 1, \ldots, n, \tag{2.5}$$

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* $i = 1, \ldots, n$, *are independent mean zero random variables with covariance matrix* $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{C} \in PDS(k)$. *In this case, the vector of unknown covariance parameters is given by*

$$\boldsymbol{\theta} = \text{vech}(\mathbf{C}) = (c_{11}, \ldots, c_{1k}, c_{22}, \ldots, c_{kk})^T \in \mathbb{R}^{\frac{1}{2}k(k+1)}. \tag{2.6}$$

The model can be obtained as a special case of (2.1), *by taking* $\mathbf{X}_i = \mathbf{x}_i^T \otimes \mathbf{I}_k$ *and* $\boldsymbol{\beta} = \text{vec}(\mathbf{B}^T)$, *where* $\otimes$ *denotes the Kronecker product and* $\text{vec}(\cdot)$ *is the* k^2*-vector that stacks the columns of a matrix. Clearly, the linear multiple regression model is a special case with* $k = 1$.

Example 3. *Model* (2.1) *also includes examples, for which* $\mathbf{u}_1, \ldots, \mathbf{u}_n$ *are generated from a time series. One example, is the case where* $\mathbf{u}_i$ *has a covariance matrix with elements*

$$v_{st} = \sigma^2\rho^{|s-t|}, \quad s, t = 1, \ldots, k. \tag{2.7}$$

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*

$$v_{st} = \theta_{|s-t|+1}, \quad s, t = 1, \ldots, k, \tag{2.8}$$

in which case $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T \in \mathbb{R}^k$, *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* (2.1), *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* (2.6).

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

$$\mathbf{V}(\boldsymbol{\theta}_1) = \mathbf{V}(\boldsymbol{\theta}_2) \quad \Rightarrow \quad \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2. \tag{2.9}$$

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 $\text{vech}(\mathbf{G})$ and $\text{vech}(\mathbf{R})$. For linear mixed effects models in (2.4), identifiability of $\boldsymbol{\theta} = (\sigma_0^2, \sigma_1^2, \ldots, \sigma_r^2)$ holds for particular choices of the design matrices $\mathbf{Z}_1, \ldots, \mathbf{Z}_r$.

3 Definitions

We start by representing our observations as points in $\mathbb{R}^k \times \mathbb{R}^{kq}$ in the following way. For $r = 1, \ldots, k$, let $\mathbf{x}_r^T$ denote the r-th 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 $\mathbf{s}^T = (\mathbf{y}^T, \mathbf{x}_1^T, \ldots, \mathbf{x}_k^T)$. In this way our observations can be represented as $\mathbf{s}_1, \ldots, \mathbf{s}_n$, 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 non-decreasing 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

$$\begin{aligned} &\min_{\boldsymbol{\beta}, \boldsymbol{\theta}} \det(\mathbf{V}(\boldsymbol{\theta})) \\ &\text{subject to} \\ &\frac{1}{n} \sum_{i=1}^{n} \rho\left(\sqrt{(\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})^T \mathbf{V}(\boldsymbol{\theta})^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})} \right) \leq b_0, \end{aligned} \tag{3.1}$$

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).

The S-estimator defined by (3.1) for the setup in (2.1) includes several specific cases that have been considered in the literature. The original regression S-estimator introduced by Rousseeuw and Yohai [26] 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 [6] and Lopuhaä [16] 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 [30]. Copt and Victoria-Feser [5] and Chervoneva and Vishnyakov [3] consider S-estimators for the parameters in the linear mixed effects model (2.4).

The constant $0 < b_0 < a_0$ in (3.1) can be chosen in agreement with an assumed underlying distribution. For the multivariate regression model in [30], it is assumed that $\mathbf{y}_i \mid \mathbf{X}_i$ has an elliptically contoured density of the form

$$f_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\mathbf{y}) = \det(\boldsymbol{\Sigma})^{-1/2} h\left((\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right), \tag{3.2}$$

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 [5], it is assumed that $\mathbf{y}_i \mid \mathbf{X}_i$ has a multivariate normal distribution, which is a special case of (3.2) with $h(t) = (2\pi)^{-k/2} \exp(-t/2)$. When the underlying distribution corresponds to a density of the form (3.2), then a natural choice is $b_0 = \mathbb{E}_{\mathbf{0}, \mathbf{I}} \rho(\|\mathbf{z}\|)$, where $\mathbf{z}$ has density (3.2) 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 4), as well as its limiting variance (see Corollary 6). 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 estimator in location-scale models (see [27]) and to the least median of squares estimator in linear regression models (see [28]). 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

$$\mathcal{C}(\boldsymbol{\beta}, \boldsymbol{\theta}, c_0) = \left\{ (\mathbf{y}, \mathbf{X}) \in \mathbb{R}^k \times \mathbb{R}^{kq} : (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}(\boldsymbol{\theta})^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \leq c_0^2 \right\} \tag{3.3}$$

that contains at least $n - nb_0$ points.

Remark 3.1. *Clearly, the definition of the S-estimator in* (3.1) *has great similarities with the S-estimator for multivariate location and covariance (see [6] and [16]), defined as the solution* $(\mathbf{t}_n, \mathbf{C}_n)$ *to the minimization problem*

$$\begin{aligned} &\min_{\mathbf{t},\mathbf{C}} \det(\mathbf{C}) \\ &\textit{subject to} \\ &\frac{1}{n}\sum_{i=1}^{n} \rho\left(\sqrt{(\mathbf{y}_i - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{y}_i - \mathbf{t})}\right) \leq b_0, \end{aligned} \tag{3.4}$$

where the minimum is taken over all $\mathbf{t} \in \mathbb{R}^k$ *and* $\mathbf{C} \in PDS(k)$*. 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, 4]. However, there is a subtle, but important difference between minimization problems* (3.4) *and* (3.1)*. The important difference is that in* (3.4) *we minimize over* all *positive definite symmetric* $k \times k$ *matrices* $\mathbf{C}$*, whereas in* (3.1)*, 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:*

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

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 [5], see Corollary 6.*

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:

$$\begin{aligned} &\min_{\boldsymbol{\beta},\boldsymbol{\theta}} \det(\mathbf{V}(\boldsymbol{\theta})) \\ &\text{subject to} \\ &\int \rho\left(\sqrt{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}(\boldsymbol{\theta})^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}\right) \mathrm{d}P(\mathbf{y}, \mathbf{X}) \leq b_0, \end{aligned} \tag{3.5}$$

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 (3.1) will follow from general results on the existence and the continuity of solutions to (3.5).

The definition of the S-functionals for the multivariate location-scale model given in Lopuhaä [16] and for the multivariate regression model given by Van Aelst and Willems [30] can be obtained as special cases of (3.5), by choosing $\mathbf{X}$, $\boldsymbol{\beta}$ and $\mathbf{V}(\boldsymbol{\theta})$ as in Examples 4 and 2, respectively. Copt and Victoria-Feser [5] do not pay attention to S-functionals or the influence function in the linear mixed effects model (2.4). However, S-functionals for linear mixed effects models can be also be obtained as a special case of (3.5), 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 (3.5), under particular conditions on the probability measure P. As a consequence, this will also yield the existence of

the S-estimator, defined by (3.1). 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 [5]. In view of this, define $\mathcal{X} \subset \mathbb{R}^{kq}$ as the subset with the lowest dimension $p = \dim(\mathcal{X}) \leq kq$ satisfying

$$P(\mathbf{X} \in \mathcal{X}) = 1. \tag{4.1}$$

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 of smaller than $k + kq$.

The first condition 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 $P(K_\epsilon) \geq r + \epsilon$.

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 $\delta \geq 0$, we define a strip $H(\boldsymbol{\alpha}, \ell, \delta)$ as follows:

$$H(\boldsymbol{\alpha}, \ell, \delta) = \left\{\mathbf{s} \in \mathbb{R}^k \times \mathbb{R}^{kq} : \ell - \delta/2 \leq \boldsymbol{\alpha}^T \mathbf{s} \leq \ell + \delta/2\right\}. \tag{4.2}$$

Defined in this way, a strip is the area between two parallel hyperplanes which are symmetric around the hyperplane $H(\boldsymbol{\alpha}, \ell, 0) = \left\{\mathbf{s} \in \mathbb{R}^k \times \mathbb{R}^{kq} : \boldsymbol{\alpha}^T \mathbf{s} = \ell\right\}$. 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 (4.2) has width δ. We require the following condition

(C2$_\epsilon$) The value

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

is strictly positive.

According to (4.1), 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 $\mathcal{S}_n = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subset \mathbb{R}^k \times \mathcal{X}$. 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*, i.e., 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 [16]. The reason that (C1$_\epsilon$) slightly deviates from [16], is to handle the presence of $\mathbf{X}$ in minimization problem (3.5).

Remark 4.1. *Note that condition* (C2$_\epsilon$) *is equivalent with*

$$\omega_\epsilon = \inf_{P(J) \geq \epsilon} \inf_{\|\boldsymbol{\alpha}\| = 1} \inf_{\ell \in \mathbb{R}} \sup_{\mathbf{s} \in J} |\boldsymbol{\alpha}^T \mathbf{s} - \ell| > 0, \tag{4.3}$$

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 [19].

To establish existence of the S-functional, we follow the reasoning in [16]. The idea is to argue that one can restrict oneself to a compact set for finding solutions to (3.5). When the object function in (3.5) is continuous, this immediately yields existence of a solution of (3.5). 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 $\mathcal{V} = \{\mathbf{V}(\boldsymbol{\theta}) \in \mathrm{PDS}(k) : \boldsymbol{\theta} \in \mathbb{R}^l\}$ 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 [16] and its proof can be found in [19].

Lemma 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 (3.5), 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 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 (3.5). To establish that possible solutions $(\boldsymbol{\beta}, \boldsymbol{\theta})$ of (3.5) are in a compact set, we need that the pre-image $\{\boldsymbol{\theta} \in \mathbb{R}^l : \mathbf{V}(\boldsymbol{\theta}) \in K\}$ 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 pre-image 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 pre-image 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 1. *Consider minimization problem (3.5) with ρ satisfying (R1)-(R2). Suppose that P satisfies $(C1_\epsilon)$ and $(C2_\epsilon)$, 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 (3.5).*

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

$$\kappa(\mathcal{S}_n) = \text{maximal number of points of } \mathcal{S}_n \text{ lying on the same hyperplane in } \mathbb{R}^k \times \mathcal{X}. \tag{4.4}$$

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 1. *Consider minimization problem (3.1) 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 the minimization problem (3.1).*

Copt and Victoria-Feser [5] consider S-estimators for the linear mixed effects model (2.4). Despite their Proposition 1 about the asymptotic behavior of solutions to their S-minimization problem [5, equation (7)], the actual existence of such a solution is not established. However, this now follows from our Corollary 1. 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 bi-weight ρ-function proposed in [5] 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 1 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 & Willems [30] do not explicitly prove existence of the S-estimator. Since in their case, $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{C} \in \text{PDS}(k)$ satisfies (V1)-(V3) and the conditions imposed in [30] on the ρ-function satisfy (R1)-(R2), the existence of their S-estimator now also follows from Corollary 1, 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, i.e., $P(\partial C) = 0$, where ∂C denotes the boundary of C.

Corollary 2. *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* (3.5) *with probability measure P_t has at least one solution $\boldsymbol{\xi}(P_t)$.*

Condition (C3) is needed to apply (A.2). 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, i.e., $\boldsymbol{\xi}_n = \boldsymbol{\xi}(\mathbb{P}_n) \to \boldsymbol{\xi}(P)$, 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

$$d^2(\mathbf{s}, \boldsymbol{\xi}) = d^2(\mathbf{s}, \boldsymbol{\beta}, \boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}(\boldsymbol{\theta})^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{5.1}$$

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

Theorem 2. *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* (3.5) *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* (3.5) *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 2 is an extension of Theorem 3.1 in [16] on the continuity of S-functionals for multivariate location and scale. Continuity of S-functionals for multiple regression has been investigated in [9].

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 (3.1). Recall that $\boldsymbol{\xi}_n = \boldsymbol{\xi}(\mathbb{P}_n)$, so that we can use Theorem 2 to establish consistency of the S-estimator.

Corollary 3. *Let $\boldsymbol{\xi}_n$ be a solution to minimization problem* (3.1)*. 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* (3.5) *is unique, then*

$$\lim_{n\to\infty} \boldsymbol{\xi}_n = \boldsymbol{\xi}(P),$$

with probability one.

Theorem 2 and Corollary 3 require that $\boldsymbol{\xi}(P)$ is the unique solution to minimization problem (3.5). 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 (3.2). This situation is very similar to that of multivariate location-scale S-estimators, for which Davies [6, Theorem 1] shows that the corresponding S-minimization problem (3.5) has a unique solution. The next theorem is a direct consequence of that result. Its proof can be found in [19].

Theorem 3. *Suppose that $\rho : \mathbb{R} \to [0, \infty)$ satisfies (R1)-(R2) and suppose that the probability distribution P of $(\mathbf{y}, \mathbf{X})$ is such that $\mathbf{y} \mid \mathbf{X}$ has an elliptically contoured density $f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ from* (3.2)*, with $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}_0$ and $\boldsymbol{\Sigma} = \mathbf{V}(\boldsymbol{\theta}_0)$. Suppose that h in* (3.2) *is non-increasing and such that the functions $-\rho$ and h have at least one common point of decrease $d_0 > 0$, i.e.,*

$$\rho(s) < \rho(d_0) < \rho(t) \quad \textit{and} \quad h(s) > h(d_0) > h(t)$$

for all $s, t \geq 0$, such that $s < d_0 < t$. If $\mathbf{X}^T\mathbf{X}$ is non-singular with probability one, then the minimization problem

$$\begin{aligned} &\min_{\boldsymbol{\beta},\boldsymbol{\theta}} det(\mathbf{V}(\boldsymbol{\theta})) \\ &\textit{subject to} \\ &\int \rho\left(\sqrt{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}(\boldsymbol{\theta})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}\right) f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{y})\mathrm{d}\mathbf{y} \leq b_0, \end{aligned} \tag{5.2}$$

where the minimum is taken over all $\boldsymbol{\beta} \in \mathbb{R}^q$ and $\boldsymbol{\theta} \in \mathbb{R}^l$, such that $\mathbf{V}(\boldsymbol{\theta}) \in PDS(k)$, has the unique solution $(\boldsymbol{\beta}, \boldsymbol{\theta}) = (\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$ with probability one.

Minimization problem (5.2) seems to be slightly different from the one in (3.5). However, note that when P is such that $\mathbf{X}$ is equal to a single value with probability one, both minimization problems are identical. This situation was considered, e.g., in [5].

An elliptically contoured density for $\mathbf{y}_i \mid \mathbf{X}_i$ in the context of S-estimators for specific cases of the model (2.1) has been assumed in [6] for the multivariate location-scale model of Example 4, in [30] for the multivariate regression model of Example 2, and in [5] for the linear mixed effects model (2.4). More precisely, in [5] 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 h in (3.2) satisfies all the conditions of Theorem 3.

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 (3.1). 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 [8]) 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,

$$\epsilon_n^*(\boldsymbol{\beta}_n, \mathcal{S}_n) = \min_{1\leq m\leq n} \left\{ \frac{m}{n} : \sup_{\mathcal{S}'_m} \|\boldsymbol{\beta}_n(\mathcal{S}_n) - \boldsymbol{\beta}_n(\mathcal{S}'_m)\| = \infty \right\}, \tag{6.1}$$

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 $\mathbb{R}^k \times \mathcal{X}$.

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

$$\epsilon_n^*(\boldsymbol{\theta}_n, \mathcal{S}_n) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{\mathcal{S}'_m} \operatorname{dist}(\mathbf{V}(\boldsymbol{\theta}_n(\mathcal{S}_n))), \mathbf{V}(\boldsymbol{\theta}_n(\mathcal{S}'_m)) = \infty \right\}, \tag{6.2}$$

with $\operatorname{dist}(\cdot,\cdot)$ defined as $\operatorname{dist}(\mathbf{A},\mathbf{B}) = \max\left\{ |\lambda_1(\mathbf{A}) - \lambda_1(\mathbf{B})|, \left|\lambda_k(\mathbf{A})^{-1} - \lambda_k(\mathbf{B})^{-1}\right| \right\}$, 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 the theorem below. It extends the results for S-estimators of multivariate location and scale, see [6] and [18], and S-estimators for multivariate regression, see [30]. For S-estimators in the linear mixed effects model considered in [5], 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 [19].

Theorem 4. *Consider the minimization problem* (3.1) *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* $0 < r \le (n - \kappa(\mathcal{S}_n))/(2n)$, *where* $\kappa(\mathcal{S}_n)$ *is defined by* (4.4). *Then for any solution* $(\boldsymbol{\beta}_n, \boldsymbol{\theta}_n)$ *of minimization problem* (3.5),

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

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, 4], 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 [6, Theorem 6]).

Remark 6.1. *Van Aelst & Willems [30] also take into account the case* $r > (n - \kappa(\mathcal{S}_n))/(2n)$. *For this case, by replacing* $\lceil n - nr \rceil - \kappa(\mathcal{S}_n)$ *points, a specific solution to the S-minimization problem is constructed that breaks down. However, since there may be multiple solutions to the S-minimization problem, this does not necessarily mean that all solutions break down. In the proof of our Theorem 4, for the case* $r \le (n - \kappa(\mathcal{S}_n))/(2n)$, *we show that all solutions to* (3.1) *do not break down, when replacing at most* $\lceil nr \rceil - 1$ *points, and that the covariance part of all solutions do break down, when replacing* $\lceil nr \rceil$ *points. For the case* $r > (n - \kappa(\mathcal{S}_n))/(2n)$, *we can show that all solutions to* (3.1) *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 (3.1) and (3.5). 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 (3.5). 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 ξ_P. It follows that besides the constraint in (3.5), the pair $(\boldsymbol{\xi}_P, \lambda_P)$ satisfies

$$\begin{aligned} \int u(d)\mathbf{X}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,\mathrm{d}P(\mathbf{s}) &= \mathbf{0} \\ \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\right) + \frac{\lambda}{2}\int u(d)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,\mathrm{d}P(\mathbf{s}) &= 0, \end{aligned} \tag{7.1}$$

for $j = 1, \ldots, l$, where $u(s) = \rho'(s)/s$ and $d = d(\mathbf{s}, \boldsymbol{\xi})$ is defined by (5.1), and where we abbreviate $\mathbf{V}(\boldsymbol{\theta})$ by $\mathbf{V}$. To solve λ_P from the second set of equations, we multiply the j-th equation by θ_j and then sum over $j = 1, \ldots, l$. This leads to

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

which is solved by

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

When we insert this back into the second equation in (7.1), we find

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

or briefly

$$\int u(d)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\mathbf{H}_j\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,\mathrm{d}P(\mathbf{s}) = 0, \quad j = 1, \ldots, l, \tag{7.2}$$

where

$$\mathbf{H}_j = \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\right)\left(\sum_{t=1}^{l}\theta_t\frac{\partial\mathbf{V}}{\partial\theta_t}\right) - \mathrm{tr}\left(\mathbf{V}^{-1}\sum_{t=1}^{l}\theta_t\frac{\partial\mathbf{V}}{\partial\theta_t}\right)\frac{\partial\mathbf{V}}{\partial\theta_j}. \tag{7.3}$$

Because $\sum_{j=1}^{l} \theta_j \mathbf{H}_j = \mathbf{0}$, the system of equations (7.2) is linearly dependent. Similar to [16] 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 (7.2). We then find that any solution $\boldsymbol{\xi}_P$ of (3.5) satisfies the following equation

$$\int \Psi(\mathbf{s}, \boldsymbol{\xi}) \, \mathrm{d}P(\mathbf{s}) = \mathbf{0}, \tag{7.4}$$

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

$$\begin{aligned} \Psi_{\boldsymbol{\beta}}(\mathbf{s}, \boldsymbol{\xi}) &= u(d)\mathbf{X}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \Psi_{\boldsymbol{\theta},j}(\mathbf{s}, \boldsymbol{\xi}) &= u(d)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\mathbf{H}_j\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\right)(\rho(d) - b_0), \end{aligned} \tag{7.5}$$

for $j = 1, \ldots, l$, where $\mathbf{H}_j$ and $d = d(\mathbf{s}, \boldsymbol{\xi})$ are defined in (7.3) and (5.1), 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 (7.4) coincides with the one for the regression S-estimator in the linear mixed effects model (2.4) considered in [5] (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 [30] (see equation (2.2) in [29]). Similarly, the empirical score equation for $\Psi_{\boldsymbol{\beta}}$ coincides with the one for the location S-estimator of Example 4 considered in [16].

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 (7.1) and subtracted the S-constraint, leading to score equation (7.4) with Ψ given in (7.5). The fact that this was done in a specific way has the following reason. In cases where $\mathbf{V}(\boldsymbol{\theta})$ is linear, say

$$\mathbf{V}(\boldsymbol{\theta}) = \sum_{j=1}^{l} \theta_j \mathbf{L}_j, \tag{7.6}$$

the function $\Psi_{\boldsymbol{\theta}}$ simplifies a lot and can also be related to the covariance psi-function in [16]. 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 from Example 2. But also the multivariate location-scale model from Example 4 and the time series model (2.8) from Example 3 have linear covariance structures.

When $\mathbf{V}$ is of the form (7.6), 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, (7.3) simplifies to $\mathbf{H}_j = \operatorname{tr}\left(\mathbf{V}^{-1}\mathbf{L}_j\right)\mathbf{V} - k\mathbf{L}_j$, and $\Psi_{\boldsymbol{\theta},j}$ in (7.5) becomes

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

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

$$v(s) = u(s)s^2 - \rho(s) + b_0. \tag{7.7}$$

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

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

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 [16]. For our purposes we define

$$\Psi_{\mathbf{V}}(\mathbf{s},\boldsymbol{\xi}) = ku(d(\mathbf{s},\boldsymbol{\xi}))(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T - v(d(\mathbf{s},\boldsymbol{\xi}))\mathbf{V}. \tag{7.8}$$

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

$$\operatorname{vec}\left(\mathbf{V}^{-1}\mathbf{L}_j\mathbf{V}^{-1}\right) = \left(\mathbf{V}^{-1}\otimes\mathbf{V}^{-1}\right)\operatorname{vec}\left(\mathbf{L}_j\right)$$

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

$$\mathbf{L} = \left[\begin{array}{ccc} \operatorname{vec}(\mathbf{L}_1) & \cdots & \operatorname{vec}(\mathbf{L}_l) \end{array}\right]. \tag{7.9}$$

Then, the column vector $\Psi_{\boldsymbol{\theta}} = (\Psi_{\boldsymbol{\theta},1},\ldots,\Psi_{\boldsymbol{\theta},l})$ 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 (7.8) and $\mathbf{L}$ in (7.9). 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 (3.5) satisfies (7.4), where $\Psi = (\Psi_{\boldsymbol{\beta}},\Psi_{\boldsymbol{\theta}})$, with

$$\begin{aligned} \Psi_{\boldsymbol{\beta}}(\mathbf{s},\boldsymbol{\xi}) &= u(d)\mathbf{X}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ \Psi_{\boldsymbol{\theta}}(\mathbf{s},\boldsymbol{\xi}) &= -\mathbf{L}^T\left(\mathbf{V}^{-1}\otimes\mathbf{V}^{-1}\right)\operatorname{vec}\left(\Psi_{\mathbf{V}}(\mathbf{s},\boldsymbol{\xi})\right) \end{aligned} \tag{7.10}$$

where $d = d(\mathbf{s},\boldsymbol{\xi})$ is defined in (5.1), 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 $(\mathcal{D}_k\mathcal{D}_k)^{-1}\mathcal{D}_k^T\operatorname{vec}(\mathbf{C}) = \operatorname{vech}(\mathbf{C})$ (e.g., see [20, Ch. 3, Sec. 8]). Because $\mathbf{V}$ has full rank, it follows that equation (7.4) 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 $i = 1,\ldots,n$, are then equivalent with the ones found in [30].

For the linear mixed effects model (2.4), the covariance matrix $\mathbf{V}(\boldsymbol{\theta})$ has a linear structure with the vector $\boldsymbol{\theta} = (\sigma_0^2,\ldots,\sigma_r^2)$ 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 (7.4), by removing the factor $\mathbf{L}^T\left(\mathbf{V}^{-1}\otimes\mathbf{V}^{-1}\right)$ 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 $i = 1,\ldots,n$, is the same as the one obtained in [5]. The corresponding score equation for $\Psi_{\boldsymbol{\theta}}$ differs slightly from the one in [5], because the authors do not subtract a term with $\rho(d) - b_0$ to remove the linear dependency of the equations (7.2).

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

$$\operatorname{IF}(\mathbf{s};\boldsymbol{\xi},P) = \lim_{h\downarrow 0}\frac{\boldsymbol{\xi}((1-h)P + h\delta_{\mathbf{s}}) - \boldsymbol{\xi}(P)}{h}, \tag{8.1}$$

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 [11]). 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 [16]. Under the assumption that the limit in (8.1) exists and P has an elliptical contoured density (3.2), Van Aelst and Willems [30] 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 [5], 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 (8.1) 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 2 and Theorem 2, it follows that there exist solutions $\boldsymbol{\xi}(P_{h,\mathbf{s}_0})$ and $\boldsymbol{\xi}(P)$ to minimization problems (3.5) 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 5. *Let $\boldsymbol{\xi}(P_{h,\mathbf{s}_0})$ and $\boldsymbol{\xi}(P)$ be solutions to minimization problems (3.5) 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 (7.5) and suppose that*

$$\Lambda(\boldsymbol{\xi}) = \int \Psi(\mathbf{s}, \boldsymbol{\xi}) \, \mathrm{d}P(\mathbf{s}), \tag{8.2}$$

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

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

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

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

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

Corollary 4. *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 [19].

8.2 Elliptically contoured densities

When P is such that $\mathbf{y} \mid \mathbf{X}$ has an elliptically contoured density (3.2) 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 (8.2), is continuously differentiable. Clearly, condition (V5) implies former conditions (V4) and (V1).

Suppose that P is such that $\mathbf{y} \mid \mathbf{X}$ has an elliptically contoured density $f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ from (3.2), 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 $k \times k$ non-singular 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 (2.4) considered in [5] 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 ($\mathrm{C1}_{\epsilon'}$) and ($\mathrm{C2}_{\epsilon}$), for any $0 < \epsilon' < \epsilon \leq 1 - r$. When ρ and $\mathbf{V}$ satisfy (R1)-(R3) and (V1)-(V3), it follows from Theorem 1 and Corollary 2 that $\boldsymbol{\xi}(P)$ and $\boldsymbol{\xi}(P_{h,\mathbf{s}})$ exist, for h sufficiently small. If h in (3.2) is non-increasing and not constant on $[0, c_0^2]$, then $\boldsymbol{\xi}(P)$ is unique, according to Theorem 3, so that $\boldsymbol{\xi}(P_{h,\mathbf{s}}) \to \boldsymbol{\xi}(P)$, as $h \downarrow 0$. Hence, in order to apply Theorem 5, it remains to show that Λ in (8.2) is continuously differentiable with a non-singular derivative at $\boldsymbol{\xi}(P)$. As a first step we obtain that the derivative of Λ is a block matrix.

Lemma 2. *Suppose that P is such that $\mathbf{y} \mid \mathbf{X}$ has an elliptically contoured density $f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ from* (3.2) *and $\mathbb{E}\|\mathbf{X}\|^2 < \infty$. Suppose that $\boldsymbol{\xi}(P)$ is a solution to the corresponding minimization problem* (3.5)*, 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* (7.6)*. Let Λ be defined in* (8.2) *with Ψ defined in* (7.10)*. 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

$$\frac{\partial \Lambda_{\boldsymbol{\beta}}(\boldsymbol{\xi}(P))}{\partial \boldsymbol{\beta}} = -\alpha \mathbb{E}\left[\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X}\right], \tag{8.3}$$

with

$$\alpha = \mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\left(1 - \frac{1}{k}\right)\frac{\rho'(\|\mathbf{z}\|)}{\|\mathbf{z}\|} + \frac{1}{k}\rho''(\|\mathbf{z}\|)\right], \tag{8.4}$$

and

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

where $\mathbf{L} = \partial \mathrm{vec}(\mathbf{V}(\boldsymbol{\theta}(P)))/\partial \boldsymbol{\theta}^T$ is the $k^2 \times l$ matrix given in (7.9) *and*

$$\begin{aligned} \gamma_1 &= \frac{\mathbb{E}_{0,\mathbf{I}_k}\left[\rho''(\|\mathbf{z}\|)\|\mathbf{z}\|^2 + (k+1)\rho'(\|\mathbf{z}\|)\|\mathbf{z}\|\right]}{k+2} \\ \gamma_2 &= \frac{\mathbb{E}_{0,\mathbf{I}_k}\left[2\rho''(\|\mathbf{z}\|)\|\mathbf{z}\|^2 + k\rho'(\|\mathbf{z}\|)\|\mathbf{z}\|\right]}{2k(k+2)}, \end{aligned} \tag{8.5}$$

The proof is tedious, but straightforward, and can be found in [19].

Remark 8.1. *The proof of Lemma 2 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 B.3 in [19]. Furthermore, Lemma 2 is obtained for a linear covariance structure. However, with some additional technicalities, this result can also be

shown to hold for Ψ *defined in* (7.5) *corresponding to general covariance structures. For general covariance structures one still obtains* (8.3)*, and that*

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

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* (8.5)*, and where* $\mathbf{H}_j$ *is defined in* (7.3)*.*

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 [19].

Corollary 5. *Suppose that* P *is such that* $\mathbf{y} \mid \mathbf{X}$ *has an elliptically contoured density* $f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ *from* (3.2)*, 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* (3.5) *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* (7.6)*. Let* α*,* γ_1*, and* γ_2 *be defined in* (8.4) *and* (8.5)*, and suppose that* $\mathbb{E}_{0,\mathbf{I}_k}[\rho''(\|\mathbf{z}\|)] > 0$*. If* $\mathbf{X}$ *has full rank with probability one, then*

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

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* (7.9)*, has full rank, then* $\mathrm{IF}(\mathbf{s}_0, \boldsymbol{\theta}, P)$ *is given by*

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

Note that since $\mathbf{L}\boldsymbol{\theta}(P) = \mathrm{vec}(\mathbf{V}(\boldsymbol{\theta}(P))) = \mathrm{vec}(\boldsymbol{\Sigma})$, we can immediately obtain the influence function for the covariance functional $\mathbf{C}(P) = \mathbf{V}(\boldsymbol{\theta}(P))$. From Theorem 5 it immediately follows that $\mathrm{IF}(\mathbf{s}_0, \mathrm{vec}(\mathbf{C}), P)$ is given by

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

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 $\mathrm{IF}(\mathbf{s}, \mathrm{vec}(\mathbf{C}), P)$ 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 (2.4) with normal errors considered in [5], the influence function is not available. The expression can now be obtained from Corollary 5. The expression for $\mathrm{IF}(\mathbf{s}, \boldsymbol{\beta}, P)$ in Corollary 5 coincides with the one found for the multivariate regression S-functional in [30], where $\alpha > 0$ is the same constant as the one in the expression of the influence function for the location S-functional in [16]. Furthermore, for the multivariate regression model, one has $\boldsymbol{\theta} = \mathrm{vech}(\mathbf{C})$ 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

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

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 [30] and in [16].

9 Asymptotic normality

To establish asymptotic normality of the S-estimators, we use the score equations obtained from differentiation of the Lagrangian corresponding to the minimization problem (3.1). In the same way as before, we obtain score equation (7.4), with P equal to the empirical measure $\mathbb{P}_n$ corresponding to observations $\mathbf{s}_1, \ldots, \mathbf{s}_n$, with $\mathbf{s}_i = (\mathbf{y}_i, \mathbf{X}_i) \in \mathbb{R}^k \times \mathbb{R}^{kq}$. From (7.4), we see that any solution $\boldsymbol{\xi}_n = \boldsymbol{\xi}(\mathbb{P}_n)$ to the S-minimization problem (3.1) must satisfy

$$\int \Psi(\mathbf{s}, \boldsymbol{\xi}_n)\, \mathrm{d}\mathbb{P}_n(\mathbf{s}) = \mathbf{0}, \tag{9.1}$$

where $\Psi = (\Psi_{\boldsymbol{\beta}}, \Psi_{\boldsymbol{\theta}})$ is defined in (7.5).

9.1 General case

Writing $\boldsymbol{\xi}_P = \boldsymbol{\xi}(P)$, we decompose (9.1) as follows

$$\begin{aligned} 0 = \int \Psi(\mathbf{s}, \boldsymbol{\xi}_n)\, \mathrm{d}P(\mathbf{s}) + \int \Psi(\mathbf{s}, \boldsymbol{\xi}_P)\, \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) \\ + \int \left(\Psi(\mathbf{s}, \boldsymbol{\xi}_n) - \Psi(\mathbf{s}, \boldsymbol{\xi}_P)\right)\, \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}). \end{aligned} \tag{9.2}$$

The essential step in establishing asymptotic normality of $\boldsymbol{\xi}_n$, is to show that the third term on the right hand side of (9.2) is of the order $o_P(n^{-1/2})$. To this end we will apply results from empirical process theory as developed in Pollard [23]. This leads to the following theorem.

Theorem 6. *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* (3.1) *and* (3.5)*, and suppose that $\boldsymbol{\xi}_n \to \boldsymbol{\xi}(P)$ in probability. Suppose that Λ, as defined in* (8.2) *with Ψ defined in* (7.5)*, is continuously differentiable with a non-singular 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*

$$\mathbf{D}(P)^{-1}\mathbb{E}\left[\Psi(\mathbf{s}, \boldsymbol{\xi}(P))\Psi(\mathbf{s}, \boldsymbol{\xi}(P))^T\right]\mathbf{D}(P)^{-1}.$$

Theorem 6 is similar to Theorem 4.1 in [16]. Note that Theorem 6 confirms the well know 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 5,

$$\mathbf{D}(P)^{-1}\mathbb{E}\left[\Psi(\mathbf{s}, \boldsymbol{\xi}(P))\Psi(\mathbf{s}, \boldsymbol{\xi}(P))^T\right]\mathbf{D}(P)^{-1} = \mathbb{E}\left[\mathrm{IF}(\mathbf{s}, \boldsymbol{\xi}, P)\mathrm{IF}(\mathbf{s}, \boldsymbol{\xi}, P)^T\right]. \tag{9.3}$$

Van Aelst and Willems [30] 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 [5] consider asymptotic normality for S-estimators in the linear mixed effects model (2.4) 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 (3.2), 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 6. *Suppose that P is such that $\mathbf{y} \mid \mathbf{X}$ has an elliptically contoured density $f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ from (3.2), 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 (3.1) and (3.5), 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 (7.6). Let α, γ_1, and γ_2 be defined in (8.4) and (8.5), and suppose that $\mathbb{E}_{0,\mathbf{I}_k}[\rho''(\|\mathbf{z}\|)] > 0$. 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*

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

If $\gamma_1 > 0$ and the $k^2 \times l$ matrix $\mathbf{L}$, as defined in (7.9), has full rank, then $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}(P))$ is asymptotically normal with mean zero and covariance matrix

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

where

$$\sigma_1 = \frac{k(k+2)\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[u(\|\mathbf{z}\|)^2\|\mathbf{z}\|^4\right]}{\left(\mathbb{E}_{0,\mathbf{I}_k}\left[\rho''(\|\mathbf{z}\|)\|\mathbf{z}\|^2 + (k+1)\rho'(\|\mathbf{z}\|)\|\mathbf{z}\|\right]\right)^2}$$
$$\sigma_2 = -\frac{2}{k}\sigma_1 + \frac{4\mathbb{E}_{\mathbf{0},\mathbf{I}_k}[(\rho(\|\mathbf{z}\|) - b_0)^2]}{\left(\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\rho'(\|\mathbf{z}\|)^2\right]\right)^2}$$

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 6 it follows that

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

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

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

Corollary 6 is a direct consequence of Theorem 6. Its proof, in particular the derivations of the expressions for the limiting covariances, can be found in [19]. Note that the constants $\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\rho'(\|\mathbf{z}\|)^2\right]/(k\alpha^2)$, σ_1 and σ_2, are the same as the ones found in [16] for the location and covariance S-estimators, respectively. In fact, Corollary 6 is an extension of Corollary 5.1 in [16] 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 6. These estimators have been considered in [30], but asymptotic normality has not been established. Under the assumption that the heuristic (9.3) holds, asymptotic relative efficiencies are computed on the basis of this heuristic. Indeed, now that Corollary 6 has been established, one may check that (9.3) 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 6 differ from the ones found in [5] for the linear mixed effects model (2.4) with $\mathbf{X}_i = \mathbf{X}$, for $i = 1, \ldots, n$. The results in [5] are obtained by re-parameterizing $\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu}$ and interpreting the model as a multivariate location-scale model. Then building on the results in [16] for S-estimators of multivariate location-scale, the limiting covariances in [5] are found by application of the delta method. However, in view of Remark 3.1 this does not seem to be a correct approach.

Remark 9.1. *Although our expressions for the limiting covariances in Corollary 6 differ from the ones found in Proposition 1 in [5], somewhat surprisingly, they yield the same matrices for the example discussed in Section 5.1 in [5]. 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 6 yield different matrices as the ones found in [5]. Moreover, the corresponding confidence regions based on the expressions in Corollary 6 can be substantially smaller than the ones based on the expressions found in [5]. See the simulation in Section 10.*

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 6 and the ones in Copt and Victoria-Feser [5]. To this end we will study the behavior of the estimators for samples generated from a model that is close to the one in [5]:

$$\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta} + \gamma_i \mathbf{Z} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{10.1}$$

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 [5], 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 [5], we set $\beta_1 = \beta_2 = 1$ and $\theta_1 = \theta_2 = 1$.

In [5], the authors took $\mathbf{Z} = (1, 1, 1, 1)^T$ and $\mathbf{R} = \mathbf{I}_4$. With these choices the expression

$$\mathrm{Var}_{\mathrm{CVF}}(\boldsymbol{\beta}_n) = \frac{\mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[\rho'(\|\mathbf{z}\|)^2\right]}{k\alpha^2} (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma} \mathbf{X} (\mathbf{X}^T\mathbf{X})^{-1} \tag{10.2}$$

found in [5] for the limiting covariance matrix of $\sqrt{n}(\boldsymbol{\beta}_n - \boldsymbol{\beta})$ (see (14) in [5]), is equal to our expression

$$\mathrm{Var}_{\mathrm{LGRG}}(\boldsymbol{\beta}_n) = \frac{\mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[\rho'(\|\mathbf{z}\|)^2\right]}{k\alpha^2} (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}, \tag{10.3}$$

found in Corollary 6, 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 (10.2) and (10.3), 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 (10.1) and computed the value of S-estimators $\boldsymbol{\beta}_n$ and $\boldsymbol{\theta}_n$ by means of Tukey's bi-weight

$$\rho_{\mathrm{B}}(s; c) = \begin{cases} s^2/2 - s^4/(2c^2) + s^6/(6c^4), & |s| \le c \\ c^2/6 & |s| > c. \end{cases} \tag{10.4}$$

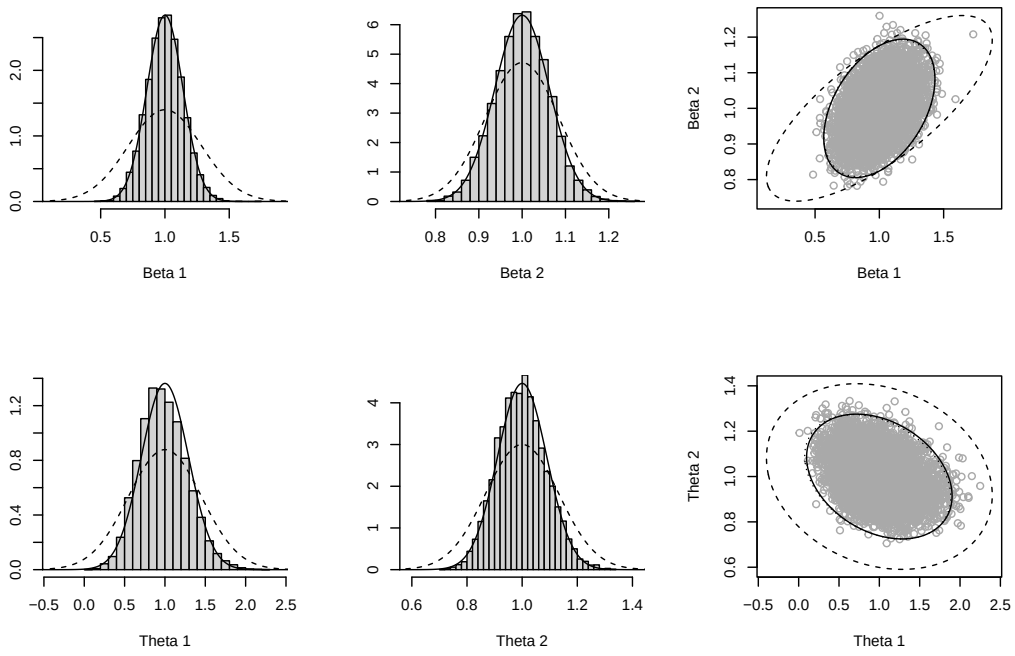

Figure 1: Empirical marginal and joined distributions together with limiting marginal and joined distributions of $\sqrt{n}(\boldsymbol{\beta}_n - \boldsymbol{\beta})$ (first row) and $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta})$ (second row).

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 4, this corresponds to (asymptotic) breakdown point 50%.

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 the densities and 95% contourlines of the theoretical limiting marginal and joined normal distributions using the covariance matrix in (10.2). The solid curves correspond to the marginal and joined normal distributions using the covariance matrix in (10.3). The empirical contourlines based on the sample mean and sample covariance of the 10 000 estimates are plotted in dotted lines, but they almost indistinguishable from the solid contourlines. We find

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

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% contourline 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 [5], the limiting covariance matrix was given by (see (15) in [5]) $(\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T\mathbf{V}_{\boldsymbol{\Sigma}}\mathbf{L}(\mathbf{L}^T\mathbf{L})^{-1}$, where $\mathbf{V}_{\boldsymbol{\Sigma}} = \sigma_1(\mathbf{I}_{k^2} + \mathbf{K}_{k,k})(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) + \sigma_2 \mathrm{vec}(\boldsymbol{\Sigma})\mathrm{vec}(\boldsymbol{\Sigma})^T$, see Corollary 5.1 in [16]. Because

$$(\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T(\mathbf{I}_{k^2} + \mathbf{K}_{k,k}) = 2(\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T,$$
$$(\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T\mathrm{vec}(\boldsymbol{\Sigma}) = \boldsymbol{\theta}(P),$$

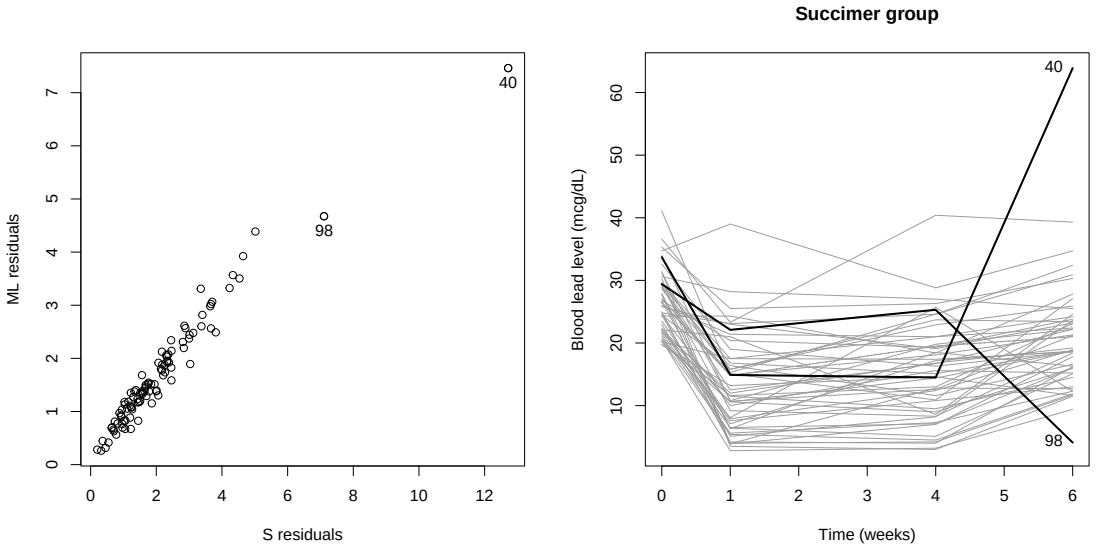

Figure 2: Left picture: standardized residuals for the S-estimates (horizontal axis) and the ML estimates (vertical axis). Right picture: observations for the subjects in the treatment group.

the expression given in [5] becomes

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

This differs from our Corollary 6, which gives

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

For the choices of $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{R}$ in [5], both expressions are equal. However, for the alternative choice for $\mathbf{R}$ made in alternative 2, one finds

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

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 dataset is discussed in [10] 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 [10] 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 $i = 1, \ldots, 100$ and $j = 1, \ldots, 4$, where $(t_1, \ldots, t_4) = (0, 1, 4, 6)$ 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 i-th subject is in the placebo group and $\delta_i = 1$, otherwise. The random effects $\boldsymbol{\gamma}_i = (\gamma_{1i}, \gamma_{2i}, \gamma_{3i})$, $i = 1, \ldots, 100$, are assumed to be independent mean zero normal random vectors with a diagonal covariance matrix consisting of variances $\sigma^2_{\gamma_1}$, $\sigma^2_{\gamma_2}$ and $\sigma^2_{\gamma_3}$, respectively. The measurement errors $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{i4})$, $i = 1, \ldots, 100$, are assumed to be independent mean zero random vectors with covariance matrix $\sigma^2_\epsilon\mathbf{I}_4$, also being independent of the random effects. In this way we are fitting a balanced linear

mixed effects model with unknown parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_6)$ and $\boldsymbol{\theta} = (\sigma^2_{\gamma_1}, \sigma^2_{\gamma_2}, \sigma^2_{\gamma_3}, \sigma^2_{\epsilon})$, and a linear covariance structure.

We estimated $(\boldsymbol{\beta}, \boldsymbol{\theta})$ by means of maximum likelihood and by means of the S-estimator corresponding to Tukey's bi-weight defined in (10.4). 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 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.

A Proofs and technical lemmas

Proof of Theorem 1

Proof. Let $(\boldsymbol{\beta}, \boldsymbol{\theta}) \in \mathbb{R}^q \times \mathbb{R}^l$ satisfy the S-constraint in (3.5). Then from (R1)-(R2) it follows that

$$P(\mathcal{C}(\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\theta}), c_0)) \geq 1 - \frac{1}{a_0}\int \rho\left(\sqrt{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}(\boldsymbol{\theta})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}\right) \mathrm{d}P(\mathbf{s}) \geq 1 - r. \tag{A.1}$$

Since $1 - r \geq \epsilon$, Lemma 1(i) then implies that $\lambda_k(\mathbf{V}(\boldsymbol{\theta})) \geq a_1 > 0$. Because

$$\lim_{m\to\infty}\int \rho\left(\|\mathbf{y}\|/m\right) \mathrm{d}P(\mathbf{y}, \mathbf{X}) = 0,$$

we can find $m_0 > 0$, such that $\int \rho(\|\mathbf{y}\|/m_0)\,\mathrm{d}P(\mathbf{y}, \mathbf{X}) \leq b_0$. Lemma 1(ii) then yields that $\lambda_1(\mathbf{V}(\boldsymbol{\theta})) \leq a_2 < \infty$. Application of Lemma 1(iii), with $a = 1 - r$ together with $(\mathrm{C1}_\epsilon)$, implies that $\|\boldsymbol{\beta}\| \leq M < \infty$. It follows that $\boldsymbol{\beta}$ is in a compact subset of $\mathbb{R}^q$ and $\mathbf{V}(\boldsymbol{\theta})$ is in a compact set $K \subset \mathbb{R}^{k\times k}$.

According to (2.9), the mapping $\boldsymbol{\theta} \mapsto \mathbf{V}(\boldsymbol{\theta})$ is one-to-one, so that we can restrict $\boldsymbol{\theta}$ to the pre-image $\mathbf{V}^{-1}(K)$. Then with conditions (V1) and (V3) it follows that also $\mathbf{V}^{-1}(K)$ is compact in $\mathbb{R}^l$. We conclude that for solving minimization problem (3.5), we can restrict ourselves to a compact set $K' \subset \mathbb{R}^q \times \mathbb{R}^l$. As $\det(\mathbf{V}(\boldsymbol{\theta}))$ is a continuous function of $(\boldsymbol{\beta}, \boldsymbol{\theta})$, due to condition (V1), it must attain a minimum on K'. □

Proof of Corollary 1

Proof. Let $\mathbb{P}_n$ be the empirical measure corresponding to the collection $\mathcal{S}_n$. Then $\mathbb{P}_n$ satisfies $(\mathrm{C1}_\epsilon)$ for any $0 < \epsilon \leq 1 - r$ and satisfies $(\mathrm{C2}_\epsilon)$, for $\epsilon = (\kappa(\mathcal{S}_n) + 1)/n$. Clearly $0 < \epsilon \leq 1 - r$, where $r = b_0/a_0$, so according to Theorem 1 there exists at least one solution to (3.5) with $P = \mathbb{P}_n$. This means that there exists at least one solution to (3.1). □

Proof of Corollary 2

Proof. First note there exists $0 < \eta < \epsilon' - \epsilon$. According to Ranga Rao [24, Theorem 4.2] we have

$$\sup_{C\in\mathfrak{C}} |P_t(C) - P(C)| \to 0, \quad \text{as } t \to \infty. \tag{A.2}$$

Because strips $H(\boldsymbol{\alpha}, \ell, \delta) \in \mathfrak{C}$, property (A.2) implies that every strip with $P_t(H(\boldsymbol{\alpha}, \ell, \delta)) \geq \epsilon + \eta$, for t sufficiently large, must also satisfy $P(H(\boldsymbol{\alpha}, \ell, \delta)) \geq \epsilon$. This means that

$$\inf \{\delta : P_t(H(\boldsymbol{\alpha}, \ell, \delta)) \geq \epsilon + \eta\} \geq \inf \{\delta : P(H(\boldsymbol{\alpha}, \ell, \delta)) \geq \epsilon\} > 0.$$

It follows that, for t sufficiently large, P_t satisfies condition (C2$_{\epsilon+\eta}$). Next, consider the compact set K from (C1$_{\epsilon'}$). Without loss of generality we may assume that it belongs to $\mathfrak{C}$. Therefore, as $P(K) \geq r + \epsilon'$, for t sufficiently large $P_t(K) \geq r + \epsilon + \eta$. It follows that, for t sufficiently large, P_t satisfies condition (C1$_{\epsilon+\eta}$). Since $\epsilon + \eta < 1 - r$, according to Theorem 1 at least one solution $\boldsymbol{\xi}(P_t)$ exists, for t sufficiently large. □

Proof of Theorem 2

Proof. First note that there exists $0 < \eta < \epsilon' - \epsilon$. Denote $\boldsymbol{\xi}(P_t) = \boldsymbol{\xi}_t = (\boldsymbol{\beta}_t, \boldsymbol{\theta}_t)$. Similar to (A.1) we find that $P_t(\mathcal{C}(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t, c_0)) \geq 1 - r$. Therefore, as $\mathcal{C}(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t, c_0) \in \mathfrak{C}$ and $1 - r > 1 - r - \eta$, it follows from (A.2) that

$$P(\mathcal{C}(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t, c_0)) \geq P_t(\mathcal{C}(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t, c_0)) - \sup_{C \in \mathfrak{C}} |P_t(C) - P(C)| \geq 1 - r - \eta, \tag{A.3}$$

for t sufficiently large. Since $1 - r - \eta > \epsilon$, this means that, according to Lemma 1(i), there exists $a_1 > 0$ only depending on c_0 and P, such that for t sufficiently large,

$$\lambda_k(\mathbf{V}(\boldsymbol{\theta}_t)) \geq a_1 > 0.$$

Denote $\boldsymbol{\xi}(P) = \boldsymbol{\xi}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$ and let $d_0(\mathbf{s}) = d(\mathbf{s}, \boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$. Then according to Lemma B.1 in [19], for any $\sigma > -1$,

$$\int \rho\left(\frac{d_0(\mathbf{s})}{1+\sigma}\right) \mathrm{d}P_t(\mathbf{s}) \to \int \rho\left(\frac{d_0(\mathbf{s})}{1+\sigma}\right) \mathrm{d}P(\mathbf{s}),$$

as $t \to \infty$. As the limit is strictly decreasing at $\sigma = 0$, and $\boldsymbol{\xi}_0$ satisfies the constraint in (3.5), we find that for all $\sigma > 0$,

$$\int \rho\left(\frac{d_0(\mathbf{s})}{1+\sigma}\right) \mathrm{d}P_t(\mathbf{s}) \leq b_0,$$

for t sufficiently large. Hence, similar to the proof of Lemma 1(ii) we conclude that for any possible solution $\boldsymbol{\xi}_t = (\boldsymbol{\beta}_t, \boldsymbol{\theta}_t)$, it must hold that

$$\det(\mathbf{V}(\boldsymbol{\theta}_t)) \leq (1+\sigma)^{2k} \det(\mathbf{V}(\boldsymbol{\theta}_0)),$$

for t sufficiently large. As $\sigma > 0$ can be taken arbitrarily small, we conclude that

$$\limsup_{t \to \infty} \det(\mathbf{V}(\boldsymbol{\theta}_t)) \leq \det(\mathbf{V}(\boldsymbol{\theta}_0)), \tag{A.4}$$

and we find that $\lambda_1(\mathbf{V}(\boldsymbol{\theta}_t)) \leq \det(\mathbf{V}(\boldsymbol{\theta}_0))/a_1^{k-1} < \infty$, for t large sufficiently large. Finally, let K be the compact set from (C1$_{\epsilon'}$), so that $P(K) \geq r + \epsilon' > r + \epsilon + \eta$. Then, according to (A.3), it follows from Lemma 1(iii) with $a = 1 - r - \eta$, that there exists $0 < M < \infty$ such that $\|\boldsymbol{\beta}_t\| \leq M$, for t sufficiently large. This means that there exists a compact set K', such that for t sufficiently large the sequence $\{(\boldsymbol{\beta}_t, \mathbf{V}(\boldsymbol{\theta}_t))\} \subset K'$. Then, similar to the second part of the proof of Theorem 1, the conditions on the mapping $\boldsymbol{\theta} \mapsto \mathbf{V}(\boldsymbol{\theta})$ yield that there exists a compact set $K'' \subset \mathbb{R}^{q+l}$, such that for t sufficiently large, the sequence $\{\boldsymbol{\xi}_t\} \subset K''$.

Consider a convergent subsequence $\{\boldsymbol{\xi}_{t_j}\}$ with $\boldsymbol{\xi}_{t_j} \to \boldsymbol{\xi}_L$. With Lemma B.1 in [19] and the fact that $\boldsymbol{\xi}_{t_j}$ satisfies the S-constraint in (3.5) at $P = P_{t_j}$, we find

$$\int \rho(d(\mathbf{s}, \boldsymbol{\xi}_L)) \, \mathrm{d}P(s) = \lim_{j \to \infty} \int \rho(d(s, \boldsymbol{\xi}_{t_j})) \, \mathrm{d}P_{t_j}(s) \leq b_0.$$

Hence, $\boldsymbol{\xi}_L$ satisfies the S-constraint in (3.5), which has solution $\boldsymbol{\xi}_0$. This means that $\det(\mathbf{V}(\boldsymbol{\theta}_L)) \geq \det(\mathbf{V}(\boldsymbol{\theta}_0))$. But then from (A.4), it follows $\det(\mathbf{V}(\boldsymbol{\theta}_L)) = \det(\mathbf{V}(\boldsymbol{\theta}_0))$. Uniqueness of $\boldsymbol{\xi}_0$ together with identifiability (2.9) then implies that $\boldsymbol{\xi}_L = \boldsymbol{\xi}_0$. Because $\{\boldsymbol{\xi}_t\}$ eventually stays in a compact set, this means that we must have $\lim_{t \to \infty} \boldsymbol{\xi}_t = \boldsymbol{\xi}_0$. □

Proof of Corollary 3

Proof. We apply Theorem 2 to the sequence $\mathbb{P}_n$, $n = 1, 2, \ldots$, of probability measures, where $\mathbb{P}_n$ is the empirical measure corresponding to $(\mathbf{y}_1, \mathbf{X}_1), \ldots, (\mathbf{y}_n, \mathbf{X}_n)$. According to the Portmanteau Theorem (e.g., see Theorem 2.1 in [2]), $\mathbb{P}_n$ converges weakly to P, with probability one. The corollary then follows from Theorem 2. □

Proof of Theorem 5

Proof. Denote $\boldsymbol{\xi}_{h,\mathbf{s}_0} = \boldsymbol{\xi}(P_{h,\mathbf{s}_0})$. This solution satisfies the score equation (7.4) for the regression S-functional at $P_{h,\mathbf{s}_0}$, that is

$$\int \Psi(\mathbf{s}, \boldsymbol{\xi}_{h,\mathbf{s}_0}) \, \mathrm{d}P_{h,\mathbf{s}_0}(\mathbf{s}) = \mathbf{0}.$$

We decompose as follows

$$\begin{aligned}
\mathbf{0} &= \int \Psi(\mathbf{s}, \boldsymbol{\xi}_{h,\mathbf{s}_0}) \, \mathrm{d}P_{h,\mathbf{s}_0}(\mathbf{s}) \\
&= (1-h) \int \Psi(\mathbf{s}, \boldsymbol{\xi}_{h,\mathbf{s}_0}) \, \mathrm{d}P(\mathbf{s}) + h\Psi(\mathbf{s}_0, \boldsymbol{\xi}_{h,\mathbf{s}_0}) \\
&= (1-h)\Lambda(\boldsymbol{\xi}_{h,\mathbf{s}_0}) + h\Big(\Psi(\mathbf{s}_0, \boldsymbol{\xi}_{h,\mathbf{s}_0}) - \Psi(\mathbf{s}_0, \boldsymbol{\xi}(P))\Big) + h\Psi(\mathbf{s}_0, \boldsymbol{\xi}(P)).
\end{aligned}$$

We first determine the order of $\boldsymbol{\xi}_{h,\mathbf{s}_0} - \boldsymbol{\xi}(P)$, as $h \downarrow 0$. Because $\boldsymbol{\xi} \mapsto \Psi(\mathbf{s}_0, \boldsymbol{\xi})$ is continuous, it follows that

$$\Psi(\mathbf{s}_0, \boldsymbol{\xi}_{h,\mathbf{s}_0}) = \Psi(\mathbf{s}_0, \boldsymbol{\xi}(P)) + o(1), \qquad \text{as } h \downarrow 0.$$

Furthermore, because $\boldsymbol{\xi} \mapsto \Lambda(\boldsymbol{\xi})$ is continuously differentiable at $\boldsymbol{\xi}(P)$, we have that

$$\Lambda(\boldsymbol{\xi}_{h,\mathbf{s}_0}) = \Lambda(\boldsymbol{\xi}(P)) + \mathbf{D}(P)(\boldsymbol{\xi}_{h,\mathbf{s}_0} - \boldsymbol{\xi}(P)) + o(\|\boldsymbol{\xi}_{h,\mathbf{s}_0} - \boldsymbol{\xi}(P)\|).$$

Since $\boldsymbol{\xi}(P)$ is the S-functional at P, it is a zero of the corresponding score equation, i.e., $\Lambda(\boldsymbol{\xi}(P)) = 0$. It follows that

$$\mathbf{0} = (1-h)\mathbf{D}(P)(\boldsymbol{\xi}_{h,\mathbf{s}_0} - \boldsymbol{\xi}(P)) + o(\|\boldsymbol{\xi}_{h,\mathbf{s}_0} - \boldsymbol{\xi}(P)\|) + o(h) + h\Psi(\mathbf{s}_0, \boldsymbol{\xi}(P)).$$

Because $\mathbf{D}(P)$ is non-singular and $\Psi(\mathbf{s}_0, \boldsymbol{\xi}(P))$ is fixed, this implies $\boldsymbol{\xi}_{h,\mathbf{s}_0} - \boldsymbol{\xi}(P) = O(h)$. After inserting this in the previous equality, it follows that

$$\begin{aligned}
\mathbf{0} &= (1-h)\mathbf{D}(P)(\boldsymbol{\xi}_{h,\mathbf{s}_0} - \boldsymbol{\xi}(P)) + h\Psi(\mathbf{s}_0, \boldsymbol{\xi}(P)) + o(h) \\
&= \mathbf{D}(P)(\boldsymbol{\xi}_{h,\mathbf{s}_0} - \boldsymbol{\xi}(P)) + h\Psi(\mathbf{s}_0, \boldsymbol{\xi}(P)) + o(h).
\end{aligned}$$

We conclude

$$\frac{\boldsymbol{\xi}_{h,\mathbf{s}_0} - \boldsymbol{\xi}(P)}{h} = -\mathbf{D}(P)^{-1}\Psi(\mathbf{s}_0, \boldsymbol{\xi}(P)) + o(1), \qquad \text{as } h \downarrow 0.$$

This means that the limit of the left hand side exists and

$$\mathrm{IF}(\mathbf{s}_0; \boldsymbol{\xi}, P) = \lim_{h \downarrow 0} \frac{\boldsymbol{\xi}((1-h)P + h\delta_{\mathbf{s}_0}) - \boldsymbol{\xi}(P)}{h} = -\mathbf{D}(P)^{-1}\Psi(\mathbf{s}_0, \boldsymbol{\xi}(P)).$$

Next, consider the covariance functional $\mathbf{C}(P) = \mathbf{V}(\boldsymbol{\theta}(P))$. By definition

$$\mathrm{IF}(\mathbf{s}_0; \mathrm{vec}(\mathbf{C}), P) = \lim_{h \downarrow 0} \frac{\mathrm{vec}(\mathbf{C}(P_{h,\mathbf{s}_0})) - \mathrm{vec}(\mathbf{C}(P))}{h}.$$

Due to (V4), by applying the chain rule (e.g., see [20, Theorem 12, page 108]), we find

$$\begin{aligned}
\mathrm{vec}(\mathbf{C}(P_{h,\mathbf{s}_0})) - \mathrm{vec}(\mathbf{C}(P)) &= \frac{\partial \mathrm{vec}(\mathbf{V}(\boldsymbol{\theta}(P)))}{\partial \boldsymbol{\theta}^T}\Big(\boldsymbol{\theta}(P_{h,\mathbf{s}_0}) - \boldsymbol{\theta}(P)\Big) + o(\|\boldsymbol{\theta}(P_{h,\mathbf{s}_0}) - \boldsymbol{\theta}(P)\|) \\
&= \left(\frac{\partial \mathrm{vec}(\mathbf{V}(\boldsymbol{\theta}(P)))}{\partial \boldsymbol{\theta}^T}\right)\Big(\boldsymbol{\theta}(P_{h,\mathbf{s}_0}) - \boldsymbol{\theta}(P)\Big) + o(h).
\end{aligned}$$

Dividing by h and letting $h \downarrow 0$, finishes the proof. □

Proof of Theorem 6

Proof. Write $\boldsymbol{\xi}_P = \boldsymbol{\xi}(P)$. Then from (9.2) and Lemma B.8 in [19], it follows that

$$\mathbf{0} = \Lambda(\boldsymbol{\xi}_n) + \int \Psi(\mathbf{s}, \boldsymbol{\xi}_P)\, \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) + o_P(n^{-1/2}). \tag{A.5}$$

From Theorem 2, we know that $\boldsymbol{\xi}_n \to \boldsymbol{\xi}_P$ with probability one. For the first term on the right hand side we have that

$$\Lambda(\boldsymbol{\xi}_n) = \mathbf{D}(P)(\boldsymbol{\xi}_n - \boldsymbol{\xi}_P) + o_P(\|\boldsymbol{\xi}_n - \boldsymbol{\xi}_P\|) = \mathbf{D}(P)(\boldsymbol{\xi}_n - \boldsymbol{\xi}_P) + o_P(\|\boldsymbol{\xi}_n - \boldsymbol{\xi}_P\|),$$

using that $\Lambda(\boldsymbol{\xi}_P) = \mathbf{0}$, due to the fact that $\boldsymbol{\xi}_P$ is the solution of (3.5), see also (7.4). From Lemma B.2 in [19] we have that $\|\Psi_{\boldsymbol{\beta}}\| \leq C_1\|\mathbf{X}\|$ and $\|\Psi_{\boldsymbol{\theta}}\| \leq C_2$, for universal constants $0 < C_1, C_2 < \infty$. Hence, from the conditions of the theorem, it follows that $\mathbb{E}\|\Psi(\mathbf{s}, \boldsymbol{\xi}(P))\|^2 < \infty$. This means that for the second term on the right hand side of (A.5),

$$\int \Psi(\mathbf{s}, \boldsymbol{\xi}_P)\, \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) = \frac{1}{n}\sum_{i=1}^{n} (\Psi(\mathbf{s}_i, \boldsymbol{\xi}_P) - \mathbb{E}\Psi(\mathbf{s}, \boldsymbol{\xi}_P)) = O_P(n^{-1/2}),$$

according to the central limit theorem. It follows that

$$\mathbf{0} = \mathbf{D}(P)(\boldsymbol{\xi}_n - \boldsymbol{\xi}_P) + o_P(\|\boldsymbol{\xi}_n - \boldsymbol{\xi}_P\|) + O_P(n^{-1/2}),$$

so that $\|\boldsymbol{\xi}_n - \boldsymbol{\xi}_P\| = O_P(n^{-1/2})$. If we insert this is in (A.5), we obtain

$$\mathbf{0} = \mathbf{D}(P)(\boldsymbol{\xi}_n - \boldsymbol{\xi}_P) + \frac{1}{n}\sum_{i=1}^{n} (\Psi(\mathbf{s}_i, \boldsymbol{\xi}_P) - \mathbb{E}\Psi(\mathbf{s}, \boldsymbol{\xi}_P)) + o_P(n^{-1/2}),$$

from which it follows that

$$\sqrt{n}(\boldsymbol{\xi}_n - \boldsymbol{\xi}_P) = -\mathbf{D}(P)^{-1}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} (\Psi(\mathbf{s}_i, \boldsymbol{\xi}_P) - \mathbb{E}\Psi(\mathbf{s}, \boldsymbol{\xi}_P))\right) + o_P(1).$$

After application of the central limit theorem, we conclude that $\sqrt{n}(\boldsymbol{\xi}_n - \boldsymbol{\xi}_P)$ is asymptotically normal with mean zero and covariance matrix

$$\mathbf{D}(P)^{-1}\mathbb{E}\left[\Psi(\mathbf{s}, \boldsymbol{\xi}_P)\Psi(\mathbf{s}, \boldsymbol{\xi}_P)^T\right]\mathbf{D}(P)^{-1},$$

where we use that $\mathbb{E}\Psi(\mathbf{s}, \boldsymbol{\xi}_P) = \Lambda(\boldsymbol{\xi}_P) = \mathbf{0}$. This proves the theorem. □

References

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

[2] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, Inc., New York-London-Sydney, 1968.

[3] I. Chervoneva and M. Vishnyakov. Generalized s-estimators for linear mixed effects models. *Statistica Sinica*, 24(3):1257–1276, 2014.

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

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

[6] P. L. Davies. Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *Ann. Statist.*, 15(3):1269–1292, 1987.

[7] Eugene Demidenko. *Mixed models*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2013. Theory and applications with R.

[8] David Donoho and Peter J. Huber. The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser., pages 157–184. Wadsworth, Belmont, CA, 1983.

[9] María V. Fasano, Ricardo A. Maronna, Mariela Sued, and Víctor J. Yohai. Continuity and differentiability of regression M functionals. *Bernoulli*, 18(4):1284–1309, 2012.

[10] Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware. *Applied longitudinal analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2011.

[11] Frank R. Hampel. The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, 69:383–393, 1974.

[12] H. O. Hartley and J. N. K. Rao. Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54:93–108, 1967.

[13] Stephane Heritier, Eva Cantoni, Samuel Copt, and Maria-Pia Victoria-Feser. *Robust methods in biostatistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2009.

[14] Robert I. Jennrich and Mark D. Schluchter. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42(4):805–820, 1986.

[15] Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.

[16] Hendrik P. Lopuhaä. On the relation between S-estimators and M-estimators of multivariate location and covariance. *Ann. Statist.*, 17(4):1662–1683, 1989.

[17] Hendrik P. Lopuhaä. Asymptotic expansion of S-estimators of location and covariance. *Statist. Neerlandica*, 51(2):220–237, 1997.

[18] Hendrik P. Lopuhaä and Peter J. Rousseeuw. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Statist.*, 19(1):229–248, 1991.

[19] H.P. Lopuhaä, V. Gares, and A. Ruiz-Gazen. Supplement to "S-estimation in linear models with structured covariance matrices". 2022.

[20] Jan R. Magnus and Heinz Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 1988.

[21] Deborah Nolan and David Pollard. U-processes: rates of convergence. *Ann. Statist.*, 15(2):780–799, 1987.

[22] José C. Pinheiro, Chuanhai Liu, and Ying Nian Wu. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *J. Comput. Graph. Statist.*, 10(2):249–276, 2001.

[23] David Pollard. *Convergence of stochastic processes*. Springer Series in Statistics. Springer-Verlag, New York, 1984.

[24] R. Ranga Rao. Relations between weak and uniform convergence of measures with applications. *Ann. Math. Statist.*, 33:659–680, 1962.

[25] C. Radhakrishna Rao. Estimation of variance and covariance components in linear models. *J. Amer. Statist. Assoc.*, 67:112–115, 1972.

[26] P. Rousseeuw and V. Yohai. Robust regression by means of S-estimators. In *Robust and nonlinear time series analysis (Heidelberg, 1983)*, volume 26 of *Lect. Notes Stat.*, pages 256–272. Springer, New York, 1984.

[27] Peter Rousseeuw. Multivariate estimation with high breakdown point. In *Mathematical statistics and applications, Vol. B (Bad Tatzmannsdorf, 1983)*, pages 283–297. Reidel, Dordrecht, 1985.

[28] Peter J. Rousseeuw. Least median of squares regression. *J. Amer. Statist. Assoc.*, 79(388):871–880, 1984.

[29] Stefan Van Aelst and Gert Willems. Multivariate regression S-estimators for robust estimation and inference, 2004, preprint received by personal communication.

[30] Stefan Van Aelst and Gert Willems. Multivariate regression S-estimators for robust estimation and inference. *Statist. Sinica*, 15(4):981–1001, 2005.

B Supplemental Material

For later use we first define two important matrix norms and mention some useful properties. For $m \times n$ real-valued matrices $\mathbf{A}$, we define the Euclidean norm or Frobenius norm as

$$\|\mathbf{A}\| = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}.$$

and the spectral norm by

$$\|\mathbf{A}\|_2 = \sup_{\|\mathbf{u}\|=1} \|\mathbf{A}\mathbf{u}\|.$$

Recall that for real-valued $\mathbf{A}$, the largest eigenvalue is defined by

$$\lambda_1(\mathbf{A}) = \sup_{\|\mathbf{u}\|=1} \mathbf{u}^T\mathbf{A}\mathbf{u}. \tag{B.1}$$

This means that $\|\mathbf{A}\|_2^2 = \lambda_1(\mathbf{A}^T\mathbf{A})$. Other useful properties are

$$\|\mathbf{A}\|_2 \leq \|\mathbf{A}\| \leq \sqrt{\min(m,n)}\|\mathbf{A}\|_2 \tag{B.2}$$

and

$$\|\mathbf{A}\|^2 = \mathrm{tr}(\mathbf{A}^T\mathbf{A}) = \mathrm{tr}(\mathbf{A}\mathbf{A}^T), \tag{B.3}$$

and

$$\|\mathbf{u}\mathbf{v}^T\| = \|\mathbf{u}\|\|\mathbf{v}\| \tag{B.4}$$

for $\mathbf{u} \in \mathbb{R}^m$ and $\mathbf{v} \in \mathbb{R}^n$. When $\mathbf{A}$ is symmetric then $\lambda_1(\mathbf{A}^T\mathbf{A}) = \lambda_1(\mathbf{A}^2) = \lambda_1(\mathbf{A})^2$. In that case

$$|\lambda_1(\mathbf{A})| = \|\mathbf{A}\|_2 \leq \|\mathbf{A}\| \leq \sqrt{\min(m,n)}\|\mathbf{A}\|_2 = \sqrt{\min(m,n)}\,|\lambda_1(\mathbf{A})|\,. \tag{B.5}$$

Finally, note that both matrix norms are submultiplicative, that is

$$\|\mathbf{A}\mathbf{B}\|_2 \leq \|\mathbf{A}\|_2\|\mathbf{B}\|_2, \tag{B.6}$$

and

$$\|\mathbf{A}\mathbf{B}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|. \tag{B.7}$$

B.1 Proofs of Section 4

Proof of Lemma 1

Proof. Note that cylinder $\mathcal{C}(\boldsymbol{\beta}, \boldsymbol{\theta}, c)$ is contained in some strip $H(\boldsymbol{\alpha}, \ell, 2c\sqrt{\lambda_k(\mathbf{V}(\boldsymbol{\theta}))})$. It then follows from (C2$_\epsilon$) that $\lambda_k(\mathbf{V}(\boldsymbol{\theta})) \geq \delta_\epsilon^2/4c^2$. This proves (i). Let $\boldsymbol{\theta}_0$ be such that the pair $(\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\theta}_0)) = (\mathbf{0}, m_0^2\mathbf{I}_k)$ satisfies the S-constraint in (3.5), which is possible due to condition (V2). It then follows that for any solution $(\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\theta}))$ of (3.5), one must have $\det(\mathbf{V}(\boldsymbol{\theta})) \leq m_0^{2k}$. The lower bound on the smallest eigenvalue then implies $\lambda_1(\mathbf{V}(\boldsymbol{\theta})) \leq m_0^{2k}/a_1^{k-1} < \infty$, which proves (ii). Note that

$$\gamma_\epsilon = \inf_{P(J)\geq\epsilon} \inf_{\|\boldsymbol{\gamma}\|=1} \sup_{\mathbf{s}\in J} \|\mathbf{X}\boldsymbol{\gamma}\| > 0, \tag{B.8}$$

where the infima are taken over all subsets $J \subset \mathbb{R}^k \times \mathcal{X}$ with $P(J) \geq \epsilon$ and all unit vectors $\boldsymbol{\gamma} \in \mathbb{R}^q$. This can be seen as follows. Take $\boldsymbol{\alpha}^T = (\mathbf{0}^T, \boldsymbol{\gamma}^T, \ldots, \boldsymbol{\gamma}^T)/k$, where $\mathbf{0}$ is a k-vector of zeros, so $\boldsymbol{\alpha} \in \mathbb{R}^k \times \mathbb{R}^{kq}$ and $\|\boldsymbol{\alpha}\| = 1$. Then, with $\ell = 0$, we have $\boldsymbol{\alpha}^T\mathbf{s} - \ell = (\boldsymbol{\gamma}^T\mathbf{x}_1 + \cdots + \boldsymbol{\gamma}^T\mathbf{x}_k)/k$. Note that $\|\mathbf{X}\boldsymbol{\gamma}\|^2 = \sum_{j=1}^k (\mathbf{x}_j^T\boldsymbol{\gamma})^2$. Therefore, if $\gamma_\epsilon = 0$, then $\boldsymbol{\gamma}^T\mathbf{x}_j = 0$, for all $j = 1, \ldots, k$, which means that $\boldsymbol{\alpha}^T\mathbf{s} - \ell = 0$. This would be in contradiction with (4.3), which is equivalent to (C2$_\epsilon$), according to Remark 4.1. Now, note that

$$P(\mathcal{C}(\boldsymbol{\beta}, \boldsymbol{\theta}, c) \cap K) \geq P(\mathcal{C}(\boldsymbol{\beta}, \boldsymbol{\theta}, c)) - P(K^c) \geq a - 1 + 1 - a + \epsilon = \epsilon.$$

Hence, according to (B.8), there exists $\mathbf{s}_0 = (\mathbf{y}_0, \mathbf{X}_0) \in \mathcal{C}(\boldsymbol{\beta}, \boldsymbol{\theta}, c) \cap K$, such that $\|\mathbf{X}_0\boldsymbol{\gamma}\| \geq \gamma_\epsilon > 0$, for all $\boldsymbol{\gamma}$ with $\|\boldsymbol{\gamma}\| = 1$. Because $\mathbf{s}_0 \in \mathcal{C}(\boldsymbol{\beta}, \boldsymbol{\theta}, c)$, it holds

$$\|\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta}\|^2 \leq (\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})^T\mathbf{V}(\boldsymbol{\theta})^{-1}(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})\lambda_1(\mathbf{V}(\boldsymbol{\theta})) \leq c\lambda_1(\mathbf{V}(\boldsymbol{\theta})) \leq ca_2,$$

due to part (ii). Because $\mathbf{s}_0 = (\mathbf{y}_0, \mathbf{X}_0) \in K$, this means that

$$\|\mathbf{X}_0\boldsymbol{\beta}\| \leq \|\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta}\| + \|\mathbf{y}_0\| \leq \sqrt{ca_2} + \sup_{(\mathbf{y},\mathbf{X})\in K} \|\mathbf{y}\|.$$

Because $\mathbf{s}_0 \in \mathcal{C}(\boldsymbol{\beta}, \boldsymbol{\theta}, c)$, together with (B.8) we conclude that

$$\|\boldsymbol{\beta}\| = \|\mathbf{X}_0\boldsymbol{\beta}\| \times \frac{\|\boldsymbol{\beta}\|}{\|\mathbf{X}_0\boldsymbol{\beta}\|} \leq \frac{1}{\gamma_\epsilon}\|\mathbf{X}_0\boldsymbol{\beta}\| \leq \frac{1}{\gamma_\epsilon}\left(\sqrt{ca_2} + \sup_{(\mathbf{y},\mathbf{X})\in K} \|\mathbf{y}\|\right) < \infty.$$

This proves part (iii). □

Proof of Remark 4.1

Proof. Suppose that (C2$_\epsilon$) holds and suppose that $\omega_\epsilon = 0$. Then there exists a sequence $(J_n, \boldsymbol{\alpha}_n, \ell_n)$, with $J_n \subset \mathbb{R}^k \times \mathcal{X}$, $\boldsymbol{\alpha}_n \in \mathbb{R}^{k+kq}$, $\|\boldsymbol{\alpha}_n\| = 1$, and $\ell_n \in \mathbb{R}$, such that $P(J_n) \geq \epsilon$, for all $n = 1, 2, \ldots$, and

$$\sup_{\mathbf{s}\in J_n} |\boldsymbol{\alpha}_n^T\mathbf{s} - \ell_n| \to 0.$$

This means that for n sufficiently large $\delta_n = 2\sup_{\mathbf{s}\in J_n} |\boldsymbol{\alpha}_n^T\mathbf{s} - \ell_n| < \delta_\epsilon$. Then, there exists a strip $H(\boldsymbol{\alpha}_n, \ell_n, \delta_n)$ containing J_n, such that $\delta_n < \delta_\epsilon$ and $P(H(\boldsymbol{\alpha}_n, \ell_n, \delta_n) \cap (\mathbb{R}^k \times \mathcal{X})) \geq P(J_n) \geq \epsilon$. This would be in contradiction with the definition of δ_ϵ in (C2$_\epsilon$). On the other hand, suppose that (4.3) holds and suppose that $\delta_\epsilon = 0$. That means that we can find a sequence $(\boldsymbol{\alpha}_n, \ell_n, \delta_n)$, with $\boldsymbol{\alpha}_n \in \mathbb{R}^k \times \mathbb{R}^{kq}$, $\|\boldsymbol{\alpha}_n\| = 1$, and $\ell_n \in \mathbb{R}$, such that $\delta_n \downarrow 0$ and $P(H(\boldsymbol{\alpha}_n, \ell_n, \delta_n) \cap (\mathbb{R}^k \times \mathcal{X})) \geq \epsilon$, for all $n = 1, 2, \ldots$. This means that for n sufficiently large $\delta_n < 2\omega_\epsilon$ and then

$$\sup_{\mathbf{s}\in H(\boldsymbol{\alpha}_n,\ell_n,\delta_n)\cap\mathbb{R}^k\times\mathcal{X}} |\boldsymbol{\alpha}_n^T\mathbf{s} - \ell_n| = \delta_n/2 < \omega_\epsilon,$$

which would contradict definition (4.3). □

Lemma B.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 $t \to \infty$. Let $\boldsymbol{\xi}_t = (\boldsymbol{\beta}_t, \boldsymbol{\theta}_t)$, $t \geq 0$, be a sequence in $\mathbb{R}^q \times \mathbb{R}^l$, such that $\boldsymbol{\xi}_t \to \boldsymbol{\xi}_L$, as $t \to \infty$. If $g(\mathbf{s}, \boldsymbol{\xi}) = \rho(d(\mathbf{s}, \boldsymbol{\xi})/\alpha)$, for some $\alpha > 0$ fixed, where ρ satisfies (R2)-(R3) and $\mathbf{V}$ satisfies (V1), then*

$$\lim_{t\to\infty} \int g(\mathbf{s}, \boldsymbol{\xi}_t)\, \mathrm{d}P_t(\mathbf{s}) = \int g(\mathbf{s}, \boldsymbol{\xi}_L)\, \mathrm{d}P(\mathbf{s}).$$

Proof. Let $g_t(\mathbf{s}) = g(\mathbf{s}, \boldsymbol{\xi}_t)$ and $g_L(\mathbf{s}) = g(\mathbf{s}, \boldsymbol{\xi}_L)$. Then for every sequence $\{\mathbf{s}_t\}$, such that $\mathbf{s}_t \to \mathbf{s}$, we have

$$\lim_{t\to\infty} g_t(\mathbf{s}_t) = g_L(\mathbf{s}).$$

Now, apply Theorem 5.5 from [2]. Let $\Gamma : [0, \infty) \to [0, \infty)$ be the function

$$\Gamma(u) = u1_{[0,a_0]}(u) + a_0 1_{(a_0,\infty)}(u),$$

which is bounded and uniformly continuous. Then as a consequence of $P_t \to P$ weakly, we have

$$\lim_{t\to\infty} \int g(\mathbf{s}, \boldsymbol{\xi}_t)\, \mathrm{d}P_t(\mathbf{s}) = \lim_{t\to\infty} \int \Gamma(g_t(\mathbf{s}))\, \mathrm{d}P_t(\mathbf{s}) = \int \Gamma(g_L(\mathbf{s}))\, \mathrm{d}P(\mathbf{s}) = \int g(\mathbf{s}, \boldsymbol{\xi}_L)\, \mathrm{d}P(\mathbf{s}).$$

□

B.2 Proofs of Section 5

Proof of Theorem 3

Proof. Davies [6] defines location-scale S-estimators by means of a function $\kappa : [0, \infty) \to [0, 1]$. It relates to our ρ-function as $\rho(d) = a_0(1 - \kappa(d^2))$. The S-minimization problem considered in [6] can be formulated in our notation as follows

$$\begin{aligned} &\min_{\boldsymbol{\alpha}, \mathbf{A}} \det(\mathbf{A}) \\ &\text{subject to} \\ &\int \rho\left(\sqrt{(\mathbf{y} - \boldsymbol{\alpha})^T \mathbf{A}^{-1} (\mathbf{y} - \boldsymbol{\alpha})}\right) f_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\mathbf{y}) \, \mathrm{d}\mathbf{y} \leq b_0, \end{aligned} \tag{B.9}$$

where the minimum is taken over all $\boldsymbol{\alpha} \in \mathbb{R}^k$ and $\mathbf{A} \in \mathrm{PDS}(k)$. The conditions (R1)-(R2) imply the conditions on $\kappa(s) = 1 - \rho(\sqrt{s})/a_0$ imposed in [6], and κ and h have a common point of decrease. It then follows from Theorem 1 in [6] that (B.9) has a unique solution

$$(\boldsymbol{\alpha}^*, \mathbf{A}^*) = (\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{X}\boldsymbol{\beta}_0, \mathbf{V}(\boldsymbol{\theta}_0)).$$

Since this solution is unique, candidate solutions to (B.9) must be of the form $(\boldsymbol{\alpha}, \mathbf{A}) = (\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\theta}))$, for some $\boldsymbol{\beta} \in \mathbb{R}^q$ and $\boldsymbol{\theta} \in \mathbb{R}^l$. It follows that minimization problem (B.9) is equivalent to minimization problem (5.2). As a consequence, minimization problem (5.2) has a unique solution $(\boldsymbol{\beta}^*, \boldsymbol{\theta}^*)$ for which $\mathbf{X}\boldsymbol{\beta}^* = \mathbf{X}\boldsymbol{\beta}_0$ and $\mathbf{V}(\boldsymbol{\theta}^*) = \mathbf{V}(\boldsymbol{\theta}_0)$. Since $\mathbf{X}^T\mathbf{X}$ is non-singular we can multiply $\mathbf{X}\boldsymbol{\beta}^* = \mathbf{X}\boldsymbol{\beta}_0$ from the left by $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. It then follows that $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$. Finally, from (2.9) we find that $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$. This proves the theorem. □

B.3 Proofs for Section 6

Proof of Theorem 4

Proof. Without loss of generality we may assume that $c_0 = 1$ and $\sup \rho = 1$, so that $r = b_0$. The first step is to show that for both estimators $\epsilon_n^* \geq \lceil nr \rceil / n$. To this end, consider a collection $\mathcal{S}'_m$ obtained from the original collection $\mathcal{S}_n$ by replacing at most $m = \lceil nr \rceil - 1$ number of points in $\mathbb{R}^k \times \mathcal{X}$. We must show that at least one solution $(\boldsymbol{\beta}_n(\mathcal{S}'_m), \boldsymbol{\theta}_n(\mathcal{S}'_m))$ to the S-minimization problem (3.1) exists for the corrupted collection $\mathcal{S}'_m$, and that all possible solutions do not break down.

Denote a possible solution to the S-minimization problem (3.1) corresponding to the corrupted collection $\mathcal{S}'_m$, by

$$\boldsymbol{\xi}'_m = (\boldsymbol{\beta}'_m, \boldsymbol{\theta}'_m) = (\boldsymbol{\beta}_n(\mathcal{S}'_m), \boldsymbol{\theta}_n(\mathcal{S}'_m)),$$

and consider the corresponding cylinder $\mathcal{C}(\boldsymbol{\beta}'_m, \boldsymbol{\theta}'_m, 1)$ defined by (3.3). We apply Lemma 1 to the empirical measure $\mathbb{P}'_m$ corresponding to the corrupted collection $\mathcal{S}'_m$ of n points. Because $\boldsymbol{\xi}'_m = (\boldsymbol{\beta}'_m, \boldsymbol{\theta}'_m)$ must satisfy the S-constraint in (3.1), one can argue as in (A.1),

$$\begin{aligned} \mathbb{P}'_m(\mathcal{C}(\boldsymbol{\beta}'_m, \boldsymbol{\theta}'_m, 1)) &= \frac{1}{n} \sum_{\mathbf{s}_i \in S'_m} \mathbb{1}\left\{\mathbf{s}_i \in \mathcal{C}(\boldsymbol{\beta}'_m, \boldsymbol{\theta}'_m, 1)\right\} \\ &\geq 1 - \frac{1}{n} \sum_{\mathbf{s}_i \in S'_m} \rho\left(d(\mathbf{s}_i, \boldsymbol{\xi}'_m)\right) \geq 1 - b_0, \end{aligned}$$

where $d(\mathbf{s}_i, \boldsymbol{\xi}'_m)$ is defined in (5.1). It follows that the cylinder $\mathcal{C}(\boldsymbol{\beta}'_m, \boldsymbol{\theta}'_m, 1)$ must contain at least $\lceil n - nb_0 \rceil = \lceil n - nr \rceil$ number of points from the corrupted collection $\mathcal{S}'_m$. Furthermore, since $r \leq (n - \kappa(\mathcal{S}_n))/(2n)$, for any such subset of $\mathcal{S}'_m$ it holds that it contains

$$\lceil n - nr \rceil - m = n - \lfloor nr \rfloor - \lceil nr \rceil + 1 \geq \kappa(\mathcal{S}_n) + 1 \tag{B.10}$$

points of the original collection $\mathcal{S}_n$. It follows that the measure $\mathbb{P}'_m$ satisfies condition (C2$_\epsilon$), for $\epsilon = (\kappa(\mathcal{S}_n)+1)/n$ and with the value $\delta_\epsilon > 0$ only depending on the original collection $\mathcal{S}_n$. Moreover, we also have that $\mathbb{P}'_m(\mathcal{C}(\boldsymbol{\beta}'_m, \boldsymbol{\theta}'_m, 1)) \geq \epsilon$, for $\epsilon = (\kappa(\mathcal{S}_n)+1)/n > 0$. According to Lemma 1(i), it then follows that $\lambda_k(\mathbf{V}(\boldsymbol{\theta}'_m)) \geq a_1 > 0$, where a_1 only depends on the original collection $\mathcal{S}_n$.

Because $nb_0 - m = nr - \lceil nr \rceil + 1 > 0$ and

$$\lim_{R\to\infty} \sum_{(\mathbf{y}_i,\mathbf{X}_i)\in\mathcal{S}_n} \rho\left(\frac{\|\mathbf{y}_i\|}{R}\right) = 0,$$

we can find an $R_0 > 0$, only depending on the original collection $\mathcal{S}_n$, such that

$$\sum_{(\mathbf{y}_i,\mathbf{X}_i)\in\mathcal{S}_n} \rho\left(\frac{\|\mathbf{y}_i\|}{R_0}\right) \leq nb_0 - m.$$

The collection $\mathcal{S}'_m$ contains $n-m$ points of the original collection $\mathcal{S}_n$. Consider the smallest $M > 0$, such that

$$\sum_{(\mathbf{y}_i,\mathbf{X}_i)\in\mathcal{S}'_m\cap\mathcal{S}_n} \rho\left(\frac{\|\mathbf{y}_i\|}{M}\right) \leq nb_0 - m.$$

Because $\mathcal{S}'_m \cap \mathcal{S}_n$ has less points than $\mathcal{S}_n$, it holds that $M \leq R_0$. It follows that

$$\begin{aligned}
\int \rho\left(\frac{\|\mathbf{y}\|}{R_0}\right) \mathrm{d}\mathbb{P}'_m(\mathbf{s}) &= \frac{1}{n} \sum_{(\mathbf{y}_i,\mathbf{X}_i)\in\mathcal{S}'_m} \rho\left(\frac{\|\mathbf{y}_i\|}{R_0}\right) \\
&\leq \frac{1}{n} \sum_{(\mathbf{y}_i,\mathbf{X}_i)\in S'_m} \rho\left(\frac{\|\mathbf{y}_i\|}{M}\right) \\
&\leq \frac{1}{n}\left(\sum_{(\mathbf{y}_i,\mathbf{X}_i)\in S'_m\cap S_n} \rho\left(\frac{\|\mathbf{y}_i\|}{M}\right) + m\right) \leq b_0.
\end{aligned}$$

According to Lemma 1(ii), it then follows that $\lambda_1(\mathbf{V}(\boldsymbol{\theta}'_m)) \leq a_2 < \infty$. where a_2 only depends on a_1 and the collection $\mathcal{S}_n$.

To show that the estimate $\boldsymbol{\beta}'_m$ stays bounded, recall that the cylinder $\mathcal{C}(\boldsymbol{\beta}'_m, \boldsymbol{\theta}'_m, 1)$ contains a subset J_0 of $\kappa(\mathcal{S}_n)+1$ points from the original collection $\mathcal{S}_n$, according to (B.10). By definition, $\kappa(\mathcal{S}_n)+1$ original points cannot be on the same hyperplane, so that

$$\gamma_n = \inf_{J\subset\mathcal{S}_n} \inf_{\|\boldsymbol{\gamma}\|=1} \max_{\mathbf{s}\in J} \|\mathbf{X}\boldsymbol{\gamma}\| > 0.$$

where the first infimum runs over all subsets $J \subset \mathcal{S}_n$ of $\kappa(\mathcal{S}_n)+1$ points. By definition of γ_n, there exists an original point $\mathbf{s}_0 \in J_0 \subset \mathcal{S}_n \cap \mathcal{C}(\boldsymbol{\beta}'_m, \boldsymbol{\theta}'_m, 1)$, such that

$$\|\boldsymbol{\beta}'_m\| = \|\mathbf{X}_0\boldsymbol{\beta}'_m\| \times \frac{\|\boldsymbol{\beta}'_m\|}{\|\mathbf{X}_0\boldsymbol{\beta}'_m\|} \leq \frac{1}{\gamma_n}\|\mathbf{X}_0\boldsymbol{\beta}'_m\|.$$

Because $\mathbf{s}_0 \in \mathcal{C}(\boldsymbol{\beta}'_m, \boldsymbol{\theta}'_m, 1)$, similar to the proof of Lemma 1(iii), it follows that $\|\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta}'_m\|^2 \leq a_2$, and because $\mathbf{s}_0 \in \mathcal{S}_n$, we have that

$$\|\mathbf{X}_0\boldsymbol{\beta}'_m\| \leq \sqrt{a_2} + \max_{(\mathbf{y}_i,\mathbf{X}_i)\in\mathcal{S}_n} \|\mathbf{y}_i\| < \infty.$$

We conclude that there exists a compact set K_n, only depending on the original collection $\mathcal{S}_n$, that contains the pair $(\boldsymbol{\beta}'_m, \mathbf{V}(\boldsymbol{\theta}'_m))$. Similar to the reasoning in the proof of Theorem 1, it follows that at least one solution $(\boldsymbol{\beta}_n(\mathcal{S}'_m), \boldsymbol{\theta}_n(\mathcal{S}'_m))$ to the S-minimization problem (3.1) exists for the collection $\mathcal{S}'_m$, and that all possible solutions $\boldsymbol{\beta}_n(\mathcal{S}'_m)$ and $\boldsymbol{\theta}_n(\mathcal{S}'_m)$ do not break down.

We continue by showing $\epsilon_n^*(\boldsymbol{\theta}_n) \leq \lceil nr \rceil / n$. Replace $m = \lceil nr \rceil$ points of $\mathcal{S}_n$ to obtain a corrupted collection $\mathcal{S}'_m$ of n points. Suppose that a solution $\boldsymbol{\xi}'_m = (\boldsymbol{\beta}'_m, \boldsymbol{\theta}'_m) = (\boldsymbol{\beta}_n(\mathcal{S}'_m), \boldsymbol{\theta}_n(\mathcal{S}'_m))$ exists to the S-minimization problem (3.1) corresponding to the corrupted collection $\mathcal{S}'_m$. We must show that the estimate $\boldsymbol{\theta}'_m$ breaks down. Note that the estimates $\boldsymbol{\beta}'_m$ and $\boldsymbol{\theta}'_m$ satisfy the S-constraint in (3.1) for the corrupted collection,

$$\sum_{\mathbf{s}_i \in \mathcal{S}'_m} \rho\left(d(\mathbf{s}_i, \boldsymbol{\xi}'_m)\right) \leq nr. \tag{B.11}$$

If all $m = \lceil nr \rceil$ replaced points are outside the cylinder $\mathcal{C}(\boldsymbol{\beta}'_m, \boldsymbol{\theta}'_m, 1)$, then

$$\sum_{\mathbf{s}_i \in \mathcal{S}'_m} \rho\left(d(\mathbf{s}_i, \boldsymbol{\xi}'_m)\right) = \sum_{\mathbf{s}_i \in \mathcal{S}'_m \cap \mathcal{S}_n} \rho\left(d(\mathbf{s}_i, \boldsymbol{\xi}'_m)\right) + \lceil nr \rceil > nr,$$

when $nr \notin \mathbb{N}$. When $nr \in \mathbb{N}$, then by assumption $n - m = n - \lceil nr \rceil \geq \kappa(\mathcal{S}_n) + 1$. Hence, there is at least one point $\mathbf{s}_i \in \mathcal{S}_n$, for which $d(\mathbf{s}_i, \boldsymbol{\xi}'_m) > 0$. Because ρ is strictly increasing on $[0, 1]$, this implies

$$\sum_{\mathbf{s}_i \in \mathcal{S}'_m} \rho\left(d(\mathbf{s}_i, \boldsymbol{\xi}'_m)\right) > nr.$$

We conclude that at least one replaced point $\mathbf{s}'_i = (\mathbf{y}'_i, \mathbf{X}'_i) \in \mathcal{C}(\boldsymbol{\beta}'_m, \boldsymbol{\theta}'_m, 1)$. Similarly, if all original points in $\mathcal{S}'_m$ are outside $\mathcal{C}(\boldsymbol{\beta}'_m, \boldsymbol{\theta}'_m, 1)$, then

$$\sum_{\mathbf{s}_i \in \mathcal{S}'_m} \rho\left(d(\mathbf{s}_i, \boldsymbol{\xi}'_m)\right) \geq n - \lceil nr \rceil \geq n - \lceil n - nr \rceil + \kappa(\mathcal{S}_n) > nr,$$

which is in contradiction with (B.11). Therefore, the cylinder $\mathcal{C}(\boldsymbol{\beta}'_m, \boldsymbol{\theta}'_m, 1)$ must contain a point $\mathbf{s}_0 = (\mathbf{y}_0, \mathbf{X}_0)$ from the original collection $\mathcal{S}_n$ as well as one replaced point $\mathbf{s}'_i = (\mathbf{y}'_i, \mathbf{X}'_i)$.

Note that for each point $\mathbf{s} = (\mathbf{y}, \mathbf{X}) \in \mathcal{C}(\boldsymbol{\beta}'_m, \boldsymbol{\theta}'_m, 1)$ it holds that

$$\frac{|\mathbf{q}_1^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}'_m)|^2}{\lambda_1(\mathbf{V}(\boldsymbol{\theta}'_m))} + \cdots + \frac{|\mathbf{q}_k^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}'_m)|^2}{\lambda_k(\mathbf{V}(\boldsymbol{\theta}'_m))} \leq 1, \tag{B.12}$$

where $\lambda_j(\mathbf{V}(\boldsymbol{\theta}'_m)) > 0$, for $j = 1, \ldots, k$, are the eigenvalues of $\mathbf{V}(\boldsymbol{\theta}'_m)$ and $\mathbf{q}_1, \ldots, \mathbf{q}_k$ are the corresponding orthonormal eigenvectors. Now, replace all m points by $\mathbf{s}'_i = (\mathbf{y}'_i, \mathbf{X}'_i) = (\mathbf{z}, \mathbf{0})$, where $\mathbf{0}$ is a $k \times q$ matrix of zeros and $\mathbf{z} = t \sum_{j=1}^k \mathbf{q}_j$, so that $\mathbf{q}_j^T \mathbf{z} = t$, for each $j = 1, \ldots, k$. By sending $t \to \infty$ and the fact that at least one replaced point $(\mathbf{z}, \mathbf{0})$ satisfies (B.12), it follows that $\lambda_j(\mathbf{V}(\boldsymbol{\theta}'_m)) \to \infty$, for each $j = 1, \ldots, k$. This means $\boldsymbol{\theta}'_m$ breaks down.

The upper bound for $\epsilon_n^*(\boldsymbol{\beta}_n, \mathcal{S}_n)$ follows from the fact that $\boldsymbol{\beta}_n$ is regression equivariant. Similar to Theorem 2 in [18] it can be shown that the maximal breakdown point of regression equivariant estimators is $\lfloor (n+1)/2 \rfloor / n$. This proves the theorem. □

B.4 Proofs for Section 8

Lemma B.2. *Suppose that ρ satisfies (R2), (R4) and $\mathbf{V}$ satisfies (V4). Let $\Psi = (\Psi_{\boldsymbol{\beta}}, \Psi_{\boldsymbol{\theta}})$, as defined in (7.5). Then there exist $0 < C_1 < \infty$ and $0 < C_2 < \infty$, only depending on P, such that $\|\Psi_{\boldsymbol{\beta}}(\mathbf{s}, \boldsymbol{\xi}(P))\| \leq C_1 \|\mathbf{X}\|$ and $\|\Psi_{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\xi}(P))\| \leq C_2$.*

Proof. Consider the expression of $\Psi_{\boldsymbol{\beta}}$ in (7.5). Consecutively, we apply (B.3), (B.1), (B.5), (B.6), and (B.2). This gives

$$\begin{aligned}
\left\|\mathbf{X}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\|^2 &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\mathbf{X}\mathbf{X}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
&\leq \lambda_1(\mathbf{V}^{-1/2}\mathbf{X}\mathbf{X}^T\mathbf{V}^{-1/2})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
&= d^2 \left\|\mathbf{V}^{-1/2}\mathbf{X}\mathbf{X}^T\mathbf{V}^{-1/2}\right\|_2 \\
&\leq d^2 \left\|\mathbf{V}^{-1/2}\right\|_2^2 \left\|\mathbf{X}\mathbf{X}^T\right\|_2 \\
&\leq d^2 \|\mathbf{X}\|^2 \lambda_1(\mathbf{V}^{-1}),
\end{aligned} \tag{B.13}$$

where $d = d(\mathbf{s}, \boldsymbol{\xi})$ as defined by (5.1), and where we abbreviate $\mathbf{V}(\boldsymbol{\theta})$ by $\mathbf{V}$. This means that

$$\|\Psi_{\boldsymbol{\beta}}(\mathbf{s}, \boldsymbol{\xi}(P))\| \leq |u(d)d| \times \|\mathbf{X}\| \times \sqrt{\lambda_1(\mathbf{V}(\boldsymbol{\theta}(P))^{-1})}.$$

From (R2) and (R4) it follows that $u(s)s = \rho'(s)$ is bounded. This means that there exists a universal constant $0 < C_1 < \infty$, such that

$$\|\Psi_{\boldsymbol{\beta}}(\mathbf{s}, \boldsymbol{\xi}(P))\| \leq C_1 \|\mathbf{X}\|.$$

For $\Psi_{\boldsymbol{\theta}} = (\Psi_{\boldsymbol{\theta},1}, \cdots, \Psi_{\boldsymbol{\theta},l})$, we have

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

for $j = 1, \ldots, l$, where we write $\mathbf{V}$ instead of $\mathbf{V}(\boldsymbol{\theta})$. Recall that

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

To bound $\mathbf{H}_j$, we first obtain a bound on

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \theta_j} \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{B.14}$$

Note that $\partial \mathbf{V}/\partial \theta_j$ is symmetric, but not necessarily positive definite. When (B.14) is positive, then application of (B.1) gives

$$\begin{aligned} 0 < (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \theta_j} \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) &\leq d^2 \lambda_1 \left(\mathbf{V}^{-1/2} \frac{\partial \mathbf{V}}{\partial \theta_j} \mathbf{V}^{-1/2}\right) \\ &\leq d^2 \left\|\mathbf{V}^{-1/2}\right\|^2 \left\|\frac{\partial \mathbf{V}}{\partial \theta_j}\right\| \\ &\leq d^2 \left\|\frac{\partial \mathbf{V}}{\partial \theta_j}\right\| \lambda_1(\mathbf{V}^{-1}), \end{aligned}$$

according to (B.7) and (B.5). When (B.14) is negative, then similarly

$$\begin{aligned} 0 < (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} \left(-\frac{\partial \mathbf{V}}{\partial \theta_j}\right) \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) &\leq d^2 \lambda_1 \left(\mathbf{V}^{-1/2} \left(-\frac{\partial \mathbf{V}}{\partial \theta_j}\right) \mathbf{V}^{-1/2}\right) \\ &\leq d^2 \left\|\frac{\partial \mathbf{V}}{\partial \theta_j}\right\| \lambda_1(\mathbf{V}^{-1}). \end{aligned}$$

It follows that

$$\left|(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \theta_j} \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right| \leq d^2 \left\|\frac{\partial \mathbf{V}}{\partial \theta_j}\right\| \lambda_1(\mathbf{V}^{-1}). \tag{B.15}$$

Furthermore, according to (V4), the mapping $\boldsymbol{\theta} \mapsto \mathbf{V}(\boldsymbol{\theta})$ is continuously differentiable. This means that there exists a universal constant $0 < M_1 < \infty$, such that

$$\max_{1 \leq j \leq l} \left\|\frac{\partial \mathbf{V}(\boldsymbol{\theta}(P))}{\partial \theta_j}\right\| \leq M_1. \tag{B.16}$$

Because $\partial \mathbf{V}/\partial \theta_j$ is symmetric, according to (B.5) and (B.6)

$$\left|\lambda_1 \left(\mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \theta_j}\right)\right| \leq \left\|\mathbf{V}^{-1}\right\|_2 \left\|\frac{\partial \mathbf{V}}{\partial \theta_j}\right\|_2 \leq \sqrt{k} \lambda_1 \left(\mathbf{V}^{-1}\right) \left\|\frac{\partial \mathbf{V}}{\partial \theta_j}\right\|.$$

Together with (B.16), we find

$$\left|\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\right)\right| \leq k\left|\lambda_1\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\right)\right| \leq k^{3/2} M_1 \lambda_1\left(\mathbf{V}^{-1}\right), \tag{B.17}$$

where we abbreviate $\mathbf{V}(\boldsymbol{\theta}(P))$ by $\mathbf{V}$. It then follows there exists a constant $0 < M_2 < \infty$, only depending on P, such that at $\boldsymbol{\theta}(P)$,

$$\max_{1\leq j\leq l}\left|(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\mathbf{H}_j\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right| \leq d^2 M_2. \tag{B.18}$$

From (R2) and (R4) it follows that $u(s)s^2 = \rho'(s)s$ and $\rho(s) - b_0$ are bounded. Together with (B.17), it follows that there exists a universal constant $0 < C_2 < \infty$, such that

$$\|\Psi_{\boldsymbol{\theta},j}(\mathbf{s}, \boldsymbol{\xi}(P))\| \leq C_2,$$

for all $j = 1, \ldots, l$. This finishes the proof. □

Proof of Corollary 4

Proof. Take $\mathbf{s} = (\mathbf{y}, \mathbf{X})$ fixed and consider $\mathrm{IF}(\mathbf{s}; \boldsymbol{\xi}, P)$. Since $\mathbf{D}(P)$ does not depend on $\mathbf{s}$, from Theorem 5 and Lemma B.2, it follows immediately that $\mathrm{IF}(\mathbf{s}; \boldsymbol{\xi}, P)$ remains bounded in $\mathbf{y}$, but not necessarily in $\mathbf{X}$. □

Lemma B.3. *Consider Λ as defined by (8.2) with Ψ defined in (7.5). Suppose that ρ satisfies (R2) and (R5) and $\mathbf{V}$ satisfies (V5). Furthermore, suppose that $\mathbb{E}\|\mathbf{X}\|^2 < \infty$. Let $\boldsymbol{\xi}(P)$ be a solution to (3.5) and let N be an open neighborhood of $\boldsymbol{\xi}(P)$. Then, Λ is continuous differentiable at $\boldsymbol{\xi}(P)$ and for all $\boldsymbol{\xi} \in N$,*

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

Proof. Write $\partial\Lambda/\partial\boldsymbol{\xi}$ as the block matrix

$$\frac{\partial \Lambda(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = \begin{pmatrix} \dfrac{\partial \Lambda_{\boldsymbol{\beta}}(\boldsymbol{\xi})}{\partial \boldsymbol{\beta}} & \dfrac{\partial \Lambda_{\boldsymbol{\beta}}(\boldsymbol{\xi})}{\partial \boldsymbol{\theta}} \\ \dfrac{\partial \Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi})}{\partial \boldsymbol{\beta}} & \dfrac{\partial \Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi})}{\partial \boldsymbol{\theta}} \end{pmatrix}, \tag{B.19}$$

where

$$\Lambda_{\boldsymbol{\beta}}(\boldsymbol{\xi}) = \int \Psi_{\boldsymbol{\beta}}(\mathbf{s}, \boldsymbol{\xi})\, \mathrm{d}P(\mathbf{s}),$$
$$\Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}) = \int \Psi_{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\xi})\, \mathrm{d}P(\mathbf{s}).$$

We prove the lemma for each block separately. Consider $\partial\Lambda_{\boldsymbol{\beta}}/\partial\boldsymbol{\beta}$. We have

$$\frac{\partial \Psi_{\boldsymbol{\beta}}(\mathbf{s}, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}} = -\frac{u'(d)}{d}\mathbf{X}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\mathbf{X} - u(d)\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}, \tag{B.20}$$

where $d = d(\mathbf{s}, \boldsymbol{\xi})$ is defined by (5.1) and where we abbreviate $\mathbf{V}(\boldsymbol{\theta})$ by $\mathbf{V}$. First note that, according to (R5), $\boldsymbol{\xi} \mapsto u'(d(\mathbf{s}, \boldsymbol{\xi}))/d(\mathbf{s}, \boldsymbol{\xi})$ is continuous at $\boldsymbol{\xi}(P)$, for each $\mathbf{s}$ fixed such that $d(\mathbf{s}, \boldsymbol{\xi}(P)) \neq 0$. Together with (V1), this means that for such $\mathbf{s}$ fixed, $\boldsymbol{\xi} \mapsto \partial\Psi_{\boldsymbol{\beta}}(\mathbf{s}, \boldsymbol{\xi})/\partial\boldsymbol{\beta}$ is continuous at $\boldsymbol{\xi}(P)$. For the first term on the right hand side of (B.20), we apply (B.4) and (B.13). This gives

$$\left\|\mathbf{X}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\mathbf{X}\right\| = \left\|\mathbf{X}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\|^2 \leq d^2\|\mathbf{X}\|^2\lambda_1(\mathbf{V}^{-1}).$$

Similarly, for the second term on the right hand side of (B.20), after application of (B.7) and (B.5), we get

$$\left\|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right\| \leq \|\mathbf{X}\|^2 \left\|\mathbf{V}^{-1}\right\| \leq \sqrt{q}\|\mathbf{X}\|^2\lambda_1(\mathbf{V}^{-1}).$$

Since $\lambda_1(\mathbf{V}^{-1})$ is bounded uniformly on the neighborhood N of $\boldsymbol{\xi}(P)$ and because $u(s)$ and $u'(s)s = \rho''(s) - u(s)$ are bounded, due to (R2), it follows that there exists a constant $0 < C_1 < \infty$, only depending on P, such that

$$\left\|\frac{\partial\Psi_{\boldsymbol{\beta}}(\mathbf{s},\boldsymbol{\xi})}{\partial\boldsymbol{\beta}^T}\right\| \leq C_1\|\mathbf{X}\|^2.$$

Since $\mathbb{E}\|\mathbf{X}\|^2 < \infty$, it follows by dominated convergence that for $\boldsymbol{\xi}$ in the neighborhood N of $\boldsymbol{\xi}(P)$, it holds that

$$\frac{\partial\Lambda_{\boldsymbol{\beta}}(\boldsymbol{\xi})}{\partial\boldsymbol{\beta}} = \int \frac{\partial\Psi_{\boldsymbol{\beta}}(\mathbf{s},\boldsymbol{\xi})}{\partial\boldsymbol{\beta}}\,\mathrm{d}P(\mathbf{s}), \tag{B.21}$$

and that $\partial\Lambda_{\boldsymbol{\beta}}/\partial\boldsymbol{\beta}$ is continuous at $\boldsymbol{\xi}(P)$.

Next consider $\partial\Psi_{\boldsymbol{\beta}}/\partial\boldsymbol{\theta}$. For each $j = 1,\ldots,l$ fixed, we have

$$\begin{aligned}\frac{\partial\Psi_{\boldsymbol{\beta}}(\mathbf{s},\boldsymbol{\xi})}{\partial\theta_j} = \frac{u'(d)}{2d}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\cdot\mathbf{X}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\\ + u(d)\cdot\mathbf{X}^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}).\end{aligned} \tag{B.22}$$

First note that, similar to (B.20), $\boldsymbol{\xi}\mapsto\partial\Psi_{\boldsymbol{\beta}}(\mathbf{s},\boldsymbol{\xi})/\partial\theta_j$ is continuous at $\boldsymbol{\xi}(P)$, for each $\mathbf{s}$ fixed such that $d(\mathbf{s},\boldsymbol{\xi}(P)) \neq 0$. Consider the first term on the right hand side of (B.22). From (B.13), we have

$$\|\mathbf{X}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\| \leq d\|\mathbf{X}\|\sqrt{\lambda_1(\mathbf{V}^{-1})}.$$

Moreover, similar to the reasoning in (B.15), we find

$$\left|(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right| \leq d^2\left\|\frac{\partial\mathbf{V}}{\partial\theta_j}\right\|\lambda_1(\mathbf{V}^{-1}). \tag{B.23}$$

For the second term on the right hand side of (B.22), similar to the reasoning in (B.13), we have

$$\begin{aligned}\left\|\mathbf{X}^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\|^2 &= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{V}^{-1}\mathbf{X}\mathbf{X}^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\\ &\leq d^2\lambda_1\left(\mathbf{V}^{-1/2}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{V}^{-1}\mathbf{X}\mathbf{X}^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{V}^{-1/2}\right)\\ &\leq d^2\left\|\mathbf{V}^{-1/2}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{V}^{-1}\mathbf{X}\mathbf{X}^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{V}^{-1/2}\right\|_2\\ &\leq d^2\|\mathbf{X}\|^2\left\|\frac{\partial\mathbf{V}}{\partial\theta_j}\right\|^2\lambda_1(\mathbf{V}^{-1})^3.\end{aligned}$$

According to (V4), the mapping $\mathbf{V}(\boldsymbol{\theta})$ is continuously differentiable. This means that $\|\partial\mathbf{V}/\partial\theta_j\|$ is bounded on the neighborhood N of $\boldsymbol{\xi}(P)$. Since $\lambda_1(\mathbf{V}^{-1})$ is bounded uniformly on N and because $u(s)s = \rho'(s)$ and $u'(s)s^2 = \rho''(s)s - \rho'(s)$ are bounded, it follows that there exists a constant $0 < C_2 < \infty$, only depending on P, such that for $j = 1,\ldots,l$,

$$\left\|\frac{\partial\Psi_{\boldsymbol{\beta}}(\mathbf{s},\boldsymbol{\xi})}{\partial\theta_j}\right\| \leq C_2\|\mathbf{X}\|^2.$$

As before, it follows by dominated convergence that for $\boldsymbol{\xi}$ in the neighborhood N of $\boldsymbol{\xi}(P)$, it holds that

$$\frac{\partial\Lambda_{\boldsymbol{\beta}}(\boldsymbol{\xi})}{\partial\boldsymbol{\theta}} = \int \frac{\partial\Psi_{\boldsymbol{\beta}}(\mathbf{s},\boldsymbol{\xi})}{\partial\boldsymbol{\theta}}\,\mathrm{d}P(\mathbf{s}), \tag{B.24}$$

and that $\partial\Lambda_{\boldsymbol{\beta}}/\partial\boldsymbol{\theta}$ is continuous at $\boldsymbol{\xi}(P)$.

Next consider $\partial\Psi_{\boldsymbol{\theta},j}/\partial\boldsymbol{\beta}$, for $j = 1, \ldots, l$. We have

$$\begin{aligned}\frac{\partial\Psi_{\boldsymbol{\theta},j}(\mathbf{s},\boldsymbol{\xi})}{\partial\boldsymbol{\beta}} = \frac{u'(d)}{d}&\mathbf{X}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\cdot(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\mathbf{H}_j\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\\ &- u(d)\cdot 2\mathbf{X}^T\mathbf{V}^{-1}\mathbf{H}_j\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\\ &+ \operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\right)u(d)\mathbf{X}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}),\end{aligned} \tag{B.25}$$

where $\mathbf{H}_j$ is defined in (7.3). As before, $\boldsymbol{\xi}\mapsto\partial\Psi_{\theta_j}(\mathbf{s},\boldsymbol{\xi})/\partial\boldsymbol{\beta}$ is continuous at $\boldsymbol{\xi}(P)$, for each $\mathbf{s}$ fixed such that $d(\mathbf{s},\boldsymbol{\xi}(P))\neq 0$. Consider the first term on the right hand side of (B.25). From (B.18),

$$\left|(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\mathbf{H}_j\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right| \leq d^2 M_2. \tag{B.26}$$

Because $u'(s)s^2$ is bounded, together with (B.13), the norm of the first term on the right hand side of (B.25) is bounded by a constant times $\|\mathbf{X}\|\lambda_1(\mathbf{V}^{-1})^{1/2}$. Similar to (B.13), for the second term on the right hand side of (B.25),

$$\begin{aligned}\|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{H}_j\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\|^2 &\leq d^2\left\|\mathbf{V}^{-1/2}\mathbf{H}_j\mathbf{V}^{-1}\right\|_2^2\left\|\mathbf{X}\mathbf{X}^T\right\|_2^2\\ &\leq kd^2\|\mathbf{X}\|^2\|\mathbf{H}_j\|^2\lambda_1(\mathbf{V}^{-1})^3,\end{aligned}$$

and for the third term on the right hand side of (B.25), we can use (B.13) and (B.17). As before, since $u'(s)s = \rho''(s) - u(s)$ and $u(s)s^2 = \rho'(s)s$ are bounded, it follows that there exists a constant $0 < C_3 < \infty$, only depending on P, such that for $j = 1, \ldots, l$,

$$\left\|\frac{\partial\Psi_{\boldsymbol{\theta},j}(\mathbf{s},\boldsymbol{\xi})}{\partial\boldsymbol{\beta}}\right\| \leq C_3\|\mathbf{X}\|.$$

Since $\mathbb{E}\|\mathbf{X}\| < \infty$, if follows by dominated convergence that for $\boldsymbol{\xi}$ in the neighborhood N of $\boldsymbol{\xi}(P)$, it holds that

$$\frac{\partial\Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi})}{\partial\boldsymbol{\beta}} = \int\frac{\partial\Psi_{\boldsymbol{\theta}}(\mathbf{s},\boldsymbol{\xi})}{\partial\boldsymbol{\beta}}\,\mathrm{d}P(\mathbf{s}), \tag{B.27}$$

and that $\partial\Lambda_{\boldsymbol{\theta}}/\partial\boldsymbol{\beta}$ is continuous at $\boldsymbol{\xi}(P)$.

Finally, consider $\partial\Psi_{\boldsymbol{\theta},j}/\partial\theta_t$, for $j, t = 1, \ldots, l$. We find

$$\begin{aligned}\frac{\partial\Psi_{\boldsymbol{\theta},j}}{\partial\theta_t} =& -\frac{u'(d)}{2d}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_t}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\\ &\qquad\cdot(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\mathbf{H}_j\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\\ &- u(d)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_t}\mathbf{V}^{-1}\cdot\mathbf{H}_j\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\\ &+ u(d)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\frac{\partial\mathbf{H}_j}{\partial\theta_t}\cdot\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\\ &- u(d)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\mathbf{H}_j\cdot\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_t}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\\ &+ \operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_t}\mathbf{V}^{-1}\cdot\frac{\partial\mathbf{V}}{\partial\theta_j}\right)(\rho(d)-b_0)\\ &- \operatorname{tr}\left(\mathbf{V}^{-1}\cdot\left(\frac{\partial^2\mathbf{V}}{\partial\theta_j\theta_t}\right)\right)(\rho(d)-b_0)\\ &+ \operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\right)\cdot\frac{u(d)}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_t}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}).\end{aligned} \tag{B.28}$$

As before, together with (V5), $\boldsymbol{\xi}\mapsto\partial\Psi_{\theta_j}(\mathbf{s},\boldsymbol{\xi})/\partial\theta_t$ is continuous at $\boldsymbol{\xi}(P)$, for each $\mathbf{s}$ fixed such that $d(\mathbf{s},\boldsymbol{\xi}(P))\neq 0$. From (B.15) and (B.18), it follows that the first term on the right hand side

of (B.28) is bounded by

$$\frac{|u'(d)d^3|}{2}\left\|\frac{\partial \mathbf{V}}{\partial \theta_t}\right\|\lambda_1(\mathbf{V}^{-1})M_2.$$

Because $u'(s)s^3 = \rho''(s)s^2 - \rho'(s)s$ is bounded, together with (B.16), we conclude this is bounded on the neighborhood N of $\boldsymbol{\xi}(P)$. Similar to (B.13), the second term on the right hand side of (B.28) is bounded by

$$|u(d)|d^2\left\|\mathbf{V}^{-1/2}\frac{\partial \mathbf{V}}{\partial \theta_t}\mathbf{V}^{-1}\mathbf{H}_j\mathbf{V}^{-1/2}\right\|_2^2 \leq |u(d)|d^2\left\|\frac{\partial \mathbf{V}}{\partial \theta_t}\right\|^2\|\mathbf{H}_j\|^2\lambda_1\left(\mathbf{V}^{-1}\right)^4.$$

Since $u(s)s^2 = \rho'(s)s$ is bounded, together with (B.17), we again find that this is bounded on the neighborhood N of $\boldsymbol{\xi}(P)$. The same holds for the fourth term on the right hand side of (B.28).

We continue with the third term on the right hand side of (B.28). Similar to (B.13), this is bounded by

$$|u(d)|d^2\left\|\mathbf{V}^{-1/2}\frac{\partial \mathbf{H}_j}{\partial \theta_t}\mathbf{V}^{-1/2}\right\|_2^2 \leq |u(d)|d^2\lambda_1\left(\mathbf{V}^{-1}\right)^2\left\|\frac{\partial \mathbf{H}_j}{\partial \theta_t}\right\|.$$

We have that

$$\begin{aligned}
\frac{\partial \mathbf{H}_j}{\partial \theta_t} = &\left\{\operatorname{tr}\left(-\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_t}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\right) + \operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial^2 \mathbf{V}}{\partial \theta_j\theta_t}\right)\right\}\left(\sum_{s=1}^{l}\theta_s\frac{\partial \mathbf{V}}{\partial \theta_s}\right)\\
&+\operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\right)\left\{\left(\frac{\partial \mathbf{V}}{\partial \theta_t}\right) + \left(\sum_{s=1}^{l}\theta_s\frac{\partial^2 \mathbf{V}}{\partial \theta_s\theta_t}\right)\right\}\\
&-\left\{\operatorname{tr}\left(-\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_t}\mathbf{V}^{-1}\sum_{s=1}^{l}\theta_s\frac{\partial \mathbf{V}}{\partial \theta_s}\right) + \operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_t}\right) + \operatorname{tr}\left(\mathbf{V}^{-1}\sum_{s=1}^{l}\theta_s\frac{\partial^2 \mathbf{V}}{\partial \theta_s\theta_t}\right)\right\}\frac{\partial \mathbf{V}}{\partial \theta_j}\\
&-\operatorname{tr}\left(\mathbf{V}^{-1}\sum_{s=1}^{l}\theta_s\frac{\partial \mathbf{V}}{\partial \theta_s}\right)\left(\frac{\partial^2 \mathbf{V}}{\partial \theta_j\theta_t}\right).
\end{aligned}$$

With (B.5), (B.6), and (B.16), we find

$$\begin{aligned}
\left|\operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_t}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_s}\right)\right| &\leq k\left\|\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_t}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_s}\right\|_2\\
&\leq k\left\|\mathbf{V}^{-1}\right\|_2^2\left\|\frac{\partial \mathbf{V}}{\partial \theta_t}\right\|\left\|\frac{\partial \mathbf{V}}{\partial \theta_s}\right\|\\
&\leq k\lambda_1(\mathbf{V}^{-1})M_1^2,
\end{aligned} \tag{B.29}$$

which is uniformly bounded on the neighborhood N of $\boldsymbol{\xi}(P)$, and similar to (B.17) we find

$$\left|\operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial^2 \mathbf{V}}{\partial \theta_j\theta_t}\right)\right| \leq k^{3/2}\lambda_1(\mathbf{V}^{-1})\left\|\frac{\partial^2 \mathbf{V}}{\partial \theta_j\theta_t}\right\|. \tag{B.30}$$

Because, according to (V5), the mapping $\boldsymbol{\theta} \mapsto \mathbf{V}(\boldsymbol{\theta})$ is twice continuously differentiable, it follows that $\left\|\partial^2\mathbf{V}/\partial\theta_j\theta_t\right\|$ is uniformly bounded on the neighborhood N of $\boldsymbol{\xi}(P)$. Together with the fact that with (B.16),

$$\left\|\sum_{s=1}^{l}\theta_s\frac{\partial \mathbf{V}}{\partial \theta_s}\right\| \leq \sum_{s=1}^{l}\|\theta_s\|\left\|\frac{\partial \mathbf{V}}{\partial \theta_s}\right\| \leq M_1\sum_{s=1}^{l}\|\theta_s\|,$$

it follows that the first term of $\partial\mathbf{H}_j/\partial\theta_t$ is bounded on the neighborhood N of $\boldsymbol{\xi}(P)$. The traces in the other terms can be handled in the same way, which yields that $\|\partial\mathbf{H}_j/\partial\theta_t\|$ is bounded on the neighborhood N of $\boldsymbol{\xi}(P)$. Because $u(s)s^2 = \rho'(s)s$ is bounded, it follows that the third term on the right hand side of (B.28) is bounded.

Next, consider the fifth term on the right hand side of (B.28). From (B.29),

$$\left|\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_t}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\right)\right| \leq k\lambda_1(\mathbf{V}^{-1})M_1^2,$$

which is uniformly bounded on the neighborhood N of $\boldsymbol{\xi}(P)$. Because $\rho(s)$ is bounded, it follows that the fifth term on the right hand side of (B.28) is bounded. For the sixth term on the right hand side of (B.28), from (B.30) we have

$$\left|\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial^2 \mathbf{V}}{\partial \theta_j \theta_t}\right)\right| \leq k^{3/2}\lambda_1(\mathbf{V}^{-1})\left\|\frac{\partial^2 \mathbf{V}}{\partial \theta_j \theta_t}\right\|.$$

Because $\mathbf{V}(\boldsymbol{\theta})$ is twice continuously differentiable and $\rho(s)$ is bounded, we conclude that the sixth term on the right hand side of (B.28) is bounded. Finally, from (B.17) and (B.23) together with the fact that $u(s)s^2 = \rho'(s)s$ is bounded, it also follows that the last term on the right hand side of (B.28) is bounded. By putting everything together, it follows that there exists a constant $0 < C_4 < \infty$, only depending on P, such that for $j, t = 1, \ldots, l$,

$$\left\|\frac{\partial \Psi_{\boldsymbol{\theta},j}(\mathbf{s},\boldsymbol{\xi})}{\partial \theta_t}\right\| \leq C_4.$$

It follows by dominated convergence that for $\boldsymbol{\xi}$ in the neighborhood N of $\boldsymbol{\xi}(P)$, it holds

$$\frac{\partial \Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi})}{\partial \boldsymbol{\theta}} = \int \frac{\partial \Psi_{\boldsymbol{\theta}}(\mathbf{s},\boldsymbol{\xi})}{\partial \boldsymbol{\theta}}\,\mathrm{d}P(\mathbf{s}), \tag{B.31}$$

and that $\partial\Lambda_{\boldsymbol{\theta}}/\partial\boldsymbol{\theta}$ is continuous at $\boldsymbol{\xi}(P)$. This finishes the proof. □

For convenience we state the following result from [16] about spherically contoured densities, see Lemma 5.1 in [16]. This lemma uses the commutation matrix $\mathbf{K}_{k,k}$, which is the $k^2 \times k^2$ block matrix with the (i,j)-block being equal to the $k \times k$ matrix $\boldsymbol{\Delta}_{ji}$ consisting of zero's except a 1 at entry (j,i). A useful property (e.g., see [20, Section 3.7]) is that for any $k \times k$ matrix $\mathbf{A}$, it holds that

$$\mathbf{K}_{k,k}\mathrm{vec}(\mathbf{A}) = \mathrm{vec}(\mathbf{A}^T). \tag{B.32}$$

Lemma B.4. *Suppose that* $\mathbf{z}$ *has a* k*-variate elliptical contoured density defined in* (3.2)*, with parameters* $\boldsymbol{\mu} = \mathbf{0}$ *and* $\boldsymbol{\Sigma} = \mathbf{I}_k$*. Then* $\mathbf{u} = \mathbf{z}/\|\mathbf{z}\|$ *is independent of* $\|\mathbf{z}\|$*, has mean zero and covariance matrix* $(1/k)\mathbf{I}_k$*. Furthermore,* $\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\mathbf{u}\mathbf{u}^T\mathbf{u} = \mathbf{0}$ *and*

$$\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\mathrm{vec}(\mathbf{u}\mathbf{u}^T)\mathrm{vec}(\mathbf{u}\mathbf{u}^T)^T = \sigma_1(\mathbf{I}_{k^2} + \mathbf{K}_{k,k}) + \sigma_2\mathrm{vec}(\mathbf{I}_k)\mathrm{vec}(\mathbf{I}_k)^T,$$

where $\sigma_1 = \sigma_2 = (k(k+2))^{-1}$*.*

Proof of Lemma 2

Proof. Write $\partial\Lambda/\partial\boldsymbol{\xi}$ as in (B.19). We determine each block separately and apply Lemma B.3. Let $\boldsymbol{\xi}_P = (\boldsymbol{\beta}_P, \boldsymbol{\theta}_P) = (\boldsymbol{\beta}(P), \boldsymbol{\theta}(P))$. When we also write $\mathbf{V}_P$ instead of $\mathbf{V}(\boldsymbol{\theta}(P))$, then according to Lemma B.3 we have

$$\begin{aligned}
\frac{\partial \Lambda_{\boldsymbol{\beta}}(\boldsymbol{\xi}_P)}{\partial \boldsymbol{\beta}} &= \int \frac{\partial \Psi_{\boldsymbol{\beta}}(\mathbf{s},\boldsymbol{\xi}_P)}{\partial \boldsymbol{\beta}}\,\mathrm{d}P(\mathbf{s}) \\
&= -\mathbb{E}\left[\frac{u'(d)}{d}\mathbf{X}^T\mathbf{V}_P^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_P)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_P)^T\mathbf{V}_P^{-1}\mathbf{X} + u(d)\cdot\mathbf{X}^T\mathbf{V}_P^{-1}\mathbf{X}\right] \\
&= -\mathbb{E}\left[\mathbb{E}\left[\frac{u'(d)}{d}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\mathbf{X} + u(d)\cdot\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}\,\middle|\,\mathbf{X}\right]\right],
\end{aligned}$$

where $d^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_P)^T \mathbf{V}_P^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_P)$. The inner expectation on the right hand side is the conditional expectation of $\mathbf{y} \mid \mathbf{X}$, which has the same distribution as $\boldsymbol{\Sigma}^{1/2}\mathbf{z} + \boldsymbol{\mu}$, where $\mathbf{z}$ has a spherical density $f_{\mathbf{0},\mathbf{I}_k}$. This implies that the inner expectation on the right hand side is equal to

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

From Lemma B.4, we find

$$\begin{aligned}
\mathbb{E}_{\mathbf{0},\mathbf{I}_k} \left[\frac{u'(\|\mathbf{z}\|)}{\|\mathbf{z}\|} \mathbf{z}\mathbf{z}^T + u(\|\mathbf{z}\|)\mathbf{I}_k \right] &= \mathbb{E}_{\mathbf{0},\mathbf{I}_k} \left[u'(\|\mathbf{z}\|)\|\mathbf{z}\| \frac{\mathbf{z}\mathbf{z}^T}{\|\mathbf{z}\|^2} + u(\|\mathbf{z}\|)\mathbf{I}_k \right] \\
&= \mathbb{E}_{\mathbf{0},\mathbf{I}_k} \left[u'(\|\mathbf{z}\|)\|\mathbf{z}\| \right] \mathbb{E}_{\mathbf{0},\mathbf{I}_k} \left[\mathbf{u}\mathbf{u}^T \right] + \mathbb{E}_{\mathbf{0},\mathbf{I}_k} \left[u(\|\mathbf{z}\|) \right] \mathbf{I}_k \\
&= \mathbb{E}_{\mathbf{0},\mathbf{I}_k} \left[u'(\|\mathbf{z}\|)\|\mathbf{z}\| \right] \frac{1}{k} \mathbf{I}_k + \mathbb{E}_{\mathbf{0},\mathbf{I}_k} \left[u(\|\mathbf{z}\|) \right] \mathbf{I}_k \\
&= \alpha \mathbf{I}_k,
\end{aligned}$$

where $\mathbf{u} = \mathbf{z}/\|\mathbf{z}\|$ and

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

It follows that

$$\frac{\partial \Lambda_{\boldsymbol{\beta}}(\boldsymbol{\xi}_P)}{\partial \boldsymbol{\beta}} = -\alpha \mathbb{E} \left[\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X} \right].$$

Next, for $j = 1, \ldots, l$, consider

$$\begin{aligned}
\frac{\partial \Lambda_{\boldsymbol{\beta}}(\boldsymbol{\xi}_P)}{\partial \theta_j} &= \int \frac{\partial \Psi_{\boldsymbol{\beta}}(\mathbf{s}, \boldsymbol{\xi}_P)}{\partial \theta_j} \, \mathrm{d}P(\mathbf{s}) \\
&= \mathbb{E} \left[\mathbb{E} \left[\frac{u'(d)}{2d} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_P)^T \mathbf{V}_P^{-1} \frac{\partial \mathbf{V}_P}{\partial \theta_j} \mathbf{V}_P^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_P) \cdot \mathbf{X}^T \mathbf{V}_P^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_P) \,\middle|\, \mathbf{X} \right] \right] \\
&\quad + \mathbb{E} \left[\mathbb{E} \left[u(d) \cdot \mathbf{X}^T \mathbf{V}_P^{-1} \frac{\partial \mathbf{V}_P}{\partial \theta_j} \mathbf{V}_P^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_P) \,\middle|\, \mathbf{X} \right] \right].
\end{aligned} \tag{B.33}$$

According to Lemma B.4, the inner conditional expectation of the first term on the right hand side of (B.33) can be written as

$$\begin{aligned}
&\mathbf{X}^T \mathbb{E}_{\mathbf{0},\mathbf{I}_k} \left[\frac{u'(\|\mathbf{z}\|)}{2\|\mathbf{z}\|} \mathbf{z}^T \boldsymbol{\Sigma}^{-1/2} \frac{\partial \mathbf{V}_P}{\partial \theta_j} \boldsymbol{\Sigma}^{-1/2} \mathbf{z} \boldsymbol{\Sigma}^{-1/2} \mathbf{z} \right] \\
&= \mathbf{X}^T \mathbb{E}_{\mathbf{0},\mathbf{I}_k} \left[\frac{u'(\|\mathbf{z}\|)\|\mathbf{z}\|^2}{2} \right] \mathbb{E}_{\mathbf{0},\mathbf{I}_k} \left[\mathbf{u}^T \boldsymbol{\Sigma}^{-1/2} \frac{\partial \mathbf{V}_P}{\partial \theta_j} \boldsymbol{\Sigma}^{-1/2} \mathbf{u} \boldsymbol{\Sigma}^{-1/2} \mathbf{u} \right],
\end{aligned}$$

Since the second term on the right hand side is the expectation with respect to a spherical density of an odd function of $\mathbf{u}$, this expectation is equal to zero due to Lemma B.4. Similarly, the second term on the right hand side of (B.33) has inner conditional expectation

$$\mathbf{X}^T \mathbb{E}_{\mathbf{0},\mathbf{I}_k} \left[u(\|\mathbf{z}\|) \boldsymbol{\Sigma}^{-1} \frac{\partial \mathbf{V}_P}{\partial \theta_j} \boldsymbol{\Sigma}^{-1/2} \mathbf{z} \right] = \mathbf{X}^T \mathbb{E}_{\mathbf{0},\mathbf{I}_k} \left[u(\|\mathbf{z}\|)\|\mathbf{z}\| \right] \boldsymbol{\Sigma}^{-1} \frac{\partial \mathbf{V}_P}{\partial \theta_j} \boldsymbol{\Sigma}^{-1/2} \mathbb{E}_{\mathbf{0},\mathbf{I}_k} \left[\mathbf{u} \right] = \mathbf{0},$$

due to Lemma B.4. It follows that

$$\frac{\partial \Lambda_{\boldsymbol{\beta}}(\boldsymbol{\xi}_P)}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

Next, using Lemma B.3 and (B.25), for all $j = 1, \ldots, l$, consider

$$\begin{aligned}
\frac{\partial \Lambda_{\boldsymbol{\theta},j}}{\partial \boldsymbol{\beta}^T} &= \int \frac{\partial \Psi_{\boldsymbol{\theta},j}(\mathbf{s}, \boldsymbol{\xi}_P)}{\partial \boldsymbol{\beta}^T} \, \mathrm{d}P(\mathbf{s}) \\
&= -\mathbb{E}\left[\mathbb{E}\left[\frac{u'(d)}{d}\mathbf{X}^T\mathbf{V}_P^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_P) \cdot (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_P)^T\mathbf{V}_P^{-1}\mathbf{H}_j\mathbf{V}_P^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_P) \,\middle|\, \mathbf{X}\right]\right] \\
&\quad - \mathbb{E}\left[\mathbb{E}\left[u(d) \cdot 2\mathbf{X}^T\mathbf{V}_P^{-1}\mathbf{H}_j\mathbf{V}_P^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_P) \,\middle|\, \mathbf{X}\right]\right] \\
&\quad - \operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\right)\mathbb{E}\left[\mathbb{E}\left[\frac{\rho'(d)}{d}\mathbf{X}^T\mathbf{V}_P^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_P) \,\middle|\, \mathbf{X}\right]\right].
\end{aligned} \tag{B.34}$$

According to Lemma B.4, the first term on the right hand side of (B.34) has inner conditional expectation

$$\begin{aligned}
&\mathbf{X}^T\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\frac{u'(\|\mathbf{z}\|)}{\|\mathbf{z}\|}\boldsymbol{\Sigma}^{-1/2}\mathbf{z}\mathbf{z}^T\boldsymbol{\Sigma}^{-1/2}\mathbf{H}_j\boldsymbol{\Sigma}^{-1/2}\mathbf{z}\right] \\
&\qquad = \mathbf{X}^T\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[u'(\|\mathbf{z}\|)\|\mathbf{z}\|^2\right]\boldsymbol{\Sigma}^{-1/2}\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\mathbf{u}\mathbf{u}^T\boldsymbol{\Sigma}^{-1/2}\mathbf{H}_j\boldsymbol{\Sigma}^{-1/2}\mathbf{u}\right].
\end{aligned}$$

Again, the second term on the right hand side is the expectation with respect to a spherical density of an odd function of $\mathbf{u}$, and is therefore equal to zero. Similarly, the inner expectation of the second term on the right hand side of (B.34) is equal to

$$2\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{H}_j\boldsymbol{\Sigma}^{-1/2}\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[u(\|\mathbf{z}\|)\mathbf{z}\right] = 2\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{H}_j\boldsymbol{\Sigma}^{-1/2}\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[u(\|\mathbf{z}\|)\|\mathbf{z}\|\right]\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\mathbf{u}\right] = \mathbf{0},$$

and the inner expectation of the third term on the right hand side of (B.34) is equal to

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

It follows that

$$\frac{\partial \Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}_P)}{\partial \boldsymbol{\beta}} = \mathbf{0}.$$

Finally, to determine $\partial \Lambda_{\boldsymbol{\theta},j}(\boldsymbol{\xi}_P)/\partial \theta_s$, note that when $\mathbf{V}$ is linear, we can write

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

where $\Psi_{\mathbf{V}}$ is defined in (7.8) and has the property that

$$\int \Psi_{\mathbf{V}}(\mathbf{s}, \boldsymbol{\xi}_P) \, \mathrm{d}P(\mathbf{s}) = \mathbf{0},$$

when $\mathbf{y} \mid \mathbf{X}$ has an elliptically contoured density $f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ with parameters $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}_P$ and $\boldsymbol{\Sigma} = \mathbf{V}_P$. This means that for each $j, s = 1, \ldots, l$, we have

$$\frac{\partial \Lambda_{\boldsymbol{\theta},j}(\boldsymbol{\xi}_P)}{\partial \theta_s} = \int \frac{\partial \Psi_{\boldsymbol{\theta},j}(\mathbf{s}, \boldsymbol{\xi}_P)}{\partial \theta_s} \, \mathrm{d}P(\mathbf{s}) = -\operatorname{vec}(\boldsymbol{\Sigma}^{-1}\mathbf{L}_j\boldsymbol{\Sigma}^{-1})^T \operatorname{vec}\left(\int \frac{\partial \Psi_{\mathbf{V}}(\mathbf{s}, \boldsymbol{\xi}_P)}{\partial \theta_s} \, \mathrm{d}P(\mathbf{s})\right)$$

where $\Psi_{\mathbf{V}}$ is defined in (7.8). As before, we find

$$\begin{aligned}
\int \frac{\partial \Psi_{\mathbf{V}}(\mathbf{s}, \boldsymbol{\xi}_P)}{\partial \theta_s} \, \mathrm{d}P(\mathbf{s}) = &-\mathbb{E}\left[\mathbb{E}\left[\frac{ku'(d)}{2d}\mathbf{z}^T\boldsymbol{\Sigma}^{-1/2}\mathbf{L}_s\boldsymbol{\Sigma}^{-1/2}\mathbf{z} \cdot \boldsymbol{\Sigma}^{1/2}\mathbf{z}\mathbf{z}^T\boldsymbol{\Sigma}^{1/2} \,\middle|\, \mathbf{X}\right]\right] \\
&+ \mathbb{E}\left[\mathbb{E}\left[\frac{v'(d)}{2d}\mathbf{z}^T\boldsymbol{\Sigma}^{-1/2}\mathbf{L}_s\boldsymbol{\Sigma}^{-1/2}\mathbf{z} \cdot \boldsymbol{\Sigma} \,\middle|\, \mathbf{X}\right]\right] \\
&- \mathbb{E}\left[\mathbb{E}\left[v(d)\mathbf{L}_s \,\middle|\, \mathbf{X}\right]\right].
\end{aligned} \tag{B.35}$$

The first term on the right hand side of (B.35) is equal to

$$\begin{aligned}
&\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\frac{ku'(\|\mathbf{z}\|)}{2\|\mathbf{z}\|}\mathbf{z}^T\boldsymbol{\Sigma}^{-1/2}\mathbf{L}_s\boldsymbol{\Sigma}^{-1/2}\mathbf{z}\boldsymbol{\Sigma}^{1/2}\mathbf{z}\mathbf{z}^T\boldsymbol{\Sigma}^{1/2}\right] \\
&\quad= \mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\frac{ku'(\|\mathbf{z}\|)\|\mathbf{z}\|^3}{2}\right]\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\mathbf{u}^T\boldsymbol{\Sigma}^{-1/2}\mathbf{L}_s\boldsymbol{\Sigma}^{-1/2}\mathbf{u}\boldsymbol{\Sigma}^{1/2}\mathbf{u}\mathbf{u}^T\boldsymbol{\Sigma}^{1/2}\right].
\end{aligned}$$

Furthermore, we can write

$$\begin{aligned}
&\mathrm{vec}(\boldsymbol{\Sigma}^{-1}\mathbf{L}_j\boldsymbol{\Sigma}^{-1})^T\mathrm{vec}\left(\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\mathbf{u}^T\boldsymbol{\Sigma}^{-1/2}\mathbf{L}_s\boldsymbol{\Sigma}^{-1/2}\mathbf{u}\boldsymbol{\Sigma}^{1/2}\mathbf{u}\mathbf{u}^T\boldsymbol{\Sigma}^{1/2}\right]\right) \\
&= \mathrm{vec}(\mathbf{L}_j)^T\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\mathrm{vec}\left(\boldsymbol{\Sigma}^{1/2}\mathbf{u}\mathbf{u}^T\boldsymbol{\Sigma}^{1/2}\right)\mathbf{u}^T\boldsymbol{\Sigma}^{-1/2}\mathbf{L}_s\boldsymbol{\Sigma}^{-1/2}\mathbf{u}\right] \\
&= \mathrm{vec}(\mathbf{L}_j)^T\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\left(\boldsymbol{\Sigma}^{1/2}\otimes\boldsymbol{\Sigma}^{1/2}\right)\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\mathrm{vec}\left(\mathbf{u}\mathbf{u}^T\right)\mathrm{vec}\left(\mathbf{u}\mathbf{u}^T\right)^T\right]\mathrm{vec}\left(\boldsymbol{\Sigma}^{-1/2}\mathbf{L}_s\boldsymbol{\Sigma}^{-1/2}\right) \\
&= \mathrm{vec}\left(\boldsymbol{\Sigma}^{-1/2}\mathbf{L}_j\boldsymbol{\Sigma}^{-1/2}\right)^T\frac{1}{k(k+2)}\Big(\mathbf{I}_{k^2}+\mathbf{K}_{k,k}+\mathrm{vec}(\mathbf{I}_k)\mathrm{vec}(\mathbf{I}_k)^T\Big)\mathrm{vec}\left(\boldsymbol{\Sigma}^{-1/2}\mathbf{L}_s\boldsymbol{\Sigma}^{-1/2}\right),
\end{aligned}$$

using Lemma B.4. Application of property (B.32) and the fact that $\mathrm{vec}(\mathbf{A})^T\mathrm{vec}(\mathbf{B}) = \mathrm{tr}(\mathbf{AB})$, yields

$$\begin{aligned}
&\mathrm{vec}(\boldsymbol{\Sigma}^{-1}\mathbf{L}_j\boldsymbol{\Sigma}^{-1})^T\mathrm{vec}\left(\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\mathbf{u}^T\boldsymbol{\Sigma}^{-1/2}\mathbf{L}_s\boldsymbol{\Sigma}^{-1/2}\mathbf{u}\boldsymbol{\Sigma}^{1/2}\mathbf{u}\mathbf{u}^T\boldsymbol{\Sigma}^{1/2}\right]\right) \\
&= \frac{1}{k(k+2)}\left(2\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{L}_j\boldsymbol{\Sigma}^{-1}\mathbf{L}_s)+\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{L}_j)\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{L}_s)\right).
\end{aligned}$$

It follows that the first term on the right hand side of (B.35) leads to a first term in $\partial\Lambda_{\boldsymbol{\theta},j}(\boldsymbol{\xi}_P)/\partial\theta_s$, which is equal to

$$\frac{\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[u'(\|\mathbf{z}\|)\|\mathbf{z}\|^3\right]}{2(k+2)}\Big(2\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{L}_j\boldsymbol{\Sigma}^{-1}\mathbf{L}_s)+\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{L}_j)\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{L}_s)\Big). \tag{B.36}$$

The second term on the right hand side of (B.35) is equal to

$$\begin{aligned}
&\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\frac{v'(\|\mathbf{z}\|)}{2\|\mathbf{z}\|}\mathbf{z}^T\boldsymbol{\Sigma}^{-1/2}\mathbf{L}_s\boldsymbol{\Sigma}^{-1/2}\mathbf{z}\boldsymbol{\Sigma}\right] \\
&= \mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\frac{v'(\|\mathbf{z}\|)\|\mathbf{z}\|}{2}\right]\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\mathbf{u}^T\boldsymbol{\Sigma}^{-1/2}\mathbf{L}_s\boldsymbol{\Sigma}^{-1/2}\mathbf{u}\right]\boldsymbol{\Sigma} \\
&= \mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\frac{v'(\|\mathbf{z}\|)\|\mathbf{z}\|}{2}\right]\mathrm{vec}\left(\boldsymbol{\Sigma}^{-1/2}\mathbf{L}_s\boldsymbol{\Sigma}^{-1/2}\right)^T\mathrm{vec}\left(\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\mathbf{u}\mathbf{u}^T\right]\right)\boldsymbol{\Sigma} \\
&= \mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\frac{v'(\|\mathbf{z}\|)\|\mathbf{z}\|}{2}\right]\mathrm{vec}\left(\boldsymbol{\Sigma}^{-1/2}\mathbf{L}_s\boldsymbol{\Sigma}^{-1/2}\right)^T\mathrm{vec}\left(\frac{1}{k}\mathbf{I}_k\right)\boldsymbol{\Sigma} \\
&= \frac{\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[v'(\|\mathbf{z}\|)\|\mathbf{z}\|\right]}{2k}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{L}_s\right)\boldsymbol{\Sigma},
\end{aligned}$$

using Lemma B.4. This leads to a second term in $\partial\Lambda_{\boldsymbol{\theta},j}(\boldsymbol{\xi}_P)/\partial\theta_s$, which is equal to

$$\begin{aligned}
&-\frac{\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[v'(\|\mathbf{z}\|)\|\mathbf{z}\|\right]}{2k}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{L}_s\right)\mathrm{vec}\left(\boldsymbol{\Sigma}^{-1}\mathbf{L}_j\boldsymbol{\Sigma}^{-1}\right)^T\mathrm{vec}(\boldsymbol{\Sigma}) \\
&= -\frac{\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[v'(\|\mathbf{z}\|)\|\mathbf{z}\|\right]}{2k}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{L}_s\right)\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{L}_j\right).
\end{aligned} \tag{B.37}$$

The third term on the right hand side of (B.35) leads to a third term in $\partial\Lambda_{\boldsymbol{\theta},j}(\boldsymbol{\xi}_P)/\partial\theta_s$, which is equal to

$$\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[v(\|\mathbf{z}\|)\right]\mathrm{vec}\left(\boldsymbol{\Sigma}^{-1}\mathbf{L}_j\boldsymbol{\Sigma}^{-1}\right)^T\mathrm{vec}(\mathbf{L}_s) = \mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[v(\|\mathbf{z}\|)\right]\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{L}_j\boldsymbol{\Sigma}^{-1}\mathbf{L}_s\right). \tag{B.38}$$

We conclude that $\partial\Lambda_{\boldsymbol{\theta},j}(\boldsymbol{\xi}_P)/\partial\theta_s$ consists of three terms given in (B.36), (B.37) and (B.38). This means that $\partial\Lambda_{\boldsymbol{\theta},j}(\boldsymbol{\xi}_P)/\partial\theta_s$ has a term $\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{L}_j\boldsymbol{\Sigma}^{-1}\mathbf{L}_s)$ with coefficient

$$\frac{\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[u'(\|\mathbf{z}\|)\|\mathbf{z}\|^3\right]}{(k+2)} + \mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[v(\|\mathbf{z}\|)\right] = \gamma_1,$$

and a term $\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{L}_s)\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{L}_j)$ with coefficient

$$\frac{\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[u'(\|\mathbf{z}\|)\|\mathbf{z}\|^3\right]}{2(k+2)} - \frac{\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[v'(\|\mathbf{z}\|)\|\mathbf{z}\|\right]}{2k} = -\gamma_2,$$

where γ_1 and γ_2 are defined in (8.5), and where we use that $u'(s)s^3 = \rho''(s)s^2 - \rho'(s)s$ and $v(s) = \rho'(s)s - \rho(s) + b_0$. Finally, from the definition of $\mathbf{L}$ in (7.9) it follows that the $l \times l$ matrix with entries

$$\begin{aligned}
&\gamma_1\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{L}_j\boldsymbol{\Sigma}^{-1}\mathbf{L}_s) - \gamma_2\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{L}_s)\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{L}_j)\\
&= \gamma_1\mathrm{vec}(\mathbf{L}_j)^T\left(\boldsymbol{\Sigma}^{-1/2}\otimes\boldsymbol{\Sigma}^{-1/2}\right)\left(\boldsymbol{\Sigma}^{-1/2}\otimes\boldsymbol{\Sigma}^{-1/2}\right)\mathrm{vec}(\mathbf{L}_s)\\
&\quad - \gamma_2\mathrm{vec}(\mathbf{L}_j)^T\left(\boldsymbol{\Sigma}^{-1/2}\otimes\boldsymbol{\Sigma}^{-1/2}\right)\mathrm{vec}(\mathbf{I}_k)\mathrm{vec}(\mathbf{I}_k)^T\left(\boldsymbol{\Sigma}^{-1/2}\otimes\boldsymbol{\Sigma}^{-1/2}\right)\mathrm{vec}(\mathbf{L}_s)\\
&= \gamma_1\mathbf{L}^T\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\mathbf{L} - \gamma_2\mathrm{vec}(\mathbf{L}_j)^T\mathrm{vec}(\boldsymbol{\Sigma}^{-1})\mathrm{vec}(\boldsymbol{\Sigma}^{-1})^T\mathrm{vec}(\mathbf{L}_s),
\end{aligned}$$

is the matrix

$$\gamma_1\mathbf{L}^T\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\mathbf{L} - \gamma_2\mathbf{L}^T\mathrm{vec}(\boldsymbol{\Sigma}^{-1})\mathrm{vec}(\boldsymbol{\Sigma}^{-1})^T\mathbf{L}.$$

This proves the lemma. □

Lemma B.5. *Suppose that ρ satisfies (R3)-(R4). Let γ_1 and γ_2 defined in* (8.5) *and suppose that $\gamma_1 > 0$. Then the inverse of $\partial\Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}(P))/\partial\boldsymbol{\theta}^T$ exists and is given by*

$$\begin{aligned}
&a\left(\mathbf{L}^T\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\mathbf{L}\right)^{-1}\\
&+ b\left(\mathbf{L}^T\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\mathbf{L}\right)^{-1}\mathbf{L}^T\mathrm{vec}(\boldsymbol{\Sigma}^{-1})\mathrm{vec}(\boldsymbol{\Sigma}^{-1})^T\mathbf{L}\left(\mathbf{L}^T\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\mathbf{L}\right)^{-1}
\end{aligned}$$

where $a = 1/\gamma_1$ and $b = \gamma_2/(\gamma_1(\gamma_1 - k\gamma_2))$.

Proof. Together with Lemma 2, first write

$$\begin{aligned}
\frac{\partial\Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}(P))}{\partial\boldsymbol{\theta}^T} &= \gamma_1\mathbf{L}^T\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\mathbf{L} - \gamma_2\mathbf{L}^T\mathrm{vec}(\boldsymbol{\Sigma}^{-1})\mathrm{vec}(\boldsymbol{\Sigma}^{-1})^T\mathbf{L}\\
&= \gamma_1\mathbf{E}^T\mathbf{E} - \gamma_2\mathbf{E}^T\mathrm{vec}(\mathbf{I}_k)\mathrm{vec}(\mathbf{I}_k)^T\mathbf{E},
\end{aligned}$$

where $\mathbf{E} = \left(\boldsymbol{\Sigma}^{-1/2}\otimes\boldsymbol{\Sigma}^{-1/2}\right)\mathbf{L}$. Since $\mathbf{V}(\boldsymbol{\theta}(P)) = \boldsymbol{\Sigma}$, by definition of $\mathbf{L}$, it follows that for $\boldsymbol{\theta}(P) = (\theta_1,\ldots,\theta_l)^T$,

$$\mathbf{L}\,\boldsymbol{\theta}(P) = \sum_{j=1}^{l}\theta_j\mathrm{vec}(\mathbf{L}_j) = \mathrm{vec}\left(\sum_{j=1}^{l}\theta_j\mathbf{L}_j\right) = \mathrm{vec}(\boldsymbol{\Sigma}).$$

This means that

$$\mathbf{E}\,\boldsymbol{\theta}(P) = \left(\boldsymbol{\Sigma}^{-1/2}\otimes\boldsymbol{\Sigma}^{-1/2}\right)\mathbf{L}\,\boldsymbol{\theta}(P) = \left(\boldsymbol{\Sigma}^{-1/2}\otimes\boldsymbol{\Sigma}^{-1/2}\right)\mathrm{vec}(\boldsymbol{\Sigma}) = \mathrm{vec}(\mathbf{I}_k). \tag{B.39}$$

Since $\mathbf{L}$ has full rank, also $\mathbf{E}$ has full rank. This means that $(\mathbf{E}^T\mathbf{E})^{-1}$ exists, and satisfies

$$(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathrm{vec}(\mathbf{I}_k) = \boldsymbol{\theta}(P). \tag{B.40}$$

Now, write

$$\begin{aligned}&\gamma_1\mathbf{E}^T\mathbf{E} - \gamma_2\mathbf{E}^T\text{vec}(\mathbf{I}_k)\text{vec}(\mathbf{I}_k)^T\mathbf{E}\\&= \left(\gamma_1\mathbf{I}_k - \gamma_2\mathbf{E}^T\text{vec}(\mathbf{I}_k)\text{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\right)(\mathbf{E}^T\mathbf{E}).\end{aligned}$$

When we multiply the first matrix from the left with a matrix of the same type,

$$\begin{aligned}&\left(a\mathbf{I}_k + b\mathbf{E}^T\text{vec}(\mathbf{I}_k)\text{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\right)\\&\qquad\left(\gamma_1\mathbf{I}_k - \gamma_2\mathbf{E}^T\text{vec}(\mathbf{I}_k)\text{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\right),\end{aligned} \tag{B.41}$$

then we find four terms. A term $\mathbf{I}_k$ with coefficient $a\gamma_1$, two terms

$$\mathbf{E}^T\text{vec}(\mathbf{I}_k)\text{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1},$$

with coefficient $-a\gamma_2 + b\gamma_1$, and the term

$$\mathbf{E}^T\text{vec}(\mathbf{I}_k)\text{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\text{vec}(\mathbf{I}_k)\text{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}$$

with coefficient $-b\gamma_2$. Consider the scalar valued inner product in the middle

$$\text{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\text{vec}(\mathbf{I}_k) = \text{vec}(\mathbf{I}_k)^T\mathbf{E}\,\boldsymbol{\theta}(P) = \text{vec}(\mathbf{I}_k)^T\text{vec}(\mathbf{I}_k) = k, \tag{B.42}$$

by application of (B.40) and then (B.39). It follows that the term with coefficient $-b\gamma_2$ reduces to

$$k\mathbf{E}^T\text{vec}(\mathbf{I}_k)\text{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}.$$

Hence the matrix product in (B.41) is equal to

$$a\gamma_1\mathbf{I}_k + (-a\gamma_2 + b\gamma_1 - kb\gamma_2)\mathbf{E}^T\text{vec}(\mathbf{I}_k)\text{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1} \tag{B.43}$$

When we multiply the same matrix from the right,

$$\left(\gamma_1\mathbf{I}_k - \gamma_2\mathbf{E}^T\text{vec}(\mathbf{I}_k)\text{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\right)\left(a\mathbf{I}_k + b\mathbf{E}^T\text{vec}(\mathbf{I}_k)\text{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\right),$$

we find the same result (B.43). This matrix is equal to $\mathbf{I}_k$ if and only if $a\gamma_1 = 1$ and $-a\gamma_2 + b\gamma_1 - kb\gamma_2 = 0$, or equivalently

$$\begin{aligned}a &= 1/\gamma_1\\b &= \frac{a\gamma_2}{\gamma_1 - k\gamma_2} = \frac{\gamma_2}{\gamma_1(\gamma_1 - k\gamma_2)},\end{aligned} \tag{B.44}$$

where we use that $\gamma_1 > 0$ and $\gamma_1 - k\gamma_2 = \mathbb{E}_{0,\mathbf{I}_k}\left[\rho'(\|\mathbf{z}\|)\right]/2 > 0$, due to (R3)-(R4). We conclude that the inverse of $\gamma_1\mathbf{I}_k - \gamma_2\mathbf{E}^T\text{vec}(\mathbf{I}_k)\text{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}$ exists and is equal to

$$a\mathbf{I}_k + b\mathbf{E}^T\text{vec}(\mathbf{I}_k)\text{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}$$

with a and b given in (B.44). Hence, the inverse of the matrix $\gamma_1\mathbf{E}^T\mathbf{E} - \gamma_2\mathbf{E}^T\text{vec}(\mathbf{I}_k)\text{vec}(\mathbf{I}_k)^T\mathbf{E}$ is equal to

$$\begin{aligned}&\left(\mathbf{E}^T\mathbf{E}\right)^{-1}\left(\gamma_1\mathbf{I}_k - \gamma_2\mathbf{E}^T\text{vec}(\mathbf{I}_k)\text{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\right)^{-1}\\&= \left(\mathbf{E}^T\mathbf{E}\right)^{-1}\left(a\mathbf{I}_k + b\mathbf{E}^T\text{vec}(\mathbf{I}_k)\text{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\right)\\&= a(\mathbf{E}^T\mathbf{E})^{-1} + b(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\text{vec}(\mathbf{I}_k)\text{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}.\end{aligned}$$

After inserting $\mathbf{E} = \left(\boldsymbol{\Sigma}^{-1/2}\otimes\boldsymbol{\Sigma}^{-1/2}\right)\mathbf{L}$, this finishes the proof. □

Proof of Corollary 5

Proof. Since ρ is strictly increasing on $[0, c_0]$, the function $u(s) = \rho'(s)/s > 0$, for $0 < s \leq c_0$ and zero for $s > c_0$. This means that

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

Furthermore, since $\mathbf{X}$ has full rank, the inverse of $\mathbb{E}\left[\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}\right]$ exists. It follows that the matrix $\partial\Lambda_{\boldsymbol{\beta}}(\boldsymbol{\xi}(P))/\partial\boldsymbol{\beta}$ in (8.3) is non-singular. According to Lemma B.5, also $\partial\Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}(P))/\partial\boldsymbol{\theta}$ is non-singular. Together, with Lemma B.3 and Lemma 2, we conclude that $\partial\Lambda/\partial\boldsymbol{\xi}$ is continuously differentiable with a non-singular derivative at $\boldsymbol{\xi}(P)$, so that Theorem 5 applies. Together with Lemma 2, this implies that

$$\begin{aligned}\mathrm{IF}(\mathbf{s}_0, \boldsymbol{\beta}, P) &= -\left(\frac{\partial\Lambda_{\boldsymbol{\beta}}(\boldsymbol{\xi}(P))}{\partial\boldsymbol{\beta}}\right)^{-1}\Psi_{\boldsymbol{\beta}}(\mathbf{s}_0, \boldsymbol{\xi}(P)) \\ &= \frac{u(d_0)}{\alpha}\left(\mathbb{E}\left[\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}\right]\right)^{-1}\mathbf{X}_0^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})\end{aligned}$$

where $d_0^2 = (\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})$. From Theorem 5, together with Lemma 2, it also follows that

$$\begin{aligned}&\mathrm{IF}(\mathbf{s}_0, \boldsymbol{\theta}, P) = -\left(\frac{\partial\Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}(P))}{\partial\boldsymbol{\theta}}\right)^{-1}\Psi_{\boldsymbol{\theta}}(\mathbf{s}_0, \boldsymbol{\xi}(P)) \\ &= \left(\frac{\partial\Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}(P))}{\partial\boldsymbol{\theta}}\right)^{-1}\mathbf{L}^T(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\times \\ &\qquad\qquad \times \mathrm{vec}\left(ku(d_0)(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})^T - v(d_0)\boldsymbol{\Sigma}\right) \qquad \text{(B.45)} \\ &= ku(d_0)\left(\frac{\partial\Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}(P))}{\partial\boldsymbol{\theta}}\right)^{-1}\mathbf{L}^T(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\mathrm{vec}\left((\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})^T\right) \\ &\quad - v(d_0)\left(\frac{\partial\Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}(P))}{\partial\boldsymbol{\theta}}\right)^{-1}\mathbf{L}^T(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\mathrm{vec}(\boldsymbol{\Sigma}).\end{aligned}$$

Consider the first term on the right hand side of (B.45). We have that

$$\begin{aligned}&(\boldsymbol{\Sigma}^{-1/2}\otimes\boldsymbol{\Sigma}^{-1/2})\mathrm{vec}\left((\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})^T\right) \\ &\qquad = \mathrm{vec}\left(\boldsymbol{\Sigma}^{-1/2}(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1/2}\right)\end{aligned}$$

and from Lemma B.5,

$$\begin{aligned}&\left(\frac{\partial\Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}(P))}{\partial\boldsymbol{\theta}}\right)^{-1}\mathbf{L}^T(\boldsymbol{\Sigma}^{-1/2}\otimes\boldsymbol{\Sigma}^{-1/2}) \\ &\quad = a(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T + b(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathrm{vec}(\mathbf{I}_k)\mathrm{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T,\end{aligned} \qquad \text{(B.46)}$$

where $\mathbf{E} = (\boldsymbol{\Sigma}^{-1/2}\otimes\boldsymbol{\Sigma}^{-1/2})\mathbf{L}$. This implies

$$\begin{aligned}&\left(\frac{\partial\Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}(P))}{\partial\boldsymbol{\theta}}\right)^{-1}\mathbf{L}^T(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\mathrm{vec}\left((\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})^T\right) \\ &\quad = a(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathrm{vec}\left(\boldsymbol{\Sigma}^{-1/2}(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1/2}\right) \\ &\qquad + b(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathrm{vec}(\mathbf{I}_k)\mathrm{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\times \\ &\qquad\qquad \times \mathrm{vec}\left(\boldsymbol{\Sigma}^{-1/2}(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1/2}\right).\end{aligned}$$

The first term on the right hand side is equal to

$$a\Big(\mathbf{L}^T(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\mathbf{L})\Big)^{-1}\mathbf{L}^T(\boldsymbol{\Sigma}^{-1/2}\otimes\boldsymbol{\Sigma}^{-1/2})\times \\ \times \operatorname{vec}\Big(\boldsymbol{\Sigma}^{-1/2}(\mathbf{y}_0-\mathbf{X}_0\boldsymbol{\beta})(\mathbf{y}_0-\mathbf{X}_0\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1/2}\Big)$$

and, with (B.39) and (B.40), the second term on the right hand side is equal to

$$\begin{aligned} &b\boldsymbol{\theta}(P)\boldsymbol{\theta}(P)^T\mathbf{E}^T\operatorname{vec}\Big(\boldsymbol{\Sigma}^{-1/2}(\mathbf{y}_0-\mathbf{X}_0\boldsymbol{\beta})(\mathbf{y}_0-\mathbf{X}_0\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1/2}\Big)\\ &= b\boldsymbol{\theta}(P)\operatorname{vec}(\mathbf{I}_k)^T\operatorname{vec}\Big(\boldsymbol{\Sigma}^{-1/2}(\mathbf{y}_0-\mathbf{X}_0\boldsymbol{\beta})(\mathbf{y}_0-\mathbf{X}_0\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1/2}\Big)\\ &= b\boldsymbol{\theta}(P)\operatorname{tr}\Big(\boldsymbol{\Sigma}^{-1/2}(\mathbf{y}_0-\mathbf{X}_0\boldsymbol{\beta})(\mathbf{y}_0-\mathbf{X}_0\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1/2}\Big)\\ &= bd_0^2\boldsymbol{\theta}(P). \end{aligned}$$

It follows that the first term on the right hand side of (B.45) is equal to

$$aku(d_0)\Big(\mathbf{L}^T(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\mathbf{L})\Big)^{-1}\mathbf{L}^T(\boldsymbol{\Sigma}^{-1/2}\otimes\boldsymbol{\Sigma}^{-1/2})\times \\ \times \operatorname{vec}\Big(\boldsymbol{\Sigma}^{-1/2}(\mathbf{y}_0-\mathbf{X}_0\boldsymbol{\beta})(\mathbf{y}_0-\mathbf{X}_0\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1/2}\Big) \\ + bku(d_0)d_0^2\boldsymbol{\theta}(P).$$

Next consider the second term on the right hand side of (B.45). We have that

$$(\boldsymbol{\Sigma}^{-1/2}\otimes\boldsymbol{\Sigma}^{-1/2})\operatorname{vec}(\boldsymbol{\Sigma}) = \operatorname{vec}(\mathbf{I}_k),$$

and with (B.46), together with (B.39) and (B.40),

$$\begin{aligned} &\left(\frac{\partial\Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}(P))}{\partial\boldsymbol{\theta}}\right)^{-1}\mathbf{L}^T(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\operatorname{vec}(\boldsymbol{\Sigma})\\ &= a(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\operatorname{vec}(\mathbf{I}_k) + b(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\operatorname{vec}(\mathbf{I}_k)\operatorname{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\operatorname{vec}(\mathbf{I}_k)\\ &= a\boldsymbol{\theta}(P) + b\boldsymbol{\theta}(P)\operatorname{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\operatorname{vec}(\mathbf{I}_k)\\ &= a\boldsymbol{\theta}(P) + b\boldsymbol{\theta}(P)\operatorname{vec}(\mathbf{I}_k)^T\mathbf{E}\boldsymbol{\theta}(P)\\ &= a\boldsymbol{\theta}(P) + b\boldsymbol{\theta}(P)\operatorname{vec}(\mathbf{I}_k)^T\operatorname{vec}(\mathbf{I}_k)\\ &= (a+bk)\boldsymbol{\theta}(P). \end{aligned}$$

It follows that the second term on the right hand side of (B.45) is equal to

$$-v(d_0)(a+bk)\boldsymbol{\theta}(P).$$

Putting things together, we find that $\mathrm{IF}(\mathbf{s}_0,\boldsymbol{\theta},P)$ is equal to

$$aku(d_0)\Big(\mathbf{L}^T(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\mathbf{L})\Big)^{-1}\mathbf{L}^T(\boldsymbol{\Sigma}^{-1/2}\otimes\boldsymbol{\Sigma}^{-1/2})\times \\ \times \operatorname{vec}\Big(\boldsymbol{\Sigma}^{-1/2}(\mathbf{y}_0-\mathbf{X}_0\boldsymbol{\beta})(\mathbf{y}_0-\mathbf{X}_0\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1/2}\Big) \\ + (bku(d_0)d_0^2 - av(d_0) - bkv(d_0))\boldsymbol{\theta}(P).$$

Since $v(d_0) = u(d_0)d_0^2 - \rho(d_0) + b_0$, we have that

$$\begin{aligned} bku(d_0)d_0^2 - av(d_0) - bkv(d_0) &= -\left(\frac{v(d_0)}{\gamma_1} - \frac{k\gamma_2}{\gamma_1(\gamma_1-k\gamma_2)}(\rho(d_0)-b_0)\right)\\ &= -\frac{u(d_0)d_0^2}{\gamma_1} + \left(\frac{1}{\gamma_1} + \frac{k\gamma_2}{\gamma_1(\gamma_1-k\gamma_2)}\right)(\rho(d_0)-b_0)\\ &= -\frac{u(d_0)d_0^2}{\gamma_1} + \frac{\rho(d_0)-b_0}{\gamma_1-k\gamma_2} \end{aligned}$$

We conclude that $\mathrm{IF}(\mathbf{s}_0, \boldsymbol{\theta}, P)$ is given by

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

This proves the corollary. □

B.5 Proofs of Section 9

As preparation we first establish the following lemma, which is similar to Lemma 22 in [21].

Lemma B.6. *Let $\rho(\cdot)$ be a real-valued function of bounded variation on $\mathbb{R}^+$. The class of all functions on $\mathbb{R}^p$ of the form*

$$\mathbf{s}=(\mathbf{y},\mathbf{X})\mapsto\rho(\|\mathbf{A}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\|)$$

with $\mathbf{A}$ ranging over all $k\times k$ matrices and $\boldsymbol{\beta}$ ranging over $\mathbb{R}^q$, has polynomial discrimination.

Proof. Consider the class of functions

$$g_{\mathbf{A},\boldsymbol{\beta}}(\mathbf{s},t)=\|\mathbf{A}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\|^2-\rho^{-1}(t)^2.$$

According to Lemma 18 in [21], it suffices to show that the functions $g_{\mathbf{A},\boldsymbol{\beta}}(\cdot,\cdot)$ span a finite-dimensional vector space. In order to do so, write

$$\begin{aligned}\|\mathbf{A}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\|^2&=\sum_{i=1}^k\sum_{j=1}^k(\mathbf{A}^T\mathbf{A})_{ij}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})_i(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})_j\\&=\sum_{i=1}^k\sum_{j=1}^k\sum_{s=1}^k a_{is}a_{sj}\left(y_i-\sum_{r=1}^q x_{ir}\beta_r\right)\left(y_j-\sum_{w=1}^q x_{jw}\beta_w\right)\end{aligned}$$

This is a polynomial in $\mathbf{s}=(y_1,\ldots,y_k,x_{11},\ldots,x_{kq})$, with coefficients in $\mathbb{R}$. This means that the class of functions $g_{\mathbf{A},\boldsymbol{\beta}}(\mathbf{s},t)=\|\mathbf{A}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\|^2-\rho^{-1}(t)^2$ forms a finite dimensional vector space. □

A useful first step is the following lemma.

Lemma B.7. *Let $u:\mathbb{R}\to\mathbb{R}$ be a function of bounded variation. Let $\mathbf{s}=(\mathbf{y},\mathbf{X})=(s_1,\ldots,s_p)\in\mathbb{R}^p$, and define*

$$g(\mathbf{s},\boldsymbol{\beta},\mathbf{V})=u(\|\mathbf{V}^{-1/2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\|)\quad\text{for }\boldsymbol{\beta}\in\mathbb{R}^q,\mathbf{V}\in PDS(k).$$

Consider the classes of functions

$$\begin{aligned}\mathcal{F}&=\{g(\mathbf{s},\boldsymbol{\beta},\mathbf{V}):\boldsymbol{\beta}\in\mathbb{R}^q,\mathbf{V}\in PDS(k)\},\\\mathcal{F}_a&=\{g(\mathbf{s},\boldsymbol{\beta},\mathbf{V})s_a:g\in\mathcal{F}\},\\\mathcal{F}_{ab}&=\{g(\mathbf{s},\boldsymbol{\beta},\mathbf{V})s_as_b:g\in\mathcal{F}\},\end{aligned}$$

for $a,b=1,\ldots,p$. Denote by $\mathcal{G}$, $\mathcal{G}_a$, and $\mathcal{G}_{ab}$, the corresponding classes of graphs of the functions in $\mathcal{F}$, $\mathcal{F}_a$, and $\mathcal{F}_{ab}$, respectively. Then $\mathcal{G}$, $\mathcal{G}_a$, and $\mathcal{G}_{ab}$, all have polynomial discrimination for $a,b=1,\ldots,p$.

Proof. Because the function $u(\cdot)$ is of bounded variation, it follows from Lemma B.6 that the class $\mathcal{G}$ has polynomial discrimination. To show the same for the class $\mathcal{G}_a$, suppose that $\mathcal{G}_a$ is not of polynomial discrimination. This means that for every integer N there exists a set $V\subset\mathbb{R}^{p+1}$ of N points, such that all subsets of V can be written as $D_a\cap V$ for some $D_a\in\mathcal{G}_a$. Let $V=\{\mathbf{v}_1,\ldots,\mathbf{v}_N\}$, where $\mathbf{v}_i=(\mathbf{s}_i,t_i)$, for $i=1,\ldots,N$. Then for every $(\mathbf{s},t)\in V$ with $\mathbf{s}=(s_1,\ldots,s_a,\ldots,s_p)$, it must

hold that $s_a \neq 0$, otherwise this point cannot be separated from the other points by an element of the class $\mathcal{G}_a$. Define the set $V_a = \{(\mathbf{s}, t/s_a) : (\mathbf{s}, t) \in V\}$. Note that for $D_a \in \mathcal{G}_a$, we have

$$\begin{aligned}(\mathbf{s},t) \in D_a \quad &\Leftrightarrow \quad 0 \leq t \leq g(\mathbf{s},\boldsymbol{\beta},\mathbf{V})s_a \text{ or } g(\mathbf{s},\boldsymbol{\beta},\mathbf{V})s_a \leq t \leq 0 \\ &\Leftrightarrow \quad 0 \leq \frac{t}{s_a} \leq g(\mathbf{s},\boldsymbol{\beta},\mathbf{V}) \text{ or } g(\mathbf{s},\boldsymbol{\beta},\mathbf{V}) \leq \frac{t}{s_a} \leq 0 \\ &\Leftrightarrow \quad (\mathbf{s}, t/s_a) \in D,\end{aligned}$$

where $D \in \mathcal{G}$. This implies that every subset of V_a can be written as $D \cap V_a$, for some $D \in \mathcal{G}$. However, this is in contradiction with the fact that $\mathcal{G}$ has polynomial discrimination. We conclude that also $\mathcal{G}_a$ has polynomial discrimination. A similar argument yields that $\mathcal{G}_{ab}$ has polynomial discrimination. □

Lemma B.7 is comparable to Lemma 3 in [17], for similar classes of functions built from functions $g(\mathbf{y}, \mathbf{t}, \mathbf{C}) = u(\|\mathbf{C}^{-1/2}(\mathbf{y} - \mathbf{t})\|)$, where $\mathbf{t} \in \mathbb{R}^k$ and $\mathbf{C} \in \text{PDS}(k)$. With Lemma B.7 we can establish suitable bounds on the third term of (9.2). This is provided by the following key lemma. Once having established Lemma B.8, asymptotic normality can be derived easily from (9.2).

Lemma B.8. *Let $\Psi = (\Psi_{\boldsymbol{\beta}}, \Psi_{\boldsymbol{\theta}})$ be defined in (7.5) and let $\boldsymbol{\xi}_n = \boldsymbol{\xi}(\mathbb{P}_n)$ and $\boldsymbol{\xi}_P = \boldsymbol{\xi}(P)$ be the solutions to minimization problems (3.1) and (3.5). Suppose that ρ satisfies (R1)-(R4), such that $u(s)$ is of bounded variation, and suppose that $\mathbf{V}$ satisfies (V4). Suppose that $\boldsymbol{\xi}_n \to \boldsymbol{\xi}_P$, in probability, and that $\mathbb{E}\|\mathbf{s}\|^2 < \infty$. Then*

$$\int (\Psi(\mathbf{s}, \boldsymbol{\xi}_n) - \Psi(\mathbf{s}, \boldsymbol{\xi}_P)) \, \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) = o_P(1/\sqrt{n}). \tag{B.47}$$

Proof. First write $\Psi_{\boldsymbol{\theta},j}(\mathbf{s}, \boldsymbol{\xi}) = \Psi_{2,j}(\mathbf{s}, \boldsymbol{\xi}) - \Psi_{3,j}(\mathbf{s}, \boldsymbol{\xi})$, for $j = 1, \ldots, l$, where

$$\begin{aligned}\Psi_{2,j}(\mathbf{s}, \boldsymbol{\xi}) &= u(d)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} \mathbf{H}_j \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \Psi_{3,j}(\mathbf{s}, \boldsymbol{\xi}) &= \text{tr}\left(\mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \theta_j}\right) (\rho(d) - b_0),\end{aligned} \tag{B.48}$$

where $\mathbf{H}_j$ and $d = d(\mathbf{s}, \boldsymbol{\xi})$ are defined in (7.3) and (5.1). It suffices to show that

$$\int (\Psi_{\boldsymbol{\beta}}(\mathbf{s}, \boldsymbol{\xi}_n) - \Psi_{\boldsymbol{\beta}}(\mathbf{s}, \boldsymbol{\xi}_P)) \, \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) = o_P(1/\sqrt{n}), \tag{B.49}$$

$$\int (\Psi_{2,j}(\mathbf{s}, \boldsymbol{\xi}_n) - \Psi_{2,j}(\mathbf{s}, \boldsymbol{\xi}_P)) \, \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) = o_P(1/\sqrt{n}), \tag{B.50}$$

$$\int (\Psi_{3,j}(\mathbf{s}, \boldsymbol{\xi}_n) - \Psi_{3,j}(\mathbf{s}, \boldsymbol{\xi}_P)) \, \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) = o_P(1/\sqrt{n}), \tag{B.51}$$

for $j = 1, \ldots, l$.

To obtain (B.51), first write $\mathbf{M}_n = \mathbf{V}(\boldsymbol{\theta}_n)^{-1}$ and $\mathbf{M}_P = \mathbf{V}(\boldsymbol{\theta}_P)^{-1}$, so that $\mathbf{M}_n \to \mathbf{M}_P$, in probability, according to condition (V1). Decompose as follows

$$\begin{aligned}&\Psi_{3,j}(\mathbf{s}, \boldsymbol{\xi}_n) - \Psi_{3,j}(\mathbf{s}, \boldsymbol{\xi}_P) \\ &\quad = \text{tr}\left(\mathbf{M}_n^{-1} \frac{\partial \mathbf{V}(\boldsymbol{\theta}_n)}{\partial \theta_j}\right) (\rho(d(\mathbf{s}, \boldsymbol{\xi}_n)) - \rho(d(\mathbf{s}, \boldsymbol{\xi}_P))) \\ &\qquad + \left\{\text{tr}\left(\mathbf{M}_n^{-1} \frac{\partial \mathbf{V}(\boldsymbol{\theta}_n)}{\partial \theta_j}\right) - \text{tr}\left(\mathbf{M}_P^{-1} \frac{\partial \mathbf{V}(\boldsymbol{\theta}_P)}{\partial \theta_j}\right)\right\} (\rho(d(\mathbf{s}, \boldsymbol{\xi}_P)) - b_0).\end{aligned} \tag{B.52}$$

After integration with respect to $\mathbb{P}_n - P$, the first term on the right hand side of (B.52) becomes

$$\text{tr}\left(\mathbf{M}_n^{-1} \frac{\partial \mathbf{V}(\boldsymbol{\theta}_n)}{\partial \theta_j}\right) \int (\rho(d(\mathbf{s}, \boldsymbol{\xi}_n)) - \rho(d(\mathbf{s}, \boldsymbol{\xi}_P))) \, \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}), \tag{B.53}$$

where with (V4),

$$\operatorname{tr}\left(\mathbf{M}_n^{-1}\frac{\partial \mathbf{V}(\boldsymbol{\theta}_n)}{\partial \theta_j}\right) \to \operatorname{tr}\left(\mathbf{M}_P^{-1}\frac{\partial \mathbf{V}(\boldsymbol{\theta}_P)}{\partial \theta_j}\right) \tag{B.54}$$

in probability. Furthermore, note that all functions $\rho(d(\cdot,\boldsymbol{\xi}))$, for $\boldsymbol{\xi} \in \mathbb{R}^q \times \mathbb{R}^l$ are members of the class

$$\mathcal{F} = \left\{\rho(\|\mathbf{V}^{-1/2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\|) : \boldsymbol{\beta} \in \mathbb{R}^q, \mathbf{V} \in \mathrm{PDS}(k)\right\}.$$

From (R1)-(R2) it follows that the function ρ is of bounded variation (being the sum of two monotone functions). According to Lemma B.7, the class $\mathcal{G}$, consisting of subgraphs of functions in the class $\mathcal{F}$, has polynomial discrimination. Moreover, the class $\mathcal{F}$ has a constant envelope. Then as a result of the empirical process theory developed in Pollard [23] (e.g., see Theorem 1 in [17]), it follows that for every $\delta > 0$,

$$\sup_{f_1,f_2\in[\delta]} \sqrt{n}\left|\int (f_1(\mathbf{s}) - f_2(\mathbf{s}))\ \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s})\right| \to 0$$

in probability, where

$$[\delta] = \left\{(f_1,f_2) : f_1, f_2 \in \mathcal{F} \text{ and } \int (f_1-f_2)^2\,\mathrm{d}P \le \delta^2\right\}.$$

Because $\boldsymbol{\xi}_n \to \boldsymbol{\xi}_P$ in probability one, the pair of functions $f_1(\mathbf{s}) = \rho(d(\mathbf{s},\boldsymbol{\xi}_n)$ and $f_2(\mathbf{s}) = \rho(d(\mathbf{s},\boldsymbol{\xi}_P)$ are in the set $[\delta]$, for sufficiently large n, with probability tending to one. It follows that

$$\begin{aligned}&\sqrt{n}\left|\int (\rho(d(\mathbf{s},\boldsymbol{\xi}_n) - \rho(d(\mathbf{s},\boldsymbol{\xi}_P))\ \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s})\right| \\ &\quad\le \sup_{f_1,f_2\in[\delta]} \sqrt{n}\left|\int (f_1(\mathbf{s}) - f_2(\mathbf{s}))\ \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s})\right| \to 0,\end{aligned} \tag{B.55}$$

in probability. Together with the fact that $\operatorname{tr}(\mathbf{M}_P^{-1}\partial\mathbf{V}(\boldsymbol{\theta}_P)/\partial\theta_j)$ is bounded, this proves that (B.53) is of the order $o_P(1/\sqrt{n})$. For the second term on the right hand side of (B.52), we have that according to the central limit theorem

$$\int (\rho(d(\mathbf{s},\boldsymbol{\xi}_P)) - b_0)\,\mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) = O_P(1/\sqrt{n}).$$

Together with (B.54) this implies that, after integration with respect to $\mathbb{P}_n - P$, the second term on the right hand side of (B.52) is of the order $o_P(1/\sqrt{n})$. This proves (B.51).

To obtain (B.49), decompose as follows

$$\begin{aligned}&\Psi_{\boldsymbol{\beta}}(\mathbf{s},\boldsymbol{\xi}_n) - \Psi_{\boldsymbol{\beta}}(\mathbf{s},\boldsymbol{\xi}_P) \\ &\quad= \left\{u(d(\mathbf{s},\boldsymbol{\xi}_n))\mathbf{X}^T\mathbf{M}_n\mathbf{y} - u(d(\mathbf{s},\boldsymbol{\xi}_P))\mathbf{X}^T\mathbf{M}_P\mathbf{y}\right\} \\ &\qquad- \left\{u(d(\mathbf{s},\boldsymbol{\xi}_n))\mathbf{X}^T\mathbf{M}_n\mathbf{X}\boldsymbol{\beta}_n - u(d(\mathbf{s},\boldsymbol{\xi}_P))\mathbf{X}^T\mathbf{M}_P\mathbf{X}\boldsymbol{\beta}_P\right\}.\end{aligned} \tag{B.56}$$

We will treat both terms on the right hand side separately. For the first term on the right hand side of (B.56), we write

$$\left\{u(d(\mathbf{s},\boldsymbol{\xi}_n)) - u(d(\mathbf{s},\boldsymbol{\xi}_P))\right\}\mathbf{X}^T\mathbf{M}_n\mathbf{y} + u(d(\mathbf{s},\boldsymbol{\xi}_P)\mathbf{X}^T(\mathbf{M}_n - \mathbf{M}_P)\mathbf{y}. \tag{B.57}$$

First consider the first term in (B.57). We consider each single element of the vector $\mathbf{X}^T\mathbf{M}_n\mathbf{y} \in \mathbb{R}^q$ separately. For $i = 1,\ldots,q$ fixed, write

$$\left(\mathbf{X}^T\mathbf{M}_n\mathbf{y}\right)_i = \sum_{j=1}^k x_{ji}\left(\mathbf{M}_n\mathbf{y}\right)_j = \sum_{j=1}^k x_{ji}\sum_{s=1}^k m_{n,js}y_s = \sum_{j=1}^k\sum_{s=1}^k x_{ji}y_s m_{n,js}.$$

Hence, after integration with respect to $\mathbb{P}_n - P$, the i-th component of the first term in (B.57) can be written as a finite sum with summands

$$\int \Big(u(d(\mathbf{s}, \boldsymbol{\xi}_n)) - u(d(\mathbf{s}, \boldsymbol{\xi}_P))\Big) x_{ji} y_s \, \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) m_{n,js},$$

for $i = 1, \ldots, q$ and $j, s \in \{1, \ldots, k\}$ fixed, where $m_{n,js} \to m_{P,js}$, Because u is of bounded variation and since $\mathbf{s} = (\mathbf{y}, \mathbf{X}) = (y_1, \ldots, y_k, x_{11}, \ldots, x_{kq})$, all functions $u(d(\mathbf{s}, \boldsymbol{\xi}))x_{ji}y_s$, for $\boldsymbol{\xi} \in \mathbb{R}^q \times \mathbb{R}^l$, are members of the class

$$\mathcal{F}_{ab} = \Big\{ u(\|\mathbf{V}^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|) s_a s_b : \boldsymbol{\beta} \in \mathbb{R}^q, \mathbf{V} \in \mathrm{PDS}(k) \Big\}.$$

According to Lemma B.7, the corresponding class of subgraphs has polynomial discrimination, so similar to (B.55) it follows that for each $i = 1, \ldots, q$ and $j, s \in \{1, \ldots, k\}$ fixed,

$$\int \Big(u(d(\mathbf{s}, \boldsymbol{\xi}_n)) - u(d(\mathbf{s}, \boldsymbol{\xi}_P))\Big) x_{ji} y_s \, \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) = o_P(n^{-1/2}),$$

which means that

$$\int \Big\{u(d(\mathbf{s}, \boldsymbol{\xi}_n)) - u(d(\mathbf{s}, \boldsymbol{\xi}_P))\Big\} \mathbf{X}^T \mathbf{M}_P \mathbf{y} \, \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) = o_P(n^{-1/2}). \tag{B.58}$$

Next, consider the second term on the right hand side of (B.57). Similar to the first term, after integration with respect to $\mathbb{P}_n - P$, the i-th component of the second term on the right hand side of (B.57) can be written as a finite sum of summands

$$\int u(d(\mathbf{s}, \boldsymbol{\xi}_P)) x_{ji} y_s \, \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s})(m_{n,js} - m_{P,js}),$$

for $i = 1, \ldots, q$, and $j, s \in \{1, \ldots, k\}$ fixed. According to the central limit theorem, the integral is of the order $O_P(n^{-1/2})$, and since $m_{n,js} \to m_{P,js}$, in probability, it follows that the product is of the order $o_P(n^{-1/2})$. We conclude that

$$\int u(d(\mathbf{s}, \boldsymbol{\xi}_P)) \mathbf{X}^T (\mathbf{M}_n - \mathbf{M}_P) \mathbf{y} \, \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) = o_P(n^{-1/2}). \tag{B.59}$$

Putting together (B.58) and (B.59), it follows for the first term on the right hand side of (B.56) that

$$\int \Big\{u(d(\mathbf{s}, \boldsymbol{\xi}_n)) \mathbf{X}^T \mathbf{M}_n \mathbf{y} - u(d(\mathbf{s}, \boldsymbol{\xi}_P)) \mathbf{X}^T \mathbf{M}_P \mathbf{y}\Big\} \, \mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) = o_P(n^{-1/2}). \tag{B.60}$$

For the second term on the right hand side of (B.56), we write

$$\begin{aligned} & u(d(\mathbf{s}, \boldsymbol{\xi}_n)) \mathbf{X}^T \mathbf{M}_n \mathbf{X} \boldsymbol{\beta}_n - u(d(\mathbf{s}, \boldsymbol{\xi}_P)) \mathbf{X}^T \mathbf{M}_P \mathbf{X} \boldsymbol{\beta}_P \\ &= \Big(u(d(\mathbf{s}, \boldsymbol{\xi}_n)) - u(d(\mathbf{s}, \boldsymbol{\xi}_P))\Big) \mathbf{X}^T \mathbf{M}_n \mathbf{X} \boldsymbol{\beta}_n \\ &\quad + u(d(\mathbf{s}, \boldsymbol{\xi}_P) \Big(\mathbf{X}^T \mathbf{M}_n \mathbf{X} \boldsymbol{\beta}_n - \mathbf{X}^T \mathbf{M}_P \mathbf{X} \boldsymbol{\beta}_P\Big). \end{aligned} \tag{B.61}$$

Consider the first term on the right hand side of (B.61). For $i = 1, \ldots, q$ fixed, write

$$\begin{aligned} \big(\mathbf{X}^T \mathbf{M}_n \mathbf{X} \boldsymbol{\beta}_n\big)_i &= \sum_{j=1}^k x_{ji} (\mathbf{M}_n \mathbf{X} \boldsymbol{\beta}_n)_j = \sum_{j=1}^k x_{ji} \sum_{s=1}^k m_{n,js} (\mathbf{X} \boldsymbol{\beta}_n)_s \\ &= \sum_{j=1}^k x_{ji} \sum_{s=1}^k m_{n,js} \sum_{t=1}^k x_{st} \boldsymbol{\beta}_{n,t} = \sum_{j=1}^k \sum_{s=1}^k \sum_{t=1}^k x_{ji} x_{st} m_{n,js} \boldsymbol{\beta}_{n,t}. \end{aligned}$$

We see that, after integration with respect to $\mathbb{P}_n - P$, the i-th component of the first term on the right hand side of (B.61) can be written as a finite summation of summands

$$\int \Big(u(d(\mathbf{s},\boldsymbol{\xi}_n)) - u(d(\mathbf{s},\boldsymbol{\xi}_P))\Big) x_{ji}x_{st}\,\mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) m_{n,js}\boldsymbol{\beta}_{n,t},$$

for $i = 1,\ldots,q$ and $s,t,j \in \{1,\ldots,k\}$, where $m_{n,js} \to m_{P,js}$ and $\boldsymbol{\beta}_{n,t} \to \boldsymbol{\beta}_{P,t}$, in probability. All functions $u(d(\mathbf{s},\boldsymbol{\xi}))x_{ji}y_s$, for $\boldsymbol{\xi} \in \mathbb{R}^q \times \mathbb{R}^l$, are members of the class

$$\mathcal{F}_{ab} = \Big\{u(\|\mathbf{V}^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|)s_a s_b : \boldsymbol{\beta} \in \mathbb{R}^q, \mathbf{V} \in \mathrm{PDS}(k)\Big\},$$

and u is of bounded variation. According to Lemma B.7, the corresponding class of subgraphs has polynomial discrimination, so similar to (B.55) it follows that

$$\int \Big(u(d(\mathbf{s},\boldsymbol{\xi}_n)) - u(d(\mathbf{s},\boldsymbol{\xi}_P))\Big) x_{ji}x_{st}\,\mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) = o_P(n^{-1/2}),$$

which means that

$$\int \Big(u(d(\mathbf{s},\boldsymbol{\xi}_n)) - u(d(\mathbf{s},\boldsymbol{\xi}_P))\Big) \mathbf{X}^T\mathbf{M}_n\mathbf{X}\boldsymbol{\beta}_n\,\mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) = o_P(n^{-1/2}). \tag{B.62}$$

Next, consider the second term on the right hand side of (B.61). We then have to deal with summands of the form

$$\int u(d(\mathbf{s},\boldsymbol{\xi}_P))x_{ji}x_{st}\,\mathrm{d}(\mathbb{P}_n - P)(\mathbf{s})(m_{n,js}\boldsymbol{\beta}_{n,t} - m_{P,js}\boldsymbol{\beta}_{P,t}),$$

for $i = 1,\ldots,q$, and $s,t,j \in \{1,\ldots,k\}$, where $m_{n,js} \to m_{P,js}$ and $\boldsymbol{\beta}_{n,t} \to \boldsymbol{\beta}_{P,t}$, in probability. According to the central limit theorem, the integral is of the order $O_P(n^{-1/2})$. Because, $m_{n,is}\boldsymbol{\beta}_{n,t} \to m_{P,is}\boldsymbol{\beta}_{P,t}$, in probability, it follows that the product is of the order $o_P(n^{-1/2})$. We conclude,

$$\int u(d(\mathbf{s},\boldsymbol{\xi}_P)\Big(\mathbf{X}^T\mathbf{M}_n\mathbf{X}\boldsymbol{\beta}_n - \mathbf{X}^T\mathbf{M}_P\mathbf{X}\boldsymbol{\beta}_P\Big)\,\mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) = o_P(n^{-1/2}). \tag{B.63}$$

Putting together (B.60) and (B.63), proves (B.49)

Finally, consider $\Psi_{2,j}$ in (B.48), with $\mathbf{H}_j$ defined (7.3). Write

$$\begin{aligned}\mathbf{M}_n &= \mathbf{V}(\boldsymbol{\theta}_n)^{-1}\mathbf{H}_j(\boldsymbol{\theta}_n)\mathbf{V}(\boldsymbol{\theta}_n)^{-1}\\ \mathbf{M}_P &= \mathbf{V}(\boldsymbol{\theta}_P)^{-1}\mathbf{H}_j(\boldsymbol{\theta}_P)\mathbf{V}(\boldsymbol{\theta}_P)^{-1},\end{aligned}$$

so that $\mathbf{M}_n \to \mathbf{M}_P$, in probability, according to condition (V4). Decompose $\Psi_{2,j}(\mathbf{s},\boldsymbol{\xi}_n) - \Psi_{2,j}(\mathbf{s},\boldsymbol{\xi}_P)$ as follows

$$\begin{aligned}&\Big\{u(d(\mathbf{s},\boldsymbol{\xi}_n)) - u(d(\mathbf{s},\boldsymbol{\xi}_P))\Big\}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_n)^T\mathbf{M}_n(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_n)\\ &\quad + u(d(\mathbf{s},\boldsymbol{\xi}_P))\Big\{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_n)^T\mathbf{M}_n(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_n) - (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_P)^T\mathbf{M}_P(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_P)\Big\}.\end{aligned} \tag{B.64}$$

The first term in (B.64), can be written as the trace of the matrix

$$\Big\{u(d(\mathbf{s},\boldsymbol{\xi}_n)) - u(d(\mathbf{s},\boldsymbol{\xi}_P))\Big\}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_n)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_n)^T\mathbf{M}_n,$$

where $\boldsymbol{\beta}_n \to \boldsymbol{\beta}_P$ and $\mathbf{M}_n \to \mathbf{M}_P$, in probability. As before, we consider each single entry of this $k \times k$ matrix. The (i,j)-th element of $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_n)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_n)^T$ is equal to

$$\begin{aligned}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_n)_i(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_n)_j = y_iy_j &- \sum_{s=1}^k y_jx_{is}\beta_{n,s} - \sum_{t=1}^k y_ix_{jt}\beta_{n,t}\\ &+ \sum_{s=1}^k\sum_{t=1}^k x_{is}x_{jt}\beta_{n,s}\beta_{n,t}.\end{aligned} \tag{B.65}$$

We see that the (i,j)-th entry of

$$\Big\{u(d(\mathbf{s},\boldsymbol{\xi}_n)) - u(d(\mathbf{s},\boldsymbol{\xi}_P))\Big\}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_n)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_n)^T$$

is a combination of four summations. The last of these summations arising from (B.65), after integration with respect to $\mathbb{P}_n - P$, has summands

$$\int \Big(u(d(\mathbf{s},\boldsymbol{\xi}_n)) - u(d(\mathbf{s},\boldsymbol{\xi}_P))\Big)x_{is}x_{jt}\,\mathrm{d}(\mathbb{P}_n - P)(\mathbf{s})\,\beta_{n,s}\beta_{n,t},$$

where $\beta_{n,s}\beta_{n,t} \to \beta_{P,s}\beta_{P,t}$, in probability. All functions $u(d(\mathbf{s},\boldsymbol{\xi}))x_{ji}y_s$, for $\boldsymbol{\xi} \in \mathbb{R}^q \times \mathbb{R}^l$, are members of the class

$$\mathcal{F}_{ab} = \Big\{u(\|\mathbf{V}^{-1/2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\|)s_a s_b : \boldsymbol{\beta} \in \mathbb{R}^q, \mathbf{V} \in \mathrm{PDS}(k)\Big\},$$

where u is of bounded variation According to Lemma B.7, the corresponding class of subgraphs has polynomial discrimination, so similar to (B.55) it follows that

$$\int \Big\{u(d(\mathbf{s},\boldsymbol{\xi}_n)) - u(d(\mathbf{s},\boldsymbol{\xi}_P))\Big\}x_{is}x_{jt}\,\mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) = o_P(n^{-1/2}).$$

The other three summations that arise from the right hand side of (B.65) can be handled in the same way. It follows, that for each $i,j \in \{1,\ldots,k\}$ fixed,

$$\int \Big\{u(d(\mathbf{s},\boldsymbol{\xi}_n)) - u(d(\mathbf{s},\boldsymbol{\xi}_P))\Big\}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_n)_i(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_n)_j\,\mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) = o_P(n^{-1/2}),$$

which means that

$$\int \Big\{u(d(\mathbf{s},\boldsymbol{\xi}_n)) - u(d(\mathbf{s},\boldsymbol{\xi}_P))\Big\}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_n)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_n)^T\mathbf{M}_n\,\mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) = o_P(n^{-1/2}).$$

After taking traces, we conclude that

$$\int \Big\{u(d(\mathbf{s},\boldsymbol{\xi}_n)) - u(d(\mathbf{s},\boldsymbol{\xi}_P))\Big\}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_n)^T\mathbf{M}_n(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_n)\,\mathrm{d}(\mathbb{P}_n - P)(\mathbf{s}) = o_P(n^{-1/2}). \tag{B.66}$$

Next, consider the second term on the right hand side of (B.64). First, note that

$$\begin{aligned}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{M}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = &\sum_{i=1}^k\sum_{j=1}^k y_i y_j m_{ij} + \sum_{i=1}^k\sum_{j=1}^k y_i \sum_{s=1}^k x_{js}\beta_s m_{ij} \\ &+ \sum_{i=1}^k\sum_{j=1}^k y_j \sum_{t=1}^k x_{it}\beta_t m_{ij} + \sum_{i=1}^k\sum_{j=1}^k\sum_{s=1}^k\sum_{t=1}^k x_{is}x_{jt}\beta_s\beta_t m_{ij}.\end{aligned}$$

This means that

$$\begin{aligned}&(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_n)^T\mathbf{M}_n(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_n) - (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_P)^T\mathbf{M}_P(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_P) \\ &= \sum_{i=1}^k\sum_{j=1}^k y_i y_j\,(m_{n,ij} - m_{P,ij}) + \sum_{i=1}^k\sum_{j=1}^k\sum_{s=1}^k y_i x_{js}\,(\beta_{n,s}m_{n,ij} - \beta_{P,s}m_{P,ij}) \\ &\quad + \sum_{i=1}^k\sum_{j=1}^k\sum_{t=1}^k y_j x_{it}\,(\beta_{n,t}m_{n,ij} - \beta_{P,t}m_{P,ij}) \\ &\quad + \sum_{i=1}^k\sum_{j=1}^k\sum_{s=1}^k\sum_{t=1}^k x_{is}x_{jt}\,(\beta_{n,s}\beta_{n,t}m_{n,ij} - \beta_{P,s}\beta_{P,t}m_{P,ij})\,.\end{aligned} \tag{B.67}$$

We see that the (i,j)-th entry of

$$u(d(\mathbf{s},\boldsymbol{\xi}_P))\Big\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_n)^T\mathbf{M}_n(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_n)-(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_P)^T\mathbf{M}_P(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_P)\Big\}$$

can be written as the combination of four summations. The last of the summations arising from (B.67), after integration with respect to $\mathbb{P}_n - P$, has summands

$$\int u(d(\mathbf{s},\boldsymbol{\xi}_P))x_{is}x_{jt}\,\mathrm{d}(\mathbb{P}_n-P)(\mathbf{s})\,(\beta_{n,s}\beta_{n,t}m_{n,ij}-\beta_{P,s}\beta_{P,t}m_{P,ij})\,,$$

where $\beta_{n,s}\beta_{n,t}m_{n,ij} \to \beta_{P,s}\beta_{P,t}m_{P,ij}$, in probability. According to the central limit theorem, the integral is of the order $O_P(n^{-1/2})$, whereas the second term tends to zero. All functions $u(d(\mathbf{s},\boldsymbol{\xi}))x_{is}x_{jt}$, for $\boldsymbol{\xi}\in\mathbb{R}^q\times\mathbb{R}^l$, are members of the class

$$\mathcal{F}_{ab}=\Big\{u(\|\mathbf{V}^{-1/2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\|)s_as_b:\boldsymbol{\beta}\in\mathbb{R}^q,\mathbf{V}\in\mathrm{PDS}(k)\Big\}\,,$$

where u is of bounded variation According to Lemma B.7, the corresponding class of subgraphs has polynomial discrimination, so similar to (B.55) it follows that

$$\int u(d(\mathbf{s},\boldsymbol{\xi}_P))x_{is}x_{jt}\,\mathrm{d}(\mathbb{P}_n-P)(\mathbf{s})\,(\beta_{n,s}\beta_{n,t}m_{n,ij}-\beta_{P,s}\beta_{P,t}m_{P,ij})=o_P(n^{-1/2}).$$

The other three summations that arise from the right hand side of (B.67) can be handled in the same way, so that

$$\begin{aligned}&\int u(d(\mathbf{s},\boldsymbol{\xi}_P))\Big\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_n)^T\mathbf{M}_n(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_n)-(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_P)^T\mathbf{M}_P(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_P)\Big\}\,\mathrm{d}(\mathbb{P}_n-P)(\mathbf{s})\\&\quad=o_P(n^{-1/2}).\end{aligned}\tag{B.68}$$

Putting together (B.66) and (B.68), proves (B.50) for each $j=1,\ldots,l$. This finishes the proof of Lemma B.8. □

Proof of Corollary 6

Proof. As in the proof of Corollary 5, it follows that $\partial\Lambda/\partial\boldsymbol{\xi}$ is continuously differentiable with a non-singular derivative at $\boldsymbol{\xi}(P)$, so that Theorem 6 applies. According to Theorem 6 and Lemma 2, it follows that $\sqrt{n}(\boldsymbol{\beta}_n-\boldsymbol{\beta}(P))$ is asymptotically normal with mean zero and covariance matrix

$$\frac{1}{\alpha^2}\left(\mathbb{E}\left[\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}\right]\right)^{-1}\mathbb{E}\left[\Psi_{\boldsymbol{\beta}}(\mathbf{s},\boldsymbol{\xi}_P)\Psi_{\boldsymbol{\beta}}(\mathbf{s},\boldsymbol{\xi}_P)^T\right]\left(\mathbb{E}\left[\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}\right]\right)^{-1}$$

where $\Psi_{\boldsymbol{\beta}}$ is defined in (7.10). We find that

$$\mathbb{E}\left[\Psi_{\boldsymbol{\beta}}(\mathbf{s},\boldsymbol{\xi}_P)\Psi_{\boldsymbol{\beta}}(\mathbf{s},\boldsymbol{\xi}_P)^T\right]=\mathbb{E}\left[\mathbf{X}^T\mathbb{E}\left[u(d)^2\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\middle|\mathbf{X}\right]\mathbf{X}\right],$$

where $d^2=(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$ and $u(s)=\rho'(s)/s$. As before, with $\mathbf{z}=\boldsymbol{\Sigma}^{-1/2}(\mathbf{y}-\boldsymbol{\mu})$ and $\mathbf{u}=\mathbf{z}/\|\mathbf{z}\|$, according to Lemma B.4, the inner conditional expectation can be written as

$$\boldsymbol{\Sigma}^{-1/2}\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[u(\|\mathbf{z}\|)^2\|\mathbf{z}\|^2\right]\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\mathbf{u}\mathbf{u}^T\right]\boldsymbol{\Sigma}^{-1/2}=\frac{\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[u(\|\mathbf{z}\|)^2\|\mathbf{z}\|^2\right]}{k}\boldsymbol{\Sigma}^{-1}.$$

This implies that the asymptotic covariance of $\sqrt{n}(\boldsymbol{\beta}_n-\boldsymbol{\beta}(P))$ is given by

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

Again, according to Theorem 6 and Lemma 2, it follows that $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}(P))$ is asymptotically normal with mean zero and covariance matrix

$$\left(\frac{\partial \Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}(P))}{\partial \boldsymbol{\theta}}\right)^{-1} \mathbb{E}\left[\operatorname{vec}\left(\Psi_{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\xi}_P)\right) \operatorname{vec}\left(\Psi_{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\xi}_P)\right)^T\right] \left(\frac{\partial \Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}(P))}{\partial \boldsymbol{\theta}}\right)^{-1}$$

where $\Psi_{\boldsymbol{\theta}}$ is defined in (7.10). We have

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{vec}\left(\Psi_{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\xi}_P)\right) \operatorname{vec}\left(\Psi_{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\xi}_P)\right)^T\right] \\
&= \mathbf{E}^T \mathbb{E}\left[\operatorname{vec}\left(\boldsymbol{\Sigma}^{-1/2} \Psi_{\mathbf{V}}(\mathbf{s}, \boldsymbol{\xi}_P) \boldsymbol{\Sigma}^{-1/2}\right) \operatorname{vec}\left(\boldsymbol{\Sigma}^{-1/2} \Psi_{\mathbf{V}}(\mathbf{s}, \boldsymbol{\xi}_P) \boldsymbol{\Sigma}^{-1/2}\right)^T\right] \mathbf{E}
\end{aligned}$$

where $\Psi_{\mathbf{V}}$ is defined in (7.8) and $\mathbf{E} = \left(\boldsymbol{\Sigma}^{-1/2} \otimes \boldsymbol{\Sigma}^{-1/2}\right) \mathbf{L}$ and

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{vec}\left(\boldsymbol{\Sigma}^{-1/2} \Psi_{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\xi}_P) \boldsymbol{\Sigma}^{-1/2}\right) \operatorname{vec}\left(\boldsymbol{\Sigma}^{-1/2} \Psi_{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\xi}_P) \boldsymbol{\Sigma}^{-1/2}\right)^T\right] \\
&= k^2 \mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[u(\|\mathbf{z}\|)^2 \|\mathbf{z}\|^4\right] \mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[\operatorname{vec}\left(\mathbf{u}\mathbf{u}^T\right) \operatorname{vec}\left(\mathbf{u}\mathbf{u}^T\right)^T\right] \\
&\quad - k \mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[u(\|\mathbf{z}\|) v(\|\mathbf{z}\|) \|\mathbf{z}\|^2\right] \mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[\operatorname{vec}\left(\mathbf{u}\mathbf{u}^T\right) \operatorname{vec}\left(\mathbf{I}_k\right)^T\right] \\
&\quad - k \mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[u(\|\mathbf{z}\|) v(\|\mathbf{z}\|) \|\mathbf{z}\|^2\right] \mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[\operatorname{vec}\left(\mathbf{I}_k\right) \operatorname{vec}\left(\mathbf{u}\mathbf{u}^T\right)^T\right] \\
&\quad + \mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[v(\|\mathbf{z}\|)^2\right] \mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[\operatorname{vec}\left(\mathbf{I}_k\right) \operatorname{vec}\left(\mathbf{I}_k\right)^T\right].
\end{aligned}$$

From Lemma B.4, the first term on the right hand side is equal to

$$\frac{k \mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[u(\|\mathbf{z}\|)^2 \|\mathbf{z}\|^4\right]}{k+2} \left(\mathbf{I}_{k^2} + \mathbf{K}_{k,k} + \operatorname{vec}(\mathbf{I}_k) \operatorname{vec}(\mathbf{I}_k)^T\right).$$

This leads to one term $\mathbf{I}_{k^2} + \mathbf{K}_{k,k}$ with coefficient

$$\frac{k \mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[u(\|\mathbf{z}\|)^2 \|\mathbf{z}\|^4\right]}{k+2}$$

and using that, according to Lemma B.4, $\mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[\mathbf{u}\mathbf{u}^T\right] = (1/k)\mathbf{I}_k$, we find a second term $\operatorname{vec}(\mathbf{I}_k)\operatorname{vec}(\mathbf{I}_k)^T$ with coefficient

$$\frac{k \mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[u(\|\mathbf{z}\|)^2 \|\mathbf{z}\|^4\right]}{k+2} - 2 \mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[u(\|\mathbf{z}\|) v(\|\mathbf{z}\|) \|\mathbf{z}\|^2\right] + \mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[v(\|\mathbf{z}\|)^2\right].$$

Since $v(s) = u(s)s^2 - \rho(s) + b_0$, we have that

$$\frac{k}{k+2} u(s)^2 s^4 - 2u(s)v(s)s^2 + v(s)^2 = -\frac{2}{k+2} u(s)^2 s^4 + (\rho(s) - b_0)^2.$$

This means that

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{vec}\left(\boldsymbol{\Sigma}^{-1/2} \Psi_{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\xi}_P) \boldsymbol{\Sigma}^{-1/2}\right) \operatorname{vec}\left(\boldsymbol{\Sigma}^{-1/2} \Psi_{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\xi}_P) \boldsymbol{\Sigma}^{-1/2}\right)^T\right] \\
&= \delta_1 \left(\mathbf{I}_{k^2} + \mathbf{K}_{k,k}\right) + \delta_2 \operatorname{vec}(\mathbf{I}_k) \operatorname{vec}(\mathbf{I}_k)^T
\end{aligned}$$

where

$$\begin{aligned}
\delta_1 &= \frac{k \mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[u(\|\mathbf{z}\|)^2 \|\mathbf{z}\|^4\right]}{k+2} \\
\delta_2 &= -\frac{2}{k+2} \mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[u(\|\mathbf{z}\|)^2 \|\mathbf{z}\|^4\right] + \mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[\left(\rho(\|\mathbf{z}\|) - b_0\right)^2\right] \\
&= -\frac{2}{k} \delta_1 + \mathbb{E}_{\mathbf{0}, \mathbf{I}_k}\left[\left(\rho(\|\mathbf{z}\|) - b_0\right)^2\right]
\end{aligned} \tag{B.69}$$

Since, $\mathbf{K}_{k,k}(\mathbf{\Sigma}^{-1/2} \otimes \mathbf{\Sigma}^{-1/2}) = (\mathbf{\Sigma}^{-1/2} \otimes \mathbf{\Sigma}^{-1/2})\mathbf{K}_{k,k}$, together with (B.32), this implies that

$$\begin{aligned}\mathbb{E}\left[\mathrm{vec}\left(\Psi_{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\xi}_P)\right) \mathrm{vec}\left(\Psi_{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\xi}_P)\right)^T\right] &= \mathbf{E}^T \left(\delta_1 \left(\mathbf{I}_{k^2} + \mathbf{K}_{k,k}\right) + \delta_2 \mathrm{vec}(\mathbf{I}_k)\mathrm{vec}(\mathbf{I}_k)^T\right) \mathbf{E} \\ &= 2\delta_1 \mathbf{E}^T\mathbf{E} + \delta_2 \mathbf{E}^T \mathrm{vec}(\mathbf{I}_k)\mathrm{vec}(\mathbf{I}_k)^T\mathbf{E}.\end{aligned}$$

Furthermore, according to Lemma B.5,

$$\left(\frac{\partial \Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}(P))}{\partial \boldsymbol{\theta}}\right)^{-1} = a(\mathbf{E}^T\mathbf{E})^{-1} + b(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathrm{vec}(\mathbf{I}_k)\mathrm{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1},$$

with a and b defined in (B.44). By application of (B.39), (B.40), and (B.42), we find that

$$\begin{aligned}&\left(\frac{\partial \Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}(P))}{\partial \boldsymbol{\theta}}\right)^{-1} \mathbb{E}\left[\mathrm{vec}\left(\Psi_{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\xi}_P)\right) \mathrm{vec}\left(\Psi_{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\xi}_P)\right)^T\right] \\ &= 2a\delta_1\mathbf{I}_k + (a\delta_2 + 2b\delta_1 + bk\delta_2)(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathrm{vec}(\mathbf{I}_k)\mathrm{vec}(\mathbf{I}_k)^T\mathbf{E}\end{aligned}$$

and

$$\begin{aligned}&\left(\frac{\partial \Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}(P))}{\partial \boldsymbol{\theta}}\right)^{-1} \mathbb{E}\left[\mathrm{vec}\left(\Psi_{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\xi}_P)\right) \mathrm{vec}\left(\Psi_{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\xi}_P)\right)^T\right] \left(\frac{\partial \Lambda_{\boldsymbol{\theta}}(\boldsymbol{\xi}(P))}{\partial \boldsymbol{\theta}}\right)^{-1} \\ &= 2\sigma_1(\mathbf{E}^T\mathbf{E})^{-1} + \sigma_2(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathrm{vec}(\mathbf{I}_k)\mathrm{vec}(\mathbf{I}_k)^T\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1} \\ &= 2\sigma_1(\mathbf{E}^T\mathbf{E})^{-1} + \sigma_2\boldsymbol{\theta}(P)\boldsymbol{\theta}(P)^T\end{aligned}$$

where

$$\begin{aligned}\sigma_1 &= a^2\delta_1 \\ \sigma_2 &= 2b(2a + kb)\delta_1 + (a + kb)^2\delta_2.\end{aligned}$$

When we insert the expressions for δ_1, δ_2, a, and b given in (B.69) and (B.44), then we find

$$\begin{aligned}\sigma_1 &= \frac{k(k+2)\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[u(\|\mathbf{z}\|)^2\|\mathbf{z}\|^4\right]}{\left(\mathbb{E}_{0,\mathbf{I}_k}\left[\rho''(\|\mathbf{z}\|)\|\mathbf{z}\|^2 + (k+1)\rho'(\|\mathbf{z}\|)\|\mathbf{z}\|\right]\right)^2} \\ \sigma_2 &= -\frac{2}{k}\sigma_1 + \frac{4\mathbb{E}_{\mathbf{0},\mathbf{I}_k}[(\rho(\|\mathbf{z}\|) - b_0)^2]}{\left(\mathbb{E}_{\mathbf{0},\mathbf{I}_k}\left[\rho'(\|\mathbf{z}\|)\|\mathbf{z}\|\right]\right)^2}\end{aligned}$$

By substituting $\mathbf{E} = \left(\mathbf{\Sigma}^{-1/2} \otimes \mathbf{\Sigma}^{-1/2}\right)\mathbf{L}$, we find that the limiting covariance of $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}(P))$ is given by

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

This finishes the proof. □