KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics

Multivariate generalized S-estimators

E. Roelant, S. Van Aelst and C. Croux

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 0802

Multivariate Generalized S-estimators

Roelant E. [a,*] Van Aelst S. [a] Croux C. [b]

[a] *Department of Applied Mathematics and Computer Science, Ghent University - UGent, Krijgslaan 281-S9, B-9000 Gent, Belgium*

[b] *Katholieke Universiteit Leuven, University Centre of Statistics, Naamsestraat 69, B-3000 Leuven, Belgium*

Abstract

In this paper we introduce generalized S-estimators for the multivariate regression model. This class of estimators combines high robustness and high efficiency. They are defined by minimizing the determinant of a robust estimator of the scatter matrix of differences of residuals. In the special case of a multivariate location model, the generalized S-estimator has the important independency property, and can be used for high breakdown estimation in independent component analysis. Robustness properties of the estimators are investigated by deriving their breakdown point and the influence function. We also study the efficiency of the estimators, both asymptotically and at finite samples. To obtain inference for the regression parameters, we discuss the fast and robust bootstrap for multivariate generalized S-estimators. The method is illustrated on several real data examples.

Key words: Bootstrap, efficiency, multivariate regression, robustness.

* Corresponding author. Tel.:+32-92644756; fax: +32-92644995
Email address: Ella.Roelant@ugent.be (Roelant E.).

1 Introduction

In this paper we introduce a new class of estimators for the multivariate regression model, called Generalized S-estimators (GS). Generalized S-estimators are defined by minimizing the determinant of a robust estimator of the scatter matrix of differences of residuals. Using differences instead of the residuals themselves has several advantages. First of all, at most models this will lead to an increase in statistical efficiency, while the robustness of the estimators, as measured by their breakdown point, remains the same. The breakdown point of an estimator is the highest possible percentage of outliers than an estimator can withstand. It turns out to be possible to achieve the highest possible value for the breakdown point, 50%, even when working with differences of residuals. A second advantage is that GS-estimators allow to estimate the slope and the scatter matrix of the error terms of the multivariate regression model, without needing to estimate the intercept. Hence, the estimation procedure is "intercept free."

The multivariate regression model encompasses both the multivariate location-scale model, as a multivariate regression model with only an intercept, and the univariate regression model. While GS-estimators were already considered for univariate regression (Croux et al. 1994), they were not studied yet for the multivariate location-scale model. In the latter model, the "intercept free" property of the GS estimator translates into "location free" estimation. Hence, GS-estimators allow for estimation of scatter while not needing to estimate the location. Moreover, since the GS-estimator is based on differences, it has the independence property, meaning that when the components of a random vector are independent, the scatter matrix estimate is diagonal (Tyler et

al. 2007). This is not true for S-estimators of scatter in general. The independence property is highly important in independent component analysis (ICA). Briefly, the ICA problem consists of finding an original random vector with independent components when only an unknown linear mixture is observed (Hyvärinen et al. 2001). Oja et al. (2006) proposed a method for ICA that is based on the use of two different scatter matrices that are required to have the independence property; see also Tyler et al (2007). By using the GS-estimator, a high breakdown approach to robust ICA is obtained. Other scatter matrix estimators, based on differences of observations were proposed by Dümbgen (1998), and Sirkiä et al. (2007). They are of the M-type and their breakdown point decreases with the dimension (Dümbgen and Tyler 2005), and thus do not have a high degree of robustness.

Consider the multivariate linear regression model given by

$$\mathbf{y} = \alpha + \mathcal{B}^T\mathbf{u} + \epsilon \tag{1}$$

where $\mathbf{u}$ is the p-variate predictor, $\mathbf{y}$ the q-variate response and ϵ the q-variate error term which has center zero and a positive definite scatter matrix Σ. The unknown parameters $\theta = (\alpha, \mathcal{B}^T)^T \in \mathbb{R}^{(p+1)\times q}$ and $\Sigma \in \mathbb{R}^{q\times q}$ are to be estimated from the observations $\mathcal{Z}_n = \{\mathbf{z}_i := (\mathbf{x}_i^T, \mathbf{y}_i^T)^T = (1, \mathbf{u}_i^T, \mathbf{y}_i^T)^T, i = 1, \ldots, n\} \subset \mathbb{R}^{p+q+1}$. The classical estimator for this model is the least squares estimator, but it is well known that this estimator can be highly influenced by outliers.

In the univariate regression case a lot of research has been done to construct more robust estimators. Classes of robust estimators in this setting include M-estimators (Hampel et al. 1986), least median of squares and least trimmed squares estimators (Rousseeuw 1984), S-estimators (Rousseeuw and

Yohai 1984), MM-estimators (Yohai 1987), CM-estimators (Mendes and Tyler 1996) and τ-estimators (Yohai and Zamar 1988). Croux et al. (1994) introduced a class of regression estimators, called generalized S-estimators or GS-estimators. While an S-estimator of regression minimizes an S-estimator of scale of the residuals, a GS-estimator minimizes an S-estimator of scale applied on the pairwise differences of the residuals, instead of on the residuals themselves. It has been shown that for bounded loss functions these univariate GS-estimators have nice properties such as a high breakdown point, a higher efficiency than the original S-estimators. Moreover, they do not require the assumption of asymmetric errors (see also Hössjer et al. 1994, Berrendero and Romo 1998 and Berrendero 2002). In this paper, we extend the definition of GS-estimates to multivariate regression.

Recently, several robust estimators for multivariate regression have been introduced. Methods based on robust estimators for multivariate location and scatter applied to the joint distribution of responses and explanatory variables have been proposed by Ollila, Oja and Hettmansperger (2002) using sign covariance matrices, Ollila, Oja and Koivunen (2003) using rank covariance matrices and Rousseeuw et al. (2004) using the minimum covariance determinant estimator. An alternative approach is to define a robust regression estimator by minimizing a robust estimate of the covariance matrix of the residuals. Agulló et al. (2008) proposed the multivariate least trimmed squares estimator, Van Aelst and Willems (2005) considered multivariate regression S-estimators, while Ben, Martinez and Yohai (2006) introduced τ-estimators for multivariate regression. All these procedures, however, are not based on differences of residuals, and are not intercept or location free.

The remainder of the paper is organized as follows. In Section 2 we intro-

duce the multivariate regression GS-estimators and determine their breakdown point. Section 3 describes the algorithm for computing the GS-estimators. In Section 4 we define the functional form of the estimator. We show that the GS-functional is Fisher-consistent if the differences of the errors have an elliptical distribution. We also derive the influence function of the GS-functional. Asymptotic variances and corresponding efficiencies are given in Section 5. Section 6 discusses the fast and robust bootstrap method for GS-estimators. Section 7 presents two real data examples and Section 8 concludes. All the proofs can be found in the Appendix.

2 Definition and breakdown point

We now define Generalized S-estimators for the multivariate regression model given in (1).

Definition 1 *Let* $\mathcal{Z}_n = \{\mathbf{z}_i := (\mathbf{x}_i^T, \mathbf{y}_i^T)^T = (1, \mathbf{u}_i^T, \mathbf{y}_i^T)^T, i = 1, \ldots, n\} \subset \mathbb{R}^{p+q+1}$. *The GS-estimates of multivariate regression* $(\widehat{\mathcal{B}}_n, \widehat{\Sigma}_n)$ *minimizes among all* $(B, C) \in \mathbb{R}^{p\times q} \times PDS(q)$, *with* $PDS(q)$ *the set of positive definite symmetric* $q \times q$ *matrices, the determinant* $|C|$, *subject to the condition*

$$\binom{n}{2}^{-1} \sum_{i<j} \rho([(\mathbf{r}_i - \mathbf{r}_j)^T C^{-1} (\mathbf{r}_i - \mathbf{r}_j)]^{1/2}) = k \tag{2}$$

where $\mathbf{r}_i = \mathbf{y}_i - B^T \mathbf{u}_i - \alpha$.

Note that the objective function does not depend on the intercept α. The constant k can be chosen as $k = E_{F\times F}[\rho(\|\epsilon_1 - \epsilon_2\|)]$, which ensures consistency at the model with error distribution F (see Section 4). The choice $\rho(u) = u^2$ yields the non-robust least squares (LS) estimator. To obtain robust estimates,

we impose the following properties on the loss function ρ:

- ρ is symmetric, twice continuously differentiable and $\rho(0) = 0$
- ρ is strictly increasing on $[0, c]$ and constant on $[c, \infty)$ for some $c < \infty$.

Throughout this paper we use the well-known class of Tukey biweight ρ-functions given by:

$$\rho_c(t) = \begin{cases} \frac{t^2}{2} - \frac{t^4}{2c^2} + \frac{t^6}{6c^4}, & |t| \leq c \\ \\ \frac{c^2}{6}, & |t| \geq c \end{cases}$$

Similarly as in Lopuhaä (1989), it can be shown that definition 1 implies that multivariate GS-estimators satisfy the following first-order conditions:

$$\sum_{i<j} u(d_{ij})(\mathbf{u}_i - \mathbf{u}_j)(\mathbf{y}_i - \mathbf{y}_j - B^T(\mathbf{u}_i - \mathbf{u}_j))^T = \mathbf{0} \tag{3}$$

$$\sum_{i<j} \{qu(d_{ij})(\mathbf{y}_i - \mathbf{y}_j - B^T(\mathbf{u}_i - \mathbf{u}_j))(\mathbf{y}_i - \mathbf{y}_j - B^T(\mathbf{u}_i - \mathbf{u}_j))^T - v(d_{ij})C\} = \mathbf{0} \tag{4}$$

with $d_{ij}^2 = (\mathbf{y}_i - \mathbf{y}_j - B^T(\mathbf{u}_i - \mathbf{u}_j))^T C^{-1}(\mathbf{y}_i - \mathbf{y}_j - B^T(\mathbf{u}_i - \mathbf{u}_j))$, $u(t) = \psi(t)/t$ and $v(t) = \psi(t)t - \rho(t) + k$, where $\psi(t) = \rho'(t)$.

To study the global robustness of the multivariate GS-estimators, we derive their finite-sample breakdown point. For a given data set $\mathcal{Z}_n$, the finite-sample breakdown point ϵ_n^* of an estimator T_n is the smallest fraction of observations of $\mathcal{Z}_n$ that need to be replaced by arbitrary values to carry the estimate T_n beyond all bounds (Donoho and Huber 1983). Formally,

$$\epsilon_n^*(T_n, \mathcal{Z}_n) = \min \left\{ \frac{m}{n}; \sup_{\mathcal{Z}_n'} \|T_n(\mathcal{Z}_n) - T_n(\mathcal{Z}_n')\| = \infty \right\}$$

where the supremum is over all possible collections $\mathcal{Z}_n'$ that differ from $\mathcal{Z}_n$ in at most m data points. The breakdown point of a covariance estimator is

the smallest fraction of outliers that can make the first eigenvalue arbitrarily large or the last eigenvalue arbitrarily small. We derive the breakdown point for data sets that satisfy the following general position condition.

Condition 1 *The differences of the observations* $(\mathbf{u}_i^T, \mathbf{y}_i^T)^T$ *are in general position, meaning that no* $\binom{p+q+1}{2}$ *of the differences* $((\mathbf{u}_i - \mathbf{u}_j)^T, (\mathbf{y}_i - \mathbf{y}_j)^T)^T$ *with* $i < j$ *belong to the same hyperplane in* $\mathbb{R}^{p+q}$.

Note that if the differences of the $(\mathbf{u}_i^T, \mathbf{y}_i^T)^T$ are in general position, then the points $(\mathbf{u}_i^T, \mathbf{y}_i^T)^T$ themselves are also in general position. The latter means that no $p+q+1$ of the $(\mathbf{u}_i^T, \mathbf{y}_i^T)^T$ lie on the same hyperplane of $\mathbb{R}^{p+q}$. If the observations are sampled from a continuous distribution, then condition 1 holds with probability 1.

The breakdown point of multivariate regression GS-estimators, given next, extends the results for the univariate regression case in Croux et al. (1994).

Theorem 1 *Let* $\mathcal{Z}_n \subset \mathbb{R}^{p+q+1}$. *Denote* $r := k/\sup(\rho)$. *If* $\mathcal{Z}_n$ *satisfies condition 1 and* $\binom{n}{2}(1-r) \geq \binom{p+q+1}{2}$ *then the breakdown point of the multivariate GS-estimator is given by*

$$\epsilon_n^*(\widehat{\mathcal{B}}_n, \mathcal{Z}_n) = \epsilon_n^*(\widehat{\Sigma}_n, \mathcal{Z}_n) = \frac{1}{n}\min(\lceil n - 1/2 - \sqrt{1 + (1-r)(4n^2 - 4n)}/2 \rceil, \lceil 1/2 - p - q + \sqrt{1 + (1-r)(4n^2 - 4n)}/2 \rceil).$$

The maximal breakdown point is achieved for $r = 1 - ((n - 1 + p + q)^2 - 1)/(4n^2 - 4n)$, in which case $\epsilon_n^* = \lceil n - p - q/2 \rceil / n$. The asymptotic breakdown point $\epsilon^* = \lim_{n \to \infty} \epsilon_n^*$ equals

$$\epsilon^* = \min(1 - \sqrt{1-r}, \sqrt{1-r}).$$

Taking $r = 0.75$ yields an asymptotic breakdown point of $\epsilon^* = 0.5$. Hence, GS-estimators can attain the highest possible value for the breakdown point. In practice, if the GS-estimator needs to achieve a specified breakdown point ϵ^*, for example $\epsilon^* = 0.5$, and to have consistency at a model with error distribution F, typically the normal distribution, the constant c in Tukey's biweight function needs to be taken as the solution of $1-\sqrt{1 - E_{F\times F}[\rho_c(\|\epsilon_1 - \epsilon_2\|)]/(c^2/6)} = \epsilon^*$.

3 Algorithm

The algorithm we propose is anologous to the fast S-algorithm of Salibian-Barrera and Yohai (2006) for univariate regression. For any sequence of values $e_1, \ldots, e_{\tilde{n}}$, the corresponding scale s is given by the solution of

$$\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \rho\left(\frac{e_i}{s}\right) = k.$$

The fast S-algorithm uses local improvement steps (I-steps) to update an initial estimate of the regression coefficients. In our algorithm, the I-steps are based on the scale of the norm of the pairwise differences of the residuals $\|\mathbf{r}_i - \mathbf{r}_j\|_C = ((\mathbf{r}_i - \mathbf{r}_j)^T C^{-1} (\mathbf{r}_i - \mathbf{r}_j))^{1/2}$. The actual algorithm can be described as follows:

1. Draw N random sub-samples of size $p + q$. For each sub-sample calculate the least squares estimate $\widehat{\mathcal{B}}^0_m$, $m = 1, \ldots, N$, and the corresponding shape matrix $\widehat{\Gamma}^0_m$ of the residuals, i.e. the covariance matrix $\widehat{\Sigma}^0_m$ of the residuals is rescaled to have determinant equal to 1. Denote the residuals by $\mathbf{r}_i(\widehat{\mathcal{B}}^0_m)$, for $i = 1, \ldots, n$.
2. For each sub-sample, apply κ I-steps (e.g. $\kappa = 2$) as follows. Set $v = 1$.

a. Calculate an approximate solution of equation (2), as

$$s_v = \sqrt{s_{v-1}^2 \times \left(\sum_{i<j} \rho(\|\mathbf{r}_i(\widehat{\mathcal{B}}_m^{v-1}) - \mathbf{r}_j(\widehat{\mathcal{B}}_m^{v-1})\|_{\widehat{\Gamma}_m^{v-1}}/s_{v-1}) / \binom{n}{2} k \right)}$$

with s_0 the median absolute deviation of the norms $\|\mathbf{r}_i(\widehat{\mathcal{B}}_m^{v-1}) - \mathbf{r}_j(\widehat{\mathcal{B}}_m^{v-1})\|_{\widehat{\Gamma}_m^{v-1}}$.

b. Determine the weights $w_{ij} = u(\|\mathbf{r}_i(\widehat{\mathcal{B}}_m^{v-1}) - \mathbf{r}_j(\widehat{\mathcal{B}}_m^{v-1})\|_{\widehat{\Gamma}_m^{v-1}}/s_v)$ and calculate $\widehat{\mathcal{B}}_m^v$ as the weighted least squares fit based on the differences of the observations. Compute then $\widehat{\Sigma}_m^v = \sum_{i<j} w_{ij}(\mathbf{r}_i(\widehat{\mathcal{B}}_m^{v-1}) - \mathbf{r}_j(\widehat{\mathcal{B}}_m^{v-1}))(\mathbf{r}_i(\widehat{\mathcal{B}}_m^{v-1}) - \mathbf{r}_j(\widehat{\mathcal{B}}_m^{v-1}))^T$ with corresponding shape estimate $\widehat{\Gamma}_m^v$.

c. Calculate the pairwise differences of the residuals corresponding to $\widehat{\mathcal{B}}_m^v$.

d. Repeat steps a, b and c for $v = 2, \ldots, \kappa$.

Each sub-sample thus yields an improved estimate $(\widehat{\mathcal{B}}_m^\kappa, \widehat{\Gamma}_m^\kappa), m = 1, \ldots, N$.

3. We now select the τ best solutions (e.g. $\tau = 5$) in an efficient way. For $m = 1, \ldots, \tau$, we calculate the scale $s_m = s(\|\mathbf{r}_i(\widehat{\mathcal{B}}_m^\kappa) - \mathbf{r}_j(\widehat{\mathcal{B}}_m^\kappa)\|_{\widehat{\Gamma}_m^\kappa}), m = 1, \ldots, \tau$. For $m \geq \tau$, we denote by I_m the set containing the τ optimal solutions found after examining the first m candidates, and A_m denotes the maximum of the scales of the solutions in I_m. The next solution $(\widehat{\mathcal{B}}_{m+1}^\kappa, \widehat{\Gamma}_{m+1}^\kappa)$ will be included in I_{m+1} if and only if $s(\|\mathbf{r}_i(\widehat{\mathcal{B}}_{m+1}^\kappa) - \mathbf{r}_j(\widehat{\mathcal{B}}_{m+1}^\kappa)\|_{\widehat{\Gamma}_{m+1}^\kappa}) < A_m$ which is equivalent to

$$\frac{1}{\binom{n}{2}} \sum_{i<j} \rho(\|\mathbf{r}_i(\widehat{\mathcal{B}}_{m+1}^\kappa) - \mathbf{r}_j(\widehat{\mathcal{B}}_{m+1}^\kappa)\|_{\widehat{\Gamma}_{m+1}^\kappa}/A_m) < k. \tag{5}$$

If condition (5) holds, then we compute the scale $s(\|\mathbf{r}_i(\widehat{\mathcal{B}}_{m+1}^\kappa) - \mathbf{r}_j(\widehat{\mathcal{B}}_{m+1}^\kappa)\|_{\widehat{\Gamma}_{m+1}^\kappa})$ and we correspondingly update I_m and A_m to obtain I_{m+1} and A_{m+1}. If inequality (5) does not hold, then $I_{m+1} = I_m$ and $A_{m+1} = A_m$. Let us denote $(\widehat{\mathcal{B}}_m^B, \widehat{\Gamma}_m^B, s_m^B), m = 1, \ldots, \tau$ the τ optimal solutions and s_m^B their corresponding scales, for $m = 1, \ldots, \tau$.

4. Apply further I-steps to each of the optimal solutions $(\widehat{\mathcal{B}}_m^B, \widehat{\Gamma}_m^B, s_m^B)$, $m = 1, \ldots, \tau$, until convergence, which yields the fully iterated solutions $(\widehat{\mathcal{B}}_m^F, \widehat{\Gamma}_m^F, s_m^F)$,

$m = 1, \ldots, \tau$, where $s_m^F = s(\|\mathbf{r}_i(\widehat{\mathcal{B}}_m^F) - \mathbf{r}_j(\widehat{\mathcal{B}}_m^F)\|_{\widehat{\Gamma}_m^F})$. The final estimate is the solution $(\widehat{\mathcal{B}}_m^F, \widehat{\Gamma}_m^F)$ associated with the smallest scale s_m^F and the corresponding estimate of the covariance matrix of the residuals is obtained as $\widehat{\Sigma}_m^F = (s_m^F)^2 \widehat{\Gamma}_m^F$.

4 Fisher-consistency and influence function

Let $\mathcal{H}$ denote the class of all distributions on $\mathbb{R}^{p+q}$. We define the GS-functional **GS**: $\mathcal{H} \to (\mathbb{R}^{p\times q} \times PDS(q))$ as the solution $\mathbf{GS}(H) = (\mathcal{B}_{GS}(H), \Sigma_{GS}(H))$ of the problem of minimizing $|C|$ subject to

$$\iint \rho([(\mathbf{y}_1 - \mathbf{y}_2 - B^T(\mathbf{u}_1 - \mathbf{u}_2))^T C^{-1} (\mathbf{y}_1 - \mathbf{y}_2 - B^T(\mathbf{u}_1 - \mathbf{u}_2))]^{1/2}) dH(\mathbf{z_1}) dH(\mathbf{z}_2) = k$$

among all $(B, C) \in \mathbb{R}^{p\times q} \times PDS(q)$ and where $\mathbf{z}_l = (\mathbf{u}_l^T, \mathbf{y}_l^T)^T$ for $l = 1, 2$. It can be easily seen that the resulting GS-functional is affine equivariant.

We assume that the following two conditions are satisfied for the distribution H of $\mathbf{z} = (\mathbf{u}^T, \mathbf{y}^T)^T$ in model (1).

Condition 2 *We assume that the differences of the errors $\epsilon_i - \epsilon_j$ in model (1) have a distribution F_Σ with density $f_\Sigma(\mathbf{x}) = g(\mathbf{x}^T \Sigma^{-1} \mathbf{x})/\sqrt{|\Sigma|}$, with $\Sigma \in PDS(q)$ the scatter matrix. Furthermore, the function g is assumed to have a strictly negative derivative g'.*

Condition 2 requires that the error terms have a unimodal elliptically symmetric distribution around the origin. Note that if the error terms are independent and elliptically symmetrically, then the distribution of the differences of the errors remains elliptically symmetric (Hult and Lindskog 2002). We need another regularity condition on the model distribution H, before stating the

result on Fisher-consistency.

Condition 3 *For all $\beta \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}^q$ not both equal to zero at the same time, it holds that*

$$P_H(\beta^T(\mathbf{u}_1 - \mathbf{u}_2) + \gamma^T(\mathbf{y}_1 - \mathbf{y}_2) = 0) < 1 - r.$$

Theorem 2 *The functionals $\mathcal{B}_{GS}$ and Σ_{GS} are Fisher-consistent estimators of the parameters $\mathcal{B}$ and Σ at any model distribution H satisfying conditions 2 and 3:*

$$\mathcal{B}_{GS}(H) = \mathcal{B} \quad \text{and } \Sigma_{GS}(H) = \Sigma.$$

The influence function of a functional T at a distribution H measures the effect on T of an infinitesimal contamination at a single point (Hampel et al. 1986). If we denote a point mass distribution at $\mathbf{z} = (\mathbf{u}^T, \mathbf{y}^T)^T$ by $\Delta_{\mathbf{z}}$, and consider the contaminated distribution $H_{\varepsilon,\mathbf{z}} = (1-\varepsilon)H + \varepsilon\Delta_{\mathbf{z}}$, then the influence function is given by

$$IF(\mathbf{z}; T, H) = \lim_{\varepsilon \downarrow 0} \frac{T(H_{\varepsilon,\mathbf{z}}) - T(H)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} T(H_{\varepsilon,\mathbf{z}})_{|\varepsilon=0}.$$

Due to affine equivariance of the GS-functional, it suffices to look at model distributions H_0 that satisfy conditions 2 and 3 and for which $\mathcal{B} = 0$, and $\Sigma = I_q$. Denote $F_0 = F_{I_q}$ and let G be the distribution of $\mathbf{u}$.

Theorem 3 *For model distributions H_0 verifying the above conditions, the influence functions of the GS-estimators for multivariate regression at $\mathbf{z}_0 = (\mathbf{u}_0^T, \mathbf{y}_0^T)^T$ are given by*

$$IF(\mathbf{z}_0; \mathcal{B}_{GS}, H_0) = [Cov(\mathbf{u})]^{-1}(\mathbf{u}_0 - E_G[\mathbf{u}])\frac{\bar{\psi}(\mathbf{y}_0)^T}{\beta} \tag{6}$$

$$\begin{aligned} IF(\mathbf{z}_0; \Sigma_{GS}, H_0) &= IF(\mathbf{y}_0; \Sigma_{GS}, F_0) \\ &= \frac{2}{\gamma_1} q E_{F_0}\left[\psi(\|\mathbf{y}_1 - \mathbf{y}_0\|)\|\mathbf{y}_1 - \mathbf{y}_0\|\left(\frac{(\mathbf{y}_1 - \mathbf{y}_0)(\mathbf{y}_1 - \mathbf{y}_0)^T}{\|\mathbf{y}_1 - \mathbf{y}_0\|^2} - \frac{1}{q} I_q\right)\right] \\ &+ \frac{4E_{F_0}\left[\rho(\|\mathbf{y}_1 - \mathbf{y}_0\|) - k\right]}{\gamma_3} I_q \end{aligned} \tag{7}$$

where

$$\bar{\psi}(\mathbf{y}_0) = E_{F_0}\left[\frac{\psi(\|\mathbf{y}_0 - \mathbf{y}_1\|)}{\|\mathbf{y}_0 - \mathbf{y}_1\|}(\mathbf{y}_0 - \mathbf{y}_1)\right],$$

and $\beta = E_{F_0 \times F_0}\left[\frac{1}{q}\psi'(\|\mathbf{y}_1 - \mathbf{y}_2\|) + \left(1 - \frac{1}{q}\right) u(\|\mathbf{y}_1 - \mathbf{y}_2\|)\right]$, $\gamma_1 = E_{F_0 \times F_0}[\psi'(\|\mathbf{y}_1 - \mathbf{y}_2\|)\|\mathbf{y}_1 - \mathbf{y}_2\|^2 + (q+1)\psi(\|\mathbf{y}_1 - \mathbf{y}_2\|)\|\mathbf{y}_1 - \mathbf{y}_2\|]/(q+2)$ *and* $\gamma_3 = E_{F_0 \times F_0}[\psi(\|\mathbf{y}_1 - \mathbf{y}_2\|)\|\mathbf{y}_1 - \mathbf{y}_2\|]$.

For the model with only a constant term, the expression of the influence function of Σ_{GS} is equivalent to the influence function of the symmetrized M-estimators of multivariate scatter of Sirkiä et al. (2007). If $q = 1$, then the influence function of $\mathcal{B}_{GS}$ is identical to the influence function of the univariate GS-estimator (see Croux et al. 1994). Since $\bar{\psi}$ is a bounded function, it can be seen that the influence function of $\mathcal{B}_{GS}$ is bounded in $\mathbf{y}_0$ but unbounded in $\mathbf{u}_0$. Hence good leverage points can have a high effect on the GS-estimator, but bad leverage points will have a bounded influence.

5 Efficiency

The asymptotic variance-covariance matrix of the GS-estimator at the model distribution H_0 can be computed by means of the influence function, as

$$ASV(\mathcal{B}_{GS}, H_0) = E[IF(\mathbf{z}; \mathcal{B}_{GS}, H_0) \otimes IF(\mathbf{z}; \mathcal{B}_{GS}, H_0)^T]$$

(see Hampel et al. 1986) where $A \otimes B$ denotes the Kronecker product of a $(d_1 \times d_2)$ matrix A with a $(d_3 \times d_4)$ matrix B, which results in a $(d_1 d_3 \times d_2 d_4)$ matrix with $d_1 d_2$ blocks of size $(d_3 \times d_4)$. For $1 \leq j \leq d_1$ and $1 \leq k \leq d_2$ the (j,k)th block equals $a_{jk}B$, where a_{jk} are the elements of the matrix A. Denoting $\Sigma_{\mathbf{u}} := Cov[\mathbf{u}]$, it follows from (6) that

$$ASV(\mathcal{B}_{GS}, H_0) = K_{pq}\left(\text{diag}\left(\frac{E_{F_0}[\bar{\psi}(\mathbf{y}_0) \otimes \bar{\psi}(\mathbf{y}_0)^T]}{\beta^2}\right) \otimes \Sigma_{\mathbf{u}}^{-1}\right), \tag{8}$$

where K_{pq} is the commutation matrix, a $(pq \times pq)$ matrix consisting of pq blocks of size $(q \times p)$. For $1 \leq l \leq p$ and $1 \leq m \leq q$ the (l,m)th block of K_{pq} equals the $(q \times p)$ matrix Δ_{ml} which is 1 at entry (m,l) and 0 everywhere else.

From (6) and (8) we find that the asymptotic variance of $(\mathcal{B}_{GS})_{jk}$ is

$$ASV((\mathcal{B}_{GS})_{jk}, H_0) = (\Sigma_{\mathbf{u}}^{-1})_{jj}\frac{E_{F_0}[\bar{\psi}(\mathbf{y}_0)_k^2]}{\beta^2} \tag{9}$$

while the asymptotic covariances, for $j \neq j'$, are given by

$$ASC((\mathcal{B}_{GS})_{jk}, (\mathcal{B}_{GS})_{j'k}, H_0) = (\Sigma_{\mathbf{u}}^{-1})_{jj'}\frac{E_{F_0}[\bar{\psi}(\mathbf{y}_0)_k^2]}{\beta^2}$$

and all other asymptotic covariances (for $k \neq k'$) equal 0.

Since we assumed, w.l.o.g. due to affine equivariance, that $\Sigma_{\mathbf{u}} = I_p$ at H_0, we have that all asymptotic covariances are zero. Furthermore $ASV((\mathcal{B}_{GS})_{jk}, H_0) = E_{F_0}[\bar{\psi}(\mathbf{y}_0)_k^2]/\beta^2$ does not depend on k and j. Hence, we can compute the asymptotic relative efficiency of the GS-estimator with respect to the least-squares estimator as:

$$ARE(\mathcal{B}_{GS}, H_0) = \frac{ASV((\mathcal{B}_{LS})_{jk}, H_0)}{ASV((\mathcal{B}_{GS})_{jk}, H_0)}$$

for all $j = 1, \ldots, p$ and $k = 1, \ldots, q$. The asymptotic relative efficiency of a multivariate regression GS-estimator does not depend on the dimension p or the distribution of the carriers, but only on the dimension q and the distribution of the errors terms.

Table 1 shows the relative asymptotic efficiencies for H_0 a multivariate normal distribution, and for a multivariate Student distributions T_ν with $\nu = 3$ and 8 degrees of freedom. Results are presented for both GS- and S-estimators (see Table 3.1 in Van Aelst and Willems 2005 for the efficiencies of S-estimators), based on a Tukey biweight loss function. The reported values in Table 1 are based on numerical integration of the analytic expression in (9). From Table 1 we see that the efficiencies for the GS-estimator are high for the 25% as well as for the 50% breakdown point case. For the T_3 distribution, the GS-estimator is far more efficient than the least squares estimator. For the T_8 distribution the GS-estimator still outperforms the LS-estimator in higher dimensions. Comparing the GS-estimator with the S-estimator, we see that using the pairwise differences generally results in a higher efficiency, in particular for the 50% breakdown point estimates.

We also performed a simulation study to investigate the finite-sample efficiency of the GS-estimator. We generated $m = 1\,000$ random samples with predictors drawn from the multivariate standard normal distribution. The errors were generated from the multivariate normal distribution or from the multivariate T_3 distribution. We considered multivariate regression models with $p + 1 = 2$ and $q = 2$ and $p + 1 = 5$ and $q = 5$. The matrix $(\alpha, \mathcal{B}^T)^T$ was set to zero. For each sample we calculated both the S-estimates (including an intercept term) and GS-estimates. The Monte Carlo variance of $\widehat{\mathcal{B}}_n$ is measured

Table 1

Asymptotic relative efficiencies for S- and GS-estimators with respect to the LS estimator at normal and Student distributions.

ϵ^*	25%					50%				
GS	$q=1$	$q=2$	$q=3$	$q=5$	$q=10$	$q=1$	$q=2$	$q=3$	$q=5$	$q=10$
Φ	0.818	0.912	0.940	0.974	0.973	0.683	0.719	0.770	0.843	0.923
T_8	0.982	1.077	1.103	1.138	1.162	0.798	0.885	0.944	1.031	1.151
T_3	1.902	2.061	2.125	2.235	2.342	1.603	1.872	2.070	2.196	2.445
S	$q=1$	$q=2$	$q=3$	$q=5$	$q=10$	$q=1$	$q=2$	$q=3$	$q=5$	$q=10$
Φ	0.759	0.912	0.951	0.976	0.990	0.287	0.580	0.722	0.846	0.933
T_8	0.894	1.059	1.108	1.141	1.162	0.390	0.739	0.897	1.038	1.153
T_3	1.738	2.035	2.137	2.222	2.289	0.904	1.601	1.903	2.177	2.140

as $n \operatorname{ave}_{j,k}(\widehat{\operatorname{Var}}((\hat{\mathcal{B}}_n)_{jk}))$ for $j = 1, \ldots, p$ and $k = 1, \ldots, q$, where $\widehat{\operatorname{Var}}((\hat{\mathcal{B}}_n)_{jk})$ is the empirical variance over the m simulated estimates. The finite-sample relative efficiency is then computed as the inverse of this variance estimate for the normal distribution, and as $\nu/(\nu-2)$ divided by the variance estimate for the T_ν distribution. Table 2 lists these finite-sample relative efficiencies for the 25% breakdown S- and GS-estimator for the normal and T_3 model. The finite-sample relative efficiencies are generally slightly lower than the asymptotic relative efficiencies of Table 1. If we compare the GS-estimator with the S-estimator we see that the relative efficiencies are comparable at the normal distribution, but at the T_3 distribution the relative efficiencies of the

Table 2

Finite-sample relative efficiencies for $\widehat{\mathcal{B}}_{GS}$ and $\widehat{\mathcal{B}}_{S}$ (25% breakdown) with respect to the LS estimator at the normal and T_3 distribution

			$n = 30$	$n = 50$	$n = 100$	$n = 200$	$n = \infty$
GS	Φ	$q = 2$	0.881	0.901	0.858	0.867	0.912
		$q = 5$	0.797	0.859	0.921	0.941	0.974
	T_3	$q = 2$	1.809	1.859	1.901	2.025	2.061
		$q = 5$	1.415	1.669	1.861	1.960	2.235
S	Φ	$q = 2$	0.875	0.901	0.859	0.867	0.912
		$q = 5$	0.798	0.862	0.924	0.945	0.976
	T_3	$q = 2$	1.788	1.838	1.882	2.005	2.035
		$q = 5$	1.407	1.656	1.846	1.943	2.222

GS-estimator are always higher.

6 Robust inference

6.1 Fast and robust bootstrap

We now consider the issue of statistical inference for the regression parameter $\mathcal{B}$. We use the fast and robust bootstrap procedure introduced by Salibian-Barrera and Zamar (2002) for univariate regression MM-estimators. The bootstrap principle is to generate a large number of samples from the original data

set, and to recalculate the estimates for each of these resamples. Then, the distribution, of $\sqrt{n}(\widehat{\mathcal{B}}_n - \mathcal{B})$ can be approximated by the sample distribution of $\sqrt{n}(\widehat{\mathcal{B}}^*_n - \widehat{\mathcal{B}}_n)$ where $\widehat{\mathcal{B}}^*_n$ is the value of the resampled estimator. When there are outliers present in the data, this method can be expected to be more accurate than using the asymptotic variance. However, the standard bootstrap procedure is non-robust, as some bootstrap samples may contain a fraction of outliers that exceeds the breakdown point of the robust estimates, and computationally demanding, due to the high computation time of robust estimators. Both these problems are resolved by the fast and robust bootstrap (FRB) procedure.

For S-estimators in multivariate models, inference based on FRB has been developed by Van Aelst and Willems (2005) and Salibian-Barrera, Van Aelst and Willems (2006, 2008). The FRB procedure computes bootstrap values of $\widehat{\mathcal{B}}_n$ without explicitly calculating the actual estimate for each resample. The FRB gains a considerable amount of computation time by approximating $\widehat{\mathcal{B}}^*_n$ in each resample based on a fixed-point representation of the estimator. Because a reweighted representation of the estimator is bootstrapped, the method will be more robust since outliers downweighted in the original sample, will also be downweighted in each resample, regardless the fraction of outliers in each resample.

Suppose that an estimator of the parameter Θ can be represented by a smooth fixed-point equation $\mathbf{g}(\widehat{\Theta}_n) = \widehat{\Theta}_n$, with $\mathbf{g}$ depending on n. Then, using the smoothness of $\mathbf{g}$, we can calculate a Taylor expansion about the limiting value of the estimate $\widehat{\Theta}_n$:

$$\widehat{\Theta}_n = \mathbf{g}(\Theta) + \nabla\mathbf{g}(\Theta)(\widehat{\Theta}_n - \Theta) + R_n$$

where R_n is a remainder term and $\nabla \mathbf{g}(.)$ is the matrix of partial derivatives. Supposing that the remainder term is small, this equation can be rewritten as

$$\sqrt{n}(\widehat{\Theta}_n - \Theta) \approx [I - \nabla \mathbf{g}(\Theta)]^{-1} \sqrt{n}(\mathbf{g}(\Theta) - \Theta).$$

Taking bootstrap equivalents at both sides and estimating the matrix $[I - \nabla \mathbf{g}(\Theta)]^{-1}$ by $[I - \nabla \mathbf{g}(\widehat{\Theta}_n)]^{-1}$ yields

$$\sqrt{n}(\widehat{\Theta}_n^* - \widehat{\Theta}_n) \approx [I - \nabla \mathbf{g}(\widehat{\Theta}_n)]^{-1} \sqrt{n}(\mathbf{g}^*(\widehat{\Theta}_n) - \widehat{\Theta}_n). \tag{10}$$

For each bootstrap sample, we can calculate the right-hand side of this equation instead of the left-hand side. Hence, we approximate the actual estimate in each sample by computing the function $\mathbf{g}^*$ in $\widehat{\Theta}_n$ and then apply a linear correction given by $[I - \nabla \mathbf{g}(\widehat{\Theta}_n)]^{-1}$.

We now apply this procedure to the multivariate GS-estimator. We can rewrite the estimating equations (3) and (4) as

$$\begin{aligned}
\widehat{\mathcal{B}}_n &= \mathbf{A}_n(\widehat{\mathcal{B}}_n, \widehat{\Sigma}_n)^{-1} \mathbf{B}_n(\widehat{\mathcal{B}}_n, \widehat{\Sigma}_n) \\
\widehat{\Sigma}_n &= \mathbf{V}_n(\widehat{\mathcal{B}}_n, \widehat{\Sigma}_n) + w_n(\widehat{\mathcal{B}}_n, \widehat{\Sigma}_n) \widehat{\Sigma}_n
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{A}_n(B, C) &= \sum_{i<j} u(d_{ij})(\mathbf{u}_i - \mathbf{u}_j)(\mathbf{u}_i - \mathbf{u}_j)^T \\
\mathbf{B}_n(B, C) &= \sum_{i<j} u(d_{ij})(\mathbf{u}_i - \mathbf{u}_j)(\mathbf{y}_i - \mathbf{y}_j)^T \\
\mathbf{V}_n(B, C) &= \frac{1}{\binom{n}{2} k} \sum_{i<j} q u(d_{ij})(\mathbf{y}_i - \mathbf{y}_j - B^T(\mathbf{u}_i - \mathbf{u}_j))(\mathbf{y}_i - \mathbf{y}_j - B^T(\mathbf{u}_i - \mathbf{u}_j))^T \\
w_n(B, C) &= \frac{1}{\binom{n}{2} k} \sum_{i<j} w(d_{ij})
\end{aligned}$$

with $w(t) = \rho(t) - \rho'(t)t$. Write

$$\Theta := \begin{pmatrix} \text{vec}(\mathcal{B}) \\ \\ \text{vec}(\Sigma) \end{pmatrix}, \widehat{\Theta}_n := \begin{pmatrix} \text{vec}(\widehat{\mathcal{B}}_n) \\ \\ \text{vec}(\widehat{\Sigma}_n) \end{pmatrix},$$

and for any matrices B and C, put

$$\mathbf{g}\begin{pmatrix} \text{vec}(B) \\ \\ \text{vec}(C) \end{pmatrix} := \begin{pmatrix} \text{vec}(\mathbf{A}_n(B,C)^{-1}\mathbf{B}_n(B,C)) \\ \\ \text{vec}(\mathbf{V}_n(B,C) + w_n(B,C)C) \end{pmatrix}.$$

The expression for the matrix $\nabla\mathbf{g}(.)$ of partial derivatives can be found in the Appendix.

Now, for a bootstrap sample $\{((\mathbf{u}_i^*)^T, (\mathbf{y}_i^*)^T)^T, i = 1, \ldots, n\}$ we have that

$$\mathbf{g}^*(\widehat{\Theta}_n) = \begin{pmatrix} \text{vec}(\mathbf{A}_n^*(\widehat{\mathcal{B}}_n, \widehat{\Sigma}_n)^{-1}\mathbf{B}_n^*(\widehat{\mathcal{B}}_n, \widehat{\Sigma}_n)) \\ \\ \text{vec}(\mathbf{V}_n^*(\widehat{\mathcal{B}}_n, \widehat{\Sigma}_n) + w_n^*(\widehat{\mathcal{B}}_n, \widehat{\Sigma}_n)\widehat{\Sigma}_n) \end{pmatrix}$$

where $\mathbf{A}_n^*$, $\mathbf{B}_n^*$, $\mathbf{V}_n^*$ and w_n^* are the bootstrap versions of the quantities $\mathbf{A}_n$, $\mathbf{B}_n$, $\mathbf{V}_n$ and w_n, that is with $(\mathbf{u}_i^T, \mathbf{y}_i^T)^T$ replaced by $((\mathbf{u}_i^*)^T, (\mathbf{y}_i^*)^T)^T$. Thus, in order to get the values of $\sqrt{n}(\widehat{\Theta}_n^* - \widehat{\Theta}_n)$ for each bootstrap sample, we calculate $\mathbf{g}^*(\widehat{\Theta}_n)$, apply the linear correction given by the matrix of partial derivatives and use approximation (10). We use casewise resampling to generate the bootstrap samples, which means that we draw with replacement from the observations $\{(\mathbf{u}_i^T, \mathbf{y}_i^T)^T, i = 1, \ldots, n\}$.

We now focus on confidence intervals resulting from the FRB procedure. We investigate the robustness of the bootstrap confidence interval by deriving the breakdown point of bootstrap quantile estimates. For a statistic T_n, and

$t \in [0,1]$, let Q_t^* denote the tth quantile of the bootstrap sample distribution of T_n^*:

$$Q_t^* = \min\{x : \frac{1}{R} \times \#\{T_n^{*j} \geq x; j = 1, \ldots, R\} \leq t\}$$

where R is the number of bootstrap samples drawn. Singh (1998) defined the upper breakdown point of a statistic as the minimum proportion of asymmetric contamination that can carry the statistic over any bound. The expected upper breakdown point of the bootstrap quantile Q_t^* is defined as the minimum proportion of asymmetric contamination that is expected to be able to carry Q_t^* over any bound, where the expectation is taken over the distribution of drawing R samples with replacement. For the FRB, if we look at the pairwise differences of the observations in a bootstrap sample, then this sample of differences must contain at least p differences of two good observations. Hence, we need in the bootstrap sample at least c_p good observations such that $\binom{c_p}{2} \geq p$ to obtain at least p differences of good observations among the differences of the bootstrap sample. An easy calculation yields $c_p = \lceil \frac{1}{2} + \frac{1}{2}\sqrt{1+8p} \rceil$. Let $B(n, \delta)$ be the number of distinct non-outlying observations in a resample of size n, drawn with replacement from a sample of size n with a proportion δ of outliers.

Theorem 4 *Let $\mathcal{Z}_n \subset \mathbb{R}^{p+q+1}$ and assume that the data satisfies condition 1. Let ϵ_n^* be the breakdown point of a GS-estimate $\widehat{\mathcal{B}}_n$. Then the expected upper breakdown point of the t-th fast bootstrap quantile for any regression parameter $\mathcal{B}_{jk}, j = 1, \ldots, p; k = 1, \ldots, q$ is given by $\min(\epsilon_n^*, \epsilon_n^E)$ where*

$$\epsilon_n^E = \inf\{\delta \in [0,1] : P(B(n,\delta) < c_p) \geq t\}.$$

Table 3 lists values for ϵ_n^E for different dimensions and samples sizes, for the

Table 3

Expected upper breakdown values for FRB using maximal breakdown GS-estimators

		$p = 2,\ q = 1$				$p = 8,\ q = 2$			
	n	10	30	50	100	20	30	50	100
$Q^*_{0.05}$	ϵ_n^E	0.40	0.50	0.50	0.50	0.50	0.50	0.50	0.50
$Q^*_{0.005}$	ϵ_n^E	0.30	0.50	0.50	0.50	0.40	0.50	0.50	0.50

GS-estimator with maximal breakdown point. Two different quantiles $Q^*_{0.05}$ and $Q^*_{0.005}$ are considered, which can respectively be used to construct 90% and 99% percentile confidence intervals. We see that only for the smallest sample sizes the expected upper breakdown point for the FRB is lower than 50%, in all other cases the maximum breakdown point is reached.

We now show that the FRB converges to the same limiting distribution as the distribution of the GS-estimator does. We need the following assumptions on ρ:

(A.1) The following functions are bounded and almost everywhere continuous:

$$\frac{\rho'(x)}{x}, \frac{\rho''(x)}{x^2} - \frac{\rho'(x)}{x^3}, \frac{\rho'''(x)}{x^3} - 3\frac{\rho''(x)}{x^4} + 3\frac{\rho'(x)}{x^5}, \rho''(x) \text{ and } \frac{\rho'''(x)}{x}$$

(A.2) $E_{G\times G}[\frac{\rho'(d)}{d}(\mathbf{u}_1 - \mathbf{u}_2)(\mathbf{u}_1 - \mathbf{u}_2)^T]^{-1}$ exists.

Theorem 5 *Let ρ be a loss function satisfying (A.1). Let $(\widehat{\mathcal{B}}_n, \widehat{\Sigma}_n)$ be the multivariate GS-estimators and assume that $\widehat{\mathcal{B}}_n \xrightarrow{P} \mathcal{B}$ and $\widehat{\Sigma}_n \xrightarrow{P} \Sigma$. Then, given that assumption (A.2) is satisfied, the distributions of $\sqrt{n}(\widehat{\mathcal{B}}^*_n - \widehat{\mathcal{B}}_n)$*

and $\sqrt{n}(\widehat{\Sigma}_n^* - \widehat{\Sigma}_n)$ *converge weakly to the same limit distributions as those of* $\sqrt{n}(\widehat{\mathcal{B}}_n - \mathcal{B})$ *and* $\sqrt{n}(\widehat{\Sigma}_n - \Sigma)$ *respectively, conditional on the first* n *observations and along almost all sample sequences.*

6.2 Simulation results

We investigate the performance of confidence intervals for the regression coefficients based on FRB. Simulations were performed for sample sizes $n = 30$, 50, 100 and 200 for a multivariate regression model with $p = 4$ and $q = 5$. The predictor variables were generated from a multivariate normal distribution $N_p(\mathbf{0}, I_p)$. The true value of the parameter $\mathcal{B}$ was set to $\mathbf{1}_{p,q}$, the $p \times q$ matrix having 1 for each entry. We consider the following simulations schemes:

- normal errors: generated from $N_q(\mathbf{0}, I_q)$
- long-tailed errors: generated from a multivariate Student distribution with 3 degrees of freedom (T_3)
- vertical outliers: a proportion $1-\delta$ of the errors is generated from $N_q(\mathbf{0}, I_q)$, and a proportion δ generated from $N_q(5\sqrt{\chi^2_{q,.99}}\mathbf{1}_{q,1}, 1.5I_q)$, for $\delta = 0.15$ and $\delta = 0.25$
- bad leverage points: a proportion $1 - \delta$ of the errors is generated from $N_q(\mathbf{0}, I_q)$, and a proportion δ of the responses generated from $N_q(-10\mathbf{1}_{q,1}, 10I_q)$ with corresponding predictors replaced by predictors generated from $N_p(10\mathbf{1}_{p,1}, 10I_p)$, for $\delta = 0.15$ and $\delta = 0.25$.

We computed both the 25% and 50% GS-estimators for 1 000 data sets generated as described above and applied the FRB procedure with $B = 1\,000$ recalculated values $(\widehat{\mathcal{B}}_n^*, \widehat{\Sigma}_n^*)$.

Bootstrap confidence intervals for the components $\mathcal{B}_{jk}$ were constructed using the bias corrected and accelerated (BCA) method (see e.g. Davison and Hinkley 1997). The bootstrap intervals are compared with confidence intervals based on the asymptotic normality of the GS-estimator. The latter $100(1-\alpha)\%$ confidence intervals are of the form

$$\left[(\widehat{\mathcal{B}}_n)_{jk} - \Phi^{-1}(1-\frac{\alpha}{2})\sqrt{\widehat{V}_{jk}/n}, (\widehat{\mathcal{B}}_n)_{jk} + \Phi^{-1}(1-\frac{\alpha}{2})\sqrt{\widehat{V}_{jk}/n}\right]$$

where $\widehat{V}_{jk}$ denotes the empirical version of the asymptotic variance (EASV) of the (j,k)-th component of $\widehat{\mathcal{B}}_n$. The estimates $\hat{V}_{jk}$ are obtained by replacing Σ by $\widehat{\Sigma}_n$, replacing F_0 by the empirical distribution of the vectors $\widehat{\Sigma}_n^{-1/2}(\mathbf{y}_i - \widehat{\mathcal{B}}_n^T\mathbf{u}_i)$, and finally replacing $\Sigma_{\mathbf{u}}$ by the corresponding sample moment.

Figure 1 shows the coverage for 95% confidence intervals computed by FRB and EASV. From Figure 1 we clearly see that the coverage of the EASV-based intervals is generally lower than 95%. As the sample size grows, the EASV-based intervals converge to a 95% coverage, except in the case of bad leverage points. The FRB performs better than the EASV method. For small sample sizes the FRB is generally somewhat conservative except for bad leverage points. However, also in that case the coverage converges quickly to 95% when the sample size increases.

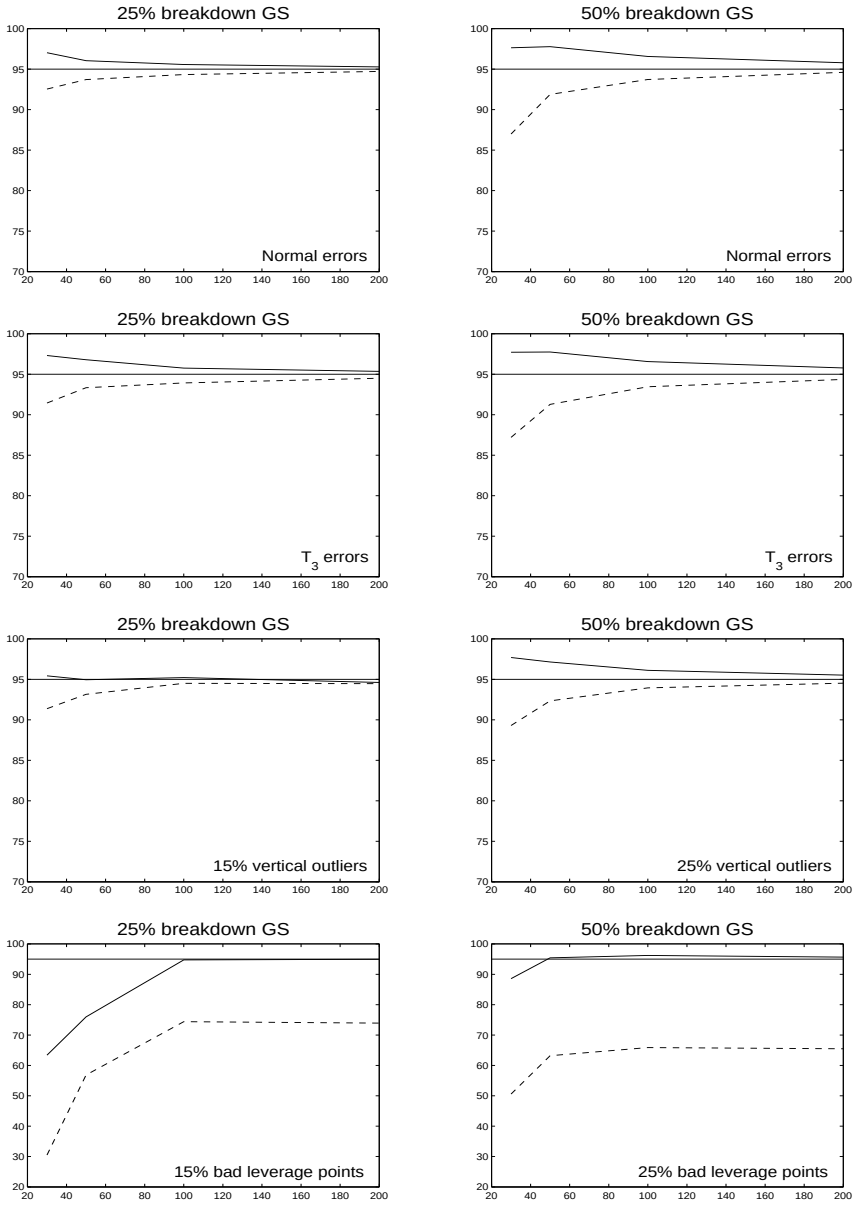

Fig. 1. Coverage for 95% confidence intervals, for FRB (–) and EASV (- -): $p = 4; q = 5$.

7 Examples

School data

This example considers data of $n = 70$ school sites in the U.S. (Charnes, Cooper and Rhodes 1981). We fit a multivariate regression model with 3 response variables: total reading score measured by the Metropolitan Achievement Test, total mathematics score measured by the Metropolitan Achievement Test and the Coopersmith self-esteem inventory. There are 5 explanatory variables: education level of mother, highest occupation of a family member, number of parental visits to the school, parent counselling concerning school-related topics and the number of teachers at the school. The model parameters were estimated with the least squares estimator and with 50% breakdown GS-estimator. We considered a model with intercept. For the GS-estimator, the intercept was estimated afterwards by applying an efficient robust estimator of multivariate location on the residuals of the GS-estimator $\mathbf{y}_i - \widehat{\mathcal{B}}_n^t \mathbf{u}_i$, for $i = 1, \ldots, n$. An appropriate choice is the M-type estimator of location of Lopuhaä (1992). This estimator is highly robust and highly efficient but requires a preliminary estimate of the scatter matrix. The GS-estimator, however, delivers a residual scatter matrix estimate of the residuals, along with the slope estimator, which we then use in the procedure of Lopuhaä (1992).

The diagnostic plots in Figure 2 show the Mahalanobis distances of the residuals versus the Mahalanobis distances of the explanatory variables (see also Rousseeuw et al. 2004). The left panel presents this plot for the least squares estimator, the right panel for the multivariate GS. For the diagnostic plot based on the robust GS, the Mahalanobis distances are computed using the

robust GS-estimator of Σ, and are therefore called robust distances. The horizontal and vertical lines correspond respectively to $\sqrt{\chi^2_{q,.975}}$ and $\sqrt{\chi^2_{p,.975}}$, and enable us to classify data points into regular observations, vertical outliers, good and bad leverage points. The least squares estimator detects one small vertical outlier and 5 small to moderate good leverage points. On the other hand, the GS-estimator reveals one very large bad leverage point (59), two moderate to large bad leverage points (35 and 44) and two moderate to large vertical outliers (12 and 21). Moreover, there are at least five good leverage points (10, 67, 1, 66, 50). The least squares estimator is thus clearly attracted by the bad leverage points. Table 4 gives 95% confidence intervals, computed with the fast and robust bootstrap discussed in Section 6, for the slope matrix based on S- and GS-estimates. The confidence limits using GS-estimates are in bold whenever this interval is shorter than the corresponding interval based on S-estimates. We see that for almost all parameters the GS-estimates yield more precise confidence intervals. This is without surprise, since GS is in general more efficient than S.

(a) (b)

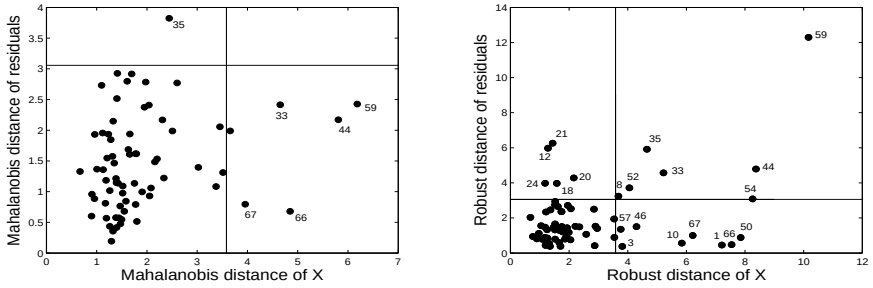

Fig. 2. Diagnostic plots for the school data; (a) Least squares estimator; (b) 50% breakdown GS-estimator

Table 4

95% confidence limits for the school data based on S- and GS-estimates

	S-estimate	lower	upper	GS-estimate	lower	upper
$\mathcal{B}_{11}$	0.109	-0.064	0.265	0.112	**-0.052**	**0.267**
$\mathcal{B}_{21}$	4.441	1.660	6.826	4.542	**1.980**	**6.980**
$\mathcal{B}_{31}$	0.056	-0.523	0.571	0.019	**-0.562**	**0.490**
$\mathcal{B}_{41}$	-0.637	-1.150	-0.202	-0.632	**-1.082**	**-0.219**
$\mathcal{B}_{51}$	-0.128	-0.591	0.107	-0.129	**-0.513**	**0.155**
$\mathcal{B}_{12}$	0.057	-0.161	0.228	0.053	**-0.158**	**0.223**
$\mathcal{B}_{22}$	4.952	2.374	7.913	5.131	2.444	8.304
$\mathcal{B}_{32}$	0.141	-0.625	0.798	0.094	**-0.639**	**0.746**
$\mathcal{B}_{42}$	-0.726	-1.295	-0.261	-0.726	**-1.190**	**-0.282**
$\mathcal{B}_{52}$	-0.147	-0.575	0.071	-0.147	**-0.522**	**0.084**
$\mathcal{B}_{13}$	-0.021	-0.070	0.027	-0.021	**-0.065**	**0.025**
$\mathcal{B}_{23}$	1.573	0.884	2.385	1.602	0.861	2.444
$\mathcal{B}_{33}$	0.270	0.099	0.476	0.258	**0.075**	**0.437**
$\mathcal{B}_{43}$	0.013	-0.240	0.232	0.018	**-0.211**	**0.223**
$\mathcal{B}_{53}$	0.041	-0.049	0.132	0.039	**-0.053**	**0.126**

Forbes data

The GS-estimator can also be used as a high-breakdown scatter estimator in a multivariate location-scale model, taking $p = 0$ in model (1). Afterwards the location vector can be estimated using the robust and efficient M-estimator of Lopuhaä (1992). We illustrate this with a data set taken from the 'The Data and Story Library' (http://lib.stat.cmu.edu/DASL/Stories/Forbes500CompaniesSales.html), which

contains several facts about 79 companies selected from the Forbes 500 list of 1986. We look at the following six variables: Assets (amount of assets in millions), Sales (amount of sales in millions), Market-value (market-value of the company in millions), Profits (profits in millions), Cash-flow (cash-flow in millions) and Employees (number of employees in thousands). Figure 3 compares the Mahalanobis distances computed with empirical mean and covariance matrix (horizontal axis) with the robust distances based on the 50% breakdown GS-estimator (vertical axis) using a distance-distance plot as proposed by Rousseeuw and Van Driessen (1999). If we draw horizontal and vertical lines at the usual cut off $\sqrt{\chi^2_{6,0.975}} = 3.8012$, the 9 outliers are detected by both estimators. However, there are 14 extra observations that have a robust distance above the cutoff while their Mahalanobis distances lie below the cutoff. Clearly the classical estimates were affected by the presence of these outliers.

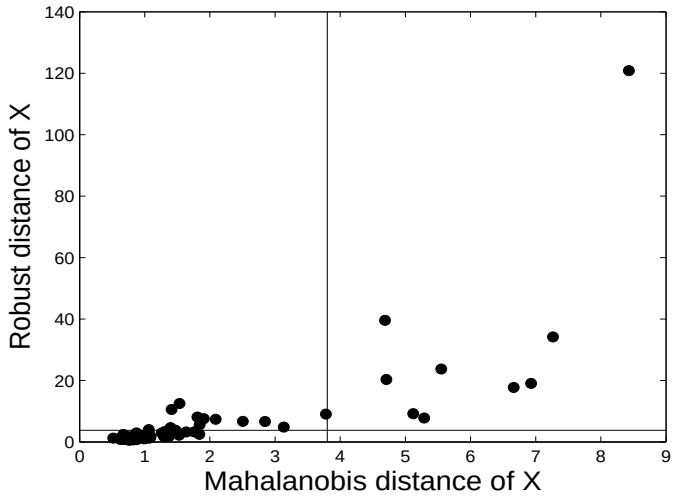

Fig. 3. Distance-distance plot for the Forbes data

8 Conclusion

In this paper, we discussed generalized S-estimators, i.e. S-estimators applied to the pairwise differences of the observations, in the multivariate regression context. We showed that they maintain the same good properties as in the univariate case, such as a high breakdown point and a higher efficiency than the multivariate regression S-estimators. To compute the GS-estimator, we constructed an algorithm based on improvement steps similar as in the fast S-algorithm for univariate regression. Furthermore we developed a fast and robust bootstrap method for the multivariate GS-estimators to obtain robust inference for the regression slopes. The examples illustrated the robustness and efficiency of the GS-estimator and its corresponding bootstrap inference.

GS-estimators estimate the regression slopes and the residual covariance matrix without needing to estimate the intercept. In the special case of the multivariate location-scale model, this implies that we can estimate the scatter matrix without needing to estimate the location of the observations. As illustrated in the examples, the intercept can easily be estimated afterwards by using the efficient and robust M-estimator of multivariate location of Lopuhaä (1992), using the residual covariance matrix of the GS-estimator as an initial estimator. In fact, similarly as for MM-estimators (Yohai 1987, Tatsuoka and Tyler 2000) one can also consider to re-estimate the regression slopes using a multivariate regression M-estimator based on an initial GS scatter matrix estimate. However, such an M-step is intended to increase the low efficiency of the initial estimator. Since GS-estimators already have a fairly high efficiency (Table 1), we do not expect that the M-step yields much further improvement.

Finally, let us stress that all the theoretical results obtained in this paper also apply to the multivariate location-scale model, the latter being a special case of the multivariate regression model. The properties of the GS-estimator were not yet investigated in the multivariate location-scale model. The major advantage of the GS-estimators of scatter with respect to most existing robust estimators of scatter is that they have the independency property. Hence, as discussed in the introduction, they are well suited for independent component analysis, and present a high breakdown alternative for the estimators considered by Sirkiä et al. (2007).

A Appendix

Proof of Theorem 1. Denote by m the number of points in the original data set of size n that are replaced by arbitrary points. This implies that the $\binom{n}{2}$ differences in the contaminated data set contain $\binom{m}{2} + m(n-m)$ contaminated differences. Since we apply the multivariate S-estimator on the set of differences, it follows from Theorem 1 in Van Aelst and Willems (2005) that the maximum number of outliers that is allowed before the estimator breaks down is given by

$$\min(\lceil \binom{n}{2} r \rceil, \lceil \binom{n}{2} - \binom{n}{2} r \rceil - h_{\mathcal{V}_n}) - 1$$

with $r = k/\sup\rho$ and $h_{\mathcal{V}_n}$ is the maximal number of differences lying on the same hyperplane. Hence breakdown because the number of contaminated differences exceeds $\lceil \binom{n}{2} r \rceil - 1$, occurs if

$$\binom{m}{2} + m(n-m) \geq \binom{n}{2} r \tag{A.1}$$

The smallest solution of the corresponding equality $-m^2+(-1+2n)m-n(n-1)r=0$ yields $m=\lceil n-\frac{1}{2}-\sqrt{1-4n(1-r)+4n^2(1-r)}/2\rceil$ (which in the limit yields $m/n<1-\sqrt{1-r}$).

We now consider breakdown because the number of contaminated differences on the same hyperplane exceeds $\lceil\binom{n}{2}-\binom{n}{2}r\rceil-h_{\mathcal{V}_n}-1$. From condition 1 it follows that the estimator can break down as soon as

$$\binom{p+q}{2}+\binom{m}{2}+m(p+q)\geq\binom{n}{2}-\binom{n}{2}r. \tag{A.2}$$

The smallest solution of the corresponding equality yields $m=\lceil\frac{1}{2}-p-q+\frac{1}{2}\sqrt{1-4n-4n^2r+4nr+4n^2}\rceil$ (which in the limit yields $m/n<\sqrt{1-r}$).

For any $C\in PDS(q)$, let $\lambda_1(C)\geq\lambda_2(C)\ldots\geq\lambda_q(C)$ denote its eigenvalues. Put

$$\begin{aligned}m=&\min(\lceil n-1/2-\sqrt{1+(1-r)(4n^2-4n)}/2\rceil,\\&\lceil 1/2-p-q+\sqrt{1+(1-r)(4n^2-4n)}/2\rceil)-1\end{aligned}$$

We first show that $\epsilon_n^*>m$ by showing that the estimator doesn't break down if we contaminate at most m observations. Formally we show that $\exists M,\alpha$ only depending on $\mathcal{Z}_n$, such that for every $\mathcal{Z}_n'=\{(1,(\mathbf{u}_i')^T,(\mathbf{y}_i')^T)^T;1\leq i\leq n\}$ obtained by replacing at most m observations from $\mathcal{Z}_n$, we have $\|\widehat{\mathcal{B}}_n(\mathcal{Z}_n')\|\leq M$ and $\lambda_1(\widehat{\Sigma}_n(\mathcal{Z}_n'))\leq\alpha$ and $\lambda_q(\widehat{\Sigma}_n(\mathcal{Z}_n'))>0$. The norm we use is

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

The inequality $\|AB\|\leq\|A\|\|B\|$ holds for any $A\in\mathbb{R}^{p\times q}$ and $B\in\mathbb{R}^{q\times r}$. Sometimes we will also use the L_2-norm $\|A\|_2=(\Sigma_{i,j}|A_{ij}|^2)^{1/2}$. Since these norms are topologically equivalent, we know that $\exists\alpha_1,\alpha_2>0$ such that $\forall B\in\mathbb{R}^{p\times q}:\alpha_1\|B\|\leq\|B\|_2\leq\alpha_2\|B\|$. For $\mathbf{w}\in\mathbb{R}^k(k\in\mathbb{N}\backslash\{0\})$, we have that

$\|\mathbf{w}\| = \|\mathbf{w}\|_2$.

Let us denote by $\mathcal{V}_n$ the set of the differences corresponding to the data set $\mathcal{Z}_n$, that is, $\mathcal{V}_n = \{((\mathbf{u}_i - \mathbf{u}_j)^T, (\mathbf{y}_i - \mathbf{y}_j)^T)^T; 1 \leq i < j \leq n\}$. Similarly, $\mathcal{V}'_n$ corresponds to the contaminated data set $\mathcal{Z}'_n$.

W.l.o.g. we assume that $c = 1$ and thus $\sup(\rho) = \rho(\infty) = 1$ such that $r = k$. Indeed we can always rescale the function ρ if necessary. Since ρ is continuous and we have that

$$\binom{n}{2}k - \binom{m}{2} - m(n-m) = \binom{n}{2}r - \binom{m}{2} - m(n-m) > 0$$

according to the reverse of (A.1), we can find a smallest radius $s > 0$ and cylinder $\mathcal{C}(\mathbf{0}, s^2 I_q) := \{(\mathbf{w}, \mathbf{v}); \|\mathbf{v}\| \leq s\}$ such that

$$\sum_{((\mathbf{u}_i - \mathbf{u}_j)^T, (\mathbf{y}_i - \mathbf{y}_j)^T)^T \in \mathcal{V}_n} \rho\left(\frac{\|\mathbf{y}_i - \mathbf{y}_j\|}{s}\right) = \binom{n}{2}k - \binom{m}{2} - m(n-m).$$

This yields the determinant $V = |s^2 I_q| = s^{2q}$. For the smallest cylinder $\mathcal{C}(\mathbf{0}, l^2 I_q) = \{(\mathbf{w}, \mathbf{v}); \|\mathbf{v}\| \leq l\}$ such that

$$\sum_{((\mathbf{u}_i - \mathbf{u}_j)^T, (\mathbf{y}_i - \mathbf{y}_j)^T)^T \in \mathcal{V}_n \cap \mathcal{V}'_n} \rho\left(\frac{\|\mathbf{y}_i - \mathbf{y}_j\|}{l}\right) = \binom{n}{2}k - \binom{m}{2} - m(n-m)$$

it then holds that $|l^2 I_q| = l^{2q} \leq V$. Moreover,

$$\begin{aligned}
&\sum_{((\mathbf{u}'_i - \mathbf{u}'_j)^T, (\mathbf{y}'_i - \mathbf{y}'_j)^T)^T \in \mathcal{V}'_n} \rho\left(\frac{\|\mathbf{y}'_i - \mathbf{y}'_j\|}{l}\right) \\
&\leq \sum_{((\mathbf{u}_i - \mathbf{u}_j)^T, (\mathbf{y}_i - \mathbf{y}_j)^T)^T \in \mathcal{V}_n \cap \mathcal{V}'_n} \rho\left(\frac{\|\mathbf{y}_i - \mathbf{y}_j\|}{l}\right) + \binom{m}{2} + m(n-m) = \binom{n}{2}k.
\end{aligned}$$

It follows that for the optimal solution $\mathcal{C}(\widehat{\mathcal{B}}_n(\mathcal{Z}'_n), \widehat{\Sigma}_n(\mathcal{Z}'_n)) = \mathcal{C}(\widehat{\mathcal{B}}_n(\mathcal{V}'_n), \widehat{\Sigma}_n(\mathcal{V}'_n)) := \{(\mathbf{w}, \mathbf{v}); (\mathbf{v} - \widehat{\mathcal{B}}_n(\mathcal{V}'_n)^T \mathbf{w})^T \widehat{\Sigma}_n(\mathcal{V}'_n)^{-1} (\mathbf{v} - \widehat{\mathcal{B}}_n(\mathcal{V}'_n)^T \mathbf{w}) \leq 1\}$ that satisfies

$$\sum_{((\mathbf{u}'_i - \mathbf{u}'_j)^T, (\mathbf{y}'_i - \mathbf{y}'_j)^T)^T \in \mathcal{V}'_n} \rho\left(d'_{ij}(\widehat{\mathcal{B}}_n(\mathcal{V}'_n), \widehat{\Sigma}_n(\mathcal{V}'_n))\right) \leq \binom{n}{2}k \tag{A.3}$$

where $d'_{ij}(\widehat{\mathcal{B}}_n(\mathcal{V}'_n), \widehat{\Sigma}_n(\mathcal{V}'_n)) = ((\mathbf{r}'_{ij})^T \widehat{\Sigma}_n(\mathcal{V}'_n)^{-1} \mathbf{r}'_{ij})^{1/2}$ with $\mathbf{r}'_{ij} = \mathbf{y}'_i - \mathbf{y}'_j - \widehat{\mathcal{B}}_n(\mathcal{V}'_n)^T(\mathbf{u}'_i - \mathbf{u}'_j)$, we must have that $|\widehat{\Sigma}_n(\mathcal{V}'_n)| \leq V$.

Condition (A.3) implies that the cylinder $\mathcal{C}(\widehat{\mathcal{B}}_n(\mathcal{V}'_n), \widehat{\Sigma}_n(\mathcal{V}'_n))$ contains a subcollection of at least $\binom{n}{2} - \binom{n}{2}r$ points of $\mathcal{V}'_n$. From the reverse of (A.2) it follows that this subcollection contains at least $\binom{n}{2} - \binom{n}{2}r - \binom{m}{2} > \binom{p+q}{2} + m(p+q)$ differences that involve at least one original data point of $\mathcal{Z}_n$. This inequality implies one of the two following cases:

- this cylinder contains at least $p+q+1$ differences between two original data points not all lying on the same hyperplane.
- the cylinder contains at most $\binom{p+q}{2}$ differences of original data points and these differences are lying on a hyperplane. The above inequality then implies that there is at least 1 contaminated point for which the differences with $p+q+1$ original data points are lying in the cylinder.

We now show that, for every $V > 0$, there exists a constant $M > 0$, only depending on $\mathcal{Z}_n$, such that $\|\widehat{\mathcal{B}}_n(\mathcal{V}'_n)\| > M$ implies that the determinant of $\widehat{\Sigma}_n(\mathcal{V}'_n)$ is larger than V.

Let $\lambda_1 \geq \ldots \geq \lambda_q$ be the eigenvalues of $\widehat{\Sigma}_n(\mathcal{V}'_n)$, then $|\widehat{\Sigma}_n(\mathcal{V}'_n)| = \lambda_1 \ldots \lambda_q$. In the first case there exists a constant $\beta > 0$ such that $\lambda_j > \beta$ for all $j = 1, \ldots, q$. (For every $\mathbf{w} \in \mathbb{R}^p$, the axes of the ellipsoid $\{\mathbf{v} | (\mathbf{v} - \widehat{\mathcal{B}}_n(\mathcal{V}'_n)^T\mathbf{w})^T \widehat{\Sigma}_n(\mathcal{V}'_n)^{-1}(\mathbf{v} - \widehat{\mathcal{B}}_n(\mathcal{V}'_n)^T\mathbf{w}) \leq 1\}$ have lengths $\sqrt{\lambda_j}; j = 1, \ldots, q$.)

For symmetric $q \times q$ matrices A, it holds that $\lambda_q(A) = \inf_{\mathbf{v}} \frac{\mathbf{v}^T A \mathbf{v}}{\mathbf{v}^T \mathbf{v}}$ from which we obtain that for $(\mathbf{w}, \mathbf{v}) \in \mathcal{C}(\widehat{\mathcal{B}}_n(\mathcal{V}'_n), \widehat{\Sigma}_n(\mathcal{V}'_n))$

$$\|\mathbf{v} - \widehat{\mathcal{B}}_n(\mathcal{V}'_n)^T\mathbf{w}\|^2 \leq (\mathbf{v} - \widehat{\mathcal{B}}_n(\mathcal{V}'_n)^T\mathbf{w})^T \widehat{\Sigma}_n(\mathcal{V}'_n)^{-1}(\mathbf{v} - \widehat{\mathcal{B}}_n(\mathcal{V}'_n)^T\mathbf{w})\lambda_1 \leq \lambda_1.$$

In particular, for $\mathbf{v} = \mathbf{0}$ we have $\|\widehat{\mathcal{B}}_n(\mathcal{V}_n')^T\mathbf{w}\|^2 \leq \lambda_1$.

Since $\mathcal{C}(\widehat{\mathcal{B}}_n(\mathcal{V}_n'), \widehat{\Sigma}_n(\mathcal{V}_n'))$ contains $p+q+1$ differences of 2 original points that are in general position, there exists a constant $d > 0$, not depending on $\widehat{\mathcal{B}}_n(\mathcal{V}_n')$ or $\widehat{\Sigma}_n(\mathcal{V}_n')$, such that $\|\mathbf{w}\| < d$ implies that $(\mathbf{w}, \mathbf{0}) \in \mathcal{C}(\widehat{\mathcal{B}}_n(\mathcal{V}_n'), \widehat{\Sigma}_n(\mathcal{V}_n'))$. If follows that $\sup_{\|\mathbf{w}\|=d} \|\widehat{\mathcal{B}}_n(\mathcal{V}_n')^T\mathbf{w}\|^2 \leq \lambda_1$, so $\|\widehat{\mathcal{B}}_n(\mathcal{V}_n')^T\|^2 \leq \frac{\lambda_1}{d^2}$.

Now consider case 2 where we have at least 1 contaminated point whose differences with $p+q+1$ original points belongs to $\mathcal{C}(\widehat{\mathcal{B}}_n(\mathcal{V}_n'), \widehat{\Sigma}_n(\mathcal{V}_n'))$. From the triangle inequality if follows that the $\binom{p+q+1}{2}$ differences of these $p+q+1$ original points belong to $\mathcal{C}_2(\widehat{\mathcal{B}}_n(\mathcal{V}_n'), \widehat{\Sigma}_n(\mathcal{V}_n')) = \{(\mathbf{w}, \mathbf{v}); (\mathbf{v} - \widehat{\mathcal{B}}_n(\mathcal{V}_n')^T\mathbf{w})^T\widehat{\Sigma}_n(\mathcal{V}_n')^{-1}(\mathbf{v} - \widehat{\mathcal{B}}_n(\mathcal{V}_n')^T\mathbf{w}) \leq 4\}$. Because $\mathcal{C}_2(\widehat{\mathcal{B}}_n(\mathcal{V}_n'), \widehat{\Sigma}_n(\mathcal{V}_n'))$ contains $\binom{p+q+1}{2}$ differences of original points which are in general position, then similarly as in case 1 it follows that there exists constants $\beta > 0$ and $d > 0$ such that $\|\widehat{\mathcal{B}}_n(\mathcal{V}_n')^T\|^2 \leq \frac{\lambda_1}{d^2}$.

Hence, in both cases we obtain that

$$\|\widehat{\mathcal{B}}_n(\mathcal{V}_n')\| \leq \frac{1}{\alpha_1}\|\widehat{\mathcal{B}}_n(\mathcal{V}_n')\|_2 = \frac{1}{\alpha_1}\|\widehat{\mathcal{B}}_n(\mathcal{V}_n')^T\|_2 \leq \frac{\alpha_2}{\alpha_1}\|\widehat{\mathcal{B}}_n(\mathcal{V}_n')^T\| \leq \frac{\alpha_2\sqrt{\lambda_1}}{\alpha_1 d}.$$

Define

$$M = \frac{\alpha_2 V^{1/2}}{\alpha_1 d\beta^{\frac{q-1}{2}}}.$$

Then we have that $\|\widehat{\mathcal{B}}_n(\mathcal{V}_n')\| > M$ implies that $|\widehat{\Sigma}_n(\mathcal{V}_n')| = \lambda_1 \cdots \lambda_q > V$.

As shown, $\|\widehat{\mathcal{B}}_n(\mathcal{V}_n')\| > M$ implies that $|\widehat{\Sigma}_n(\mathcal{V}_n')| > V$ which yields a contradiction. We have thus shown that $\|\widehat{\mathcal{B}}_n(\mathcal{V}_n')\| \leq M$. Moreover, since $|\widehat{\Sigma}_n(\mathcal{V}_n')| \leq V$ and $\lambda_j > \beta$ for all $j = 1, \ldots, q$, there exists a constant $0 < \alpha < \infty$ (depending on β and V) such that $\lambda_1(\widehat{\Sigma}_n(\mathcal{V}_n')) \leq \alpha$.

We now prove that $\epsilon_n(\widehat{\mathcal{B}}_n(\mathcal{V}_n')), \epsilon_n(\widehat{\Sigma}_n(\mathcal{V}_n')) \leq \frac{1}{n}\lceil n - 1/2 - \sqrt{1 + (1-r)(4n^2 - 4n)}/2\rceil$. Replace $\lceil n - 1/2 - \sqrt{1 + (1-r)(4n^2 - 4n)}/2\rceil$ points of $\mathcal{Z}_n$ to obtain $\mathcal{Z}_n'$, then

$\mathcal{V}'_n$ has at least $\binom{n}{2}r$ contaminated differences, call this amount m'. Let

$$\mathcal{C}(B,C) = \{(\mathbf{w},\mathbf{v}); (\mathbf{v}-B^T\mathbf{w})^T C^{-1}(\mathbf{v}-B^T\mathbf{w}) \leq 1\} \tag{A.4}$$

be a cylinder that satisfies

$$\sum_{((\mathbf{u}'_i-\mathbf{u}'_j)^T,(\mathbf{y}'_i-\mathbf{y}'_j)^T)^T \in \mathcal{V}'_n} \rho(d'_{ij}(B,C)) \leq \binom{n}{2}k = \binom{n}{2}r. \tag{A.5}$$

Now suppose that all differences where at least one contaminated point is involved, are outside $\mathcal{C}(B,C)$. Then

$$\sum_{((\mathbf{u}'_i-\mathbf{u}'_j)^T,(\mathbf{y}'_i-\mathbf{y}'_j)^T)^T \in \mathcal{V}'_n} \rho(d'_{ij}(B,C)) = \sum_{((\mathbf{u}'_i-\mathbf{u}'_j)^T,(\mathbf{y}'_i-\mathbf{y}'_j)^T)^T \in \mathcal{V}_n \cap \mathcal{V}'_n} \rho(d'_{ij}(B,C)) + m' \geq \binom{n}{2}r.$$

If $m' = \binom{n}{2}r$ then $\binom{n}{2} - m' = \binom{n}{2} - \binom{n}{2}r \geq \binom{p+q+1}{2}$, so there exists at least one difference $((\mathbf{u}'_i - \mathbf{u}'_j)^T, (\mathbf{y}'_i - \mathbf{y}'_j)^T)^T \in \mathcal{V}_n \cap \mathcal{V}'_n$ for which $d'_{ij}(B,C) > 0$. Because ρ is strictly increasing, this implies that $\sum_{\mathcal{V}'_n} \rho(d'_{ij}(B,C)) > \binom{n}{2}r$ so we have a contradiction. Hence, any cylinder of type (A.4) that satisfies (A.5) contains at least one difference involving an outlier. By letting $\|\mathbf{y}\| \to \infty$ for the contaminated points and also making sure that the distance between them is large, we have $\|\mathbf{y}_1 - \mathbf{y}_2\| \to \infty$ in all cases, hence we can make sure that at least one of the eigenvalues of C goes to infinity. Therefore, both $\widehat{\mathcal{B}}_n(Z_n)$ and $\widehat{\Sigma}_n(Z_n)$ break down in this case.

We now show that $\epsilon_n^* \leq (\lceil 1/2 - p - q + \sqrt{1 + (1-r)(4n^2-4n)}/2 \rceil)/n$. Condition 1 implies that there are at most $p+q$ original points on the same hyperplane of $\mathbb{R}^{p+q}$. Hence, $\exists \alpha \in \mathbb{R}^q, \gamma \in \mathbb{R}^p$ such that $\alpha^T\mathbf{y}_i - \gamma^T\mathbf{u}_i = 0$ for all $i \in I \subset \{1,\ldots,n\}$ with size$(I) = p+q$. If $\alpha \neq \mathbf{0}$ then $\exists B \in \mathbb{R}^{p\times q}$ such that $\gamma = B\alpha$ which implies $\alpha^T(\mathbf{y}_i - B^T\mathbf{u}_i) = 0, \forall i \in I$, so $\mathbf{y}_i - B^T\mathbf{u}_i \in S$ with S a $(q-1)$-dimensional subspace of $\mathbb{R}^q$. Take $D \in \mathbb{R}^{p\times q}$ with $\|D\| = 1$ such that

$\{D^T\mathbf{u}; \mathbf{u} \in \mathbb{R}^p\} \subset S$ (such a D always exists). Now replace $m = \lceil 1/2 - p - q + \sqrt{1 + (1-r)(4n^2 - 4n)}/2 \rceil$ observations of $\mathcal{Z}_n$, not lying on S by $((l\mathbf{u}_0)^T, ((B + tD)^T l\mathbf{u}_0)^T)^T, l = 1, \ldots, m$ for some arbitrarily chosen $\mathbf{u}_0 \in \mathbb{R}^p$ and $t \in \mathbb{R}$. For the contaminated points it then holds that the residuals $\mathbf{r}_l(B+tD)$ equal $\mathbf{0}$ and thus also the differences between residuals of two contaminated data points equal $\mathbf{0}$. For the difference of two observations with indices $i, j \in I$ we have that $(\mathbf{r}_i - \mathbf{r}_j)(B+tD) = \mathbf{y}_i - \mathbf{y}_j - B^T(\mathbf{u}_i - \mathbf{u}_j) - tD^T(\mathbf{u}_i - \mathbf{u}_j) \in S$ and for the difference of an observation with index $i \in I$ and a contaminated observation we have $(\mathbf{r}_i - \mathbf{r}_l)(B+tD) = \mathbf{r}_i(B+tD) \in S$. Denote $\{\mathbf{e}_1, \ldots, \mathbf{e}_{q-1}\}$ an orthonormal basis of S and $\mathbf{e}_q$ a normed vector orthogonal to S. Denote $P = [\mathbf{e}_1, \ldots, \mathbf{e}_q]$. Consider C of the form $C = P\Lambda P^T$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$. Then we have that $((\mathbf{r}_l - \mathbf{r}_{l'})(B + tD))^T C^{-1}(\mathbf{r}_l - \mathbf{r}_{l'})(B + tD) = 0$ for the difference of 2 outliers. For the observations satisfying $(\mathbf{r}_i - \mathbf{r}_j)(B + tD) \in S$, there exists coefficients $\zeta_1, \ldots, \zeta_q$ such that $(\mathbf{r}_i - \mathbf{r}_j)(B + tD) = \sum_{k=1}^{q-1} \zeta_k \mathbf{e}_k$. Therefore

$$
\begin{aligned}
&((\mathbf{r}_i - \mathbf{r}_j)(B + tD))^T C^{-1}(\mathbf{r}_i - \mathbf{r}_j)(B + tD) \\
&= \sum_{k=1}^{q-1} \zeta_k \mathbf{e}_k^T \left(\sum_{l=1}^{q} \lambda_l^{-1} \mathbf{e}_l \mathbf{e}_l^T \right) \sum_{k=1}^{q-1} \zeta_k \mathbf{e}_k \\
&= \left(\sum_{k=1}^{q-1} \zeta_k \lambda_k^{-1} \mathbf{e}_k^T \right) \sum_{k=1}^{q-1} \zeta_k \mathbf{e}_k = \sum_{k=1}^{q-1} \zeta_k^2 \lambda_k^{-1}.
\end{aligned}
$$

Now $\sum_{i<j} \rho((((\mathbf{r}_i - \mathbf{r}_j)(B + tD))^T C^{-1}(\mathbf{r}_i - \mathbf{r}_j)(B + tD))^{1/2}) =$

$$
\sum_{\text{diff of 2 outliers}} + \sum_{\text{number on } S} + \sum_{\text{remainder}} \leq \sum_{\text{number on } S} + \binom{n}{2} r.
$$

where number on $S = \binom{p+q}{2} + \lceil 1/2 - p - q + \sqrt{1 + (1-r)(4n^2 - 4n)}/2 \rceil (p+q)$. Hence we need

$$
\sum_{\text{number on } S} \leq 0. \tag{A.6}
$$

By letting $\lambda_1, \ldots, \lambda_{q-1} \to \infty$ we can make $((\mathbf{r}_i - \mathbf{r}_j)(B+tD))^T C^{-1}(\mathbf{r}_i - \mathbf{r}_j)(B + tD) \to 0$, such that $\sum_{\text{number on } S} \to 0$. We have that the optimal solution

$(\widehat{\mathcal{B}}_n(\mathcal{Z}_n'), \widehat{\Sigma}_n(\mathcal{Z}_n'))$ satisfies $|\widehat{\Sigma}_n(\mathcal{Z}_n')| \leq |C|$ for any $(B+tD, C)$ satisfying (A.6). Now $|C| = \lambda_1 \cdots \lambda_q$ and condition (A.6) does not depend on λ_q so we can let $\lambda_q \to 0$ yielding $|C| \to 0$. By letting $t \to \infty$, we thus obtain that both $\widehat{\mathcal{B}}_n(\mathcal{Z}_n')$ and $\widehat{\Sigma}_n(\mathcal{Z}_n')$ break down.

If $\alpha = \mathbf{0}$, then $\gamma^T \mathbf{u}_i = 0$ for all $i \in I$. We now put the $m = \lceil 1/2 - p - q + \sqrt{1 + (1-r)(4n^2 - 4n)}/2 \rceil$ outliers on the vertical hyperplane $\gamma^T \mathbf{u}_i = 0$ at infinity such that at least $\binom{n}{2} - \binom{n}{2} r$ differences are lying on the vertical hyperplane. It can easily be seen that if at least $\binom{n}{2} - \binom{n}{2} r$ points lie on a hyperplane, then this hyperplane is an optimal solution with an accompanying covariance matrix having zero determinant. In this case however, the hyperplane $\gamma^T(\mathbf{u}_i - \mathbf{u}_j) = 0$ is vertical such that $\|\widehat{\mathcal{B}}_n(\mathcal{Z}_n')\| = \infty$ and $|\widehat{\Sigma}_n(\mathcal{Z}_n'| = 0$. □

Proof of Theorem 2. Due to equivariance we may assume that $\mathcal{B} = 0$ and $\Sigma = I_q$, so $\mathbf{y}_1 - \mathbf{y}_2 = \epsilon_1 - \epsilon_2 \sim F$. It now suffices to show that $\mathcal{B}_{GS}(H) = 0$. Since the constant k can be chosen such that $k = E_F[\rho(\|\epsilon_1 - \epsilon_2\|)]$ which assures consistency at the model with F the distribution of the difference of the errors, it follows that $\Sigma_{GS}(H) = I_q$. Because $\mathcal{B}_{GS}$ is the GS-solution it satisfies the first order condition:

$$\iint u(d_H(\mathbf{r}_1 - \mathbf{r}_2))(\mathbf{u}_1 - \mathbf{u}_2)(\mathbf{y}_1 - \mathbf{y}_2 - \mathcal{B}_{GS}^T(H)(\mathbf{u}_1 - \mathbf{u}_2))^T dH(\mathbf{u}_1 - \mathbf{u}_2, \mathbf{y}_1 - \mathbf{y}_2) = 0 \qquad \text{(A.7)}$$

Now suppose that $\mathcal{B}_{GS} \neq 0$. Let $\lambda_1, \ldots, \lambda_q$ be the eigenvalues of Σ_{GS} and $\mathbf{v}_1, \ldots, \mathbf{v}_q$ the corresponding eigenvectors. There will be at least one $1 \leq j \leq q$ such that $\mathcal{B}_{GS}\mathbf{v}_j \neq 0$. Fix this j. From (A.7) it follows that we should have

$$\iint \mathbf{v}_j^T(\mathcal{B}_{GS}^T(\mathbf{u}_1 - \mathbf{u}_2))u(d_H(\mathbf{r}_1 - \mathbf{r}_2))(\mathbf{y}_1 - \mathbf{y}_2 - \mathcal{B}_{GS}^T(H)(\mathbf{u}_1 - \mathbf{u}_2))^T \mathbf{v}_j dF(\mathbf{y}_1 - \mathbf{y}_2) dG(\mathbf{u}_1 - \mathbf{u}_2) = 0$$

which can be rewritten as

$$\int_{\mathbb{R}^p} \mathbf{v}_j^T(\mathcal{B}_{GS}^T(\mathbf{u}_1-\mathbf{u}_2))I(\mathbf{u}_1-\mathbf{u}_2)dG(\mathbf{u}_1-\mathbf{u}_2)=0 \tag{A.8}$$

with

$$I(\mathbf{u}_1-\mathbf{u}_2)=\int_{\mathbb{R}^q} u(d_H(\mathbf{r}_1-\mathbf{r}_2))(\mathbf{y}_1-\mathbf{y}_2-\mathcal{B}_{GS}^T(H)(\mathbf{u}_1-\mathbf{u}_2))^T\mathbf{v}_j dF(\mathbf{y}_1-\mathbf{y}_2).$$

Fix $\mathbf{u}_1-\mathbf{u}_2$ and set $\mathbf{d}=(d_1,\ldots,d_q)^T:=\mathcal{B}_{GS}^T(\mathbf{u}_1-\mathbf{u}_2)$. Since $\mathbf{y}_1-\mathbf{y}_2$ is spherically symmetrically distributed, for computing $I(\mathbf{u}_1-\mathbf{u}_2)$ we may assume w.l.o.g. that $\Sigma_{GS}=\mathrm{diag}(\lambda_1\ldots,\lambda_q)$ as well as $\mathbf{v}_j=(1,0,\ldots,0)^T$.

Because $u(s)=\rho'(s)/s$ only differs from zero if $s\le c$. With $s=d_H(\mathbf{r}_1-\mathbf{r}_2)$ we obtain $\sqrt{\sum_{j=1}^q \frac{(y_{1j}-y_{2j}-d_j)^2}{\lambda_j}}\le c$. For every $d_1-c\sqrt{\lambda_1}\le y_{11}-y_{21}\le d_1+c\sqrt{\lambda_1}$ denote

$$\mathcal{C}(y_{11}-y_{21})=\left\{(y_{12}-y_{22},\ldots,y_{1q}-y_{2q})\in\mathbb{R}^{q-1}|\sum_{j=2}^q\frac{(y_{1j}-y_{2j}-d_j)^2}{\lambda_j}\le c^2-\frac{(y_{11}-y_{21}-d_1)^2}{\lambda_1}\right\}$$

Then we can rewrite $I(\mathbf{u}_1-\mathbf{u}_2)$ as

$$\begin{aligned}
&I(\mathbf{u}_1-\mathbf{u}_2)\\
&=\int_{d_1-c\sqrt{\lambda_1}}^{d_1+c\sqrt{\lambda_1}}\int_{\mathcal{C}(y_{11}-y_{21})} u(d_H(\mathbf{r}_1-\mathbf{r}_2))(y_{11}-y_{21}-d_1)g((y_{11}-y_{21})^2+\cdots+(y_{1q}-y_{2q})^2)d(y_{11}-y_{21})\ldots d(y_{1q}-y_{2q})\\
&=\int_{-c\sqrt{\lambda_1}}^{c\sqrt{\lambda_1}} t\int_{\mathcal{C}(d_1+t)} u(d_H(\mathbf{r}_1-\mathbf{r}_2))g((d_1+t)^2+\cdots+(y_{1q}-y_{2q})^2)d(y_{12}-y_{22})\ldots d(y_{1q}-y_{2q})dt.
\end{aligned}$$

Since $\mathcal{C}(d_1+t)=\mathcal{C}(d_1-t)$ it follows that

$$\begin{aligned}
I(\mathbf{u}_1-\mathbf{u}_2)=\int_0^{c\sqrt{\lambda_1}} t\int_{\mathcal{C}(d_1+t)} u(d_H(\mathbf{r}_1-\mathbf{r}_2))&g\left((d_1+t)^2+(y_{12}-y_{22})^2+\cdots+(y_{1q}-y_{2q})^2\right)\\
&-g\left((d_1-t)^2+(y_{12}-y_{22})^2+\cdots+(y_{1q}-y_{2q})^2\right)d(y_{12}-y_{22})\ldots d(y_{1q}-y_{2q})dt.
\end{aligned}$$

If $d_1>0$ we have $(d_1+t)^2+(y_{12}-y_{22})^2+\cdots+(y_{1q}-y_{2q})^2>(d_1-t)^2+(y_{12}-y_{22})^2+\cdots+(y_{1q}-y_{2q})^2$ (for $t>0$) and since g is strictly decreasing this implies $I(\mathbf{u}_1-\mathbf{u}_2)<0$. Similarly, we can show that $d_1<0$ implies

$I(\mathbf{u}_1 - \mathbf{u}_2) > 0$ and that $d_1 = 0$ yields $I(\mathbf{u}_1 - \mathbf{u}_2) = 0$. Hence, we have shown that $\mathbf{v}_j^T(\mathcal{B}_{GS}^T(\mathbf{u}_1 - \mathbf{u}_2)) > 0$ implies $I(\mathbf{u}_1 - \mathbf{u}_2) < 0$ and if $\mathbf{v}_j^T(\mathcal{B}_{GS}^T(\mathbf{u}_1 - \mathbf{u}_2)) < 0$, then $I(\mathbf{u}_1 - \mathbf{u}_2) > 0$. Also $\mathbf{v}_j^T(\mathcal{B}_{GS}^T(\mathbf{u}_1 - \mathbf{u}_2)) = 0$ implies $I(\mathbf{u}_1 - \mathbf{u}_2) = 0$. However, due to the regularity condition 3 on the model distribution, the latter event occurs with probability less than $1 - r$. Therefore, we obtain

$$\int_{\mathbb{R}^p} \mathbf{v}_j^T(\mathcal{B}_{GS}^T(\mathbf{u}_1 - \mathbf{u}_2)) I(\mathbf{u}_1 - \mathbf{u}_2) dG(\mathbf{u}_1 - \mathbf{u}_2) < 0$$

which contradicts (A.8), so we conclude that $\mathcal{B}_{GS} = 0$. □

Proof of Theorem 3. It can be shown that the GS-functional $\mathbf{GS}(H) = (\mathcal{B}_{GS}(H), \Sigma_{GS}(H))$ can be represented (as in Lopuhaä 1989) by the following equations

$$\iint u(d_H(\mathbf{r}_1 - \mathbf{r}_2))(\mathbf{u}_1 - \mathbf{u}_2)(\mathbf{y}_1 - \mathbf{y}_2 - \mathcal{B}_{GS}^T(H)(\mathbf{u}_1 - \mathbf{u}_2))^T dH dH = 0 \tag{A.9}$$

$$\begin{aligned} &\iint qu(d_H(\mathbf{r}_1 - \mathbf{r}_2))(\mathbf{y}_1 - \mathbf{y}_2 - \mathcal{B}_{GS}^T(H)(\mathbf{u}_1 - \mathbf{u}_2))(\mathbf{y}_1 - \mathbf{y}_2 - \mathcal{B}_{GS}^T(H)(\mathbf{u}_1 - \mathbf{u}_2))^T dH dH \\ =&\iint v(d_H(\mathbf{r}_1 - \mathbf{r}_2)) dH dH \Sigma_{GS}(H) \end{aligned} \tag{A.10}$$

with $(\mathbf{u}_1, \mathbf{y}_1)$ and $(\mathbf{u}_2, \mathbf{y}_2)$ realizations of two independent variables $\sim H$. $\mathbf{r}_i = \mathbf{y}_i - \mathcal{B}^T\mathbf{u}_i - \alpha$ so $\mathbf{r}_1 - \mathbf{r}_2 = \mathbf{y}_1 - \mathbf{y}_2 - \mathcal{B}_{GS}^T(H)(\mathbf{u}_1 - \mathbf{u}_2)$ and $d_H(\mathbf{r}_1 - \mathbf{r}_2) = ((\mathbf{r}_1 - \mathbf{r}_2)^T\Sigma_{GS}^{-1}(H)(\mathbf{r}_1 - \mathbf{r}_2))^{1/2}$.

We first derive the influence function of the slope matrix $\mathcal{B}_{GS}$ at H_0. From (A.9) it follows that

$$\frac{\partial}{\partial \epsilon} \iint u(d_{H_\epsilon}(\mathbf{r}_1 - \mathbf{r}_2))(\mathbf{u}_1 - \mathbf{u}_2)(\mathbf{y}_1 - \mathbf{y}_2 - \mathcal{B}_{GS}^T(H_\epsilon)(\mathbf{u}_1 - \mathbf{u}_2))^T dH_\epsilon dH_\epsilon|_{\epsilon=0} = 0$$

where $H_\epsilon = H_{\epsilon,\mathbf{z}_0} = (1 - \epsilon)H_0 + \epsilon\Delta_{\mathbf{z}_0}$. This yields

$$
\begin{aligned}
&\frac{\partial}{\partial\epsilon}\Big[(1-\epsilon)^2\iint u(d_{H_\epsilon}(\mathbf{r}_1-\mathbf{r}_2))(\mathbf{u}_1-\mathbf{u}_2)(\mathbf{y}_1-\mathbf{y}_2-\mathcal{B}_{GS}^T(H_\epsilon)(\mathbf{u}_1-\mathbf{u}_2))^T dH_0 dH_0\\
&+2\epsilon(1-\epsilon)\iint u(d_{H_\epsilon}(\mathbf{r}_1-\mathbf{r}_2))(\mathbf{u}_1-\mathbf{u}_2)(\mathbf{y}_1-\mathbf{y}_2-\mathcal{B}_{GS}^T(H_\epsilon)(\mathbf{u}_1-\mathbf{u}_2))^T dH_0\Delta_{\mathbf{z}_0}\\
&+\epsilon^2\iint u(d_{H_\epsilon}(\mathbf{r}_1-\mathbf{r}_2))(\mathbf{u}_1-\mathbf{u}_2)(\mathbf{y}_1-\mathbf{y}_2-\mathcal{B}_{GS}^T(H_\epsilon)(\mathbf{u}_1-\mathbf{u}_2))^T\Delta_{\mathbf{z}_0}^2\Big]\,|_{\epsilon=0}=0
\end{aligned}
$$

Differentiating with respect to ϵ and accounting for equation (A.9) yields

$$
\begin{aligned}
&\frac{\partial}{\partial\epsilon}\Big[\iint u(d_{H_\epsilon}(\mathbf{r}_1-\mathbf{r}_2))(\mathbf{u}_1-\mathbf{u}_2)(\mathbf{y}_1-\mathbf{y}_2-\mathcal{B}_{GS}^T(H_\epsilon)(\mathbf{u}_1-\mathbf{u}_2))^T dH_0 dH_0\Big]\,|_{\epsilon=0}\\
&+2\iint u(d_{H_0}(\mathbf{r}_1-\mathbf{r}_2))(\mathbf{u}_1-\mathbf{u}_2)(\mathbf{y}_1-\mathbf{y}_2-\mathcal{B}_{GS}^T(H_0)(\mathbf{u}_1-\mathbf{u}_2))^T dH_0\Delta_{\mathbf{z}_0}=0
\end{aligned}
$$

Rewriting term 1, we get

$$
\begin{aligned}
&-\iint u'(d_{H_0}(\mathbf{r}_1-\mathbf{r}_2))\frac{\partial}{\partial\epsilon}d_{H_\epsilon}(\mathbf{r}_1-\mathbf{r}_2)|_{\epsilon=0}(\mathbf{u}_1-\mathbf{u}_2)(\mathbf{y}_1-\mathbf{y}_2-\mathcal{B}_{GS}^T(H_0)(\mathbf{u}_1-\mathbf{u}_2))^T dH_0 dH_0\\
&-\iint u(d_{H_0}(\mathbf{r}_1-\mathbf{r}_2))(\mathbf{u}_1-\mathbf{u}_2)(\mathbf{u}_1-\mathbf{u}_2)^T(-IF(\mathbf{z}_0;\mathcal{B}_{GS},H_0))dH_0 dH_0\\
&=2\int u(d_{H_0}(\mathbf{r}_1-\mathbf{r}_0))(\mathbf{u}_1-\mathbf{u}_0)(\mathbf{y}_1-\mathbf{y}_0-\mathcal{B}_{GS}^T(H_0)(\mathbf{u}_1-\mathbf{u}_0))^T dH_0
\end{aligned}
$$

Since $\mathcal{B}_{GS}(H_0)=0$ and $\Sigma_{GS}(H_0)=I_q$ we have $d_{H_0}(\mathbf{r}_1-\mathbf{r}_2)=\sqrt{(\mathbf{y}_1-\mathbf{y}_2)^T(\mathbf{y}_1-\mathbf{y}_2)}=\|\mathbf{y}_1-\mathbf{y}_2\|$. Hence, we obtain

$$
\begin{aligned}
&-\iint u'(\|\mathbf{y}_1-\mathbf{y}_2\|)\frac{\partial}{\partial\epsilon}d_{H_\epsilon}(\mathbf{r}_1-\mathbf{r}_2)|_{\epsilon=0}(\mathbf{u}_1-\mathbf{u}_2)(\mathbf{y}_1-\mathbf{y}_2)^T dH_0 dH_0\\
&+\iint u(\|\mathbf{y}_1-\mathbf{y}_2\|)(\mathbf{u}_1-\mathbf{u}_2)(\mathbf{u}_1-\mathbf{u}_2)^T dH_0 dH_0 IF(\mathbf{z}_0;\mathcal{B}_{GS},H_0)\\
&=2\int u(d_{H_0}(\mathbf{r}_1-\mathbf{r}_0))(\mathbf{u}_1-\mathbf{u}_0)(\mathbf{y}_1-\mathbf{y}_0)^T dH_0 \qquad\qquad (A.11)
\end{aligned}
$$

Using that

$$
\begin{aligned}
&\frac{\partial}{\partial\epsilon}d_{H_\epsilon}(\mathbf{r}_1-\mathbf{r}_2)|_{\epsilon=0}\\
&=\frac{(-IF(\mathbf{z}_0;\mathcal{B}_{GS},H_0)^T(\mathbf{u}_1-\mathbf{u}_2))^T}{\|\mathbf{y}_1-\mathbf{y}_2\|}(\mathbf{y}_1-\mathbf{y}_2)+\frac{1}{2}\frac{(\mathbf{y}_1-\mathbf{y}_2)^T}{\|\mathbf{y}_1-\mathbf{y}_2\|}IF(\mathbf{z}_0;\Sigma_{GS}^{-1},H_0)(\mathbf{y}_1-\mathbf{y}_2)
\end{aligned}
$$

the first term of (A.11) becomes

$$
\begin{aligned}
&- \iint u'(\|\mathbf{y}_1 - \mathbf{y}_2\|) \frac{\partial}{\partial \epsilon} d_{H_\epsilon}(\mathbf{r}_1 - \mathbf{r}_2)|_{\epsilon=0} (\mathbf{u}_1 - \mathbf{u}_2)(\mathbf{y}_1 - \mathbf{y}_2)^T dH_0 dH_0 \\
&= \iint (\mathbf{u}_1 - \mathbf{u}_2)(\mathbf{u}_1 - \mathbf{u}_2)^T dG dG IF(\mathbf{z}_0; \mathcal{B}_{GS}, H_0) \iint \frac{u'(\|\mathbf{y}_1 - \mathbf{y}_2\|)}{\|\mathbf{y}_1 - \mathbf{y}_2\|} (\mathbf{y}_1 - \mathbf{y}_2)(\mathbf{y}_1 - \mathbf{y}_2)^T dF_0 dF_0 \\
&- \frac{1}{2} \iint (\mathbf{u}_1 - \mathbf{u}_2) dG dG \iint \frac{u'(\|\mathbf{y}_1 - \mathbf{y}_2\|)}{\|\mathbf{y}_1 - \mathbf{y}_2\|} (\mathbf{y}_1 - \mathbf{y}_2)^T IF(\mathbf{z}_0; \Sigma_{GS}^{-1}, H_0)(\mathbf{y}_1 - \mathbf{y}_2)(\mathbf{y}_1 - \mathbf{y}_2)^T dF_0 dF_0
\end{aligned}
$$

The last term vanishes because $E_{G\times G}[\mathbf{u}_1 - \mathbf{u}_2] = 0$. Hence equation (A.11) becomes:

$$
\begin{aligned}
&E_{G\times G}[(\mathbf{u}_1 - \mathbf{u}_2)(\mathbf{u}_1 - \mathbf{u}_2)^T] IF(\mathbf{z}_0; \mathcal{B}_{GS}, H_0) \left[\iint \frac{u'(\|\mathbf{y}_1 - \mathbf{y}_2\|)}{\|\mathbf{y}_1 - \mathbf{y}_2\|} (\mathbf{y}_1 - \mathbf{y}_2)(\mathbf{y}_1 - \mathbf{y}_2)^T dF_0 dF_0 \right. \\
&\left. + \iint u(\|\mathbf{y}_1 - \mathbf{y}_2\|) dF_0 dF_0 \right] \\
&= 2 \int u(d_{H_0}(\mathbf{r}_1 - \mathbf{r}_0))(\mathbf{u}_1 - \mathbf{u}_0)(\mathbf{y}_1 - \mathbf{y}_0)^T dH_0
\end{aligned}
$$

From symmetry it follows that $\iint \frac{u'(\|\mathbf{y}_1 - \mathbf{y}_2\|)}{\|\mathbf{y}_1 - \mathbf{y}_2\|} (\mathbf{y}_1 - \mathbf{y}_2)(\mathbf{y}_1 - \mathbf{y}_2)^T dF_0 dF_0 = \iint u'(\|\mathbf{y}_1 - \mathbf{y}_2\|) \frac{1}{q} \|\mathbf{y}_1 - \mathbf{y}_2\| dF_0 dF_0 I_q$ hence we obtain

$$
IF(\mathbf{z}_0; \mathcal{B}_{GS}, H_0) = E_{G\times G}[(\mathbf{u}_1 - \mathbf{u}_2)(\mathbf{u}_1 - \mathbf{u}_2)^T]^{-1} \frac{2 \int u(\|\mathbf{y}_1 - \mathbf{y}_0\|)(\mathbf{u}_1 - \mathbf{u}_0)(\mathbf{y}_1 - \mathbf{y}_0)^T dH_0}{E_{F_0\times F_0} \left[u'(\|\mathbf{y}_1 - \mathbf{y}_2\|) \frac{\|\mathbf{y}_1 - \mathbf{y}_2\|}{q} + u(\|\mathbf{y}_1 - \mathbf{y}_2\|) \right]}
$$

Using $u'(t)t = \psi'(t) - \psi(t)/t$ yields

$$
\begin{aligned}
&IF(\mathbf{z}_0; \mathcal{B}_{GS}, H_0) \\
&= E_{G\times G}[(\mathbf{u}_1 - \mathbf{u}_2)(\mathbf{u}_1 - \mathbf{u}_2)^T]^{-1} \frac{2 \int (\mathbf{u_1} - \mathbf{u}_0) dG \int u(\|\mathbf{y}_1 - \mathbf{y}_0\|)(\mathbf{y}_1 - \mathbf{y}_0)^T dF_0}{E_{F_0\times F_0} \left[\frac{1}{q} \psi'(\|\mathbf{y}_1 - \mathbf{y}_2\|) + \left(1 - \frac{1}{q}\right) u(\|\mathbf{y}_1 - \mathbf{y}_2\|) \right]} \\
&= [Cov(\mathbf{u})]^{-1} \frac{\int (\mathbf{u_1} - \mathbf{u}_0) dG \int u(\|\mathbf{y}_1 - \mathbf{y}_0\|)(\mathbf{y}_1 - \mathbf{y}_0)^T dF_0}{E_{F_0\times F_0} \left[\frac{1}{q} \psi'(\|\mathbf{y}_1 - \mathbf{y}_2\|) + \left(1 - \frac{1}{q}\right) u(\|\mathbf{y}_1 - \mathbf{y}_2\|) \right]}
\end{aligned}
$$

The influence function of Σ_{GS} is derived in a similar way, now by differentiating equation (A.10)

$$
\begin{aligned}
&\frac{\partial}{\partial \epsilon} \iint q u(d_{H_\epsilon}(\mathbf{r}_1 - \mathbf{r}_2))(\mathbf{y}_1 - \mathbf{y}_2 - \mathcal{B}_{GS}^T(H_\epsilon)(\mathbf{u}_1 - \mathbf{u}_2))(\mathbf{y}_1 - \mathbf{y}_2 - \mathcal{B}_{GS}^T(H_\epsilon)(\mathbf{u}_1 - \mathbf{u}_2))^T dH_\epsilon dH_\epsilon|_{\epsilon=0} \\
&= \frac{\partial}{\partial \epsilon} \iint v(d_{H_\epsilon}(\mathbf{r}_1 - \mathbf{r}_2)) dH_\epsilon dH_\epsilon|_{\epsilon=0} \Sigma_{GS}(H_0) + \iint v(d_{H_0}(\mathbf{r}_1 - \mathbf{r}_2)) dH_0 dH_0 IF(\mathbf{z}_0; \Sigma_{GS}, H_0)
\end{aligned}
$$

Differentiating and taking (A.10) into account leads to

$$
\begin{aligned}
&E_{H_0\times H_0}[v(\|\mathbf{y}_1-\mathbf{y}_2\|)]IF(\mathbf{z}_0,\Sigma_{GS},H_0)\\
&+\frac{1}{2}\iint\frac{v'(\|\mathbf{y}_1-\mathbf{y}_2\|)}{\|\mathbf{y}_1-\mathbf{y}_2\|}(\mathbf{y}_1-\mathbf{y}_2)^TIF(\mathbf{z}_0,\Sigma_{GS}^{-1},H_0)(\mathbf{y}_1-\mathbf{y}_2)dH_0dH_0I_q\\
&-\frac{q}{2}\iint\frac{u'(\|\mathbf{y}_1-\mathbf{y}_2\|)}{\|\mathbf{y}_1-\mathbf{y}_2\|}(\mathbf{y}_1-\mathbf{y}_2)^TIF(\mathbf{z}_0,\Sigma_{GS}^{-1},H_0)(\mathbf{y}_1-\mathbf{y}_2)(\mathbf{y}_1-\mathbf{y}_2)(\mathbf{y}_1-\mathbf{y}_2)^TdH_0dH_0\\
&=2\int qu(\|\mathbf{y}_1-\mathbf{y}_0\|)(\mathbf{y}_1-\mathbf{y}_0)(\mathbf{y}_1-\mathbf{y}_0)^TdH_0-2\int v(\|\mathbf{y}_1-\mathbf{y}_0\|)dH_0I_q
\end{aligned}
$$

and we rewrite this as (using $IF(\mathbf{z}_0;\Sigma_{GS}^{-1},H_0)=-IF(\mathbf{z}_0;\Sigma_{GS},H_0)$)

$$
\begin{aligned}
&E_{H_0\times H_0}[v(\|\mathbf{y}_1-\mathbf{y}_2\|)]IF(\mathbf{z}_0,\Sigma_{GS},H_0)\\
&-\frac{1}{2}\sum_{i,j=1}^{q}\iint\frac{v'(\|\mathbf{y}_1-\mathbf{y}_2\|)}{\|\mathbf{y}_1-\mathbf{y}_2\|}(\mathbf{y}_1-\mathbf{y}_2)_i^TIF(\mathbf{z}_0,(\Sigma_{GS})_{ij},H_0)(\mathbf{y}_1-\mathbf{y}_2)_jdH_0dH_0I_q\\
&+\frac{q}{2}\sum_{i,j=1}^{q}\iint\frac{u'(\|\mathbf{y}_1-\mathbf{y}_2\|)}{\|\mathbf{y}_1-\mathbf{y}_2\|}(\mathbf{y}_1-\mathbf{y}_2)_iIF(\mathbf{z}_0,(\Sigma_{GS})_{ij},H_0)(\mathbf{y}_1-\mathbf{y}_2)_j(\mathbf{y}_1-\mathbf{y}_2)(\mathbf{y}_1-\mathbf{y}_2)^TdH_0dH_0\\
&=2\int qu(\|\mathbf{y}_1-\mathbf{y}_0\|)(\mathbf{y}_1-\mathbf{y}_0)(\mathbf{y}_1-\mathbf{y}_0)^TdH_0-2\int v(\|\mathbf{y}_1-\mathbf{y}_0\|)dH_0I_q
\end{aligned}
$$

Eliminating the terms which are 0 and following Lopuhaä (1999, Lemma 2.1) it holds that

$$
\begin{aligned}
&\gamma_1IF(\mathbf{z}_0;\Sigma_{GS},H_0)-\gamma_2\mathrm{tr}IF(\mathbf{z}_0;\Sigma_{GS},H_0)I_q\\
&=2\int qu(\|\mathbf{y}_1-\mathbf{y}_0\|)(\mathbf{y}_1-\mathbf{y}_0)(\mathbf{y}_1-\mathbf{y}_0)^TdH_0-2\int v(\|\mathbf{y}_1-\mathbf{y}_0\|)dH_0I_q
\end{aligned}
$$

where

$$
\begin{aligned}
\gamma_1&=E_{F_0\times F_0}\left[v(\|\mathbf{y}_1-\mathbf{y}_2\|)\right]+\frac{1}{q+2}E_{F_0\times F_0}\left[u'(\|\mathbf{y}_1-\mathbf{y}_2\|)(\|\mathbf{y}_1-\mathbf{y}_2\|)^3\right]\\
\gamma_2&=E_{F_0\times F_0}\left[\frac{1}{2q}v'(\|\mathbf{y}_1-\mathbf{y}_2\|)\|\mathbf{y}_1-\mathbf{y}_2\|\right]-\frac{1}{2(q+2)}E_{F_0\times F_0}\left[u'(\|\mathbf{y}_1-\mathbf{y}_2)\|\mathbf{y}_1-\mathbf{y}_2\|^3\right]
\end{aligned}
$$

or rewriting this with $\gamma_3=E_{F_0\times F_0}[\psi(\|\mathbf{y}_1-\mathbf{y}_2\|)\|\mathbf{y}_1-\mathbf{y}_2\|]$ gives:

$$
\begin{aligned}
&IF(\mathbf{z}_0; \Sigma_{GS}, H_0) \\
&= \frac{2}{\gamma_1} q \int \psi(\|\mathbf{y}_1 - \mathbf{y}_0\|) \|\mathbf{y}_1 - \mathbf{y}_0\| \left[\frac{(\mathbf{y}_1 - \mathbf{y}_0)(\mathbf{y}_1 - \mathbf{y}_0)^T}{\|\mathbf{y}_1 - \mathbf{y}_0\|^2} - \frac{1}{q} I_q \right] dF_0 + \frac{4 \int (\rho(\|\mathbf{y}_1 - \mathbf{y}_0\|) - H) dF_0}{\gamma_3} I_q
\end{aligned}
$$

□

Expression for $\nabla \mathbf{g}(.)$. The matrix of partial derivatives $\nabla \mathbf{g}(.)$ is given by

	pq	qq
pq	$\frac{\partial \text{vec}(\mathbf{A}_n^{-1}\mathbf{B}_n)}{\partial \text{vec}(B)^T}$	$\frac{\partial \text{vec}(\mathbf{A}_n^{-1}\mathbf{B}_n)}{\partial \text{vec}(C)^T}$
qq	$\frac{\partial \text{vec}(\mathbf{V}_n + w_n C)}{\partial \text{vec}(B)^T}$	$\frac{\partial \text{vec}(\mathbf{V}_n + w_n C)}{\partial \text{vec}(C)^T}$

These expressions can be obtained by differentiation. Denote $\mathbf{r}_i := \mathbf{y}_i - B^T \mathbf{u}_i -$

α

1.

$$
\begin{aligned}
&\frac{\partial \text{vec}(\mathbf{A}_n^{-1}\mathbf{B}_n)}{\partial \text{vec}(B)^T} \\
&= -(I_q \otimes \mathbf{A}_n^{-1}) \sum_{i<j} \frac{u'(d_{ij})}{d_{ij}} \text{vec}((\mathbf{u}_i - \mathbf{u}_j)(\mathbf{y}_i - \mathbf{y}_j)^T) \text{vec}((\mathbf{u}_i - \mathbf{u}_j)(\mathbf{r}_i - \mathbf{r}_j)^T C^{-1})^T \\
&+ (\mathbf{B}_n \otimes I_p)^T ((\mathbf{A}_n^T)^{-1} \otimes \mathbf{A}_n^{-1}) \sum_{i<j} \frac{u'(d_{ij})}{d_{ij}} \text{vec}((\mathbf{u}_i - \mathbf{u}_j)(\mathbf{u}_i - \mathbf{u}_j)^T) \text{vec}((\mathbf{u}_i - \mathbf{u}_j)(\mathbf{r}_i - \mathbf{r}_j)^T C^{-1})^T
\end{aligned}
$$

2.

$$
\begin{aligned}
&\frac{\partial \text{vec}(\mathbf{A}_n^{-1}\mathbf{B}_n)}{\partial \text{vec}(C)^T} \\
&= -(I_q \otimes \mathbf{A}_n^{-1}) \sum_{i<j} \frac{u'(d_{ij})}{2d_{ij}} \text{vec}((\mathbf{u}_i - \mathbf{u}_j)(\mathbf{y}_i - \mathbf{y}_j)^T) \text{vec}(C^{-1}(\mathbf{r}_i - \mathbf{r}_j)(\mathbf{r}_i - \mathbf{r}_j)^T C^{-1})^T \\
&+ (\mathbf{B}_n \otimes I_p)^T ((\mathbf{A}_n^T)^{-1} \otimes \mathbf{A}_n^{-1}) \sum_{i<j} \frac{u'(d_{ij})}{2d_{ij}} \text{vec}((\mathbf{u}_i - \mathbf{u}_j)(\mathbf{u}_i - \mathbf{u}_j)^T) \text{vec}(C^{-1}(\mathbf{r}_i - \mathbf{r}_j)(\mathbf{r}_i - \mathbf{r}_j)^T C^{-1})^T
\end{aligned}
$$

3.

$$
\begin{aligned}
&\frac{\partial \text{vec}(\mathbf{V}_n + w_n C)}{\partial \text{vec}(B)^T} \\
=&-\frac{1}{\binom{n}{2}k}\sum_{i<j} qu(d_{ij})\left[(I_q \otimes (\mathbf{r}_i - \mathbf{r}_j)) + ((\mathbf{r}_i - \mathbf{r}_j) \otimes I_q)\right]((\mathbf{u}_i - \mathbf{u}_j)^T \otimes I_q)K_{pq} \\
&-\frac{1}{\binom{n}{2}k}\sum_{i<j}\frac{w'(d_{ij})}{d_{ij}}\text{vec}C\text{vec}((\mathbf{u}_i - \mathbf{u}_j)(\mathbf{r}_i - \mathbf{r}_j)^T C^{-1})^T \\
&-\frac{1}{\binom{n}{2}k}\sum_{i<j} q\frac{u'(d_{ij})}{d_{ij}}\text{vec}((\mathbf{r}_i - \mathbf{r}_j)(\mathbf{r}_i - \mathbf{r}_j)^T)\text{vec}((\mathbf{u}_i - \mathbf{u}_j)(\mathbf{r}_i - \mathbf{r}_j)^T C^{-1})^T
\end{aligned}
$$

4.

$$
\begin{aligned}
&\frac{\partial \text{vec}(\mathbf{V}_n + w_n C)}{\partial \text{vec}(C)^T} \\
=&-\frac{1}{\binom{n}{2}k}\sum_{i<j} q\frac{u'(d_{ij})}{2d_{ij}}\text{vec}((\mathbf{r}_i - \mathbf{r}_j)(\mathbf{r}_i - \mathbf{r}_j)^T)\text{vec}(C^{-1}(\mathbf{r}_i - \mathbf{r}_j)(\mathbf{r}_i - \mathbf{r}_j)^T C^{-1})^T \\
&-\frac{1}{\binom{n}{2}k}\sum_{i<j}\frac{w'(d_{ij})}{2d_{ij}}\text{vec}C\text{vec}(C^{-1}(\mathbf{r}_i - \mathbf{r}_j)(\mathbf{r}_i - \mathbf{r}_j)^T C^{-1})^T \\
&+\frac{1}{\binom{n}{2}k}\sum_{i<j} w(d_{ij})I_{qq}
\end{aligned}
$$

Lemma 1 *Let* $(\tilde{\mathbf{u}}_1^T, \tilde{y}_1)^T, \ldots, (\tilde{\mathbf{u}}_n^T, \tilde{y}_n)^T$ *be* $n \geq p$ *observations in* $\mathbb{R}^{p+1}$ *and* $w_{ij} \geq 0$ *are weights associated with the difference of observation* i *and* j*. Denote* $\mathbf{u}_i - \mathbf{u}_j = \sqrt{w_{ij}}(\tilde{\mathbf{u}}_i - \tilde{\mathbf{u}}_j)$ *and* $y_i - y_j = \sqrt{w_{ij}}(\tilde{y}_i - \tilde{y}_j)$ *such that if*

$$\mathbf{diffU_n} = [\mathbf{u}_1^T - \mathbf{u}_2^T, \mathbf{u}_1^T - \mathbf{u}_3^T, \ldots, \mathbf{u}_{n-1}^T - \mathbf{u}_n^T]^T$$

then $\mathbf{diffU_n}^T\mathbf{diffU_n}$ *has full rank. For a given* $(\tilde{\mathbf{u}}_{n+1}^T, \tilde{y}_{n+1})^T$ *let* $\hat{\beta}_{n+1}$ *be the weighted least squares regression estimate for the differences of the* $n+1$ *points. For any* $C > 0$ *and* $M > 0$ *there exists a finite constant* K *such that* $\|\hat{\beta}_{n+1}\| \leq K$ *for any* $(\tilde{\mathbf{u}}_{n+1}^T, \tilde{y}_{n+1})^T$ *with* $|y_{n+1} - y_i| \leq C + |\beta^T(\mathbf{u}_{n+1} - \mathbf{u}_i)|$ *for every difference getting a non-zero weight and for some* β *with* $\|\beta\| < M$*, and* K *only depends on the differences of the first* n *points and the constants* C *and* M*.*

Proof of Lemma 1. Let $\hat{\beta}_n$ be the weighted least squares estimate based on the differences of the first n points. Let us denote $\mathbf{diffu}_{n+1} = (\mathbf{u}_1^T - \mathbf{u}_{n+1}^T, \ldots, \mathbf{u}_n^T - \mathbf{u}_{n+1}^T)^T$, it can using Seber (1984 p.519) then be shown that

$$\hat{\beta}_{n+1} = (\mathbf{diffU}_n^T\mathbf{diffU}_n + \mathbf{diffu}_{n+1}^T\mathbf{diffu}_{n+1})^{-1}(\mathbf{diffU}_n^T\mathbf{diffY}_n + \mathbf{diffu}_{n+1}^T\mathbf{diffy}_{n+1}).$$

Denote $V = (\mathbf{diffU}_n^T\mathbf{diffU}_n)^{-1}$ which is positive definite, then

$$\begin{aligned}
&\hat{\beta}_{n+1} \\
&= [I_p - V\mathbf{diffu}_{n+1}^T(I_n + \mathbf{diffu}_{n+1}V\mathbf{diffu}_{n+1}^T)^{-1}\mathbf{diffu}_{n+1}]\hat{\beta}_n \\
&+ [V - V\mathbf{diffu}_{n+1}^T(I_n + \mathbf{diffu}_{n+1}V\mathbf{diffu}_{n+1}^T)^{-1}\mathbf{diffu}_{n+1}V]\mathbf{diffu}_{n+1}^T\mathbf{diffy}_{n+1}
\end{aligned}$$

To simplify the notation, put $U = \mathbf{diffu}_{n+1}$, $A = I_p - VU^T(I_n + UVU^T)^{-1}U$, and $B = V - VU^T(I_n + UVU^T)^{-1}UV$ such that we have

$$\hat{\beta}_{n+1} = A\hat{\beta}_n + BU^T\mathbf{diffy}_{n+1}.$$

We have to show that A and $BU^T\mathbf{diffy}_{n+1}$ are bounded for any $\tilde{\mathbf{u}}_{n+1}$ and $\tilde{y}_{n+1}$ or equivalently for every U and $\mathbf{diffy}_{n+1}$ satisfying the conditions stated above. Note that $I_n + UVU^T$ is positive definite because $\forall \mathbf{x} \neq 0 \in \mathbb{R}^n$: $\mathbf{x}^T(I_n + UVU^T)\mathbf{x} = \mathbf{x}^T I_n\mathbf{x} + \mathbf{x}^T UVU^T\mathbf{x} > 0$ since V is positive definite. Hence, $(I_n + UVU^T)^{-1}$ is also positive definite and has a bounded norm. $I + UVU^T$ has (i, j)th element given by

$$\delta_{i,j} + \sum_k \left(\sum_l u_{il}v_{lk} \right) u_{jk}$$

and is of order $\|U\|^2$. Because VU^TU is also of the order $\|U\|^2$ the expression $VU^T(I_n + UVU^T)^{-1}U$ remains bounded as $\|U\| \to \infty$. Hence, A remains bounded for any U. For $BU^T\mathbf{diffy}_{n+1}$ we have the following inequalities:

$$\begin{aligned}
\|BU^T\mathbf{diffy}_{n+1}\| &\leq \|BU^TC\| + \|BU^TU\beta\| \\
&\leq \|BU^T\||C| + \|BU^TU\|\|\beta\|
\end{aligned}$$

Note that $BU^TU = VU^T(I_n+UVU^T)^{-1}U = I_p - A$ which shows that $\|BU^TU\|$ is bounded. By assumption we have that $\|\beta\| \leq M$ such that it remains to be shown that $\|BU^T\|$ is bounded. Note that $BU^T = VU^T/\|U\|^2(\|U\|^{-2}(I_n + UVU^T))^{-1}$. The (j,k)th element of VU^T is $\sum_i v_{ji}u_{ki}$. Since

$$\frac{|u_{ki}|}{\|U\|} \leq \frac{|u_{ki}|}{\|u_k\|} \leq 1$$

which implies that

$$\frac{|u_{ki}|}{\|U\|^2} \to 0,$$

we have that $VU^T/\|U\|^2$ goes to 0 when $\|U\| \to \infty$. Moreover $(\|U\|^{-2}(I_n + UVU^T))^{-1}$ is bounded when $\|U\| \to \infty$. □

Proof of Theorem 4. We have to prove that the bootstrap estimates $\widehat{\mathcal{B}}_n^*$ and $\widehat{\Sigma}_n^*$ can only breakdown in bootstrap samples that contain less than c_p distinct good observations, which implies that there are less than p differences of two good observations. For $\widehat{\Sigma}_n^*$ only the explosion breakdown point is relevant, since implosion is not harmful for the eventual (i.e. after linear correction) fast bootstrap estimate of the parameter $\mathcal{B}$. Note that the linear correction matrix given by the partial derivatives is only computed once, based on the original sample, and it will be as robust as the original GS-estimates are. Hence it has breakdown point ϵ_n^*. Now the bootstrap estimates (without linear correction) are given by

$$\begin{aligned}\widehat{\mathcal{B}}_n^* &= \mathbf{A}_n^*(\widehat{\mathcal{B}}_n, \widehat{\Sigma}_n)^{-1}\mathbf{B}_n^*(\widehat{\mathcal{B}}_n, \widehat{\Sigma}_n)\\ \widehat{\Sigma}_n^* &= \mathbf{V}_n^*(\widehat{\mathcal{B}}_n, \widehat{\Sigma}_n) + w_n^*(\widehat{\mathcal{B}}_n, \widehat{\Sigma}_n)\widehat{\Sigma}_n.\end{aligned}$$

It can easily be seen that $\mathbf{V}_n$ and w_n^* are in any case bounded so that we only have to show that $\widehat{\mathcal{B}}_n$ remains bounded when there are at least p distinct differences of two good observations. Note that p is a strict minimum since

otherwise it might occur that $\mathbf{A}_n^*$ is singular. (Here we assume that differences of two good observations have weight $u(d_{ij}^*) > 0$.)

Now, $\widehat{\mathcal{B}}_n$ is a multivariate weighted least squares estimate and we can apply Lemma 1 since a multivariate least squares coefficient estimate essentially consists of q univariate least squares estimates. The weights $u(d_{ij}^*)$ are bounded, hence they can only have a bounded effect. Consider a bootstrap sample with $k \geq c_p$ distinct good observations and suppose that $(\mathbf{u}_{k+1}^T, \mathbf{y}_{k+1}^T)^T$ is some outlier included in the bootstrap sample. The effect of this outlier on the bootstrap estimate will be bounded as can be seen as follows. There exists some L, only depending on the original data set $\mathcal{Z}_n$ such that $\lambda_1(\widehat{\Sigma}_n) < L$ for all $\mathcal{Z}_n'$ obtained by replacing less than $\epsilon_n^* n$ observations. It then holds that $\inf_{\mathbf{v}} \frac{\mathbf{v}^T \widehat{\Sigma}^{-1} \mathbf{v}}{\mathbf{v}^T \mathbf{v}} = \lambda_q(\widehat{\Sigma}_n^{-1}) > 1/L$. Hence, in case $\|\mathbf{y}_{k+1} - \mathbf{y}_i - \widehat{\mathcal{B}}_n^T(\mathbf{u}_{k+1} - \mathbf{u}_i)\| \geq \sqrt{L}c$ (where c is the constant for which it holds that ρ is constant on $[c, \infty)$) it follows that $\sqrt{(\mathbf{y}_{k+1} - \mathbf{y}_i - \widehat{\mathcal{B}}_n^T(\mathbf{u}_{k+1} - \mathbf{u}_i))^T \widehat{\Sigma}_n^{-1} (\mathbf{y}_{k+1} - \mathbf{y}_i - \widehat{\mathcal{B}}_n^T(\mathbf{u}_{k+1} - \mathbf{u}_i))} \geq c$ and consequently the difference will obtain zero weight in the weighted least squares. In case $\|\mathbf{y}_{k+1} - \mathbf{y}_i - \widehat{\mathcal{B}}_n^T(\mathbf{u}_{k+1} - \mathbf{u}_i)\| < \sqrt{L}c$ for a certain i we have that $|y_{k+1,j} - y_{i,j} - \widehat{\mathcal{B}}_{n,j}^T(\mathbf{u}_{k+1} - \mathbf{u}_i)| < \sqrt{L}c$ for each $j = 1, \ldots, q$. And also $|y_{k+1,j} - y_{i,j}| < \sqrt{L}c + |\widehat{\mathcal{B}}_{n,j}^T(\mathbf{u}_{k+1} - \mathbf{u}_i)|$. Furthermore, from the robustness of the GS-estimator we have for all $\mathcal{Z}_n'$ that $\|\widehat{\mathcal{B}}_{n,j}\| < M$ for some M only depending on $\mathcal{Z}_n$. Because only the differences satisfying $\|\mathbf{y}_{k+1} - \mathbf{y}_i - \widehat{\mathcal{B}}_n^T(\mathbf{u}_{k+1} - \mathbf{u}_i)\| < \sqrt{L}c$ have an influence it follows from Lemma 1 that there exists a bound on the weighted least squares estimate $\|\widehat{\mathcal{B}}_{k+1,j}^{WLS}\|$ depending only on the first k observations, and on L, c and M. Hence, if we now consider all bootstrap samples with at least c_p distinct good observations we obtain a bound only depending on the original data set $\mathcal{Z}_n$. The expected upper breakdown point follows immediately. $\square$

Lemma 2 *Let* $Z_1 := (U_1, Y_1), \ldots, Z_n := (U_n, Y_n) \sim F$ *be a sequence of i.i.d. random vectors. Let* (B_n, S_n) *be consistent estimators for* $(\mathcal{B}, \Sigma)$*. Let* $\kappa : \mathbb{R} \to \mathbb{R}$ *be a function that is bounded and almost everywhere continuous. If* $\tilde{\kappa}(Z_1 - Z_2, B, S) = \kappa((Y_1 - Y_2 - B^T(U_1 - U_2))S^{-1}(Y_1 - Y_2 - B^T(U_1 - U_2)))$*, then*

$$\frac{1}{\binom{n}{2}} \sum_{i<j} \tilde{\kappa}(Z_i - Z_j, B_n, S_n) \xrightarrow{P} E_F[\tilde{\kappa}(Z_1 - Z_2, \mathcal{B}, \Sigma)]$$

Proof of Lemma 2. The proof is based on an argument used in Davies (1987, proof of Theorem 3). Denote $\tilde{\kappa}_n(\mathbf{z}_1 - \mathbf{z}_2) := \tilde{\kappa}(\mathbf{z}_1 - \mathbf{z}_2, B_n, S_n)$ and $\tilde{\kappa}(\mathbf{z}_1 - \mathbf{z}_2) := \tilde{\kappa}(\mathbf{z}_1 - \mathbf{z}_2, \mathcal{B}, \Sigma)$. For any $\mathbf{z}_1$ and $\mathbf{z}_2$ such that κ is continuous at $(\mathbf{y}_1 - \mathbf{y}_2 - \mathcal{B}^T(\mathbf{u}_1 - \mathbf{u}_2))\Sigma^{-1}(\mathbf{y}_1 - \mathbf{y}_2 - \mathcal{B}^T(\mathbf{u}_1 - \mathbf{u}_2))$, and for any sequence $(\mathbf{z}_{1n} - \mathbf{z}_{2n})_n$ such that $(\mathbf{z}_{1n} - \mathbf{z}_{2n})_n \to \mathbf{z}_1 - \mathbf{z}_2$, we have that

$$\tilde{\kappa}_n(\mathbf{z}_{1n} - \mathbf{z}_{2n}) \underset{n\to\infty}{\to} \tilde{\kappa}(\mathbf{z}_1 - \mathbf{z}_2).$$

Since κ is almost everywhere continuous, this convergence holds for almost all $\mathbf{z}_1 - \mathbf{z}_2$. Let H_n be the empirical distribution of $Z_1 - Z_2, \ldots, Z_{n-1} - Z_n$ we know that $H_n(\mathbf{z}_1 - \mathbf{z}_2) \xrightarrow{a.s.} H(\mathbf{z}_1 - \mathbf{z}_2)$ by using Theorem 4.1.1 of Révész (1968) which states the law of the large numbers for strong stationary sequences. Hence we can apply Theorem 5.5 of Billingsley (1968). Define $\gamma : \mathbb{R} \to \mathbb{R}$ by $\gamma(y) = y$ if $\inf \kappa \leq y \leq \sup \kappa$, $\gamma(y) = \sup \kappa$ if $y \geq \sup \kappa$ and $\gamma(y) = \inf \kappa$ if $y \leq \inf \kappa$. We then obtain from the theorem that

$$\int \gamma(\tilde{\kappa}_n(\mathbf{z}_1 - \mathbf{z}_2)) dH_n \to \int \gamma(\tilde{\kappa}(\mathbf{z}_1 - \mathbf{z}_2)) dH$$

since γ is bounded and uniformly continuous.

Proof of Theorem 5. We mostly follow the lines of Salibian-Barrera and Zamar (2002) and Salibian-Barrera et al. (2006).

We can write the estimating equations as follows:

$$\widehat{\mathcal{B}}_n = \mathbf{A}_n(\widehat{\mathcal{B}}_n, \widehat{\Sigma}_n)^{-1}\mathbf{B}_n(\widehat{\mathcal{B}}_n, \widehat{\Sigma}_n)$$
$$\widehat{\Sigma}_n = \mathbf{V}_n(\widehat{\mathcal{B}}_n, \widehat{\Sigma}_n) + w_n(\widehat{\mathcal{B}}_n, \widehat{\Sigma}_n)\widehat{\Sigma}_n$$

with properly defined functions $\mathbf{A}_n$, $\mathbf{B}_n$, $\mathbf{V}_n$ and w_n.

Consider the function $\mathbf{f} : \mathbb{R}^{pq+q^2} \to \mathbb{R}^{pq+q^2}$, for $B \in \mathbb{R}^{p\times q}$ and $C \in \mathbb{R}^{q\times q}$:

$$\mathbf{f}\begin{pmatrix} \mathrm{vec}(B) \\ \\ \mathrm{vec}(C) \end{pmatrix} := \begin{pmatrix} \mathrm{vec}(\mathbf{A}_n(B,C)^{-1}\mathbf{B}_n(B,C)) \\ \\ \mathrm{vec}(\mathbf{V}_n(B,C) + w_n(B,C)C) \end{pmatrix}$$

Let $\widehat{\Theta}_n := (\mathrm{vec}(\widehat{\mathcal{B}_n})^T\mathrm{vec}(\widehat{\Sigma}_n)^T)^T$. We have that $\mathbf{f}(\widehat{\Theta}_n) = \widehat{\Theta}_n$. Since ρ is sufficiently smooth, the function $\mathbf{f}$ allows a Taylor expansion around $\Theta := (\mathrm{vec}(\mathcal{B})^T\mathrm{vec}(\Sigma)^T)^T$:

$$\widehat{\Theta}_n = \mathbf{f}(\Theta) + \nabla\mathbf{f}(\Theta)(\widehat{\Theta}_n - \Theta) + \frac{1}{2}(I \otimes (\widehat{\Theta}_n - \Theta)^T)\mathbf{Hf}(\widetilde{\Theta}_n)(\widehat{\Theta}_n - \Theta) \qquad \text{(A.12)}$$

Here $\nabla\mathbf{f}(.) \in \mathbb{R}^{(pq+qq)\times(pq+qq)}$ is the Jacobian and $\mathbf{Hf}(.) \in \mathbb{R}^{(pq+q^2)^2\times(pq+q^2)}$ is the Hessian matrix of $\mathbf{f}$. The value of $\widetilde{\Theta}_n$ in the remainder term lies between $\widehat{\Theta}_n$ and Θ. The Hessian is obtained by taking the partial derivatives of the entries of the Jacobian, the matrix of the partial derivatives of $\mathbf{f}$. Straightforward calculations then yield that each entry in the Hessian is a combination of products of means. Taking into account that the derivative of ρ vanishes outside some interval, the assumptions on ρ ensure the existence of the population analogues of the means. Furthermore, Lemma 2 then guarantees that $\|\mathbf{Hf}(\widetilde{\Theta}_n)\| = O_p(1)$.

From the consistency of the estimators we have that $\|\widehat{\Theta}_n - \Theta\| = O_p(n^{-1/2})$. It follows that the remainder term is $o_p(n^{-1/2})$.

We can now rewrite (A.12) as follows:

$$\sqrt{n}(\widehat{\Theta}_n - \Theta) = [I - \nabla \mathbf{f}(\Theta)]^{-1}\sqrt{n}(\mathbf{f}(\Theta) - \Theta) + o_p(1).$$

It needs to be shown that the bootstrap distribution of the right-hand side of this equation converges to the asymptotic distribution of $\sqrt{n}(\widehat{\Theta}_n - \Theta)$. For any X_n, Y_n, by $X_n \dot{\sim} Y_n$, we denote that X_n and Y_n have the same limiting distribution. We have

$$\sqrt{n}(\widehat{\Theta}_n - \Theta) \dot{\sim} [I - \nabla \mathbf{f}(\Theta)]^{-1}\sqrt{n}(\mathbf{f}(\Theta) - \Theta). \tag{A.13}$$

Define the function $\mathbf{g} : \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times q} \times \mathbb{R}^{q\times q} \times \mathbb{R}^{q\times q} \to \mathbb{R}^{pq+q^2}$ by

$$\mathbf{g}(A, B, V, W) = (\text{vec}(A^{-1}B)^T, \text{vec}(V + W)^T)^T.$$

Denote $T = (\text{vec}(B)^T\text{vec}(C)^T)^T$ and

$$\overline{Y}_n(T) := (\mathbf{A}_n(.), \mathbf{B}_n(.), \mathbf{V}_n(.), \mathbf{W}_n(.))$$

where $\mathbf{W}_n(.) := w_n(.)C$. Note that the components are actually means. Furthermore, denote by $\mu_{Y(T)}$ the limiting values of these means. We then have $\mathbf{g}(\overline{Y}_n(T)) = \mathbf{f}(T)$ for any T and also $\mathbf{g}(\mu_{Y(\Theta)}) = \Theta$.

From here we can follow the same reasoning as in Salibian-Barrera et al. (2006) which leads to

$$\sqrt{n}(\widehat{\Theta}_n - \Theta) \dot{\sim} [I - \nabla \mathbf{f}(\widehat{\Theta}_n)]^{-1}\sqrt{n}(\mathbf{f}^*(\widehat{\Theta}_n) - \widehat{\Theta}_n))$$

The right-hand side is actually $(\text{vec}(\sqrt{n}(\widehat{\mathcal{B}}^*_n - \widehat{\mathcal{B}}_n))^T \quad \text{vec}(\sqrt{n}(\widehat{\Sigma}^*_n - \widehat{\Sigma}_n))^T)^T$, and the proof is complete. □

References

[1] J. Agulló, C. Croux, S. Van Aelst, The multivariate least-trimmed squares estimator, Journal of Multivariate Analysis 99 (2008) 311 – 338.

[2] M.G. Ben, E. Martinez, V.J. Yohai, Robust estimation for the multivariate linear model based on a τ-scale, Journal of Multivariate Analysis 97 (2006) 1600 – 1622.

[3] J.R. Berrendero, J. Romo, Stability under contamination of robust regression estimators based on differences of residuals, Journal of Statistical Planning and Inference 70 (1998) 149 – 165.

[4] J.R. Berrendero, On the global robustness of generalized S-estimators, Journal of Statistical Planning and Inference 102 (2002) 287 – 302.

[5] P. Billingsley, Convergence of Probability Measures, John Wiley and Sons, New York, 1968.

[6] A. Charnes, W.W. Cooper, E. Rhodes, Evaluating program and managerial efficiency: an application of data envelopment analysis to program follow through, Management Science 27 (1981) 668 – 697.

[7] C. Croux, P.J. Rousseeuw, O. Hössjer, Generalized S-estimators, Journal of the American Statistical Association 89 (1994) 1271 – 1281.

[8] P.L. Davies, Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices, The Annals of Statistics 15 (1987) 1269 – 1292.

[9] A.C. Davison, D.V. Hinkley, Bootstrap Methods and their Application, Cambridge Univ. Press, 1997.

[10] D.L. Donoho, P.J. Huber, The notion of breakdown point, in: P. Bickel, K. Doksum and J.L. Hodges (Eds.), A Festschrift for Erich Lehmann, Wadsworth, Belmont, CA, 1983, pp. 157 – 184.

[11] L. Dümbgen, On Tyler's M-functional of scatter in high dimension, Annals of the Institute of Statistical Mathematics 50 (1998) 471 – 491.

[12] L. Dümbgen, D.E. Tyler, On the breakdown properties of some multivariate M-functionals, Scandinavian Journal of Statistics 32 (2005) 247 – 264.

[13] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel, Robust Statistics: The Approach Based on Influence Functions, Wiley, New York, 1986.

[14] O. Hössjer, C. Croux, P.J. Rousseeuw, Asymptotics of Generalized S-estimators, Journal of Multivariate Analysis 51 (1994) 148 – 177.

[15] H. Hult, F. Lindskog, Multivariate extremes, aggregation and dependence in elliptical distributions, Advances in Applied probability 34 (2002) 587 – 608.

[16] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.

[17] H.P. Lopuhaä, On the relation between S-estimators and M-estimators of multivariate location and covariance, The Annals of Statistics 17 (1989) 1662 – 1683.

[18] H.P. Lopuhaä, Highly efficient estimators of multivariate location with high breakdown point, The Annals of Statistics 20 (1992) 398 – 413.

[19] H.P. Lopuhaä, Asymptotics of reweighted estimators of multivariate location and scatter, The Annals of Statistics 27 (1999) 1638 – 1665.

[20] B. Mendes, D.E. Tyler, Constrained M-estimation for regression, in: H. Rieder (Ed.), Robust Statistics, Data Analysis, and Computer Intensive Methods,

Lecture Notes in Statistics 109, New York: Springer-Verlang, 1996, pp. 299-320.

[21] H. Oja, S. Sirkiä, J. Eriksson, Scatter matrices and independent component analysis, Austrian Journal of Statistics 35 (2006) 175–189.

[22] E. Ollila, H. Oja, T.P. Hettmansperger, Estimates of regression coefficients based on the sign covariance matrix, Journal of the Royal Statistical Society Series B-Statistical Methodology 64 (2002) 447 – 466.

[23] E. Ollila, H. Oja, V. Koivunen, Estimates of regression coefficients based on lift rank covariance matrix. Journal of the American Statistical Association 98 (2003) 90 – 98.

[24] P. Révész, The Laws of Large Numbers, Academic Press, New York and London, 1968.

[25] P.J. Rousseeuw, Least median of squares regression, Journal of the American Statistical Association 79 (1984) 871 – 880.

[26] P.J. Rousseeuw, S. Van Aelst, K. Van Driessen, J. Agulló, Robust multivariate regression, Technometrics 46 (2004) 293 – 305.

[27] P.J. Rousseeuw, K. Van Driessen, A fast algorithm for the minimum covariance determinant estimator, Technometrics 41 (1999) 212 – 223.

[28] P.J. Rousseeuw, V.J. Yohai, Robust regression by means of S-estimators, in: J. Franke, W. Härdle, R.D. Martin (Eds.), Robust and Nonlinear Time Series Analysis. Lecture Notes in Statistics 26, New York: Springer-Verlag, 1984, pp. 256 – 272.

[29] M. Salibian-Barrera, S. Van Aelst, G. Willems, Principal components analysis based on multivariate MM-estimators with fast and robust bootstrap, Journal of the American Statistical Association 101 (2006) 1198 – 1211.

[30] M. Salibian-Barrera, S. Van Aelst, G. Willems, Fast and robust bootstrap. Statistical Methods and Applications 17 (2008) 41 – 71.

[31] M. Salibian-Barrera, V.J. Yohai, A fast algorithm for S-regression estimates, Journal of Computational and Graphical Statistics 15 (2006) 414 – 427.

[32] M. Salibian-Barrera, R.H. Zamar, Bootstrapping robust estimates of regression, The Annals of Statistics 30 (2002) 556 – 582.

[33] G.A.F. Seber, Multivariate Observations, Wiley, New York, 1984.

[34] K. Singh, Breakdown theory for bootstrap quantiles, The Annals of Statistics 26 (1998) 1719 – 1732.

[35] S. Sirkiä, S. Taskinen, H. Oja, Symmetrised M-estimators of multivariate scatter, Journal of Multivariate Analysis 98 (2007) 1611 – 1629.

[36] K.S. Tatsuoka, D.E. Tyler, On the uniqueness of S-functionals and M-functionals under nonelliptical distributions, The Annals of Statistics 28 (2000) 1219 – 1243.

[37] D.E. Tyler, F. Critchley, L. Dümbgen, H. Oja, Invariant coordinate selection, (2007) Manuscript.

[38] S. Van Aelst, G. Willems, Multivariate regression S-estimators for robust estimation and inference, Statistica Sinica 15 (2005) 981 – 1001.

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

[40] V.J. Yohai, R.H. Zamar, High breakdown-point estimates of regression by means of the minimization of an efficient scale, Journal of the American Statistical Association 83 (1988) 406 – 413.