Journal of Multivariate Analysis 100 (2009) 876–887

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Multivariate generalized S-estimators

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

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

[b] Katholieke Universiteit Leuven, Faculty of Economics and Business and Leuven Statistical Research Centre, Naamsestraat 69, B-3000 Leuven, Belgium

ARTICLE INFO

Article history:
Received 14 February 2008
Available online 7 September 2008

AMS subject classifications:
62F40
62F35
62J05

Keywords:
Bootstrap
Efficiency
Multivariate regression
Robustness

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 independence 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 a real data example.

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 the advantage that 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 that it is possible to achieve the highest possible value for the breakdown point, 50%, even when working with differences of residuals. GS-estimators estimate the slope and the scatter matrix of the error terms of the multivariate regression model without needing to estimate the intercept first. The intercept is treated as a nuisance parameter, and the resulting estimator is therefore considered to be "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 [1], 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. More important, 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 [2]. 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 [3]. Oja et al. [4] 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

* Corresponding author.
E-mail address: Ella.Roelant@ugent.be (E. Roelant).

0047-259X/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2008.09.002

also Tyler et al. [2]. 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 [5,6]. They are of the M-type and their breakdown point decreases with the dimension [7], and thus do not have a high degree of robustness. The GS-estimators are estimators that can have a 50% breakdown point and at the same time the independence property.

Consider the multivariate linear regression model given by

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

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 $\theta = (\alpha, \mathcal{B}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{(p+1)\times q}$ and $\Sigma \in \mathbb{R}^{q\times q}$ are to be estimated from the observations $\mathcal{Z}_n = \{\mathbf{z}_i := (\mathbf{x}_i^{\mathrm{T}}, \mathbf{y}_i^{\mathrm{T}})^{\mathrm{T}} = (1, \mathbf{u}_i^{\mathrm{T}}, \mathbf{y}_i^{\mathrm{T}})^{\mathrm{T}}, i = 1, \ldots, n\} \subset \mathbb{R}^{p+q+1}$. 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 [8], least median of squares and least trimmed squares estimators [9], S-estimators [10], MM-estimators [11], CM-estimators [12] and τ-estimators [13]. Croux et al. [1] introduced a class of regression estimators, called generalized S-estimators or GS-estimators. While an S-estimator of regression minimizes an M-estimator of scale of the residuals, a GS-estimator minimizes an M-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 and a higher efficiency than the original S-estimators. Moreover, they do not require the assumption of symmetric errors (see also [14–16]). In the univariate regression model we also mention the rank-based regression estimates of [17], which are also based on the differences of the residuals. 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 et al. [18] using sign covariance matrices, Ollila et al. [19] using rank covariance matrices and [20] 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. [21] proposed the multivariate least trimmed squares estimator, Van Aelst and Willems [22] considered multivariate regression S-estimators, while [23] introduced τ-estimators for multivariate regression. Also note that the idea of univariate least absolute deviation estimation has been extended to the multivariate case by [24–26]. All these procedures, however, are not based on differences of residuals.

GS-estimators generalize S-estimators in the sense that they are identical to S-estimators, but computed from differences $\mathbf{z}_i - \mathbf{z}_j$, instead of using the original observations $\mathbf{z}_i$. This is the same idea as used in the definition of generalized L-, M-, and R-estimators [27] for the location model. It should be stressed that S-estimators are not included in the class of GS-estimators. Another suitable name for GS-estimators may be "Symmetrized S-estimators".

The remainder of the paper is organized as follows. In Section 2 we introduce 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 a real data example and Section 8 concludes. Proofs are omitted and can be found in the technical report [28].

2. Definition and breakdown point

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

Definition 1. Let $\mathcal{Z}_n = \{\mathbf{z}_i := (\mathbf{x}_i^{\mathrm{T}}, \mathbf{y}_i^{\mathrm{T}})^{\mathrm{T}} = (1, \mathbf{u}_i^{\mathrm{T}}, \mathbf{y}_i^{\mathrm{T}})^{\mathrm{T}}, i = 1, \ldots, n\} \subset \mathbb{R}^{p+q+1}$. 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

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

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

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 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:

$$\rho_c(t) = \begin{cases} \dfrac{t^2}{2} - \dfrac{t^4}{2c^2} + \dfrac{t^6}{6c^4}, & |t| \le c \\ \dfrac{c^2}{6}, & |t| \ge c. \end{cases}$$

Similarly as in [29], it can be shown that Definition 1 implies that multivariate GS-estimators satisfy the following first-order conditions:

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

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

with $d_{ij}^2 = (\mathbf{y}_i - \mathbf{y}_j - B^{\mathrm{T}}(\mathbf{u}_i - \mathbf{u}_j))^{\mathrm{T}} C^{-1} (\mathbf{y}_i - \mathbf{y}_j - B^{\mathrm{T}}(\mathbf{u}_i - \mathbf{u}_j))$, $u(t) = \psi(t)/t$ 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 [30]. Formally,

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

where the supremum is over all possible collections $\mathcal{Z}'_{n,m}$ that differ from $\mathcal{Z}_n$ in at most m data points and D is a metric on the parameter space. For the regression estimators one has the usual distance $D(\mathcal{B}_1, \mathcal{B}_2) = \|\mathcal{B}_1 - \mathcal{B}_2\|$. For covariance estimators, the breakdown point is the smallest fraction of outliers that can make the largest eigenvalue tend to infinity or the smallest eigenvalue tend to zero. This motivates the choice of the metric $D(\Sigma_1, \Sigma_2) = \max\{|\lambda_1(\Sigma_1) - \lambda_1(\Sigma_2)|, |\lambda_q(\Sigma_1)^{-1} - \lambda_q(\Sigma_2)^{-1}|\}$ where $\lambda_1(\Sigma) \ge \cdots \ge \lambda_q(\Sigma)$ are the ordered eigenvalues of a matrix Σ. We derive the breakdown point for data sets that satisfy the following general position condition.

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

Note that if the differences of the $(\mathbf{u}_i^{\mathrm{T}}, \mathbf{y}_i^{\mathrm{T}})^{\mathrm{T}}$ are in general position, then the points $(\mathbf{u}_i^{\mathrm{T}}, \mathbf{y}_i^{\mathrm{T}})^{\mathrm{T}}$ themselves are also in general position. The latter means that no $p + q + 1$ of the $(\mathbf{u}_i^{\mathrm{T}}, \mathbf{y}_i^{\mathrm{T}})^{\mathrm{T}}$ 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 [1].

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* $\binom{n}{2}(1-r) \ge \binom{p+q+1}{2}$ *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 analogous to the fast S-algorithm of [31] 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 $\|\mathbf{r}_i - \mathbf{r}_j\|_C = ((\mathbf{r}_i - \mathbf{r}_j)^{\mathrm{T}} C^{-1}(\mathbf{r}_i - \mathbf{r}_j))^{1/2}$ and $\tilde{n} = n(n-1)/2$. The actual algorithm can be described as follows:

1. Draw N random sub-samples of size $p + q + 1$. (We advise $N = 1000$.) For each sub-sample calculate the least squares estimate $\widehat{\mathcal{B}}_m^0$, $m = 1, \ldots, N$, and the corresponding shape matrix $\widehat{\Gamma}_m^0$ of the residuals, i.e. the covariance matrix $\widehat{\Sigma}_m^0$ of the residuals is rescaled to have determinant equal to 1. Denote the residuals by $\mathbf{r}_i(\widehat{\mathcal{B}}_m^0)$, for $i = 1, \ldots, n$.
2. For each sub-sample, apply κ I-steps (we take $\kappa = 2$) as follows. Set $v = 1$.
   a. Calculate an approximate solution of Eq. (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 $\widehat{\Sigma}_m^v = \sum_{i<j} w_{ij}(\mathbf{r}_i(\widehat{\mathcal{B}}_m^{v-1}) - \mathbf{r}_j(\widehat{\mathcal{B}}_m^{v-1}))(\mathbf{r}_i(\widehat{\mathcal{B}}_m^{v-1}) - \mathbf{r}_j(\widehat{\mathcal{B}}_m^{v-1}))^{\mathrm{T}}$ 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 (we take $\tau = 5$) 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 $\widehat{\Sigma}_m^F = (s_m^F)^2 \widehat{\Gamma}_m^F$. In our experience, convergence is obtained rather quickly for most data configurations.

The recommended choices of N, κ and τ are in agreement with those of the fast LTS-algorithm of [32] and the fast S-algorithm of [31]. In our experience they also give satisfactory results in the multivariate regression setting. Note that there is no guarantee that the algorithm converges to a global minimum but in our experience it gives a good approximation to the GS-estimator.

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

$$\int\int \rho([(\mathbf{y}_1 - \mathbf{y}_2 - B^{\mathrm{T}}(\mathbf{u}_1 - \mathbf{u}_2))^{\mathrm{T}} C^{-1}(\mathbf{y}_1 - \mathbf{y}_2 - B^{\mathrm{T}}(\mathbf{u}_1 - \mathbf{u}_2))]^{1/2}) \mathrm{d}H(\mathbf{z}_1)\mathrm{d}H(\mathbf{z}_2) = k$$

among all $(B, C) \in \mathbb{R}^{p\times q} \times PDS(q)$ and where $\mathbf{z}_l = (\mathbf{u}_l^{\mathrm{T}}, \mathbf{y}_l^{\mathrm{T}})^{\mathrm{T}}$ 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 $\mathbf{z} = (\mathbf{u}^{\mathrm{T}}, \mathbf{y}^{\mathrm{T}})^{\mathrm{T}}$ 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 the distribution of the difference of two independent and elliptically symmetric error terms, remains elliptically symmetric [33]. We need another regularity condition on the model distribution H.

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:*

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

The influence function of a functional T at a distribution H measures the effect on T of an infinitesimal contamination at a single point [8]. 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 $\mathcal{B} = \mathbf{0}$, 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 satisfying 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*

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

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

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 [6]. If $q = 1$, then the influence function of $\mathcal{B}_{GS}$ is identical to the influence function of the univariate GS-estimator (see [1]). 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

Under regularity conditions, 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 [8]), where $\otimes$ stands for the Kronecker product. We do not give a proof of asymptotic normality in this paper, but we conjecture that the GS-estimator is asymptotic normal under the same conditions as in [29,14] and that a formal proof will go along the same lines as in these papers. Denoting $\Sigma_\mathbf{u} := \mathrm{Cov}[\mathbf{u}]$, it follows from (6) that

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

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	$q=1$	$q=2$	$q=3$	$q=5$	$q=10$	$q=1$	$q=2$	$q=3$	$q=5$	$q=10$
$N_q(\mathbf{0}, I_q)$	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	$q=1$	$q=2$	$q=3$	$q=5$	$q=10$	$q=1$	$q=2$	$q=3$	$q=5$	$q=10$
$N_q(\mathbf{0}, I_q)$	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

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 [22] 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. Note that in some cases the S-estimator can be more efficient than the GS-estimator, in particular for larger values of q, even at the normal model. This is somehow surprising, and might be explained by the following heuristical reasoning. Working with differences of residuals may lead to a gain in efficiency because the number of differences is larger than the original sample size. On the other hand, the distribution of the differences has a larger spread, leading to less precise regression estimators. For a least squares approach, both effects perfectly counterbalance, and there is no gain in working with differences. When the value of q increases, S- and GS-estimators tend to resemble more and more the classical estimators if no outliers are present, as it witnessed by their efficiency getting closer to 100%. Then it is not clear which of the two effects of working with differences will dominate.

We also performed a simulation study to investigate the finite-sample efficiency of the GS-estimator. We generated $m = 1000$ 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 $(\alpha, \mathcal{B}^{\mathrm{T}})^{\mathrm{T}}$ 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 as $n\mathrm{ave}_{j,k}(\widehat{\mathrm{Var}}((\widehat{\mathcal{B}}_n)_{jk}))$ for $j = 1, \ldots, p$ and $k = 1, \ldots, q$, where $\widehat{\mathrm{Var}}((\widehat{\mathcal{B}}_n)_{jk})$ 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 GS-estimator are always higher.

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	$N_q(\mathbf{0}, I_q)$	$q = 2$	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	$N_q(\mathbf{0}, I_q)$	$q = 2$	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

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 [34] 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 [22,35,36]. 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 $\mathbf{g}_n(\widehat{\Theta}_n) = \widehat{\Theta}_n$. We assume that $\widehat{\Theta}_n$ is a root-n consistent estimator of Θ. Then, using the smoothness of $\mathbf{g}_n$, we can calculate a Taylor expansion about the limiting value of the estimate $\widehat{\Theta}_n$:

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

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

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

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

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

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 $\mathbf{g}_n^*$ in $\widehat{\Theta}_n$ and then apply a linear correction given by $[I - \nabla\mathbf{g}_n(\widehat{\Theta}_n)]^{-1}$.

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

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

where

$$\mathbf{A}_n(B, C) = \sum_{i<j} u(d_{ij})(\mathbf{u}_i - \mathbf{u}_j)(\mathbf{u}_i - \mathbf{u}_j)^{\mathrm{T}}$$

$$\mathbf{B}_n(B, C) = \sum_{i<j} u(d_{ij})(\mathbf{u}_i - \mathbf{u}_j)(\mathbf{y}_i - \mathbf{y}_j)^{\mathrm{T}}$$

$$\mathbf{V}_n(B, C) = \frac{1}{\binom{n}{2} k}\sum_{i<j} qu(d_{ij})(\mathbf{y}_i - \mathbf{y}_j - B^{\mathrm{T}}(\mathbf{u}_i - \mathbf{u}_j))(\mathbf{y}_i - \mathbf{y}_j - B^{\mathrm{T}}(\mathbf{u}_i - \mathbf{u}_j))^{\mathrm{T}}$$

$$w_n(B, C) = \frac{1}{\binom{n}{2} k}\sum_{i<j} w(d_{ij})$$

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^E_n	0.40	0.50	0.50	0.50	0.50	0.50	0.50	0.50
$Q^*_{0.005}$	ϵ^E_n	0.30	0.50	0.50	0.50	0.40	0.50	0.50	0.50

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

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

and for any matrices B and C, put

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

The expression for the matrix $\nabla \mathbf{g}_n(.)$ of partial derivatives can be found in the technical report [28].

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

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

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 $(\mathbf{u}_i^{\mathrm{T}}, \mathbf{y}_i^{\mathrm{T}})^{\mathrm{T}}$ replaced by $((\mathbf{u}_i^*)^{\mathrm{T}}, (\mathbf{y}_i^*)^{\mathrm{T}})^{\mathrm{T}}$. Thus, in order to get the values of $\sqrt{n}(\widehat{\Theta}_n^* - \widehat{\Theta}_n)$ for each bootstrap sample, we calculate $\mathbf{g}_n^*(\widehat{\Theta}_n)$, 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 $\{(\mathbf{u}_i^{\mathrm{T}}, \mathbf{y}_i^{\mathrm{T}})^{\mathrm{T}}, i = 1, \ldots, n\}$.

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^*:

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

where R is the number of bootstrap samples drawn. Singh [37] 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 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 GS-estimator. We need the following assumptions on ρ:

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

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

(A.2) $E_{G\times G}[\frac{\rho'(d)}{d}(\mathbf{u}_1 - \mathbf{u}_2)(\mathbf{u}_1 - \mathbf{u}_2)^{\mathrm{T}}]^{-1}$ 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 1000 data sets generated as described above and applied the FRB procedure with $B = 1000$ 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. [38]). 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

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

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 $\widehat{\Sigma}_n^{-1/2}(\mathbf{y}_i - \widehat{\mathcal{B}}_n^{\mathrm{T}}\mathbf{u}_i)$, and finally replacing $\Sigma_{\mathbf{u}}$ by the corresponding sample moment.

Fig. 1 shows the coverage for 95% confidence intervals computed by FRB and EASV. From Fig. 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. Besides the coverage we also considered the average length of the intervals and found that the length of the FRB interval is comparable to the length of the intervals based on the EASV (for details see the technical report [28]).

7. Example

School data

This example considers data of $n = 70$ school sites in the US [39]. 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 counseling 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 [40]. 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 [40].

The diagnostic plots in Fig. 2 show the Mahalanobis distances of the residuals versus the Mahalanobis distances of the explanatory variables (see also [20]). 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,

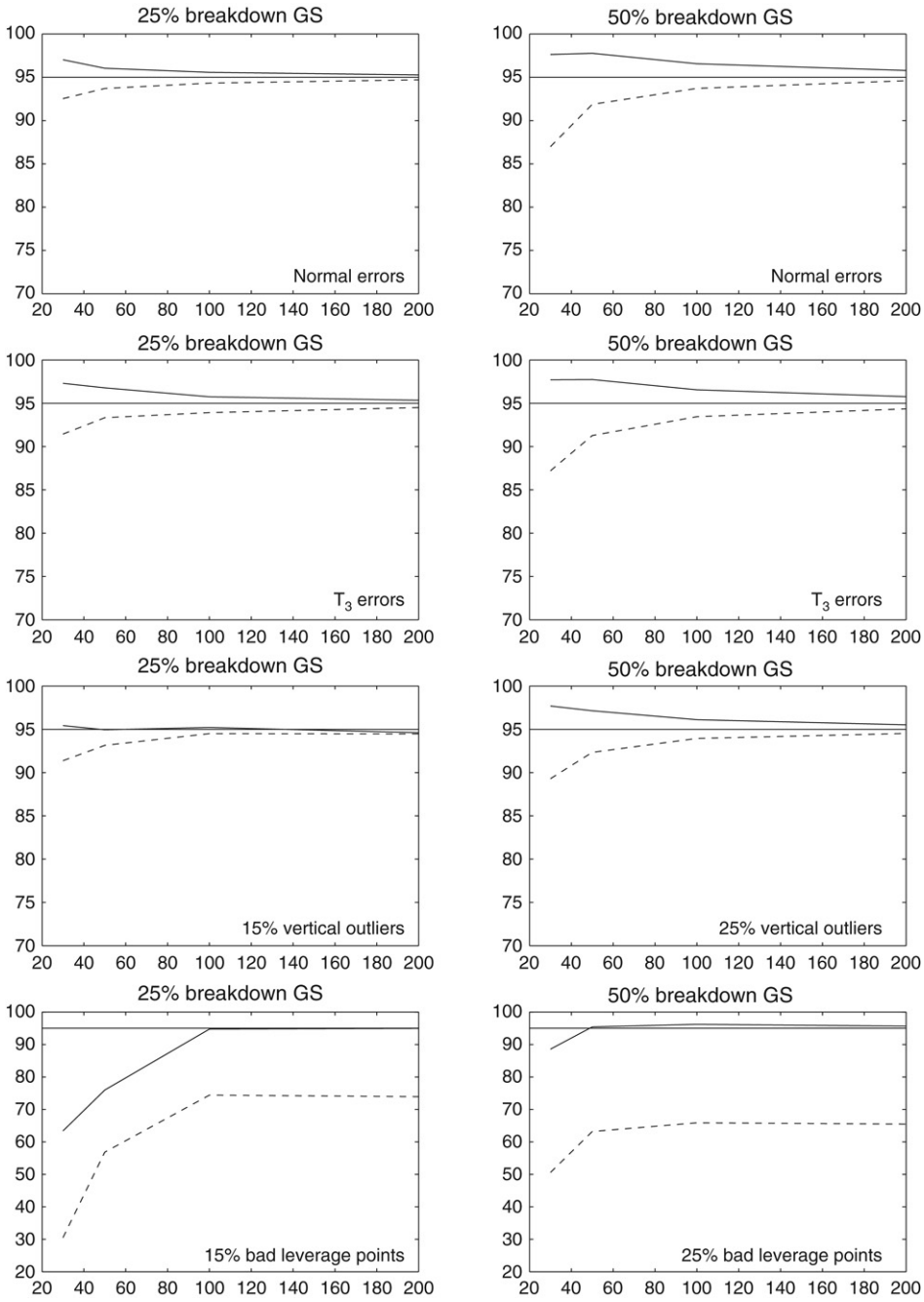

Fig. 1. Coverage for 95% confidence intervals, for FRB (–) and EASV (– –): $p = 4$; $q = 5$.

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.

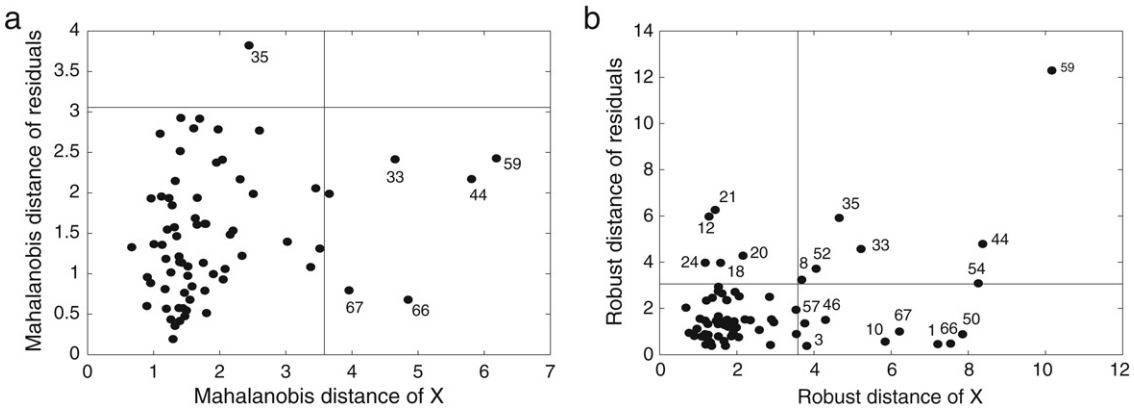

Fig. 2. Diagnostic plots for the school data; (a) Least squares estimator; (b) 50% breakdown GS-estimator.

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**

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 [40]. We illustrate this on an example of the Forbes data set in the technical report [28].

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 example 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 example, the intercept can easily be estimated afterwards by using the efficient and robust M-estimator of multivariate location of [40], using the residual covariance matrix of the GS-estimator as an initial estimator. In fact, similarly as for MM-estimators [11,41] 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 independence 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 [6]. While the GS-estimators have several advantages with respect to the regular S-estimator, in particular the independence property, they also have a shortcoming. Since the number of pairwise differences is of the order n^2, GS-estimators require much more computation time for large sample sizes. For smaller sample sizes, the computation time remains reasonable. For example, for the School data we analyzed in Section 7, computation of the GS-estimator requires 21.33 s (compared to 1.72 s for the S-estimator), which is still feasible. Constructing a faster algorithm for computing GS-estimators is a topic for further research.

Acknowledgments

We would like to thank the referees for their helpful comments and suggestions. The research of Stefan Van Aelst was supported by a grant of the Fund for Scientific Research-Flanders (FWO-Vlaanderen) and by IAP research network grant P6/03 of the Belgian government (Belgian Science Policy).

References

[1] C. Croux, P.J. Rousseeuw, O. Hössjer, Generalized S-estimators, Journal of the American Statistical Association 89 (1994) 1271–1281.
[2] D.E. Tyler, F. Critchley, L. Dümbgen, H. Oja, Invariant coordinate selection, 2007, Manuscript.
[3] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, John Wiley and Sons, New York, 2001.
[4] H. Oja, S. Sirkiä, J. Eriksson, Scatter matrices and independent component analysis, Austrian Journal of Statistics 35 (2006) 175–189.
[5] L. Dümbgen, On Tyler's M-functional of scatter in high dimension, Annals of the Institute of Statistical Mathematics 50 (1998) 471–491.
[6] S. Sirkiä, S. Taskinen, H. Oja, Symmetrised M-estimators of multivariate scatter, Journal of Multivariate Analysis 98 (2007) 1611–1629.
[7] L. Dümbgen, D.E. Tyler, On the breakdown properties of some multivariate M-functionals, Scandinavian Journal of Statistics 32 (2005) 247–264.
[8] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel, Robust Statistics: The Approach Based on Influence Functions, John Wiley and Sons, New York, 1986.
[9] P.J. Rousseeuw, Least median of squares regression, Journal of the American Statistical Association 79 (1984) 871–880.
[10] 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, in: Lecture Notes in Statistics, vol. 26, Springer-Verlag, New York, 1984, pp. 256–272.
[11] V.J. Yohai, High breakdown-point and high efficiency robust estimates for regression, The Annals of Statistics 15 (1987) 642–656.
[12] B. Mendes, D.E. Tyler, Constrained M-estimation for regression, in: H. Rieder (Ed.), Robust Statistics, Data Analysis, and Computer Intensive Methods, in: Lecture Notes in Statistics, vol. 109, Springer-Verlang, New York, 1996, pp. 299–320.
[13] 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.
[14] O. Hössjer, C. Croux, P.J. Rousseeuw, Asymptotics of generalized S-estimators, Journal of Multivariate Analysis 51 (1994) 148–177.
[15] 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.
[16] J.R. Berrendero, On the global robustness of generalized S-estimators, Journal of Statistical Planning and Inference 102 (2002) 287–302.
[17] J.T. Terpstra, J.W. McKean, Rank-based analyses of linear models using R, Journal of Statistical Software 14 (2005) 7.
[18] 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.
[19] 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.
[20] P.J. Rousseeuw, S. Van Aelst, K. Van Driessen, J. Agulló, Robust multivariate regression, Technometrics 46 (2004) 293–305.
[21] J. Agulló, C. Croux, S. Van Aelst, The multivariate least-trimmed squares estimator, Journal of Multivariate Analysis 99 (2008) 311–338.
[22] S. Van Aelst, G. Willems, Multivariate regression S-estimators for robust estimation and inference, Statistica Sinica 15 (2005) 981–1001.
[23] 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.
[24] Z.D. Bai, X.R. Chen, B.Q. Miao, C.R. Rao, Asymptotic theory of least distances estimate in multivariate linear models, Statistics 21 (1990) 503–519.
[25] B. Chakraborty, On multivariate median regression, Bernoulli 5 (1999) 683–703.
[26] B. Chakraborty, On multivariate quantile regression, Journal of Statistical Planning and Inference 110 (2003) 109–132.
[27] R.J. Serfling, Generalized L-, M-, and R-statistics, The Annals of Statistics 12 (1984) 76–86.
[28] E. Roelant, S. Van Aelst, C. Croux, Multivariate generalized S-estimators. Technical report, available at http://users.ugent.be/~svaelst/publications.html, 2008.
[29] 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.
[30] D.L. Donoho, P.J. Huber, The notion of breakdown point, in: P. Bickel, K. Doksum, J.L. Hodges (Eds.), A Festschrift for Erich Lehmann, Wadsworth, Belmont, CA, 1983, pp. 157–184.
[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] P.J. Rousseeuw, K. Van Driessen, Computing LTS regression for large data sets, Data Mining and Knowledge Discovery 12 (2006) 29–45.
[33] H. Hult, F. Lindskog, Multivariate extremes, aggregation and dependence in elliptical distributions, Advances in Applied probability 34 (2002) 587–608.
[34] M. Salibian-Barrera, R.H. Zamar, Bootstrapping robust estimates of regression, The Annals of Statistics 30 (2002) 556–582.
[35] 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.
[36] M. Salibian-Barrera, S. Van Aelst, G. Willems, Fast and robust bootstrap, Statistical Methods and Applications 17 (2008) 41–71.
[37] K. Singh, Breakdown theory for bootstrap quantiles, The Annals of Statistics 26 (1998) 1719–1732.
[38] A.C. Davison, D.V. Hinkley, Bootstrap Methods and their Application, Cambridge Univ. Press, 1997.
[39] 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.
[40] H.P. Lopuhaä, Highly efficient estimators of multivariate location with high breakdown point, The Annals of Statistics 20 (1992) 398–413.
[41] 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.