Covariance matrices of S robust regression estimators

S. Salini[a] and F. Laurini[b] and G. Morelli[b] and M. Riani[b] and A. Cerioli[b]

[a]University of Milan, Department of Economics, Management and Quantitative Methods, Via Conservatorio 7, 20122 Milano, Italy;
[b]University of Parma, Department of Economics and Management, Via J.F. Kennedy 6, 43125 Parma, Italy

ARTICLE HISTORY
Compiled September 8, 2021

ABSTRACT
Asymptotic properties of robust regression estimators are well known. However, it is not always clear what is the best strategy for confidence intervals and hypothesis testing when the sample size is not very large, since the distribution of residuals coming from robust estimates has unknown properties for small samples. In the present work we propose an analysis of various strategies for estimating the variance-covariance matrix of the S estimators at the variation of n and p, considering different ρ functions. An adaptive correction strategy is proposed. In addition to the simulation study, an example on a benchmark dataset is shown.

KEYWORDS
S-estimator; Wald-type inference; Robust Regression

1. Introduction

The residuals from linear regression, using ordinary least squares (OLS), are correlated and heteroschedastic. Given that OLS are sensitive to outliers and leverage points, robust alternatives are often used for linear regression. There are several robust estimators in the regression setting [18] and considerable efforts have been recently made in order to extend their range of applicability; see, e.g., [30] and [1] for the development of robust tests and robust generalized linear models, [12] and [2] for extensions to high-dimensional data, [7] and [13] for applications in the field of robust regression clustering, [23] and [6] for monitoring robust regression.

However, it is unknown if and how residuals are correlated and distributed for almost all robust regression estimators. [15, Section 7.6, p. 170] give some results on the covariance matrix of the estimated coefficients, but some are non-conclusive and rely on restrictions made on the size of the regression problem. Additionally, those results sometimes assume weak leverage points in the carriers and are limited to the analysis of the bias of the covariance matrix of parameter estimates. Many conjectures are obviously related to the actual lack of knowledge, for the robust regression model, of a general result of the finite sample distribution of the estimators and, consequently, lack of knowledge of the distribution of robust residuals, and their correlation.

CONTACT Silvia Salini. Email: silvia.salini@unimi.it

To perform Wald-type statistical inference [16] make several proposals and introduce corrections based on formulae of [15, Section 7.6, p. 171]. These results are based on the assumptions of balanced cases (i.e. no leverage points), symmetric error distributions and a small ratio between number of variables and sample size. Sometimes practitioners use low breakdown estimators and perform Wald-type inference, but other types of inference might be needed, and estimators with high breakdown are more often considered, for which results from the literature above give only partial support, see for example [11].

For general type of inference, the covariance matrix of the estimators is a key issue, as robust t-test might have very different distributions, when the sample size is not very large. Additionally, a further drawback is that many robust methods rely on the choice of a specific ρ function and tuning constants, making the final decision somehow arbitrary.

The main target of this work is to study the properties of covariance estimators of regression coefficients for several robust high-breakdown estimators, derived from M estimation, detailed below. Using results form [31], and without loss of generality, we investigate properties of S estimators for several ρ functions. High-breakdown regression is needed especially when the carriers are random variables that may contain leverage points, because of contamination or recording errors. Therefore, we fit regression models by conditioning on the observed vectors of the carriers, although, for simplicity, we do not make this step explicit in our notation of Section 2 below. We do not consider the case of fixed designs, where the values of the explanatory variables are known and error-free, for which alternatives to high-breakdown regression may be indeed preferable [15, Section 7.9]. The basic research question that we want to answer can thus be summarized as follows: "What properties are to be expected from alternative proposals of standard errors (and related t-tests) for some high-breakdown regression estimators that are popular in robust applied statistics, given that no unambiguous rules are available for the choice among these proposals?".

In this paper, we initially consider, in Section 3, uncontaminated data in the dependent variable to assess the efficiency of S estimators with alternative ρ functions. The simulation study is built using different sample sizes n and different numbers of parameters p. At later stage, in Section 5, we include the possibility of leverage points and different levels of contamination. This is, perhaps, crucial, as it mimics the presence of outliers in the design matrix. We report, in the paper, only one simulation scheme, but found no difference whit a different set of contamination. Additionally, in the supplementary material we have included results for heavy tailed distribution (still symmetric) and discussed thoroughly the case of Uniform error term for consistency of findings with our theoretical results.

In Section 6 an interesting example is provided by a simulated dataset with $n = 50$ and $p = 2$. The scenario without contaminated is presented first, Section 6.1, in order to show clearly a situation where inferential conclusions depend on the robust method adopted. In Section 6.2, different scenarios of data contamination are considered: point mass contamination, shift contamination only, contamination with leverage points and combination of the above schemes. The purpose of this section is to show that irrespective to the presence of different amount of outliers and/or leverage points, a correct estimation of the covariance matrix of robust regression parameters significantly increases the prediction ability of the regression model.

We leave for future research the assessment of the additional variability implied by random generation of the explanatory variables. There are several motivations for this choice. First, we want to immunize our results from the consequences of different

distributional assumptions on the carriers, in the same way as we do for the effect of σ^2 in the analysis of the variance covariance matrices. Second, the standard errors that we consider in our work were proposed without taking into account the effect of randomness in the explanatory variables and we believe that including it would make our comparisons less fair. Furthermore, [18, p. 124] show that the efficiency of regression M-type estimators does not depend on the distribution of the explanatory variables provided that they have finite variances. From this it can also be inferred that asymptotic inference on the regression parameters is the same as in the case of fixed carriers [18, p. 139], which justifies conditioning on their observed values.

The rest of the paper is structured as follows. Section 2 introduces the notation and the methodological framework. Section 3 reports the results of our simulations and provides additional insights for some special cases. Section 4 proposes a new adaptive correction. Section 5 addresses the case where leverage points are present in the design matrix and analyzes the stability of the results to different levels of contamination. In Section 6 we apply the suggested procedure to a simulated dataset and to a benchmark dataset which has been widely used in the literature. Section 7 summarizes the results and provides food for thought for further extensions.

2. Covariance matrix of robust regression coefficients

2.1. Preliminaries and notation

Here we introduce some specific notation used in the remainder of the paper. This will also link to the software implementation and to the numerical results.

Let n denote the sample size and $y = (y_1, \ldots, y_n)'$ be the vector of observations for the response variable Y. We take as our null model for uncontaminated data the classical regression relationship

$$y_i = x_i'\beta + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$ is a p-dimensional vector of unknown coefficients, $x_i = (1, x_{i1}, \ldots, x_{i(p-1)})'$, x_{ij} is the value of explanatory variable X_j, $j = 1, \ldots, p-1$, for unit i, $i = 1, \ldots, n$, and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. zero mean symmetric errors such that $\mathrm{Var}(\epsilon_i) = \sigma^2$ is finite. After estimating the parameters $\hat{\beta}$, it is possible to obtain the fitted values of the model $\hat{y} = X\hat{\beta}$ and the associated residuals, denoted by $r_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$.

We focus on properties of the fitted model (1) where the parameters are estimated with a robust regression method. More specifically, our target is to explore the finite sample properties of the covariance of robust estimators, used to build inference.

For concreteness, we take the class of regression S-estimators as our reference choice of a robust high-breakdown procedure based on a "soft trimming" approach.

This class derives from M-estimation and leads to downweighting of observations by a function ρ. For a given set of (unscaled) residuals $r_1, \ldots, r_n$ the M-estimator of scale $\hat{\sigma}$ is defined as the solution to the equation

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i}{\hat{\sigma}}\right) = \kappa, \tag{2}$$

where $0 < \kappa < \sup \rho$ and ρ is a user-supplied function that reduces the importance of observations with large residuals. In what follows we consider a number of popular ρ functions, which allow the user to achieve robustness through a high value of the breakdown point of the resulting estimator. We refer to [23] and to [31] for a detailed description of such ρ functions, all of which are currently implemented in the FSDA Toolbox of Matlab [25]. The vector $\hat{\beta}$ such that

$$\hat{\beta} = \arg\min \hat{\sigma}, \tag{3}$$

where $\hat{\sigma}$ satisfies (2), leads to the S-estimate of β and to the associated estimate of scale [29, pp. 135–143].

To achieve robustness, the ρ function must have the following properties:

(1) it is symmetric and continuously differentiable, and $\rho(0) = 0$;
(2) there exists a $c > 0$ such that ρ is strictly increasing on $[0, c]$ and constant on $[c, \infty)$;
(3) it is such that $\kappa/\rho(c) = 0.5$.

To obtain consistency of the scale estimate under the normal model (1), it is also required that

$$\kappa = E_{\Phi_{0,1}}(\rho),$$

where the expectation is taken over the standard normal distribution, whose distribution function is denoted by $\Phi_{0,1}$.

Under these conditions, if collinearities in the explanatory variables are excluded, [29] prove that the finite-sample breakdown point of the S-estimator is

$$\varepsilon = \frac{\lfloor n/2 \rfloor - p + 2}{n}, \tag{4}$$

where $\lfloor\ \rfloor$ denotes the floor function. Given a sample of size n, the finite-sample breakdown point of an estimator is defined to be the smallest fraction of contamination in the sample that can take the estimator over all bounds [29, p. 10]. It then follows from (4) that the breakdown point of the S-estimator tends to 0.5 when $n \to \infty$ and achieves a finite-sample value still very close to 0.5.

Let `bdp` denote the asymptotic value of the breakdown point of the S-estimator. If we take $0 < \texttt{bdp} \leq 0.5$, then it is possible to show that the third property of the ρ function stated above can be replaced by the more general assumption

$$\frac{\kappa}{\rho(c)} = \texttt{bdp}, \tag{5}$$

which leads to S-estimators with lower breakdown point (if $\texttt{bdp} < 0.5$) but higher statistical efficiency; see [29, p. 139]. Note that if $\rho(c)$ is normalized in such a way that $\sup \rho(c) = 1$, κ becomes exactly equal to the (asymptotic) breakdown point of the S-estimator. The relationships implied by (5) among c, `bdp` and the resulting statistical efficiency of S-estimators are studied in detail by [24]. Recent research works addressing breakdown and efficiency of "hard trimming" robust regression methods, such as least trimmed squares and its extensions, even in the presence of multicollinearity, include [26], [3], [27] and [28]. The development of ridge-type S-estimators is another interesting

and active research area [19, §5.10], that we do not pursue in this work.

In what follows we take `bdp` = 0.5. Although we acknowledge that this choice is somewhat conservative, assuming the potential presence of (almost) $n/2$ outliers, and that some efficiency will be lost in practical situations where it is reasonable to assume a lower percentage of contamination, our main goal is the comparison of alternative ways to estimate the covariance matrix of $\hat{\beta}$. For this goal `bdp` is a kind of "nuisance parameter" that must be kept constant among all the methods under comparison and not necessarily tuned to its "optimal" value. Furthermore, we may expect the difference among the compared methods to be magnified when `bdp` = 0.5 and the effect of downweighting by ρ is larger. Adaptive methods that allow the user to infer the contamination level from the data, and thus to choose the smallest value of `bdp` that ensures robustness in a given problem, are described by [23], [6] and [9, §7.4].

2.2. Details on covariance estimation

In general the covariance matrix of robust estimated parameters can be written as

$$\mathrm{Cov}(\hat{\beta}) = q^2 \times \sigma^2 \times v \times V_X^{-1}.$$

The building bricks of the above representation are:

(1) q a correction factor for the scale estimate;
(2) σ the scale parameter. In this work in order to better investigate the influence of this parameter we initially assume it is fixed and known and then we relax this assumption;
(3) v a correction factor depending on the ψ function (the derivative of the ρ function which is selected to down-weight the importance of observations with large residuals);
(4) V_X= a $p \times p$ positive semidefinite matrix (for OLS we have $V_X = X'X$).

Given that for S robust regression estimators there is a weight attached to each observation, the matrix V_X should have a structure of form $X'WX$, where W is a diagonal matrix containing the weights assigned to each observation.

In this work we study the effect of v and V_X for a variety of robust methods with the most widely used ψ functions; see Section 3. There are proposals for the "optimal" tuning of q [17] for MM estimators, but we do not address this issue here. In order to marginalize for the possible effect of the estimate of σ in the comparison among the different estimators initially we do not use this this parameter in the computation of the covariance matrix. Additionally, we check the stability of the results when this parameter is replaced by its estimate.

In our study we consider five different covariance matrix estimators, detailed in Table 1, all implemented in the common and flexible computational framework provided by the *FSDA Toolbox of Matlab* ([25]) available from *Mathworks* file exchange or from *Github*. More specifically the function which computes the covariance matrix of the different estimators is called `RobCov.m`. The HTML of this function (together with all of the other functions) can be found after installing the toolbox or on line at the web address
`http://rosa.unipr.it/FSDA/robcov.html`.
In Table 1 are listed with the five acronyms used in the paper. The robust estimates

Table 1. A summary of different strategies to estimate $\mathrm{Cov}(\hat{\beta})$ and associated formulae

Acronym	Formula
covrob	$\sigma^2 \hat{v}(X'X)^{-1}$
covrob1	$\sigma^2 K^2 \hat{v}(X'WX)^{-1}$
covrob2	$\sigma^2 K^2 \hat{v}(X'X)^{-1}$
covrob3	$\sigma^2 K \frac{n}{n-p} \frac{\sum_{i=1}^n \psi(r_i/\hat{\sigma})^2}{\sum_{i=1}^n \psi'(r_i/\hat{\sigma})} (X'WX)^{-1}$
covrob4	$\sigma^2 K^{-1} \frac{1}{n-p} \sum_{i=1}^n (\psi(r_i/\hat{\sigma}))^2 (X'WX)^{-1} X'X (X'WX)^{-1}$

proposed in Table 1 requires some "tuning constants" that are, respectively,

$$\begin{aligned} \hat{v} &= \frac{n}{n-p} n \frac{\sum_{i=1}^n \psi(r_i/\hat{\sigma})^2}{[\sum_{i=1}^n \psi'(r_i/\hat{\sigma})]^2} \\ K &= 1 + [CV(\psi')]^2 p/n \\ CV(\psi') &= \sqrt{\frac{var(\psi'(r_i/\hat{\sigma}))}{(\sum_{i=1}^n \psi'(r_i/\hat{\sigma})/n)^2}}. \end{aligned}$$

In the above expressions, the notation $CV(\psi')$ stands for the coefficient of variation of ψ', where ψ and ψ' are, respectively, the first and second derivatives of the ρ function discussed earlier in this section. The estimator called *covrob* inside Table 1 is equation (4.49) of p. 101 of [18]. The estimators called *covrob2, covrob3, covrob4* are, respectively, referred to equations (7.81) (7.82) (7.83) of p. 171 of [15] and were derived under the balanced case (i.e. $h_i = h = p/n$ where the h_i, $i = 1, \ldots, n$ is the $i-$th element of the hat matrix $H = X(X'X)^{-1}X'$). The estimator called *covrob1* is *covrob2* with $(X'X)^{-1}$ replaced by $(X'WX)^{-1}$, where W is, as stated earlier, a diagonal matrix attaching a weight to each observation.

3. Simulation framework and comparison between estimators under normality

In order to have a proper benchmark, we first analyse the empirical performance, under model (1), of all the procedures to estimate the covariance matrix of the parameters β in Table 1, considering S-estimator with Tukey's ρ function, S-estimator with Hampel's ρ function and S-estimator with Optimal ρ function, all with breakdown point `bdp` $= 0.5$. The simulation study is built using different sample sizes n and different numbers of parameters p. In particular, we consider the grid defined by the following values $n = [50, 100, 200, 300, 400, 500]$ and $p = [2, 5, 7, 10]$.

For each pair of n and p, we simulate 10000 replicates from model (1), where y is a sample from a standard normal distribution. The results that we present are based on $X_1, \ldots, X_{p-1}$ generated from the standard multivariate normal distribution.

For all replicates we calculate the difference between the empirical (Monte Carlo) covariance matrix and the estimated covariance matrices obtained with the five strategies listed in Table 1.

3.1. Benchmark with ordinary least squares based on MSE

It is important to address the first issue about accuracy of approximations used in the paper. To justify that the number of replicates (10000) is big enough, we check how far are entries in the true covariance matrix of $\hat{\beta}$ from the covariance matrices obtained with Monte Carlo, whose estimate is denoted by $\widehat{\text{Cov}}(\hat{\beta})$. Clearly, only for OLS we know the true $\text{Cov}(\hat{\beta}) = \sigma^2(X^TX)^{-1}$, but we argue heuristically that for robust estimators the same accuracy might hold.

Entries from the true covariance matrix and entries of the empirical covariance matrix, obtained by simulation, are subtracted. Subsequently, on the matrix of differences the Frobenious norm is computed, which is the Euclidean norm in many special cases. So, in our calculations, for a matrix A with dimension $r \times c$ and elements a_{ij}, we compute the norm as

$$\|A\| = \sqrt{\sum_{i=1}^{r}\sum_{j=1}^{c}|a_{ij}|^{\ 2}}\ . \tag{6}$$

The comparison works as follows. For a fixed n and p we take X, which is fixed for all 10000 replicates. The entries of $\text{Cov}(\hat{\beta})$ are known exactly (with the additional simplification that $\sigma^2 = 1$). The difference of all entries of the true covariance matrix of $\hat{\beta}$ is contrasted with the sample covariance matrix of all the $\hat{\beta}^{(j)}$, obtained for each replicate $j = 1, \ldots, 10000$. To summarize the difference of the entries we resort to the calculation of the norm induced by formula (6). The values of such norms are contrasted graphically (see Figure 1). By inspecting all panels of Figure 1 we notice that the norm of $\text{Cov}(\hat{\beta})$ (represented by a black bullet) is mostly indistinguishable from that of $\widehat{\text{Cov}}(\hat{\beta})$ (whose norm is represented by a grey cross). From such a comparison we conclude that the number of replicates used in our Monte Carlo experiment is large enough to ensure that our empirical estimates match the theoretical quantities for all combinations of n and p.

Therefore, in the sequel we claim that entries of $\text{Cov}(\hat{\beta})$ are, in practice, indistinguishable from entries of $\widehat{\text{Cov}}(\hat{\beta})$.

3.2. Graphical diagnostics for robust estimators

With simple graphical diagnostics we explore the MSE for each robust estimator. The details are the following: for each simulation $j = 1, \ldots, 10000$ we consider the robust covariance matrix estimate, say $B^{(j)} = \widehat{\text{Cov}}_R(\hat{\beta})^{(j)}$ and the benchmark covariance matrix $V = \widehat{\text{Cov}}(\hat{\beta})$, i.e. the matrix obtained through Monte Carlo. Then we compute

$$d^{(j)} = \frac{\|B^{(j)} - V\|}{p} \quad j = 1, \ldots, 10000,$$

and, by denoting with $\bar{d} = \sum_{j=1}^{10000} d^{(j)}/10000$, the corresponding MSE is given by

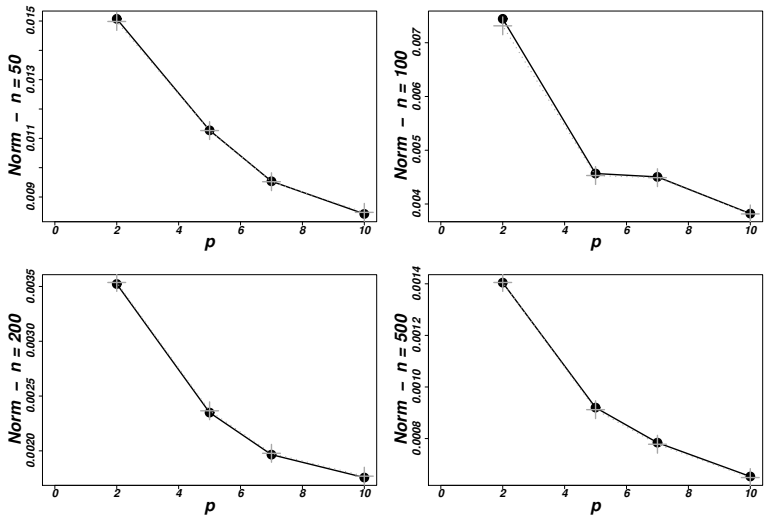

Figure 1. Distance between true covariance matrix of $\hat{\beta}$ from the empirical covariance matrix whose entries are computed after 10000 Monte Carlo replicates as function of a subset of n and p. The norm for OLS is represented by a black bullet and the norm for the Monte Carlo empirical norm is given by a grey cross

$$\text{MSE} = \frac{\sum_{j=1}^{10000} \left(d^{(j)} - \bar{d}\right)^2}{10000}.$$

It could be important on its own to compare the $\bar{d}$ for all estimators as a small $\bar{d}$ is preferable. We found consistency of results by comparing $\bar{d}$ and the MSE of the $d^{(j)}$ and prefer to report results for the MSE on which our correction factor is based. Results related to $\bar{d}$ are reported in the supplementary material. The outcomes of the MSE, for all robust estimators, are displayed in Figure 2, where each panel refers to a different value of p against the sample size n. For sake of simplicity we report results for Tukey biweight ρ function only, because with other ρ functions results were broadly consistent. For a discussion of the different ρ functions see for example [22].

The first remarkable evidence is that when $p = 2$ and $n = 50$ the two robust estimators *covrob1* and *covrob2* have a very large MSE (top left panel of Figure 2) when compared to that of the other robust methods and other choices of p and n. On the other hand these two estimators seem to be the best when p is large.

To our view, such feature was striking at first sight so that further investigation led to examine the distribution of the norm for every single simulation with $n = 50$ and $p = 2$. The boxplots of the $d^{(j)}$, in log scale, is sketched in Figure 3. There are many simulations where the norm takes large values for *covrob1* and *covrob2*, making large the associated MSE. The log scale magnifies the left tail of the distribution of the $d^{(j)}$. The rightmost boxplot of Figure 3 shows that *covrob4* has very low variability but remarkably larger bias.

Quite surprisingly, as illustrated by the boxplots in Figure 4, for the same sample size n, but much larger p, the distribution of each single norm is much less affected by "extreme" values. We have that *covrob1* and *covrob2* seem to outperform when contrasted to other robust estimators, even though the "average" performances are somehow comparable. Hence, despite a potential lack of model accuracy when $p = 10$ (with rate $n/p = 5$) the robust covariance estimators of $\hat{\beta}$ have quite nice properties

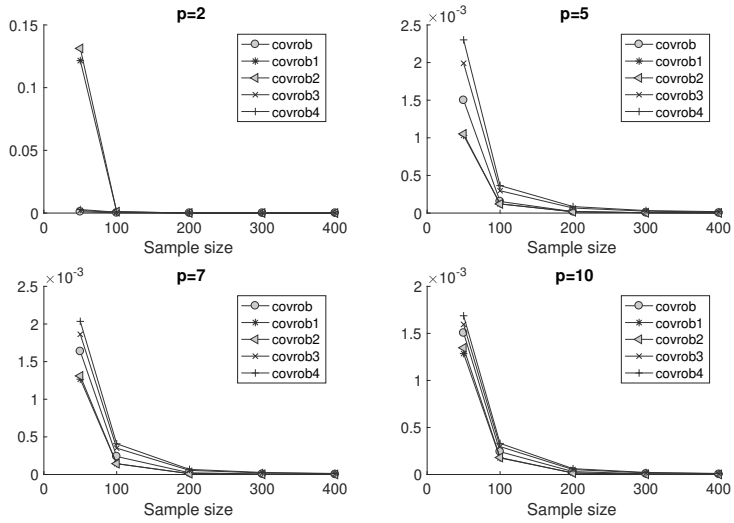

Figure 2. MSE of $d^{(j)}$ obtained with the five strategies listed in Table 1 for S-estimator using a value of breakdown point equal to 0.5 and a Tukey biweight ρ function

even for moderately small sample sizes. We tackle this issue more formally later, in Section 3.3, where some details are provided.

Similar findings hold when the error term is symmetric. Specifically, when the ϵ_i are i.i.d. following either a Student's t distribution or Uniform results are similar to those illustrated in Figure 2. Evidence of all these results can be found in the supplementary material. In order to check practical relevance to our simulation results and their stability, we also report in Figures 5 and 6 the evaluation of the final covariance matrix estimators when σ is estimated by the residual scale $\widehat{\sigma}$ from the S estimator. Figure 5 is similar to Figure 2 and shows the same pattern of the estimators. *covrob1* and *covrob2* are the worst for small p and small/moderate sample size but they become the best when n and p is large. *covrob2* seems to slightly outperform *covrob1*. Boxplots in Figure 6 are coherent with the ones in Figure 3 and they present the same behaviour of the estimators with small n and small p.

Additionally, we found comparable results when the design matrix has leverage points. The correction under these settings is postponed to Section 5.

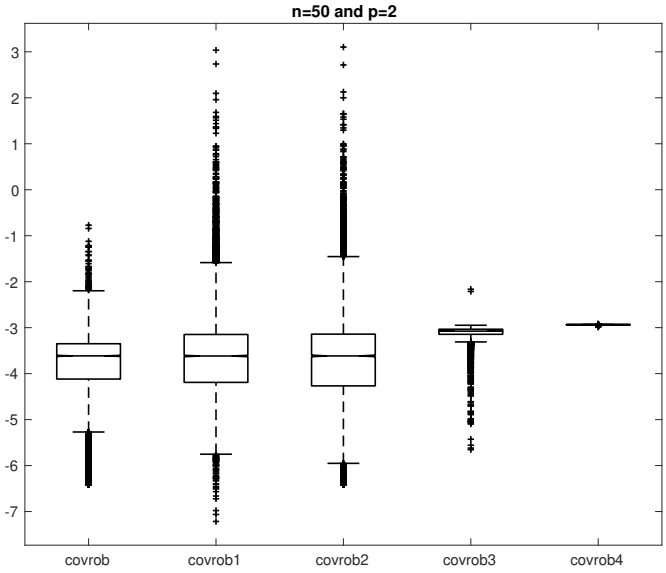

Figure 3. Boxplots of the $d^{(j)}$, in the log scale, obtained with the five strategies listed in Table 1 for S-estimator using a value of breakdown point equal to 0.5 and a Tukey biweight ρ function for $n = 50$ and $p = 2$ (the boxplots in the original scale are provided in the supplementary material)

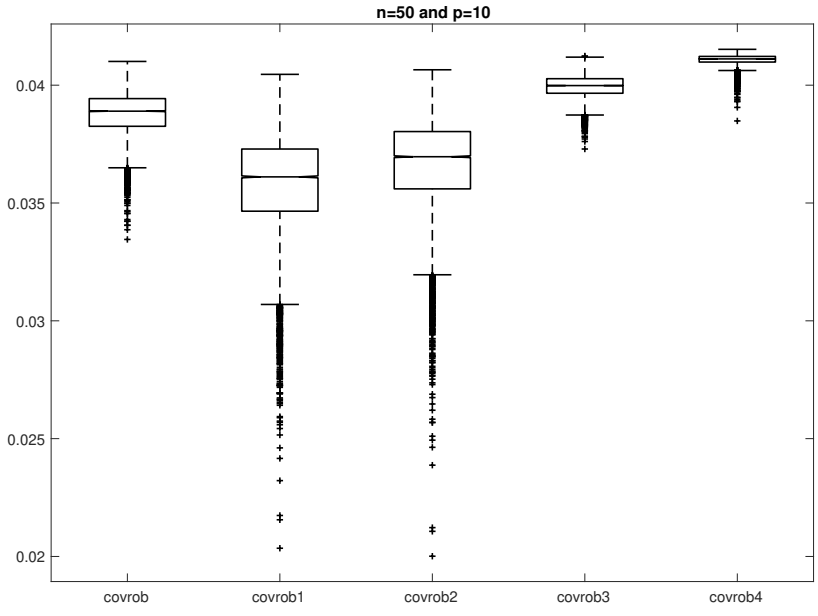

Figure 4. Boxplots of the $d^{(j)}$ obtained with the five strategies listed in Table 1 for S-estimator using a value of breakdown point equal to 0.5 and a Tukey biweight ρ function for $n = 50$ and $p = 10$

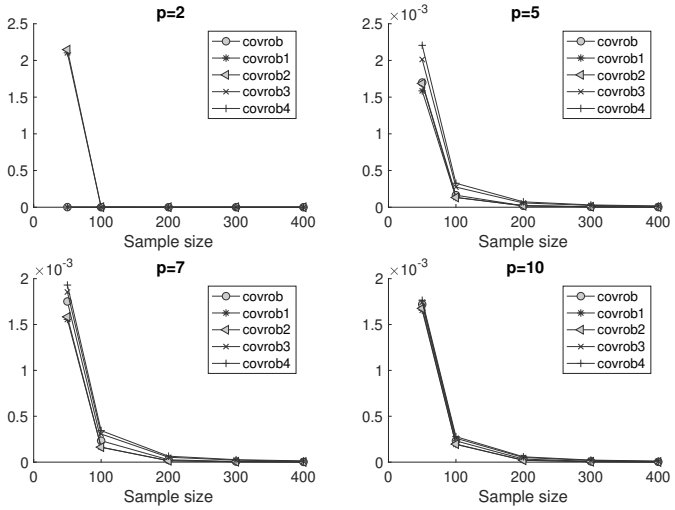

Figure 5. MSE of $d^{(j)}$ obtained with the five strategies listed in Table 1 with $\widehat{\sigma}$ for S-estimator using a value of breakdown point equal to 0.5 and a Tukey biweight ρ function

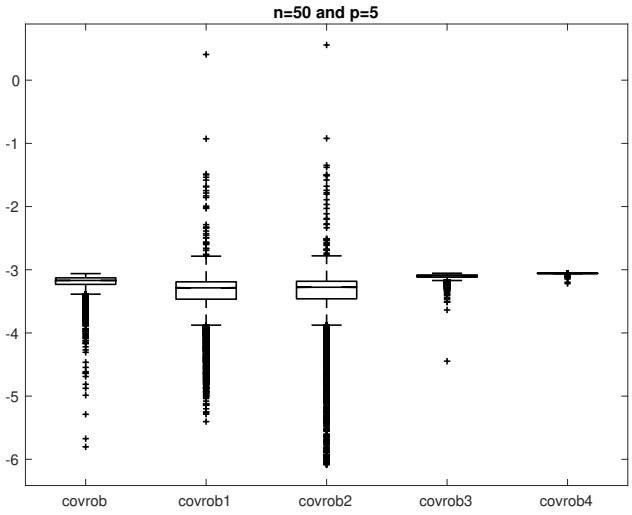

Figure 6. Boxplots of the $d^{(j)}$, in the log scale, obtained with the five strategies listed in Table 1 with $\widehat{\sigma}$ for S-estimator using a value of breakdown point equal to 0.5 and a Tukey biweight ρ function for $n = 50$ and $p = 2$

3.3. A glimpse on the motivation for the poor performance for small p and small n

In this section we tackle the problem highlighted previously, i.e. some unreliable estimates of the covariance matrix of the $\hat{\beta}$ displayed by *covrob1* and *covrob2*, when both n and p are small. We report details only for the Tukey biweight ρ function, but for other ρ functions results are similar.

Once a ρ function is selected, it is important to note that residuals get their weight through the associated ψ' function. Let us take the Tukey bi-weight ψ' function given by the mathematical expression

$$\psi'(u) = (1 - u^2)(1 - 5u^2).$$

By considering the substitution $u = x/c$, with the constant $c = 1.5476$, any value bigger than $|c|$ is set equal to zero. With these choices, the plot of the ψ' function is shown in the left panel of Figure 7. The values for which ψ' is equal to zero are given by $\pm c$ and by $c_\star = \pm 0.69211$ respectively.

To understand the unexpected outcome of some simulations carried earlier, we now explore, via Monte Carlo, some properties of the ψ' function, in terms of the mapping that it produces when a specific random variable forms its input.

Specifically, using the notation that for a given random variable U we are interested in the new variable $Z = \psi'(U)$ we want to investigate some properties of Z. In our discussion the choice of U is uniform, Gaussian and Student's $t(3)$ respectively. Beside the Gaussian and Student's $t(3)$, which are common in regression, the argument for studying the uniform relies on a simple motivation that when the sample size is relatively small ($n = 50$ in our settings) for some choices of p the set of scaled residuals r_i is consistent with a uniform distribution.

The peculiar feature of our finding is that, for small n, in some simulations the sample mean of scaled residuals, "weighted" by the ψ' function, are very close to zero. This makes the quantity which appears in the denominator of the covariance matrix (the denominator of CV) of robust estimators (see Table 1)

$$\left[\sum_{i=1}^{n} \psi'(r_i/\hat{\sigma})\right]^2$$

indistinguishable from zero and consequently blows up the standard error. The reason for the evidence is that when the input U is Uniform in the interval $[-c, c]$ then it can be easily shown that for the transform $Z = \psi'(U)$

$$E[Z] = E\left[\psi'(U)\right] = \int_{-c}^{c} \psi'(x) f_U(x; -c, c) dx = \frac{1}{2c} \int_{-c}^{c} \psi'(x) dx = 0, \tag{7}$$

as $\int_{-c}^{c} \psi'(x)dx = 0$. Such a result does not hold when U is Gaussian or Student's t. For instance, when U is standard Gaussian, then $E[\psi'(U)] \approx 0.1554$ and for U Student's t we have $E[\psi'(U)] \approx 0.1492$. The densities of these cases are sketched in the right panel of Figure 7, where a "wiggly" region around 0 is peaky for Uniform U.

As a consequence, when the sample size is small, it happens in a few cases that scaled residuals r_i are compatible with a Uniform distribution and their sample mean,

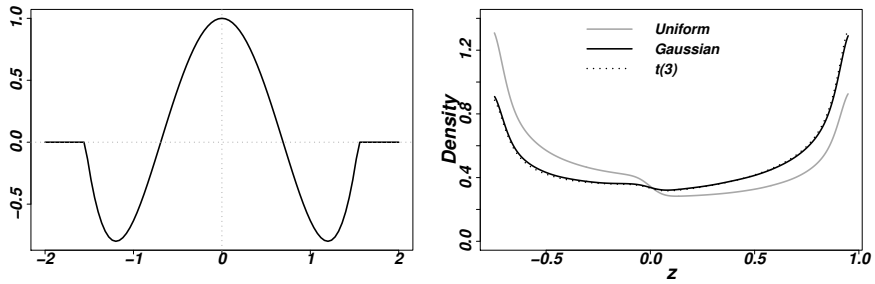

Figure 7. Left: Analytical plot of $\psi'(u)$. Right panel: Monte Carlo density of the transform $Z = \psi'(U)$ with U Uniform (grey solid), Gaussian (black solid) and Student's $t(3)$ (black dotted) respectively

when used as input in the ψ' function, might be very close to zero, making the $CV(\psi')$ unexpectedly large and the associated correction factor K (see Table 1) equally large. Note that this has strong effects on the corresponding associated hypothesis tests (or confidence intervals). This effect is overwhelming in *covrob1* and *covrob2* because in these cases the value of K appears raised to the power of two. This finding perfectly justifies the presence of large values of $d^{(j)}$ in Figure 3 for *covrob1* and *covrob2*. It also provides further evidence that M-based estimators may become unreliable, in addition to the numerical instability issues noted by [18, p.47] for the case of redescending ρ functions. When n is large then the r_i approximate better their underlying normal distribution, and such an effect vanishes.

The effect of p deserves some further discussion. For large p residuals have "heavy" tails, due to the poor accuracy of the regression model with many predictors. For this reason, when p is large, the empirical distribution of scaled residuals is far from uniform, unlike for p small or moderate. Then, when p is large, the number of residuals that get zero value (when mapped from the range of the ψ' function) is equally large. As a consequence the sample mean of mapped values onto ψ' is, on average, bigger than zero, making non-degenerate the estimate of the coefficient of variation, and reliable the estimators *covrob1* and *covrob2*.

In closing this section we highlight that when the ϵ_i are i.i.d. following a Uniform distribution, then the performance of *covrob1* and *cobrov2* is considerably deteriorated. This is consistent with theoretical results given in (7). Some empirical evidence of this is attached to the supplementary material and it can be noticed that *covrob1* and *cobrov2* perform consistently worse than other estimators for any sample size. This feature has no practical consequences for the correction proposed in this paper.

4. Proposal for a correction

In the previous section we have seen that the poor performance of *covrob1* and *covrob2* takes place in the cases in which the associated CV (that is the coefficient of variation of ψ') is very large. The purpose of this section is to find the optimal threshold of CV which tells us when the estimator of the covariance matrix based on *covrob1* or *covrob2* has to be replaced by *covrob*. The basic idea starts from exploring how critical is the estimate of the sample coefficient of variation of scaled residuals. Once a "threshold" has been identified we adopt a simple "switching" technique. Via the analysis of an average mean square error we provide practical guidelines useful to select the proper robust estimator.

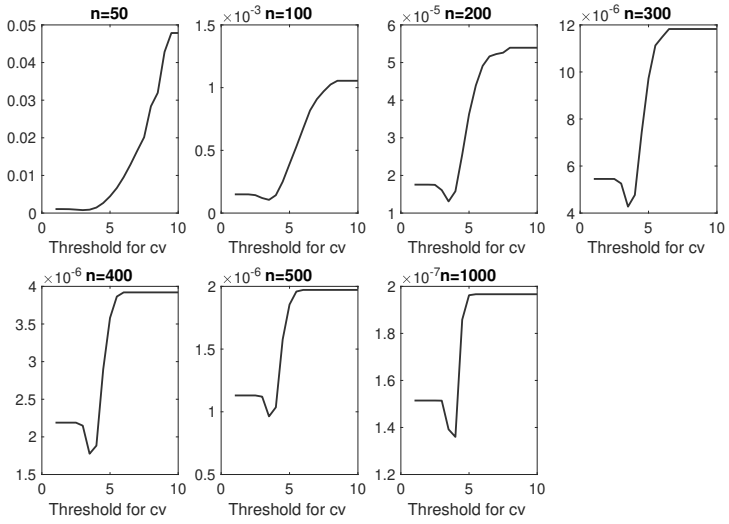

Figure 8. Monitoring of MSE for different level of CV thresholds and a variety of choices of sample size n. For all panels the level of CV is monitored for a fixed value of $p = 2$.

The algorithm used to select the "critical value" value v_0 of the CV such that a switch is required, is driven by kind of cross-validation based on the following steps:

- fix n and p and store the sample CV of residuals for all 10000 simulations, denoted $\overline{\text{CV}}^{(j)}(r)$, $j = 1, \ldots, 10000$;
- select an arbitrary v_0; for the j-th simulation, if $\overline{\text{CV}}^{(j)}(r) > v_0$ then *covrob* of Table 1 is used otherwise *covrob2* of Table 1 is selected;
- the covariance matrix of the associated estimator is contrasted with the true covariance matrix via the proper norm;
- the optimal v_0 is selected by a graphical monitoring of the norm as a function of v_0, when v_0 becomes extreme.

The application of our correction leads to "new estimator" that we call *newcovrob*, and for which we now explore the performance, to monitor the improvement obtained. An illustration of the monitoring as the sample size increases is displayed in Figure 8 where the sensitivity of the MSE is explored against the level of CV when $p = 2$ and different choices of n. The fixed value of $p = 2$ is motivated by the bad performance of the associated robust estimator for such a design input.

The different panels of Figure 8 show that the MSE of *newcovrob* starts to increase significantly after CV> 4 suggesting that that 4 is the threshold for the switching of the estimator. Our suggestion therefore is to use estimator *covrob2* in all the cases in which CV< 4 and to use *covrob* when CV is greater than 4. Note that this rule leads to a complete automatized choice of the best estimator, because CV can be easily computed for our sample.

Figure 9 shows the graphical results when our correction is used. Our correction is denoted by a solid thick black line. All other lines in Figure 9 just plot the same results displayed in Figure 2 and are re-proposed for ease of comparison.

The results obtained with our simple correction method seem to provide uniformly better results for any choice of n and p. In particular, the estimator based on our correction, significantly outperforms *covrob1* and *covrob2* for small p and small n. When

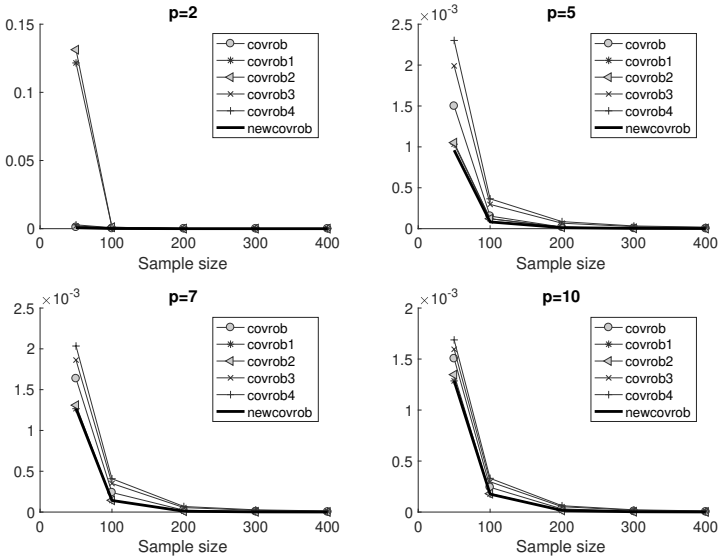

Figure 9. MSE of $d^{(j)}$ obtained with the five strategies listed in Table 1 and for the robust covariance estimator corrected, for S-estimator using a value of breakdown point equal to 0.5 and a Tukey biweight ρ function.

p grows the performance of our estimator *newcovrob* is equally very good, implying that the correction proposed shares the same large sample properties of any robust estimator.

5. Leverage points in the design matrix and different levels of contamination

In this section we check the stability of our results to departure of settings used in Section 3. In particular we consider leverage points in the design matrix, with again fixed independent variables and Gaussian error term. This is, probably, the most important violation of the assumptions, as it mimics the presence of outliers in the carriers. The contamination scheme is the following:

- percentage of contamination is 20% or 30%;
- contamination leverage for the x's,with a shift of magnitude 5, 10 or 20.

After including the contamination we perform all calculations made in previous sections and compute the MSE of all estimators, included *newcovrob*, i.e. the correction proposed in Section 4. A subset of the results are illustrated in Figure 10 and Figure 11. It is clearly visible that our correction is outperforming all existing methods even under presence of leverage points in the design matrix.

When the contamination scheme is slightly different we found no significant difference in the results, remarking that our correction is accurate even under a variety of situations. For example Figure 12 which extends Figure 8 to the case of 30 per cent contaminated data shows that the threshold of 4 for CV still applies when the data contain outliers.

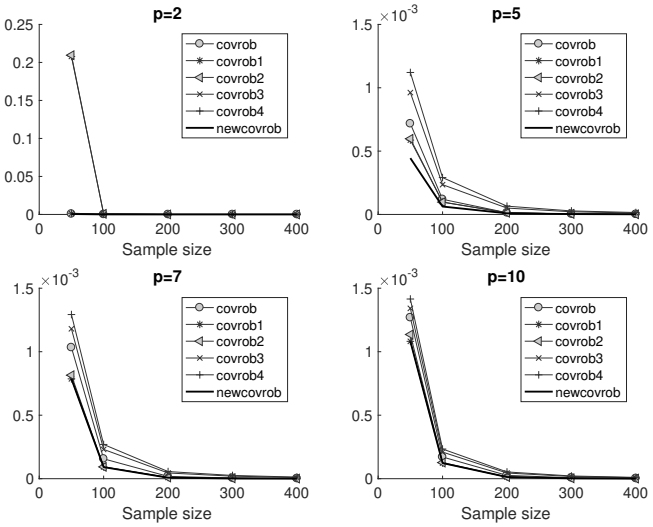

Figure 10. MSE of $d^{(j)}$ obtained with the five strategies listed in Table 1 and for the correct robust covariance estimator when the design matrix has leverage points, $shift = 5$ and percentage of contamination 10%.

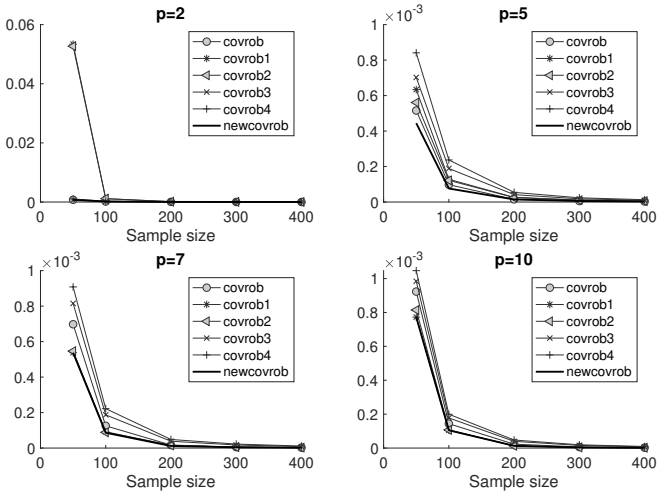

Figure 11. MSE of $d^{(j)}$ obtained with the five strategies listed in Table 1 and for the correct robust covariance estimator when the design matrix has leverage points, $shift = 20$ and percentage of contamination 30%.

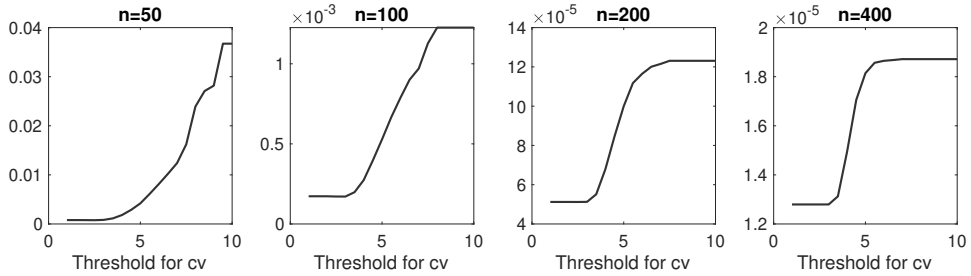

Figure 12. Monitoring of MSE for different level of CV thresholds and a variety of choices of sample size n. For all panels the level of CV is monitored for a fixed value of $p = 2$. Data contaminated with 30% of outliers.

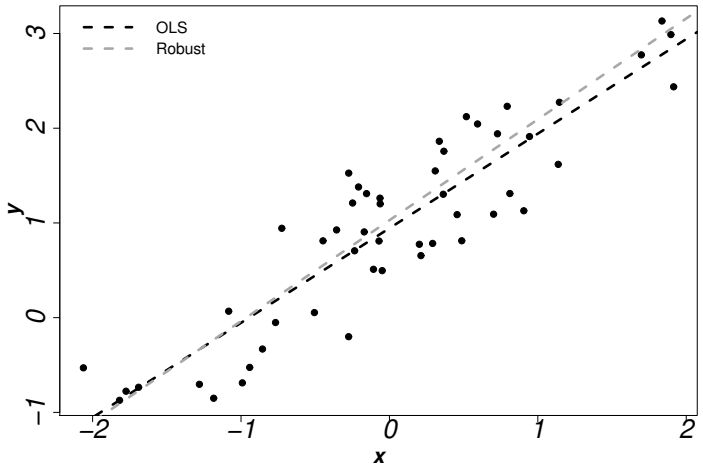

Figure 13. Simulated data with OLS and Robust fitted lines superimposed.

6. Applications

6.1. An example with simple simulated data

An interesting example is provided by a simulated dataset with $n = 50$ and $p = 2$. The plot of the original data, with superimposed the fitted OLS and the robust S estimate is illustrated in Figure 13. The fitted lines are almost identical; hence the OLS and robust residuals have similar distribution. The robust procedure does not detect any outlier. It could be argued that a robust procedure for these data might be unnecessary, but the choice of an approach cannot be data driven.

Despite the relatively simple structure the robust estimates *covrob1* and *covrob2* have some large entries in $\widehat{\mathrm{Cov}}(\hat{\beta})$, as reported in Table 2 (because the value of CV in this example is much larger than 4), whereas all other robust estimators considered in the paper have much smaller entries. This matches the findings explored in Section 3.

Table 2. Variance covariance matrix of the robust estimates *covrob1* and *covrob2* in case of $\sigma = 1$ and $\widehat{\sigma}$

	$\sigma = 1$		$\widehat{\sigma}$	
covrob1	29.6406	-0.4576	8.8986	-0.1374
	-0.4576	29.4427	-0.1374	8.8391
covrob2	29.6343	-0.1655	8.8967	-0.0497
	-0.1655	33.4254	-0.0497	10.0348

So, with the data proposed in this simple example there would be heterogeneous inferential conclusions depending on the robust method adopted. Our algorithm in *newcovrob* suggests to use the *covrob* which would hint to significant parameters estimates. We are aware that this is only a toy example, for a case of simple linear regression in the absence of outliers, but we have chosen it just to show that the robust estimators of the covariance matrices in some cases can produce very large values. Sometimes, somebody could use robust estimators, to compare them with the classic ones, in order to keep under control the eventual presence of outliers. In the event of very different empirical statistical tests for the null hypothesis on the parameters, incorrect conclusions could be drawn.

6.2. Outlier detection and inference in action

In order to show the importance of the application of the suggested approach in this section we consider different scenarios of data contamination of the dataset given in the previous section: point mass contamination, shift contamination on y, contamination with leverage points and combination of the above schemes. Figure 14 considers 4 contamination cases, and shows both fitted OLS line and the robust S estimate. In all these cases using S-estimator the slope of the regression line is correctly estimated. However, given that in all these four cases CV is greater than 4, *covrob1* and *covrob2* have some large entries in $\widehat{\text{Cov}}(\hat{\beta})$, and therefore the associated t-statistic (as shown in Table 3) point out that both the intercept and the slope are not significant. The use of our correction, estimator *newcovrob*, on the other hand, enables to correctly appreciate the significance both of the slope and of the intercept in a fully automatic way. Our approach is particularly suitable for all institutions which routinely have to analyse, in a fully automatic way, the significance of the explanatory variables. For example, the Joint Research Centre of the European Commission, with the purpose of the protection of the EU budget, on a daily bases uses robust regression for each imported product (using a eight digits TARIC classification) where y is represented by the value declared and the explanatory variable is the quantity declared ([6]). The estimated slope in this case is the so called "Outlier-Free Average Prices". Given that in the regression of value against quantity the intercept should not be significant, in this case it is important to be able to correctly appraise also the significance of this parameter. A significant intercept in this context, might be due to the presence of rounding values through the origin and or to other potential irregularities. Even if there is just a single explanatory variable, the total number of cases (products) to consider is so large (about 12 millions) that it is not possible to resort to graphical aspects, but it is necessary to have tools which enable to improve inference.

Table 3. Comparison of t-statistics using different estimators of the covariance matrix in the four cases shown in Figure 14

		case 1				case 2	
	covrob1	*covrob2*	*newcovrob*		*covrob1*	*covrob2*	*newcovrob*
Intercept	1.0379	1.002	2.1923	Intercept	1.0379	1.0241	2.2406
Slope	1.1164	1.8975	4.1518	Slope	1.1164	1.1954	2.6155
		case 3				case 4	
	covrob1	*covrob2*	*newcovrob*		*covrob1*	*covrob2*	*newcovrob*
Intercept	1.0379	1.0075	2.2043	Intercept	1.554	1.496	2.7687
Slope	1.1164	1.7325	3.7907	Slope	1.7378	1.8354	3.3969

6.3. An example with real data

In order to test the performance of the five different covariance matrix estimators, detailed in Table 1 and implemented in the common and flexible computational framework provided by the FSDA Toolbox of Matlab [25], we consider the benchmark dataset previously used by [16] and obtained from [10]. The nuclear power station dataset contains ten possible explanatory variables for the construction cost of 32 light water reactor plants built in the USA in 1967-71. This dataset has $n < 50$ cases and $p = 10$ variables. According to our simulation study, with such values of n and p, the five estimators might provide discordant results.

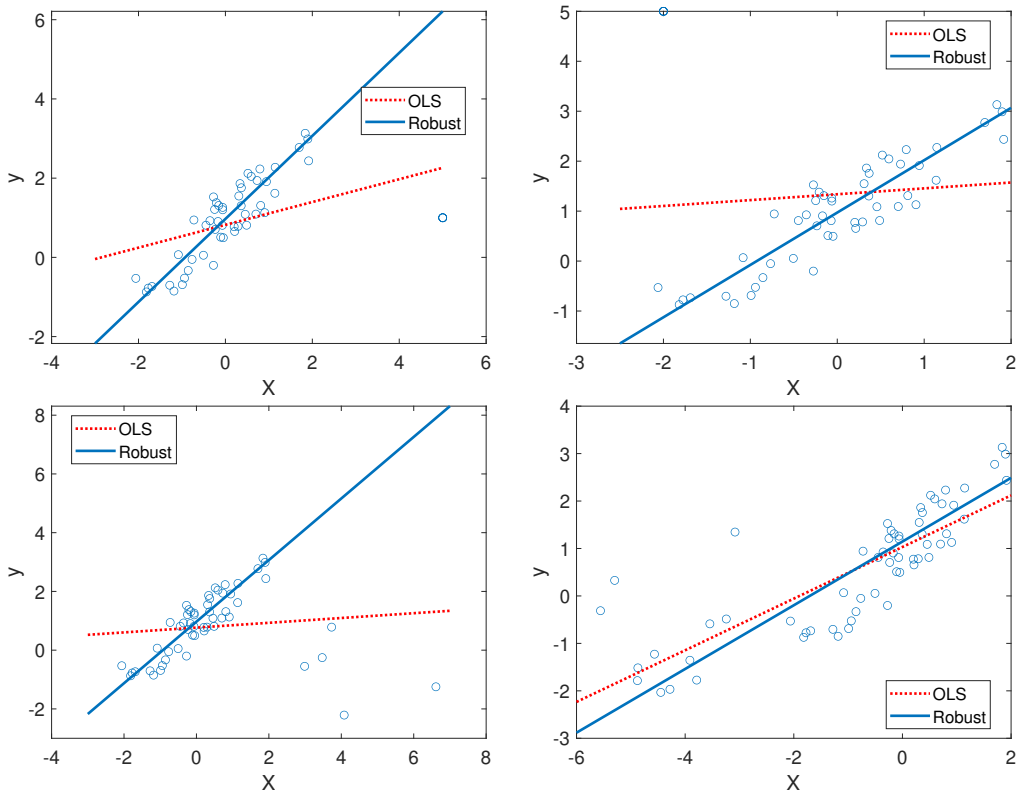

Figure 14. Comparison between robust (blue continuous line) and non robust fit (red dotted line) for 4 different cases of contamination. 10% point mass contamination (top panels), 10% contamination with bad leverage points, 20% contamination with leverage points

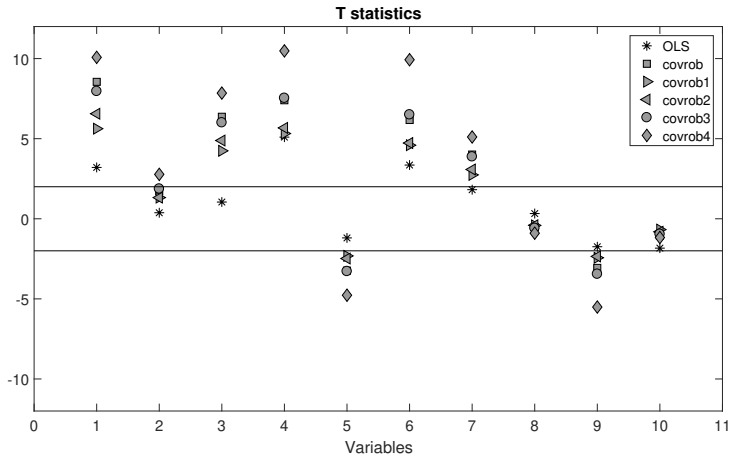

Figure 15. *t*-statistics for OLS and S estimator for the ten explanatory variables of the nuclear power plan data. For the S estimator, the standard errors are obtained with the five different strategies listed in Table 1

Figure 15 shows the *t*-statistics for each explanatory variable (on the horizontal axis) calculated with the classical OLS and contrasted with the robust procedures, i.e. the S-estimators with 0.25 breakdown point and a Tukey biweight ρ function. The five different strategies, listed in Table 1, were used to estimate the covariance matrix of the S robust estimators. Obviously the *t*-statistic of OLS estimator is different from the *t*-statistic of the S estimators because the former is not robust. More relevant for the purpose of our work is the estimate of the covariance matrix of the S-estimators. The covariance matrix estimation affects the significance of the variable. In particular, for the second variable, the *t*-statistic depends on the procedure selected for the standard error of the robust estimator.

Following the approach of [16], we display in Figure 16 the estimated 95% confidence intervals for the OLS and for the S estimator using the five different robust strategies. The estimated confidence interval of the S estimator, obtained through *covrob4*, does not include 0, suggesting that the variable *t1* is significant. With other methods the value 0 is instead included, suggesting the opposite conclusion.

According to our simulation the more suitable procedure is given by the estimator (7.81) of p. 171 of [15]. This in not in contrast with the conclusions of [16] where it is stated that the differences in the widths of the intervals are small compared to the difference of the parameter estimates. We agree that the location of the estimation has a larger impact than the estimate of uncertainty, but in this analysis we also show that the selection of the correct robust procedure for estimating the uncertainty is a key relevant issue. Notice that our *newcovrob* estimator does not show up explicitly and the motivation is simple: in this example the CV is always below the threshold of 4 and consequently the correction is unnecessary.

7. Conclusions, discussion and further research

In this work we have studied the covariance matrices of regression estimators obtained from popular high-breakdown robust methods when no outliers are present in the data.

We considered five different covariance matrices based on [18] and [15]. The performance of these five estimators are compared with the empirical covariance matrix

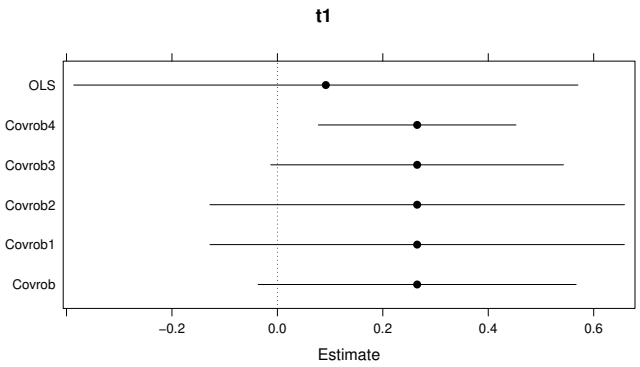

Figure 16. 95% confidence intervals for OLS estimator and S estimator for the second explanatory variable *t1* of the nuclear power plan data. The estimated covariance matrices obtained for the S estimator are obtained using the five strategies listed in Table 1

obtained through Monte Carlo simulations. We found that the performance depends on the size of the regression problem. In particular for small n and small p the estimator (7.81) of p. 171 of [15] and the same with $(X'X)^{-1}$ replaced by $(X'WX)^{-1}$ that consider the coefficient of variation of ψ' present some problems, ,due to the shape that could assume the residuals distribution. On the other hand, especially when n increases, the two mentioned estimators perform better than the others. These considerations are valid independently from the ρ function considered. For these cases we propose a correction that is based on the switch from one estimator to an other depending on the "critical value" of the coefficient of variation of ψ'. More specifically we recommend the use of estimator *covrob2* in all cases in which the coefficient of variation of ψ' is smaller than 4 and the use of estimator *covrob* when CV> 4.

Results of our work extend to several directions. In particular we investigated the properties of the estimators for some non Gaussian distribution and studied the sensitivity in presence of leverage points in the design matrix. We remark the strength of our correction which is suitable under a wide variety of practical scenarios.

Although the results are not reported, a breief preliminary study has been done by the authors also for MM estimators. In this case the estimator (4.49) of p. 101 of [18] seems always better that the others. This is somehow expected given the higher efficiency of the MM estimators. Nevertheless, we have noted that if the level of efficiency increases, for some combination of p and n, the estimators (7.82) and (7.83) of p. 171 in [15], where K is not squared, have an higher performance. This leads us to think that some better proposals than the one in the literature are possible. We will work on this and on MM estimators in the future.

In the future it will be also interesting to compare our robust estimators with quantile regression. As mentioned in [34], where in Theorem 2 and Proposition 1 a consistent estimator of $\hat{\beta}$ covariance is obtained, when the distribution of the errors in the model is heavy tailed or the data contain some outliers, it is well known that median regression, a special case of quantile regression, is more robust than mean regression.

It is important to notice that when talking about standard error for robust regression estimator, researchers in the econometric community would argue that these standard errors should be also robust to heteroskedasticity and possibly autocorrelation. In this context the use of functions of residuals in weighted regression, including the weights based on squared residuals [33], is discussed by many authors, including [5], who ex-

plore slight misspecifications in the skedastic model. Further results are sketched by [32]. The references to trimming are more recent. [8] uses the model of [14], combined with the forward search and a trimmed likelihood for robust heteroskedastic regression. [20] use trimming in robust estimation of a general quasi-likelihood model, including the special case of heteroskedastic regression. The final aim is to have a method of robust heteroskedastic regression which is also robust to the specification of the form of heteroskedasticity. [4] show how poorly this heteroskedastic robust procedure based on squared residuals can perform when compared with a model with correctly specified heteroskedasticity. More importantly, these authors demonstrate that an heteroskedastic method which assumes the model of [14] is never less efficient than OLS even when the skedastic relationship is incorrectly specified. Therefore these results show that in presence of potential heteroskedasticity it is much better to assume a basic functional form (as the one of Harvey) rather then try to estimate it using (robust) squared residuals because OLS is never more efficient than fitting Harvey's model.

References

[1] Agostinelli C, Locatelli I, Marazzi A, Yohai VJ (2017) Robust estimators of accelerated failure time regression with generalized log-gamma errors. Computational Statistics & Data Analysis 107:92–106

[2] Alfons A, Croux C, Gelper S (2013) Sparse least trimmed squares regression for analyzing high-dimensional large data sets. The Annals of Applied Statistics 7:226–248

[3] Amini, M. and Roozbeh, M. (2016). Least trimmed squares ridge estimation in partially linear regression models. *Journal of Statistical Computation and Simulation*, 86:2766–2780.

[4] Atkinson AC, Riani M, Torti F (2016) Robust methods for heteroskedastic regression. Computational Statistics and Data Analysis 1046:209–222

[5] Carroll RJ, Ruppert D (1982) A comparison between maximum likelihood and generalized least squares in a heteroscedastic linear model. Journal of the American Statistical Association 77:878–882

[6] Cerioli A, Riani M, Atkinson AC, Corbellini A (2018) The power of monitoring: how to make the most of a contaminated multivariate sample. Statistical Methods & Applications 27(4):559–587.

[7] Cerioli A, Perrotta D (2014) Robust clustering around regression lines with high density regions. Advances in Data Analysis and Classification 8:5–26

[8] Cheng TC (2011) Robust diagnostics for the heteroscedastic regression model. Computational Statistics and Data Analysis 55:1845–1866

[9] Clarke, B. R. (2018). *Robustness Theory and Applications.* Wiley, Hoboken.

[10] Cox DR, Snell EJ (1981) Applied Statistics-Principles and Examples, vol 2. CRC Press

[11] Croux C, Van Aelst S, Dehon C (2003) Bounded influence regression using high breakdown scatter matrices. Annals of the Institute of Statistical Mathematics 55(2):265–285

[12] Filzmoser P, Gschwandtner M, Todorov V (2012) Review of sparse methods in regression and classification with application to chemometrics. Journal of Chemometrics 26(3-4):42–51

[13] García-Escudero LA, Gordaliza A, Greselin F, Ingrassia S, Mayo-Iscar A (2017) Robust estimation of mixtures of regressions with random covariates, via trimming and constraints. Statistics and Computing 27(2):377–402

[14] Harvey AC (1976) Estimating regression models with multiplicative heteroscedasticity. Econometrica 44:461–465

[15] Huber PJ, Ronchetti EM (2009) Robust Regression, 2nd edn. Wiley, New York

[16] Koller M, Stahel WA (2011) Sharpening wald-type inference in robust regression for small samples. Computational Statistics & Data Analysis 55(8):2504 – 2515

[17] Maronna RA, Yohai VJ (2010) Correcting MM estimates for "fat" data sets. Computational Statistics and Data Analysis 54:3168–3173
[18] Maronna RA, Martin RD, Yohai VJ (2006) Robust Statistics. Wiley, Chichester
[19] Maronna, R. A., Martin, R. D., Yoahi, V. J., and Salibian-Barrera, M. (2019). *Robust Statistics: Theory and Methods (with R). Second Edition.* Wiley, Hoboken.
[20] Neykov NM, Filzmoser P, Neytchev PN (2012) Robust joint modeling of mean and dispersion through trimming. Computational Statistics and Data Analysis 56:34–48
[21] Perrotta, D., Checchi, E., Torti, F., Cerasa, A., and Arnes Novau, X. (2020a). Addressing price and weight heterogeneity and extreme outliers in surveillance data - the case of face masks. Technical report EUR 30431 EN, JRC122315, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-24707-4. doi:10.2760/817681.
[22] Riani, M.; Atkinson, A.C.; Corbellini, A.; Perrotta, D. (2020) Robust Regression with Density Power Divergence: Theory, Comparisons, and Data Analysis. Entropy, 22, 399.
[23] Riani M, Cerioli A, Atkinson AC, Perrotta D. (2014) Monitoring robust regression. Electronic Journal of Statistics, 8(1):646–677.
[24] Riani, M., Cerioli, A., and Torti, F. (2014). On consistency factors and efficiency of robust S-estimators. *Test*, 23:356–387.
[25] Riani M, Perrotta D, Torti F (2012) FSDA: a matlab toolbox for robust analysis and interactive data exploration. Chemometrics and Intelligent Laboratory Systems 116:17–32
[26] Roozbeh, M. (2016). Robust ridge estimator in restricted semiparametric regression models. *Journal of Multivariate Analysis*, 147:127–144.
[27] Roozbeh, M. and Arashi, M. (2017). Least-trimmed squares: asymptotic normality of robust estimator in semiparametric regression models. *Journal of Statistical Computation and Simulation*, 87:1130–1147.
[28] Roozbeh, M., Babaie-Kafaki, S., and Sadigh, A. N. (2018). A heuristic approach to combat multicollinearity in least trimmed squares regression analysis. *Applied Mathematical Modelling*, 57:105–120.
[29] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* Wiley, New York.
[30] Salibian-Barrera M, Van Aelst S, Yohai VJ (2016) Robust tests for linear regression models based on τ-estimates. Computational Statistics & Data Analysis 93:436–455
[31] Salini S, Cerioli A, Laurini F, Riani M (2016) Reliable robust regression diagnostics. International Statistical Review 84(1):99–127
[32] Welsh AH, Carroll RJ, Ruppert D (1994) Fitting heteroscedastic regression models. Journal of the American Statistical Association 89:100–116
[33] White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrica, 93:817–838
[34] Zhao W, Lian H, Liang H (2017) Quantile regression for the single-index coefficient model. Bernoulli, 23(3), 1997-2027.