

Reliable Robust Regression Diagnostics

**Silvia Salini¹, Andrea Cerioli², Fabrizio Laurini²
and Marco Riani²**

¹*Department of Economics Management and Quantitative Methods, University of Milan,
Milan, Italy*

E-mail: silvia.salini@unimi.it

²*Department of Economics, University of Parma, Parma, Italy*

Summary

Motivated by the requirement of controlling the number of false discoveries that arises in several application fields, we study the behaviour of diagnostic procedures obtained from popular high-breakdown regression estimators when no outlier is present in the data. We find that the empirical error rates for many of the available techniques are surprisingly far from the prescribed nominal level. Therefore, we propose a simulation-based approach to correct the liberal diagnostics and reach reliable inferences. We provide evidence that our approach performs well in a wide range of settings of practical interest and for a variety of robust regression techniques, thus showing general appeal. We also evaluate the loss of power that can be expected from our corrections under different contamination schemes and show that this loss is often not dramatic. Finally, we detail some possible extensions that may further enhance the applicability of the method.

Key words: Forward search; FSDA toolbox; MM-estimation; outlier detection; S-estimation; test size and power; trimming.

1 Introduction

Much of the recent work in robust statistics has focused on the attempt to reconcile the two enemy brothers of high-breakdown estimation: robustness against a large fraction of masked outliers and good statistical properties, comparable with those of classical estimators, when the normal model for all the data holds. From the point of view of estimation, the goal of this body of work has been the construction of estimators that can achieve both a high-breakdown point and a high efficiency at the normal distribution (Maronna *et al.*, 2006, Section 5.5). A non-exhaustive list of supposedly robust and efficient techniques includes MM-estimators (Maronna & Yohai, 2010; Van Aelst & Willems, 2011), tau-estimators (Van Aelst *et al.*, 2013), the reweighted version of trimmed estimators (Cizek, 2013) and the forward search (Cerioli *et al.*, 2014; Johansen & Nielsen, 2015).

From a diagnostic perspective, reaching satisfactory statistical properties under the normal model also implies good control of the number of false discoveries in situations of practical interest. There are many application fields, such as high-dimensional genomics, quality control, performance assessment and anti-fraud analysis, where such a property is highly desirable (Filzmoser *et al.*, 2008; Cerioli & Farcomeni, 2011; De Battisti & Salini, 2013; Cerioli & Perrotta, 2014). However, high-breakdown techniques tend to produce a potentially large number of spurious outliers. This tendency was first noted by Cook & Hawkins (1990), although in

a somewhat biased context, and then confirmed in subsequent studies, even when ad hoc finite-sample corrections are taken into account; see, for example, Cerioli *et al.* (2009), Fauconnier & Haesbroeck (2009), Cerioli *et al.* (2013b), Lourenco & Pires (2014) and Riani *et al.* (2014a). Cerioli *et al.* (2009), Riani *et al.* (2009) and Cerioli (2010) propose alternative strategies for overcoming this shortcoming in the multivariate framework, while Maronna & Yohai (2010) develop specific corrections for regression MM-estimates when the ratio p/n is large.

The main target of the present work is to address the diagnostic behaviour of high-breakdown techniques at the normal model from a regression perspective, by considering a wide variety of alternative estimators and by computing appropriate corrections when the null performance of the corresponding diagnostic procedures is poor. In particular, we place outlier detection in a testing scenario and develop robust regression diagnostics that are able to control empirical test sizes at a prescribed level for all the procedures that we analyse. We also evaluate the loss of power that can be expected from our corrections under different contamination schemes, and we show that this loss is often not dramatic.

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

$$y_i = x_i' \beta + \epsilon_i, \quad i = 1, \dots, n, \quad (1.1)$$

where $\beta = (\beta_0, \beta_1, \dots, \beta_{p-1})'$ is a p -dimensional vector of unknown coefficients, $x_i = (1, x_{i1}, \dots, x_{i(p-1)})'$, x_{ij} is the value of explanatory variable X_j , $j = 1, \dots, p-1$, for unit i , $i = 1, \dots, n$, and $\epsilon_1, \dots, \epsilon_n$ are i.i.d. errors such that $\epsilon_i \sim N(0, \sigma^2)$. High-breakdown regression is needed especially when $X_1, \dots, X_{p-1}$ are random variables that may contain leverage points, because of contamination or recording errors. We thus fit model (1.1) by conditioning on the observed vectors $x_1, \dots, x_n$, although, for simplicity, we do not make this step explicit in our notation. Because we are interested in the behaviour of methods under ‘good’ data structures, we mainly consider the case where the observed values of $X_1, \dots, X_{p-1}$ are a sample from a multivariate normal distribution. We also briefly address the effect of alternative distributions under the conditions described, for example, in Section 5.2 of Maronna *et al.* (2006). The specific effect of contamination on the explanatory variables is analysed towards the end of this work, when we study the diagnostic power of our corrected robust procedures.

Let

$$\hat{y}_i = x_i' \hat{\beta}, \quad i = 1, \dots, n \quad (1.2)$$

be the fitted version of Equation (1.1), where the estimated vector $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{p-1})'$ has been obtained by some (either robust or non-robust) methods. Our basic diagnostic quantities are the squared scaled residuals (Atkinson & Riani, 2000; Rousseeuw & Hubert, 2011)

$$\hat{s}_i^2 = \frac{\hat{\epsilon}_i^2}{\hat{\sigma}^2}, \quad i = 1, \dots, n, \quad (1.3)$$

where $\hat{\epsilon}_i = y_i - \hat{y}_i$ and $\hat{\sigma}^2$ is the model-based estimate of the error variance σ^2 . We compute each diagnostic $\hat{s}_i^2$ to test the null hypothesis

$$H_{0,i} : y_i \sim N(x_i' \beta, \sigma^2), \quad (1.4)$$

which states that observation y_i comes from the postulated normal regression model. A common diagnostic practice is to repeat the test of (1.4) for each observation y_i , $i = 1, \dots, n$. If

the empirical test size is close to the nominal one, say α , we should thus expect a proportion of false outliers close to α for any uncontaminated data set. Furthermore, when multiple testing is an issue, as in many of the application settings mentioned earlier, we can also use the whole set of scaled residuals (1.3) to test the hypothesis that no contamination is present in the data:

$$H_{0,\text{All}} : H_{0,1} \cap \dots \cap H_{0,n}. \quad (1.5)$$

The purpose of high-breakdown estimation is to ensure that fit (1.2) is unaffected if a fraction τ of observations in the sample is replaced by arbitrarily large values, either in the response or in the explanatory variables, provided that $\tau < 0.5$. The same property also extends to the squared scaled residuals (1.3). When $n \rightarrow \infty$ and p is fixed, it is straightforward to see that

$$\hat{\epsilon}_i - \epsilon_i \xrightarrow{p} 0 \text{ if model (1.1) holds,} \quad (1.6)$$

whenever $\hat{\beta}$ is a consistent estimator of β . Many high-breakdown regression estimators have been shown to be consistent under general conditions; see, for example, Rousseeuw & Leroy (1987), Davies (1990). Convergence (1.6), together with consistency of $\hat{\sigma}^2$, thus provides the basis for testing (1.4) and (1.5) through the asymptotic approximation

$$\hat{s}_i \simeq N(0, 1), \quad (1.7)$$

or, equivalently, through

$$\hat{s}_i^2 \simeq \chi_1^2. \quad (1.8)$$

Examples of diagnostic uses of (1.7) and (1.8), for instance, in Q - Q plots of (squared) scaled residuals and for individual outlier identification, abound in the literature; see, for example, Chapters 4 and 5 in Maronna *et al.* (2006) or Hubert *et al.* (2008). However, the reference $N(0, 1)$ or χ_1^2 distributions hold only in the limit and may provide poor approximations in small or moderate samples, thus leading to increased test sizes. Additional problems may occur because of the negative finite-sample bias typically associated with high-breakdown estimates of σ^2 (Pison *et al.*, 2002) and to the effect of inappropriate choices in the algorithm used to compute $\hat{\beta}$ and $\hat{\sigma}^2$ (Hawkins & Olive, 2002). As a result, the exact finite-sample distribution of diagnostics (1.3) is unknown, and it seems difficult to derive it analytically.

One goal of this paper is to investigate to what extent the most popular high-breakdown regression methods provide accurate testing of both (1.4) and (1.5) when the squared scaled residuals $\hat{s}_i^2$ are adopted. Because robust and efficient procedures should possess good statistical properties at the normal model, we take the performance of classical scaled ordinary least squares (OLS) residuals as our benchmark when all observations follow model (1.1), which will be assumed as the data generating process for uncontaminated data. As a by-product, with our comparisons, it is possible to assess the effect that different choices of tuning constants may have on the empirical test sizes. We will see that the empirical performance of many robust techniques is far from being satisfactory, when examined in our testing framework, especially if the focus is on the intersection hypothesis $H_{0,\text{All}}$. We investigate some of the reasons of this behaviour and show that underestimation of the scale variance σ^2 has a major effect for all robust regression techniques.

We then suggest a method for correcting the estimated scaled residuals (1.3), which leads to reliable tests, with empirical sizes of the same order of magnitude as those obtained through

OLS estimation. Our procedure is based on a combination of Monte Carlo simulation and parametric interpolation. It is simple to implement in many situations of practical interest, because we provide a method to build proper tuning coefficients for $50 \leq n \leq 500$ and $2 \leq p \leq 10$. Furthermore, extrapolation to values of n and p outside these ranges often yields satisfactory approximations. Our approach is also simple enough to be extended to other robust regression procedures that are not considered in this work. Finally, we evaluate the loss of power that can be expected from our corrections under different contamination schemes, either in the response or in the explanatory variables. We show that this loss is usually a reasonable price to pay in order to ensure control of the number of false discoveries.

The structure of the paper is as follows. In Section 2, we sketch the high-breakdown regression techniques that we consider in our work. The performance of such measures when the null model (1.1) holds is evaluated in Section 3 under different (n, p) configurations. In the same section, we also assess the adequacy of approximation (1.8) for the different forms of squared residuals, and we highlight the effect of underestimating σ^2 for several high-breakdown regression procedures. Our proposal for obtaining reliable diagnostics is described in Section 4, where we also check its robustness to model assumptions and outline some possible extensions. The diagnostic power of our corrected robust procedures is assessed in Section 5 under different contamination schemes. The paper ends with some concluding remarks in Section 6.

2 High-breakdown Regression

High-breakdown regression methods can be broadly classified into three classes, depending on the nature of the objective function that they aim at optimising (Riani *et al.*, 2014b). We call these classes soft trimming, hard trimming and adaptive hard trimming. All of them have gained much popularity over the years and are now available through a number of computer packages written in different languages. In our study, we adopt the common and flexible computational framework provided by the Forward Search for Data Analysis (FSDA) toolbox of Matlab (Riani *et al.*, 2012), available through <http://www.riani.it>.

Our first aim is to assess the behaviour of popular regression techniques belonging to the three classes. We also evaluate the effect exerted by different choices of some tuning constants required for practical implementation of the methods. In particular, we focus on breakdown point (bdp), efficiency (eff) and, for soft trimming methods, different weight functions. We refer to Riani *et al.* (2014c) for a detailed investigation of the relationships that link these tuning constants. Table 1 provides a summary of the regression procedures that we examine in our work, together with their acronyms. We take the classical OLS estimator as our benchmark under the uncontaminated normal model (1.1). The three classes of estimators are briefly described in the following. We argue that our selection of 25 robust regression procedures provides a wide list of possibilities that should be able to satisfy many practical needs. Nevertheless, we conjecture that our approach to the construction of robust and reliable diagnostics could be easily extended also to cases not considered here.

2.1 Soft Trimming

The family of soft trimming estimators, leading to downweighting of observations by a function ρ , derives from M-estimation. For a given set of residuals $\hat{\epsilon}_1, \dots, \hat{\epsilon}_n$, the M-estimator of scale is defined as the solution to the equation

$$\frac{1}{n} \sum_{i=1}^n \rho \left(\frac{\hat{\epsilon}_i}{\hat{\sigma}} \right) = K, \quad (2.1)$$

Table 1. A summary of regression procedures, for different values of bdp , eff and different weight functions ρ .

Acronym	Description
OLS	Classical ordinary least squares estimator
Sbdp050TB	S-estimator (2.2) with Tukey ρ function (2.3) and $\text{bdp} = 0.5$
Sbdp050HA	S-estimator (2.2) with Hampel ρ function (2.4) and $\text{bdp} = 0.5$
Sbdp050OP	S-estimator (2.2) with optimal ρ function (2.5) and $\text{bdp} = 0.5$
Sbdp050HY	S-estimator (2.2) with hyperbolic tangent ρ function (2.6) and $\text{bdp} = 0.5$
Sbdp025TB	S-estimator (2.2) with Tukey ρ function (2.3) and $\text{bdp} = 0.25$
Sbdp025HA	S-estimator (2.2) with Hampel ρ function (2.4) and $\text{bdp} = 0.25$
Sbdp025OP	S-estimator (2.2) with optimal ρ function (2.5) and $\text{bdp} = 0.25$
Sbdp025HY	S-estimator (2.2) with hyperbolic tangent ρ function (2.6) and $\text{bdp} = 0.25$
MMeff085TB	MM-estimator (2.7) with Tukey ρ function (2.3) and $\text{eff} = 0.85$
MMeff085HA	MM-estimator (2.7) with Hampel ρ function (2.4) and $\text{eff} = 0.85$
MMeff085OP	MM-estimator (2.7) with optimal ρ function (2.5) and $\text{eff} = 0.85$
MMeff085HY	MM-estimator (2.7) with hyperbolic tangent ρ function (2.6) and $\text{eff} = 0.85$
MMeff090TB	MM-estimator (2.7) with Tukey ρ function (2.3) and $\text{eff} = 0.90$
MMeff090HA	MM-estimator (2.7) with Hampel ρ function (2.4) and $\text{eff} = 0.90$
MMeff090OP	MM-estimator (2.7) with optimal ρ function (2.5) and $\text{eff} = 0.90$
MMeff090HY	MM-estimator (2.7) with hyperbolic tangent ρ function (2.6) and $\text{eff} = 0.90$
MMeff095TB	MM-estimator (2.7) with Tukey ρ function (2.3) and $\text{eff} = 0.95$
MMeff095HA	MM-estimator (2.7) with Hampel ρ function (2.4) and $\text{eff} = 0.95$
MMeff095OP	MM-estimator (2.7) with optimal ρ function (2.5) and $\text{eff} = 0.95$
MMeff095HY	MM-estimator (2.7) with hyperbolic tangent ρ function (2.6) and $\text{eff} = 0.95$
LTSbdp050	LTS estimator (2.8) with $\text{bdp} = 0.5$
LTSbdp025	LTS estimator (2.8) with $\text{bdp} = 0.25$
LTSrbdp050	Reweighted LTS estimator with weights (2.10) and $\text{bdp} = 0.5$
LTSrbdp025	Reweighted LTS estimator with weights (2.10) and $\text{bdp} = 0.25$
FS	Forward search regression estimator and outlier detection rule of Riani <i>et al.</i> (2009)

where $0 < K < \sup \rho$ and ρ is a function that reduces the importance of observations with large residuals. The vector $\hat{\beta}$ such that

$$\hat{\beta} = \arg \min \hat{\sigma}, \quad (2.2)$$

where $\hat{\sigma}$ satisfies (2.1), leads to the S-estimate of β and to the associated estimate of scale (Rousseeuw & Leroy, 1987).

To achieve robustness, we must consider a ρ function with 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 $K/\rho(c) = \text{bdp}$, with $0 < \text{bdp} \leq 0.5$.

Then the asymptotic breakdown point of the S-estimator tends to bdp when $n \rightarrow \infty$. As c increases, fewer observations are downweighted so that the estimate of σ^2 approaches that for least squares and $\text{bdp} \rightarrow 0$. For consistency under the normal model (1.1), we also require

$$K = E_{\Phi_{0,1}} \left[\rho \left(\frac{\hat{e}_i}{\hat{\sigma}} \right) \right],$$

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

In our study, we consider the following popular ρ functions:

- Tukey bisquare (TB):

$$\rho(u) = \begin{cases} \frac{u^2}{2} - \frac{u^4}{2c^2} + \frac{u^6}{6c^4} & \text{if } |u| \leq c \\ \frac{c^2}{6} & \text{if } |u| > c, \end{cases} \quad (2.3)$$

- Hampel (HA):

$$\rho(u) = \begin{cases} \frac{1}{2}u^2 & \text{if } |u| \leq c_1 \\ c_1|u| - \frac{1}{2}c_1^2 & \text{if } c_1 < |u| \leq c_2 \\ c_1 \frac{c_3|u| - \frac{1}{2}u^2}{c_3 - c_2} & \text{if } c_2 < |u| \leq c_3 \\ c_1(c_2 + c_3 - c_1) & \text{if } |u| > c_3. \end{cases} \quad (2.4)$$

- Optimal (OP):

$$\rho(u) = \begin{cases} 1.3846 \left(\frac{u}{c}\right)^2 & \text{if } |u| \leq \frac{2}{3}c \\ 0.5514 - 2.6917 \left(\frac{u}{c}\right)^2 + 10.7668 \left(\frac{u}{c}\right)^4 - 11.6640 \left(\frac{u}{c}\right)^6 + \\ + 4.0375 \left(\frac{u}{c}\right)^8 & \text{if } \frac{2}{3}c < |u| \leq c \\ 1 & \text{if } |u| > c. \end{cases} \quad (2.5)$$

- Hyperbolic tangent (HY):

$$\rho(u) = \begin{cases} \frac{1}{2}u^2 & \text{if } |u| \leq c_1 \\ -2\frac{c_2}{c_3} \ln \cosh \left[\frac{1}{2} \sqrt{\frac{(c_4-1)c_3^2}{c_2}} (c - |u|) \right] + \\ + \frac{\xi^2}{2} + 2\frac{c_2}{c_3} \ln \cosh \left[\frac{1}{2} \sqrt{\frac{(c_4-1)c_3^2}{c_2}} (c - \xi) \right] & \text{if } \xi \leq |u| \leq c \\ \frac{\xi^2}{2} + 2\frac{c_2}{c_3} \ln \cosh \left[\frac{1}{2} \sqrt{\frac{(c_4-1)c_3^2}{c_2}} (c - \xi) \right] & \text{if } |u| > c, \end{cases} \quad (2.6)$$

where $0 < \xi < c$ is such that

$$\xi = \sqrt{[c_2(k-1)]} \tanh \left[\frac{1}{2} \sqrt{\frac{(k-1)c_3^2}{c_2}} (c - \xi) \right],$$

c_2 and c_3 satisfy suitable conditions and c_4 is related to the bound in the change of variance curve (Hampel *et al.*, 1981). These constants are computed iteratively, by applying Newton–Raphson steps and numerical integration.

For each ρ function, the value of the breakdown point can be varied by selecting different constants in the corresponding equation.

MM-estimation is a two-step refinement of S-estimation with the aim of achieving high efficiency under model (1.1) while keeping the same breakdown point. For this purpose, in the first stage, the breakdown point is typically set at $\text{bdp} = 0.5$, and an S-estimate of β , say $\hat{\beta}^*$, is

computed. The resulting scale estimate $\hat{\sigma}^*$ is then used for computing the MM-estimate of the regression parameters, which is defined as a local minimum of

$$\frac{1}{n} \sum_{i=1}^n \rho^* \left(\frac{\hat{\epsilon}_i}{\kappa^* \hat{\sigma}^*} \right), \quad (2.7)$$

where ρ^* is a bounded ρ function, possibly different from ρ in (2.2), and κ^* is chosen to provide high efficiency. The minimum is computed iteratively starting from $\hat{\beta}^*$. A value of 0.85 for the efficiency in (2.7) is often recommended (Maronna *et al.*, 2006, p. 126), but, of course, alternative values are possible through the choice of κ^* . For simplicity, we restrict ourselves to the case where the functional form of ρ^* in (2.7) is the same as that of ρ in (2.2).

2.2 Hard Trimming

The most popular trimming method for regression, on which we focus, is least trimmed squares (LTS). In this method, the amount of trimming is determined by the choice of the trimming parameter h , $\lfloor (n + p + 1)/2 \rfloor \leq h \leq n$, where $\lfloor \cdot \rfloor$ is the floor function, which is specified in advance. The LTS estimate is intended to minimise the sum of squares of the h smallest residuals:

$$\hat{\beta} = \arg \min \sum_{i=1}^h \hat{\epsilon}_{[i]}^2, \quad (2.8)$$

where $\hat{\epsilon}_{[1]}, \dots, \hat{\epsilon}_{[n]}$ are the ordered residuals from fit (1.2). The corresponding estimate of σ^2 is

$$\hat{\sigma}^2 = k(h, n, p) \frac{1}{h} \sum_{i=1}^h \hat{\epsilon}_{[i]}^2, \quad (2.9)$$

where $k(h, n, p)$ is a correction factor, depending on h , n and p , which ensures consistency under the normal model (1.1) and also provides a finite-sample bias adjustment (Croux & Haesbroeck, 1999; Pison *et al.*, 2002).

The most resistant choice against contamination is $h = \lfloor (n + p + 1)/2 \rfloor$, which yields asymptotically $\text{bdp} = 0.5$ (Rousseeuw & Leroy, 1987). An alternative recommendation, which is often seen as a good compromise between robustness and efficiency, corresponds to choosing $\text{bdp} = 0.25$ (Hubert *et al.*, 2008, p. 95).

2.3 Adaptive Hard Trimming

The purpose of adaptive trimming methods is to select the amount of trimming h from the data. This goal requires more than one step.

In the reweighted version of the LTS estimator (LTSr), we first set $h = \lfloor (n + p + 1)/2 \rfloor$, for which preliminary and very robust estimates of β and σ , say $\hat{\beta}^*$ and $\hat{\sigma}^*$, are computed through (2.8) and (2.9). In the second stage, we obtain the final parameter estimates by applying a weighted OLS approach, with weights defined as follows:

$$w_i = \begin{cases} 1 & \text{if } \left(\frac{\hat{\epsilon}_i^*}{\hat{\sigma}^*} \right)^2 \leq \chi_{1,0.975}^2 \\ 0 & \text{if } \left(\frac{\hat{\epsilon}_i^*}{\hat{\sigma}^*} \right)^2 > \chi_{1,0.975}^2, \end{cases} \quad (2.10)$$

where $\hat{\epsilon}_i^* = y_i - \hat{y}_i^*$, $\hat{y}_i^* = x_i' \hat{\beta}^*$ and $\chi_{1,\gamma}^2$ denotes the γ -th quantile of the χ_1^2 distribution. This weighting scheme clearly relies on the adequacy of the asymptotic approximation (1.8), which is commonly used even when the error distribution is not normal.

A fully efficient iterative scheme that is not restricted to only two steps is provided by the forward search (FS). In this approach, data analysis starts from a very robust fit to a few, carefully selected observations found, for example, by LTS with $h = \lfloor (n + p + 1)/2 \rfloor$. In the initial step of the FS, we take the m_0 observations with the smallest squared residuals from the robust fit. Typically, $m_0 = p + 1$ or slightly larger. The number of observations used in fitting then increases by including those with the smallest squared residuals at the previous step, until all units are included. At step $m + 1$ of the FS, the parameter estimates are again computed through a weighted OLS approach but now with weights

$$w_i(m + 1) = \begin{cases} 1 & \text{if } \hat{\epsilon}_i(m)^2 \leq \hat{\epsilon}_{[m+1]}(m)^2 \\ 0 & \text{otherwise,} \end{cases}$$

where $\hat{\epsilon}_i(m) = y_i - \hat{y}(m)_i$, $\hat{\epsilon}_{[1]}(m)^2, \dots, \hat{\epsilon}_{[n]}(m)^2$ are the ordered squared residuals, $\hat{y}_i(m) = x_i' \hat{\beta}(m)$ and $\hat{\beta}(m)$ is the estimate of β computed at step m for which the matrix of regressors is $X(m)$, for $m = m_0, \dots, n - 1$.

The FS provides a set of $n - m_0 + 1$ residuals for each observation, starting from the initial LTS-based fit and ending with the classical OLS fit when $m = n$. Exploration of this set is very useful for diagnostic purposes (Riani *et al.*, 2014a; Riani *et al.*, 2014b) but requires a summary for precise outlier identification. While for the other robust procedures, it is possible to sequentially test all the individual outlier hypotheses (1.4) by means of the squared scaled residuals $\hat{s}_i^2$, $i = 1, \dots, n$, the sequence of steps implied by the FS would make such tests cumbersome. Therefore, we follow the rules established by Riani *et al.* (2009) that first focus on the simultaneous hypothesis (1.5) of no outliers and then move to individual outlier identification.

More precisely, let $S^*(m)$ be the subset of size m found by FS. Ordinary residuals $\hat{\epsilon}_i(m)$ can be calculated for all observations including those not in $S^*(m)$. To test for outliers, we use instead deletion residuals that are calculated for the $n - m$ observations not in $S^*(m)$. These residuals, which form the maximum likelihood tests for the outlyingness of individual observations, are

$$r_i(m) = \frac{y_i - x_i^T \hat{\beta}(m)}{\sqrt{\hat{\sigma}^2(m)\{1 + h_i(m)\}}} = \frac{\hat{\epsilon}_i(m)}{\sqrt{\hat{\sigma}^2(m)\{1 + h_i(m)\}}}, \quad (2.11)$$

where the leverage $h_i(m) = x_i^T \{X(m)^T X(m)\}^{-1} x_i$. Let the observation nearest to those forming $S^*(m)$ be $i_{\min}$, where

$$i_{\min} = \arg \min_{i \notin S^*(m)} |r_i(m)|.$$

To test whether observation $i_{\min}$ is an outlier, we use the absolute value of the minimum deletion residual

$$r_{\min}(m) = \frac{e_{i_{\min}}(m)}{\sqrt{s^2(m)\{1 + h_{i_{\min}}(m)\}}}, \quad (2.12)$$

as a test statistic. If the absolute value of (2.12) is too large, the observation $i_{\min}$ is considered to be an outlier, as well as all other observations not in $S^*(m)$. The test statistic (2.12) is the

$(m + 1)$ st ordered value of the absolute deletion residuals. We can therefore use distributional results to obtain pointwise envelopes as the subset size increases (Riani *et al.*, 2009) based on the scaled F distribution. If we had an unbiased estimator of σ^2 , the envelopes would be given by $y_{m+1,n;\gamma}$ for $m = m_0, \dots, n - 1$, the quantile of order γ of the distribution of the absolute value of (2.12) at step m . However, the estimator $s^2(m^*)$ is based on the central m observations from a normal sample—strictly the m observations with smallest squared residuals based on the parameter estimates from $S^*(m - 1)$. The variance of the truncated normal distribution containing the central m/n portion of the full distribution is

$$\sigma_T^2(m) = 1 - \frac{2n}{m} \Phi^{-1} \left(\frac{n+m}{2n} \right) \phi \left\{ \Phi^{-1} \left(\frac{n+m}{2n} \right) \right\}, \quad (2.13)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are, respectively, the standard normal density and c.d.f. (see, for example, Johnson *et al.*, 1994, pp. 156–162). Because the outlier tests we are monitoring are divided by an estimate of σ^2 that is too small, we need to scale up the values of the order statistics to obtain the envelopes

$$y_{m+1,n;\gamma}^* = y_{m+1,n;\gamma} / \sigma_T(m).$$

To be specific, in the case of the 99% envelope, $\gamma = 0.99$ corresponds to a nominal pointwise size $\alpha = 1 - \gamma$, which is equal to 1%. We expect, for the particular step m that is considered, to find exceedances of the quantile in a fraction 1% of the samples under the null normal distribution. We however require a samplewise probability of 1% of the false detection of outliers, that is, over all values of m considered in the search. In order to have a test designed to have a simultaneous size of 1%, we can proceed as follows:

- (1) Preliminary detection of signals, by looking at the diagnostic quantities (2.11). A preliminary signal is detected if (2.12) exceeds an appropriate distributional threshold for at least three subsequent values of m . Call this step $m^\dagger$.
- (2) In the second part, we superimpose envelopes for values of n from this point until the first time we introduce an observation we recognise as an outlier. More precisely, in this stage, we look at the actual squared deletion residuals of the units that enter in the fitting subset after step $m^\dagger - 1$ and compare them with thresholds derived from the reduced sample sizes $n^* = m^\dagger - 1, m^\dagger, m^\dagger + 1, \dots$, until a stopping rule is reached. The goal of this stage is to take into account the increased variability of estimates in a ‘clean’ sample of size $n^* < n$.

It is also important to emphasise that the inferential rules adopted in the FS do not involve the asymptotic distribution of the squared scaled residuals, as in (1.8), but rely on a more accurate $F_{1,n-p}$ approximation for the deletion residuals (2.11) in finite samples.

3 False Signals and Empirical Size: How Many Outliers from High-breakdown Regression?

3.1 Simulation Framework

We first analyse the empirical performance, under model (1.1), of all the procedures reported in Table 1, taking the OLS method as our benchmark. The simulation study is built using different sample sizes n and different numbers of parameters p . In particular, we consider the grid defined by

$$n = \{50, 60, 70, 80, 90, 100, 150, 200, 250, 300, 400, 500\} \text{ and } 2 \leq p \leq 10. \quad (3.1)$$

For each pair of n and p , we simulate $K = 10\,000$ replicates from model (1.1), where y is a sample from a standard normal distribution but $X_1, \dots, X_{p-1}$ are fixed for each simulation. The results that we present are based on $X_1, \dots, X_{p-1}$ generated from the standard multivariate normal distribution. We have computed the same tests by generating the explanatory variables from the Student's t and the chi-square distributions, without noticeable differences. We give some evidence of this distributional robustness later in Section 4.3. Results (not reported) for explanatory variables with a richer correlation structure are broadly consistent with our findings; hence, for ease of presentation, we consider only the simplest case where the X_j 's are mutually independent.

For each simulation and each model described in Table 1, we calculate the squared scaled residuals $\hat{s}_i^2$ defined in (1.3). We use these residuals to test the n individual hypotheses (1.4) and the simultaneous null (1.5).

3.2 Estimation of Empirical Sizes

Under model (1.1), our asymptotic threshold to test the n individual hypotheses (1.4) is given by $\chi_{1,0.99}^2$ for all the procedures listed in Table 1. For the simultaneous test of (1.5), we use instead the Bonferroni threshold $\chi_{1,1-(1-0.99/n)}^2$.

Figures 1 and 2 show the empirical (individual and simultaneous) sizes for all procedures in a small sample case, that is, when $n = 50$, for two different numbers of parameters, $p = 2$ and $p = 10$. Figures 3 and 4 provide the same output when the sample size is larger, $n = 200$. The horizontal line represents the nominal (either individual or simultaneous) size of 1%.

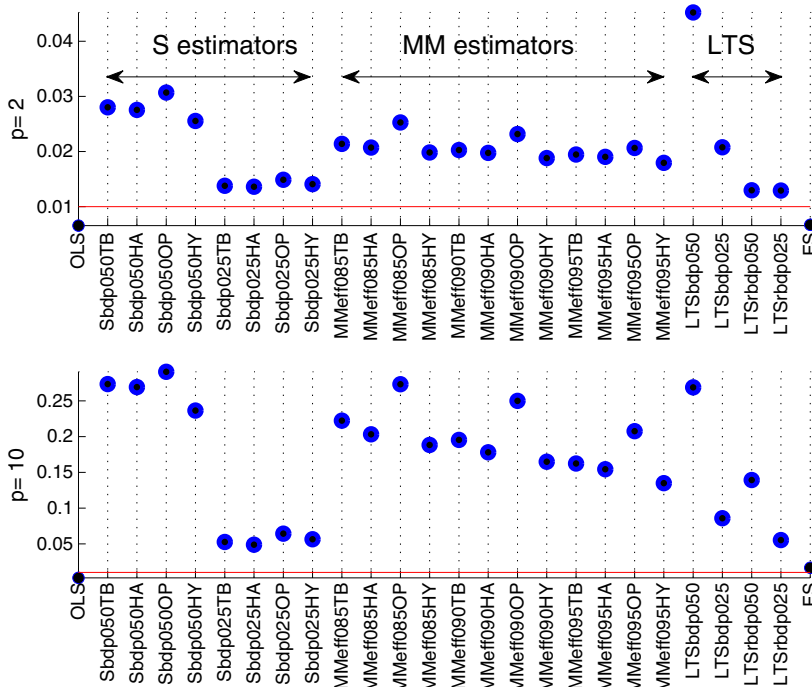


Figure 1. Empirical size of individual outlier tests for the robust procedures given in Table 1 when $n = 50$, $p = 2$ (top) and $p = 10$ (bottom).

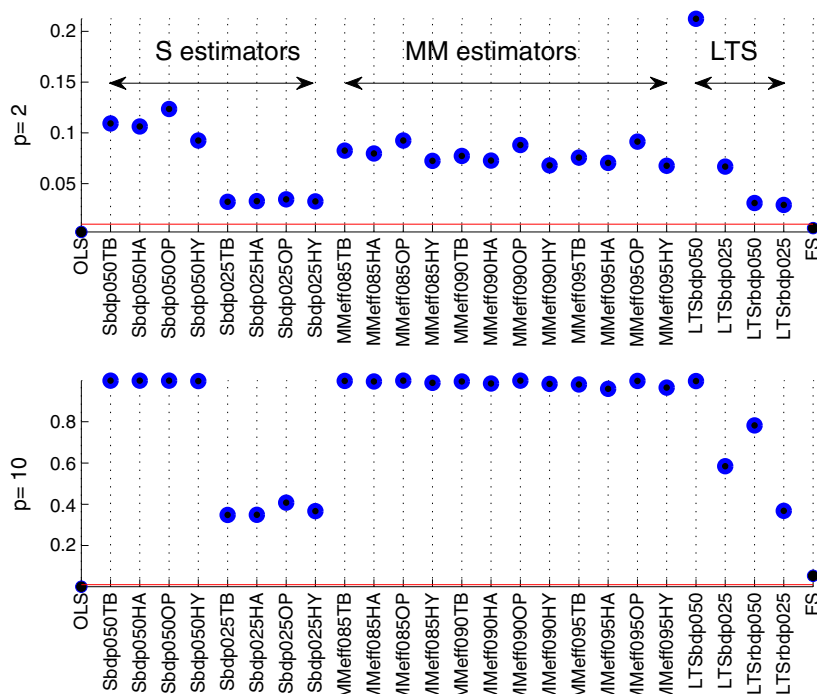


Figure 2. Empirical size of simultaneous outlier tests for the robust procedures given in Table 1 when $n = 50$, $p = 2$ (top) and $p = 10$ (bottom).

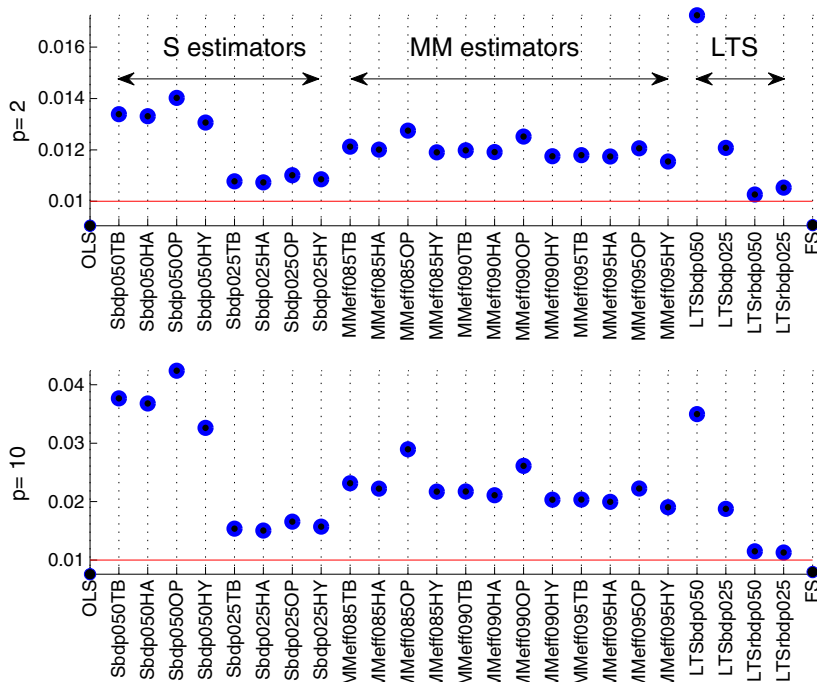


Figure 3. Empirical size of individual outlier tests for the robust procedures given in Table 1 when $n = 200$, $p = 2$ (top) and $p = 10$ (bottom).

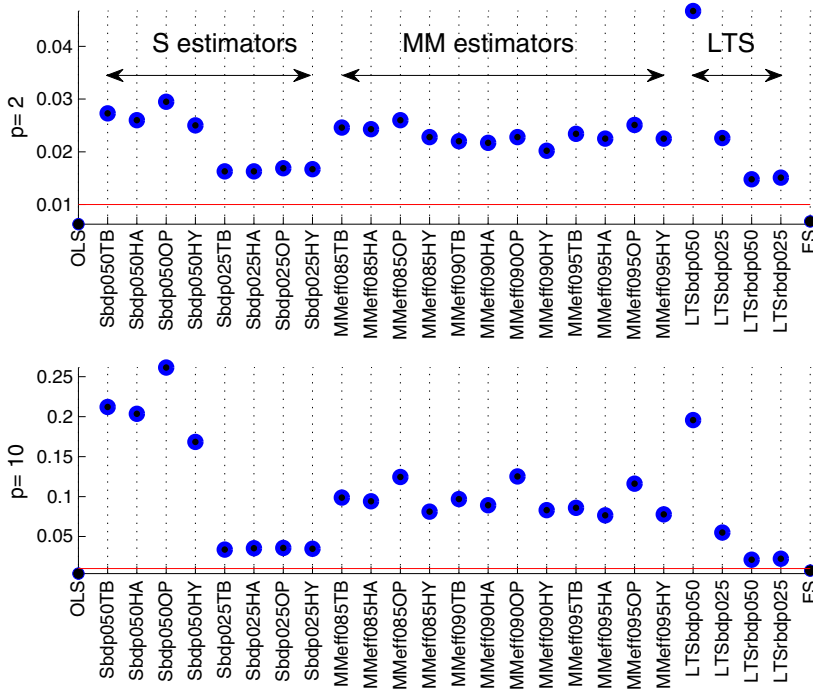


Figure 4. Empirical size of simultaneous outlier tests for the robust procedures given in Table 1 when $n = 200$, $p = 2$ (top) and $p = 10$ (bottom).

The black dot, in the bottom left part shown in Figures 1 to 4, represents the empirical size for OLS. Hence, the OLS diagnostics have empirical size close to the nominal level (under both the individual and the simultaneous tests) even for moderate and small sample sizes. The general conclusion about robust methods is that adaptive hard trimming approaches (LTSr and FS) are systematically much better than the others. In particular, the FS is the only method that provides results very close to those of OLS in any simulation setting. One motivation that justifies the improved performance of the FS over LTSr is the use of more accurate distributions in finite samples.

As expected, the behaviour of many robust diagnostic techniques is worse when testing (1.5). For simultaneous testing, an extreme percentile of the distribution is needed, and hence, the approximation to the distribution of $\hat{s}_i^2$ must be good in the tail. When n is small and p is large (e.g. $n = 50$ and $p = 10$), the empirical sizes of the test of (1.5) are close to 1 for all S-estimators with $\text{bdp} = 0.5$ and for all MM-estimators that are based on them, regardless of the efficiency level. Also, LTS with $\text{bdp} = 0.5$ leads to an estimated size very close to 1, even if the finite-sample bias adjustment of Pison *et al.* (2002) is applied to it; see Equation (2.9). One reason is that this simulation-based adjustment aims at correcting the bias in the estimate of σ^2 under model (1.1), but it is not effective in the tail of the distribution of the squared scaled residuals. When testing the individual hypotheses (1.4) and when n increases, empirical sizes provided by MM-estimators are intermediate between those of S-estimators with $\text{bdp} = 0.5$ and $\text{bdp} = 0.25$. This behaviour, combined with bad performance under (1.5), shows that it is difficult for MM-estimators to recover the efficiency that is lost at the first step, where S-estimators with $\text{bdp} = 0.5$ are computed.

The effect of bdp is evident both on S and on LTS estimators. This confirms the popular indication that $\text{bdp} = 0.5$ should be chosen only when a massive contamination is expected

(Hubert *et al.*, 2008, p. 95). When the ratio p/n is small and/or simultaneous testing is considered, both hard and soft trimming techniques should be run with $\text{bdp} = 0.25$ in order to achieve reasonable performance under the normal model. The choice of the ρ function does not seem to have a noticeable effect on test sizes. Empirical sizes are generally larger for the optimal ρ function (2.5) and smaller for the hyperbolic tangent (2.6). Moreover, the differences are less evident when n increases and p decreases, and, in general, less important than those observed for different values of bdp .

3.3 Distributional Issues

Another way to evaluate the adequacy of the squared scaled residuals (1.3) under a normal regression model is to assess the goodness-of-fit of their empirical distribution to the reference asymptotic χ_1^2 distribution. This alternative examination focuses on all estimated residuals and not only on the largest ones. Therefore, it can provide useful information for more general diagnostic checking than outlier detection, for example, for exploratory analysis through Q - Q plots. If the correct distribution of the squared scaled residuals is χ_1^2 , as implied by asymptotic theory for OLS and robust regression, then p -values will be uniformly distributed. Hence, our null distributional hypothesis is

$$H_{0,\text{GOF}} : \{P(\chi_1^2 > s_1^2) \dots P(\chi_1^2 > s_n^2)\} \text{ is an i.i.d. sample from } U(0, 1), \quad (3.2)$$

where, $s_1^2, \dots, s_n^2$ denote the observed values of $\hat{s}_1^2, \dots, \hat{s}_n^2$.

We test the uniformity hypothesis (3.2) through the Pearson chi-square statistic. We prefer this test because it is less sensitive to the effect of parameter estimation than the Kolmogorov–Smirnov statistic. See Cerioli *et al.* (2013a) for details in the multivariate setting. Indeed, our goodness-of-fit analysis is performed on the estimated residuals from fit (1.2), and the finite-sample correlation induced by estimation can lead to remarkable conservativeness in the Kolmogorov–Smirnov test of (3.2). For the Pearson chi-square, the number of classes r , used to build the contingency table, is guided by the standard Mann–Wald recommendation, that is, $r = \lceil 2n^{2/5} \rceil$, with $\lceil \cdot \rceil$ denoting the ceiling function. Other sensible choices of r give results broadly consistent with our findings discussed in the following. We have also performed the same kind of distributional test using $F_{1,n-p}$ as the reference distribution, instead of χ_1^2 . This small sample approximation accommodates the effect of estimation of the variance σ^2 , which makes the tails of the distribution of scaled residuals heavier than Gaussian. However, we have not found remarkable differences in our goodness-of-fit findings. We fix a nominal level of 1% in all our tests of (3.2) and report results in percentages (with entries smaller than 2% in bold) to improve the overall readability.

We first report our results for the case of a small sample size in Table 2, where $n = 100$, and all values of p in our simulation grid (3.1) are studied. Results from Table 2 confirm the main findings of Section 3.2. For such a sample size, the procedures for which $\text{bdp} = 0.25$ perform reasonably well, even for comparatively high values of p , with some minor exceptions for the LTS. For robust methods with $\text{bdp} = 0.5$, the uniformity hypothesis is rejected far too often, except for FS, and for some values of p in the reweighted LTS. Again, the squared scaled residuals from this adaptive trimming technique are the only high-breakdown version of (1.3) to have distributional features comparable with those of their OLS counterpart. Notice that performance for many MM-estimators quickly degenerates when p grows, showing that their asymptotic convergence is slow, regardless of the choice of the ρ function. Performance generally improves as n becomes larger, as shown in Table 3 for $n = 200$. Nevertheless, it is

Table 2. Empirical size (in percentage) of the Pearson test for testing the hypothesis of uniformity of p -values (3.2) for a nominal level of 1%.

methods ($n = 100$)	$p = 2$	$p = 3$	$p = 4$	$p = 5$	$p = 6$	$p = 7$	$p = 8$	$p = 9$	$p = 10$
OLS	0.69	0.6	0.61	0.69	0.66	0.68	0.86	0.82	0.84
Sbdp050TB	2.32	5.14	10.96	20.77	36.69	56	74.76	88.36	96.17
Sbdp050HA	2.13	4.67	10.35	19.96	35.04	55.15	73.34	88.22	95.87
Sbdp050OP	3.25	7.59	15.95	29.72	49.39	69.97	85.27	94.62	98.66
Sbdp050HY	2.06	4.52	9.76	21.87	39.53	61.12	80.24	92.88	98.19
Sbdp025TB	0.68	0.55	0.58	0.62	0.75	0.85	0.84	0.92	1.09
Sbdp025HA	0.75	0.63	0.58	0.62	0.86	1.02	1.08	1.08	1.32
Sbdp025OP	0.59	0.64	0.66	0.59	0.78	0.8	0.85	0.53	0.91
Sbdp025HY	0.65	0.46	0.6	0.74	0.98	0.92	0.78	0.9	1.14
MMeff085TB	1.36	1.47	1.72	2.16	3.32	4.79	7.41	10.12	15.93
MMeff085HA	1.09	1.33	1.68	2.23	3.37	4.82	6.97	10.61	16.53
MMeff085OP	1.24	1.53	2.23	2.45	3.26	4.55	6.73	10.15	15.96
MMeff085HY	0.92	1.08	1.32	1.44	2.14	2.95	4.15	5.49	8.4
MMeff090TB	1.31	1.45	1.71	2.1	2.81	4.04	6.22	9.54	14.98
MMeff090HA	1.24	1.36	1.5	1.91	2.94	3.77	5.85	8.96	13.84
MMeff090OP	1.19	1.63	2.22	2.73	3.87	5.95	9.1	14.25	23.09
MMeff090HY	0.99	0.92	1.1	1.27	1.69	2.24	3.43	4.37	6.47
MMeff095TB	1.31	1.45	1.71	2.1	2.81	4.04	6.22	9.54	14.98
MMeff095HA	1.24	1.36	1.5	1.91	2.94	3.77	5.85	8.96	13.84
MMeff095OP	1.19	1.63	2.22	2.73	3.87	5.95	9.1	14.25	23.09
MMeff095HY	0.99	0.92	1.1	1.27	1.69	2.24	3.43	4.37	6.47
LTSbdp050	10.5	21.64	34.41	44.61	54.66	63.27	70.81	78.3	85.62
LTSbdp025	1.14	1.51	2.17	3.1	4.49	6.02	9.26	13.46	19.67
LTSrbdp050	0.56	0.76	0.86	1.02	1.35	1.64	2.14	2.63	3.09
LTSrbdp025	0.73	0.63	0.64	0.71	0.88	0.86	0.79	0.92	0.97
FS	0.67	0.61	0.61	0.73	0.76	0.82	1.11	1.24	1.58

The sample size is $n = 100$, and all values of p in grid (3.1) are considered. Boldface entries are for percentages smaller than 2%.

remarkable that for all the values of p in the grid, the estimated size of FS becomes virtually indistinguishable from that of OLS, which represents the benchmark without contamination.

We also provide a graphical diagnostic to check the normality of S and MM residuals. We limit our attention to the case of $p = 10$ and $n = 100$, but we found similar behaviour for other choices of p/n . Our findings are displayed in Figure 5 where, from top to bottom, respectively, we consider MM- and S-estimators. Grey solid lines provide 95% envelopes from the Gaussian distribution, which have to be contrasted with the empirical 95% confidence bands of robust methods (MM and S), which have black dashed lines. In all panels, the ‘central line’ for either the normal distribution or the robust empirical method corresponds to the median. The left-hand panels of Figure 5 have the ordered index position in the abscissa, whereas the right-hand panels have the empirical quantiles. It is hard to understand what is the true underlying distribution of robust MM and S residuals, but it is clear that such residuals are far from Gaussian.

3.4 The Effect of the Scale Estimate

It is known that the main inferential problem with hard trimming procedures is the fact that $\hat{\sigma}^2$ underestimates the true residual variance σ^2 when n is small or moderate, even if consistency correction is applied to it. For instance, this negative bias leads Pison *et al.* (2002) to include a simulation-based finite-sample correction in their LTS estimate of Equation (2.9). García-Escudero & Gordaliza (2005) and García-Escudero & Gordaliza (2006) reach similar conclusions in the multivariate setting. In what follows, we further investigate the underestimation effect for hard trimming estimates of σ^2 , and we extend the results to the case of soft

Table 3. Empirical sizes (in percentage) as in Table 2 but now for $n = 200$.

Methods ($n = 200$)	$p = 2$	$p = 3$	$p = 4$	$p = 5$	$p = 6$	$p = 7$	$p = 8$	$p = 9$	$p = 10$
OLS	0.76	0.8	0.69	0.74	0.78	0.77	0.81	0.71	0.67
Sbdp050TB	1.72	2.88	4.69	8.19	12.95	20.87	31.73	45.37	60.57
Sbdp050HA	1.77	2.9	4.65	7.57	12.22	19.97	30.71	43.9	58.81
Sbdp050OP	2.45	3.94	7.67	13.11	21.29	33.85	48.82	64.76	77.76
Sbdp050HY	1.4	2.63	4.45	8.41	14.1	24.28	38.22	53.36	70.39
Sbdp025TB	0.74	0.75	0.72	0.77	0.68	0.56	0.78	0.96	1.12
Sbdp025HA	0.8	0.75	0.72	0.73	0.97	0.75	0.82	1.02	1.17
Sbdp025OP	0.61	0.68	0.56	0.77	0.77	0.71	0.87	0.91	1
Sbdp025HY	0.67	0.7	0.59	0.83	0.86	0.81	0.82	0.9	1.06
MMeff085TB	1.32	1.51	1.45	1.75	1.99	2.52	3.39	4.02	5.25
MMeff085HA	1.35	1.48	1.61	1.76	1.73	2.31	3.21	3.49	4.66
MMeff085OP	1.47	1.32	1.5	2.02	1.93	2.19	3.46	3.73	4.51
MMeff085HY	1.09	1.27	1.12	1.39	1.2	1.54	2.11	2.48	2.97
MMeff090TB	1.22	1.35	1.41	1.63	1.94	2.18	2.82	3.29	3.72
MMeff090HA	1.29	1.33	1.44	1.55	1.73	2.01	2.68	3.33	3.87
MMeff090OP	1.44	1.48	1.6	1.94	2.06	2.4	3.42	4.53	5.61
MMeff090HY	0.92	1.18	1.08	1.16	1.34	1.46	1.82	2.02	2.47
MMeff095TB	1.22	1.35	1.41	1.63	1.94	2.18	2.82	3.29	3.72
MMeff095HA	1.29	1.33	1.44	1.55	1.73	2.01	2.68	3.33	3.87
MMeff095OP	1.44	1.48	1.6	1.94	2.06	2.4	3.42	4.53	5.61
MMeff095HY	0.92	1.18	1.08	1.16	1.34	1.46	1.82	2.02	2.47
LTSbdp050	7.52	13.02	18.2	22.75	26.63	29.44	34.04	40.12	44.78
LTSbdp025	1.1	1.32	1.68	1.96	2.54	3.17	4.25	5.27	8.04
LTSrbdp050	0.62	0.79	0.84	0.88	0.63	0.67	0.87	1.1	0.97
LTSrbdp025	0.62	0.65	0.65	0.72	0.72	0.7	0.75	0.86	0.83
FS	0.75	0.81	0.73	0.74	0.78	0.81	0.9	0.82	0.79

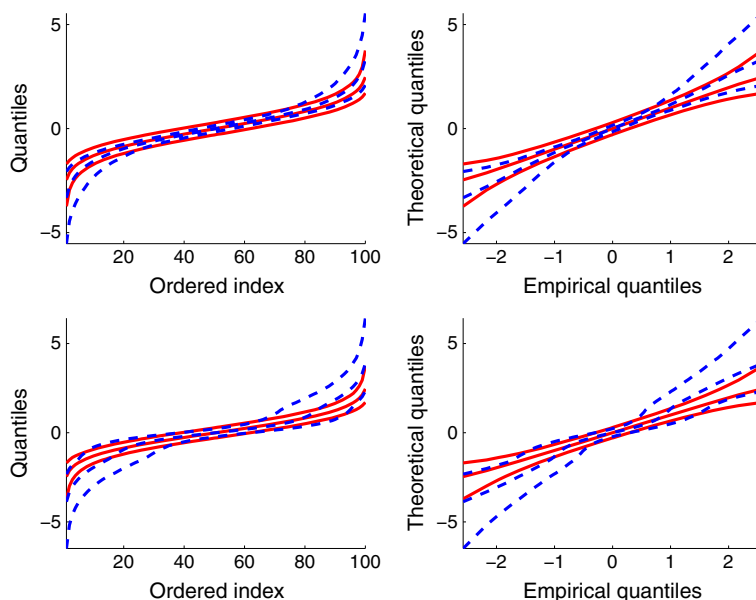


Figure 5. Left panels: empirical (dashed lines) and theoretical (solid lines) quantiles versus ordered index. Right panels: empirical quantiles versus theoretical quantiles ($Q-Q$ plot). MM-estimator *MMeff085HA* and (top) and *S*-estimator *Sbdp050TB* (bottom) for $n = 100$ and $p = 10$.

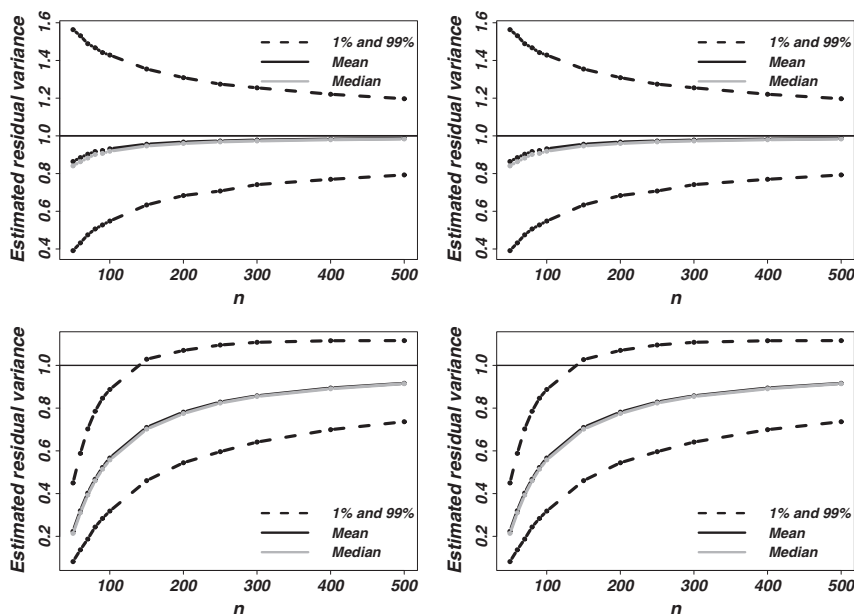


Figure 6. Monte Carlo estimate of the error variance σ^2 . The true value is $\sigma = 1$ (light black straight solid line in all panels). The top left panel is for the estimator Sbdp050OP, whereas the top right is for the estimator MMeff090OP. Both top panels refer to the case $p = 2$. The bottom left and bottom right panels refer to the same estimators but for $p = 10$.

trimming regression methods, for which we quantify the extent of such a negative bias in the estimation of σ^2 .

To assess the extent of this bias, we consider the two estimators Sbdp050OP and MMeff090OP. Figure 6 shows Monte Carlo summaries of our simulations for the two estimators: Sbdp050OP and MMeff090OP. For all panels, the dashed lines represent the 99% and 1% quantiles (from top to bottom, respectively) of the computed estimates of σ^2 with a robust method. The thin black horizontal straight line is the true value of $\sigma^2 = 1$ from which we simulated our data, which represents the benchmark. The heavy solid lines are for the sample average (black) and sample median (grey). The top panels of Figure 6 are for $p = 2$ and the bottom panels for $p = 10$. The convergence to the true value is quite slow, with higher bias when p grows. Other estimators (not shown here) display minor differences, but they all share the same common background. It is interesting to remark that the underestimation bias is still present in the LTS estimate of σ^2 even if the small sample correction factors of Pison *et al.* (2002) is applied.

Test sizes are far from the nominal level mostly because of a biased estimate of σ^2 . Using our correction factors (to be numerically introduced and theoretically motivated in Section 4) can mitigate the effect of such bias. Figure 7 shows the relationship between our correction factors and the underestimation effect on σ^2 (for ease of comparison, we show the graphical results for Sbdp050OP and MMeff090OP, but similar patterns are displayed by other robust methods). The grey lines are all referred to the sample average of $\hat{\sigma}^2$ from residuals of our simulations. The values of these averages are essentially equivalent for the robust estimators Sbdp050OP and MMeff090OP. In the top left panel of Figure 7, we have the graphical results for the individual correction factors and $p = 2$, with the solid black line associated with the MMeff090OP estimator and the dashed black line to the Sbdp050OP estimator. In the top right panel of Figure 7, it is displayed the effect when the simultaneous test is considered. The bottom panels of Figure 7 are equivalent to the top panels but now with $p = 10$. The graphical

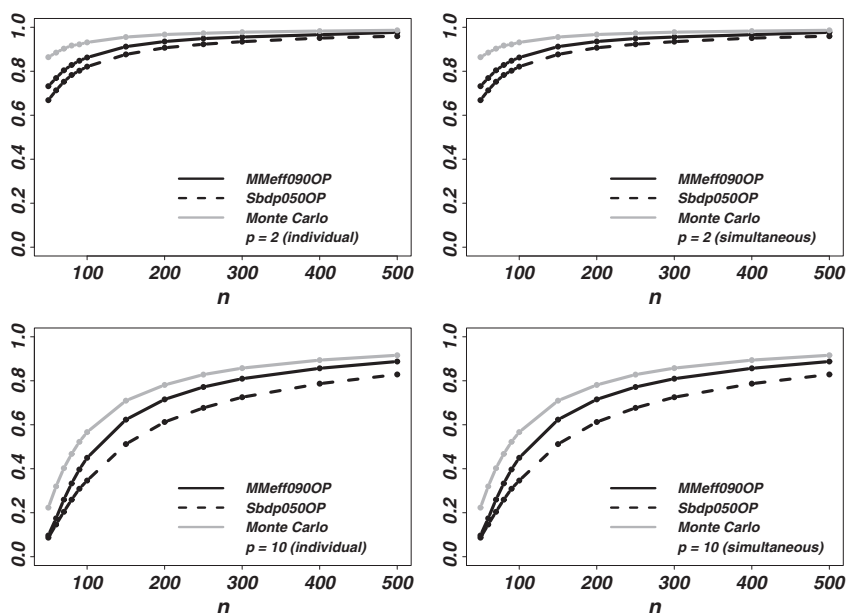


Figure 7. Relationship between our correction factors (to be introduced in Section 4) and the underestimation effect on σ^2 . Top left panel shows the results for the Sbdp050OP and MMeff090OP estimators for the individual tests and $p = 2$ (black dashed and solid lines, respectively). The grey line is the average from our simulations. Top right panel is equivalent to the top left but for simultaneous tests. The bottom panels display the same results for $p = 10$.

results from Figure 7 are suggesting that the bias in estimating σ^2 is responsible for the bad performance, in terms of low size of the test, for many robust methods. It is probably the main source for such bad performance, but clearly, it is not the only factor.

A major consequence of underestimating the residual scale is that the number of outliers flagged in each sample is typically far from the one predicted from the binomial distribution, which is the distribution for the number of outliers under the normal model (1.1) with σ^2 known. On the contrary, we see that the number of outliers detected from the FS is much closer to its theoretical counterpart, thanks to the reduced bias in the resulting estimate of σ^2 and to the use of more accurate thresholds based on the $F_{1,n-p}$ distribution, like in the case of OLS residuals.

4 Reliable Robust Diagnostics

4.1 Estimate of the Appropriate Correction Factors

It is clear from the results reported in Section 3 that the squared scaled residuals (1.3) obtained from the vast majority of high-breakdown techniques need to be adjusted in order to keep false outlier detections under control, with empirical rates close to the nominal ones. Our proposal is to scale each value $\hat{s}_i^2$, $i = 1, \dots, n$, by a fixed method-dependent constant, such that the empirical estimated size of each outlier identification rule becomes close to the nominal 1%. For this purpose, define $\varphi_{i,\gamma}$ to be the non-random value for which

$$P(\varphi_{i,\gamma} \hat{s}_i^2 \leq \chi_{1,\gamma}^2) = \gamma$$

when model (1.1) holds. Note that, asymptotically, the scaling factors $\varphi_{i,\gamma}$ are the same for all observations and converge to 1 for any $\gamma \in (0, 1)$, because of convergence in probability

reported in (1.6). Because the chi-squared threshold derived from (1.8) is also asymptotic, we base our corrections on a similar assumption of identical distribution for the squared scaled residuals. We thus suppose that there exists a value φ_γ , possibly depending on n and p , and converging to 1, as $n \rightarrow \infty$, with p fixed, for which, at least approximately,

$$P(\varphi_\gamma \hat{s}_i^2 \leq \chi_{1,\gamma}^2) = \gamma \quad i = 1, \dots, n. \quad (4.1)$$

Our aim is to find reliable estimates of φ_γ for each high-breakdown version of $\hat{s}_i^2$ and especially for those that have unsatisfactory performance in the simulation study of Section 3. To test the individual hypotheses (1.4), we fix $\gamma = 0.99$, while $\gamma = 1 - (1 - 0.99)/n$ under the simultaneous null (1.5).

It is useful to rewrite (4.1) in terms of the quantiles of the true distribution of the squared scaled residuals $\hat{s}_i^2$. Let ς_γ^2 be the value such that, for given n and p ,

$$P(\hat{s}_i^2 \leq \varsigma_\gamma^2) = \gamma,$$

neglecting again the possible finite-sample effect of individual distributional features. Because $\hat{s}_i^2$ is a continuous random variable under non-degenerate conditions, an equivalent definition of φ_γ is

$$\varphi_\gamma = \frac{\chi_{1,\gamma}^2}{\varsigma_\gamma^2}. \quad (4.2)$$

Our proposal is to obtain, for each robust technique and for the required nominal level implied by γ , a Monte Carlo estimate of ς_γ^2 , which will be plugged into (4.2). Numerically, such coefficients could be computed only for a grid of values of n and p , but in Sections 4.2 and 4.3, we show how to extend the procedure for general use. Our approach differs from that of Pison *et al.* (2002), who aim at correcting the robust LTS estimate of σ^2 under model (1.1). Being focused on the tail of the empirical distribution of the squared scaled residuals, rather than on their mean, our technique is specifically devised for the purpose of outlier detection and may be expected to improve considerably over the inadequate results for LTSbdp050 of Table 1. We are not aware of the availability of related corrections for diagnostic techniques based on soft trimming methods.

Let K denote the number of Monte Carlo simulations performed for each pair (n, p) . Furthermore, let

$$\hat{s}_{[1]}^2(k), \dots, \hat{s}_{[n]}^2(k)$$

be the ordered values of the squared scaled residuals (1.3) computed from the fit at replicate k , for $k = 1, \dots, K$. Define

$$l = \lfloor (n+1)\gamma \rfloor, \quad (4.3)$$

where $\lfloor \cdot \rfloor$ denotes the integer part. If we take

$$\hat{\varsigma}_\gamma^2(k) = \hat{s}_{[l]}^2(k),$$

it is straightforward to see that direct Monte Carlo estimation of ς_γ^2 by averaging the K estimates $\hat{\varsigma}_\gamma^2(k)$ is, in general, not feasible. From definition (4.3), we always have $l = n$ when testing (1.5) at the nominal level of 1%, while $n > 100$ is required to obtain $l < n$ when testing (1.4) at the same nominal level.

Therefore, we follow the approach of Cerioli *et al.* (2009), and we resort to a different Monte Carlo approximation of ς_γ^2 based on pooling all the $n^* = n \times K$ estimates $\hat{s}_{[i]}^2(k)$ in a single sample

$$\hat{S} = \left(\hat{s}_{[1]}^2(1), \dots, \hat{s}_{[1]}^2(K), \hat{s}_{[2]}^2(1), \dots, \hat{s}_{[2]}^2(K), \dots, \hat{s}_{[n]}^2(1), \dots, \hat{s}_{[n]}^2(K) \right)'.$$

Motivated by convergence results (here as $n^* \rightarrow \infty$) for quantiles, even in the case of non-independent observations (see, e.g. Korosok, 2008), we take the γ -th quantile of the pooled sample $\hat{S}$ as our estimate of ς_γ^2 , that is,

$$\hat{\varsigma}_\gamma^2 = \hat{S}_{[l]}, \quad (4.4)$$

where $\hat{S}_{[l]}$ is the l -th ordered value in $\hat{S}$ and l is defined as in (4.3), with n replaced by n^* . Correspondingly,

$$\hat{\phi}_\gamma = \frac{\chi_{1,\gamma}^2}{\hat{\varsigma}_\gamma^2}. \quad (4.5)$$

We compute the estimated thresholds (4.4) and the corresponding scaling factors (4.5) for all the high-breakdown regression techniques that have shown poor performance in Section 3, based on the same set of simulations that are displayed in Figures 1–4. The behaviour of the resulting outlier detection rules is investigated in detail in Sections 4.2 and 4.3. But first, we check how these corrections are affected by n , p and by the choice of *eff*, *bdp* and the ρ function. In general, we see that the estimated thresholds increase if the number of parameters p grows and decrease if the sample dimension n grows, for both the individual and the simultaneous testing frameworks. For instance, we plot in Figure 8 the value of $(\hat{\phi}_\gamma)^{-1}$ for the MM-estimator with Tukey ρ function (2.3). The three pairs of curves refer, respectively, to $p = 2$, $p = 6$ and $p = 10$. The black curves (blue in the online colour figure) refer to *eff* = 0.85 and the grey (red in the online colour figure) refer to *eff* = 0.85. It is clear that the value of *eff* has a negligible effect, when compared with that of p . Similarly, in Figure 9, we present the value of $(\hat{\phi}_\gamma)^{-1}$ for the S-estimator with *bdp* = 0.5 and for the same three values of p as before. The black curves (blue in the online colour figure) refer to the hyperbolic tangent ρ function (2.6) and the grey (red in the online colour figure) refer to the optimal ρ function (2.5). These are the ρ functions that produce the largest difference in the estimated size of outlier tests, as was shown in Section 3.2. Again, the effect connected with p is much greater than that related to the choice of the ρ function. Finally, we inspect the advantages of adaptive trimming in Figure 10, where we plot the value of $(\hat{\phi}_\gamma)^{-1}$ for the LTS estimator (2.8) (black curves) and for its reweighted version (grey curves), with *bdp* = 0.5. In this case, it is seen that the effect of reweighting is prominent for all the values of p , which further supports the adoption of an adaptive trimming approach. By comparing the plots for the two testing scenarios, it can also be noted that the relationship between $\hat{\phi}_\gamma$ and γ is not linear. As expected, the estimated thresholds converge to the theoretical chi-squared quantile for each procedure, that is, $(\hat{\phi}_\gamma)^{-1}$ converges to 1 (black horizontal line in Figures 8–10).

4.2 Null Performance of the Corrected Robust Diagnostics

For general usability of the method, our scaling factors (4.5) must be easily available for any n in the range for which standard robust diagnostic procedures behave unacceptably. Our proposal is to adopt a simple but flexible, interpolation scheme.

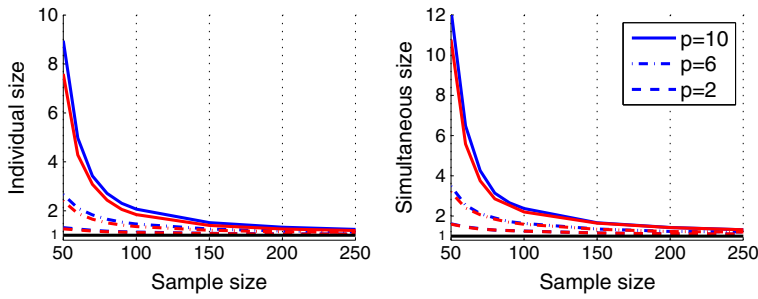


Figure 8. Plot of $(\hat{\phi}_\gamma)^{-1}$ for the MM-estimator with the Tukey p function (2.3) for different values of p and efficiency. Black: $\text{eff} = 0.85$; grey: $\text{eff} = 0.95$. Left panel: individual size; right panel, simultaneous size.

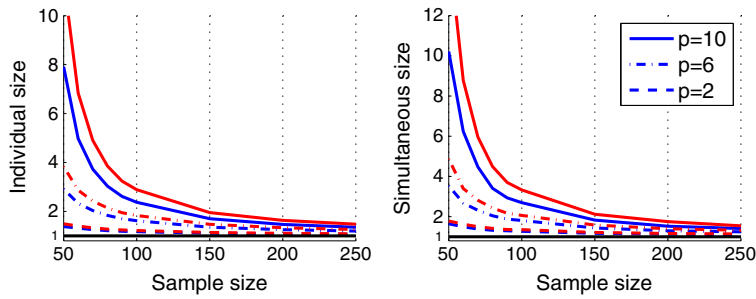


Figure 9. Plot of $(\hat{\phi}_\gamma)^{-1}$ for the S-estimator with $\text{bdp} = 0.5$ for different values of p and p functions. Black: hyperbolic tangent; Grey: optimal.

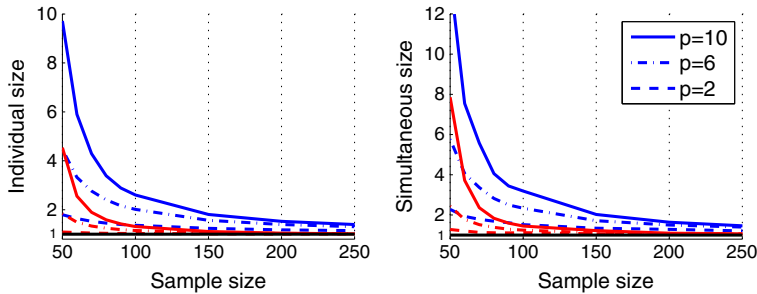


Figure 10. Plot of $(\hat{\phi}_\gamma)^{-1}$ for the LTS estimator (black curves) and for its reweighted version (grey curves) with $\text{bdp} = 0.5$, for different values of p .

First, define

$$\hat{\phi}_{\gamma,n,p} := \hat{\phi}_\gamma,$$

in order to make explicit the dependence of $\hat{\phi}_\gamma$ on both n and p . We then fix $p \leq 10$ and consider a generic sample size, say $\hat{n}$, among those not included in our simulation grid (3.1). Therefore, $50 < \hat{n} < 500$ and $\hat{n} \neq n$. The interpolated scaling factor for $\hat{n}$ is obtained as

$$\hat{\phi}_{\gamma,\hat{n},p} = \hat{\phi}_{\gamma,n_U,p} + \frac{\hat{\phi}_{\gamma,n_U,p} - \hat{\phi}_{\gamma,n_L,p}}{n_U - n_L} (\hat{n} - n_L), \quad (4.6)$$

where n_U is the smallest value of n in the grid such that $\hat{n} < n_U$ and n_L is the largest value of n in the grid such that $\hat{n} > n_L$. We emphasise that our simple linear approximation (4.6) is motivated by the monotone behaviour exhibited by all estimated factors $\hat{\phi}_{\gamma,n,p}$ as a function of n . Because the available grid of sample sizes (3.1) is rather dense, it seems unnecessary to adopt more complicated interpolation functions. If $\hat{n} < 50$, we may set

$$\hat{\phi}_{\gamma,\hat{n},p} = \hat{\phi}_{\gamma,50,p},$$

although we discourage the blind use of robust regression diagnostics (even after our correction) when the sample is very sparse, for example, when $n/p < 5$. For very small sample sizes, accurate distributional results are needed (Cerioli, 2010), while appropriate regularisation is required (Alfons *et al.*, 2013) when p approaches, or is even larger than, n . As seen in Figures 8–10, the empirical thresholds approach the theoretical ones, and the ratio (4.2) becomes close to 1 for all the techniques. Hence, for $n > 500$, the theoretical chi-squared threshold can be used.

The behaviour of our estimated corrections factors for testing both (1.4) and (1.5), when the normal regression model holds, is investigated through a new simulation study, based again on 10 000 independent replicates for each pair of n (or $\hat{n}$) and p . Table 4 reports the empirical sizes obtained for four pairs (n, p) belonging to the grid (3.1), while Table 5 refers to two values of $\hat{n}$ (namely, $\hat{n} = 74$ and $\hat{n} = 170$) not in the grid. We focus on only three procedures, those with the worst performance in Section 3.2. Our corrections provide coefficients, which are very effective for controlling the proportion of false outliers, for the prescribed nominal level of 1%, even when n (or $\hat{n}$) is small and p is large.

4.3 Robustness to Assumptions on Explanatory Variables and Extensions

We conclude our investigation of the behaviour of the proposed diagnostic procedure under the uncontaminated regression model (1.1) by considering some natural extensions and by checking their robustness to assumptions on the explanatory variables $X_1, \dots, X_{p-1}$.

Table 4. Empirical size of individual and simultaneous outlier tests for three robust procedures given in Table 1 for different values of p and n in the simulation grid, when correction (4.5) is applied.

	$p = 2$				$p = 10$			
	$n = 50$		$n = 200$		$n = 50$		$n = 200$	
	Ind	Sim	Ind	Sim	Ind	Sim	Ind	Sim
MMeff085OP	0.0100	0.0093	0.0100	0.0100	0.0100	0.0090	0.0100	0.0096
Sbdp050OP	0.0100	0.0091	0.0100	0.0100	0.0100	0.0090	0.0100	0.0095
LTSbdp050	0.0100	0.0089	0.0100	0.0100	0.0100	0.0079	0.0100	0.0097

The nominal size is 1%.

Table 5. Empirical size as in Table 4 but now for different values of $\hat{n}$ not belonging to the simulation grid.

	$p = 2$				$p = 10$			
	$\hat{n} = 74$		$\hat{n} = 170$		$\hat{n} = 74$		$\hat{n} = 170$	
	Ind	Sim	Ind	Sim	Ind	Sim	Ind	Sim
MMeff085OP	0.0102	0.0107	0.0100	0.0102	0.0099	0.0092	0.0096	0.0086
Sbdp050OP	0.0102	0.0108	0.0100	0.0105	0.0100	0.0090	0.0095	0.0083
LTSbdp050	0.0103	0.0111	0.0100	0.0097	0.0101	0.0092	0.0094	0.0074

The nominal size is 1%.

For $p > 10$, it is possible to extrapolate our correction factors (4.5) by fitting a parametric function to the estimates $\hat{\phi}_{\gamma,n,p}$. Our finding is that the simple quadratic model

$$\hat{\phi}_{\gamma,n,p} = \psi_0 + \psi_1 p + \psi_2 p^2 \tag{4.7}$$

has, on average, a satisfactory fit when applied to the different robust techniques for a given value of n and $p \leq 10$. We thus compute method-specific estimates of ψ_0 , ψ_1 and ψ_2 for each sample size n in the simulation grid and use the resulting fit $\hat{\phi}_{\gamma,n,p}$ to predict the appropriate scaling factor when $p > 10$, with some care needed for large p , because of the quadratic power in Equation 4.7. Prediction is thus straightforward when n belongs to the simulation grid. If instead the sample size $\hat{n}$ is not in the grid, from (4.7), we compute $\hat{\phi}_{\gamma,n_U,p}$ and $\hat{\phi}_{\gamma,n_L,p}$, with n_U and n_L defined as in Section 4.2. We then replace $\hat{\phi}_{\gamma,n_U,p}$ and $\hat{\phi}_{\gamma,n_L,p}$ in (4.6) with the estimates $\hat{\phi}_{\gamma,n_U,p}$ and $\hat{\phi}_{\gamma,n_L,p}$ to obtain the required model-based scaling factor $\hat{\phi}_{\gamma,\hat{n},p}$. Table 6 reports the resulting estimated empirical sizes of the tests of (1.4) and (1.5) for the three robust techniques with the worst performance in Section 3.2, for $p = 12$ and $p = 15$. The only problems occur when n is small and p is large, especially for simultaneous size, even if the size is acceptable when compared with the levels shown in Figures 1–4.

Another extension is the possibility of computing a single scaling factor (4.5) for groups of different estimators. From this grouping, we might tolerate a slight deterioration in performance but with the advantages of simplicity and practical usability of our approach. According to the results described in Section 4.2, it does not seem reasonable to cluster techniques with different bdp. However, it might be sensible to use the same threshold for soft trimming procedures adopting different ρ functions and/or varying levels of nominal efficiency. We thus consider the following possible aggregations of correction factors:

- (1) MM-estimators with the same value of `eff` and different ρ functions;
- (2) MM-estimators with different values of `eff` and different ρ functions;
- (3) S-estimators with `bdp` = 0.50 and different ρ functions.

Table 6. Empirical size of individual and simultaneous outlier tests for three robust procedures given in Table 1 and for different values of p and $\hat{n}$ outside the simulation grid, when correction (4.7) is applied.

	$p = 12$				$p = 15$			
	$\hat{n} = 74$		$\hat{n} = 170$		$\hat{n} = 74$		$\hat{n} = 170$	
	Ind	Sim	Ind	Sim	Ind	Sim	Ind	Sim
MMeff085OP	0.0165	0.0244	0.0109	0.0118	0.0288	0.0651	0.0121	0.0136
Sbdp050OP	0.0135	0.0168	0.0109	0.0098	0.0208	0.0400	0.0114	0.0118
LTSbdp050	0.0138	0.0162	0.0121	0.0139	0.0216	0.0405	0.0152	0.0191

The nominal size is 1%.

Table 7. Empirical size of individual and simultaneous outlier tests for the robust procedures given in Table 4, when the average correction in the corresponding group is applied.

Estimator	Group	$p = 2$				$p = 10$			
		$n = 50$		$n = 200$		$n = 50$		$n = 200$	
		Ind	Sim	Ind	Sim	Ind	Sim	Ind	Sim
MMeff085OP	1	0.0115	0.0122	0.0105	0.0114	0.0173	0.0219	0.0120	0.0133
MMeff085OP	2	0.0122	0.0125	0.0107	0.0115	0.0204	0.0277	0.0130	0.0140
Sbdp050OP	3	0.0115	0.0112	0.0105	0.0116	0.0154	0.0188	0.0121	0.0161

The nominal size is 1%.

We repeat the analysis of Table 4, with the method-specific scaling factors (4.5) now replaced by their group averages. Table 7 shows the results. Our corrections remain effective, with empirical sizes not far from the nominal 1%, even when we merge with respect to both efficiency level and ρ function (Group 2). We thus conclude that for soft trimming procedures, the adoption of a general scaling factor for all ρ functions and values of eff is a practical option, unless there are requirements to stick precisely to the prescribed nominal level of false detections.

Finally, we investigate to what extent our approach is robust when the observed values of the explanatory variables come from distributions different from the Gaussian. We apply our scaling factor (4.5) to samples in which $X_1, \dots, X_{p-1}$ are generated from the chi-square distribution (Table 8) and from the Student's t distribution (Table 9), both with $df = 5$. We only consider the least favourable scenario by focusing on the three robust techniques with the worst performance in Section 3.2. Under these two scenarios, the performance of our procedure is generally very good, with empirical sizes still close to the nominal 1%.

5 Detection Properties of the Robust Diagnostics

We have shown so far that our diagnostic procedure derived from high-breakdown estimation provides reliable inferences on the n individual hypotheses (1.4) and on the simultaneous null (1.5) under a wide spectrum of specifications of the normal regression model (1.1). However, for the practical usability of the method, it is also important to evaluate its performance when the standard model only holds for a proportion of the data and outliers are present. It is reasonable to expect that the adoption of our correction factors $\hat{\phi}_{\gamma, \hat{n}, p}$, which downweight the estimated residuals, will produce some loss in the power to detect contaminated observations, but the price to pay should not be too high. We thus compare the behaviour of our diagnostics with the standard robust ones under two alternative contamination schemes.

Table 8. Empirical size of individual and simultaneous outlier tests for the robust procedures given in Table 4, when the observed values of $X_1, \dots, X_{p-1}$ are generated from the χ^2_5 distribution, for different values of p and n .

	$p = 2$				$p = 10$			
	$n = 50$		$n = 200$		$n = 50$		$n = 200$	
	Ind	Sim	Ind	Sim	Ind	Sim	Ind	Sim
MMeff085OP	0.0098	0.0077	0.0100	0.0100	0.0102	0.0104	0.0100	0.0079
Sbdp050OP	0.0099	0.0104	0.0100	0.0098	0.0101	0.0109	0.0103	0.0104
LTSbdp050	0.0101	0.0121	0.0101	0.0096	0.0110	0.0103	0.0113	0.0249

The nominal size is 1%.

Table 9. Empirical size of individual and simultaneous outlier tests for the robust procedures given in Table 4, when the observed values of $X_1, \dots, X_{p-1}$ are generated from the t_5 distribution, for different values of p and n .

	$p = 2$				$p = 10$			
	$n = 50$		$n = 200$		$n = 50$		$n = 200$	
	Ind	Sim	Ind	Sim	Ind	Sim	Ind	Sim
MMeff085OP	0.0099	0.0084	0.0101	0.0086	0.0105	0.0100	0.0101	0.0081
Sbdp050OP	0.0100	0.0086	0.0102	0.0094	0.0105	0.0092	0.0102	0.0090
LTSbdp050	0.0101	0.0111	0.0106	0.0147	0.0113	0.0114	0.0112	0.0196

The nominal size is 1%.

Our first outlier framework is the contamination model

$$y_i \sim (1 - \tau)G_0 + \tau G_1 \quad i = 1, \dots, n, \quad (5.1)$$

where G_0 stands for the distribution of the random variable $N(x'_i\beta, \sigma^2)$, which defines the postulated null model, and $0 < \tau < 0.5$ is the contamination rate. We define G_1 in (5.1) as the distribution of the random variable

$$N(x'_i\beta + \lambda, \sigma^2), \quad (5.2)$$

for a given level shift λ , yielding a so-called ‘vertical outlier’. As in the previous sections, the explanatory variables $X_1, \dots, X_{p-1}$ are generated from the standard multivariate normal distribution and are then held fixed for each simulation.

Figure 11 reports the empirical performance of our reliable robust diagnostics when testing the n individual hypotheses (1.4) under contamination model (5.1), for the LTS estimator (2.8) with $\text{bdp} = 0.5$ and for its reweighted version with weights (2.10), in the case $n = 100$ and $p = 5$. Different values of the contamination rate τ and level shift λ are considered. In each simulation setting, 10 000 simulations are performed. We provide Monte Carlo estimates of the average power

$$\text{AP} = \mathbb{E} \left[\frac{N_{1|1}}{M_1} \right]$$

and of the false discovery rate

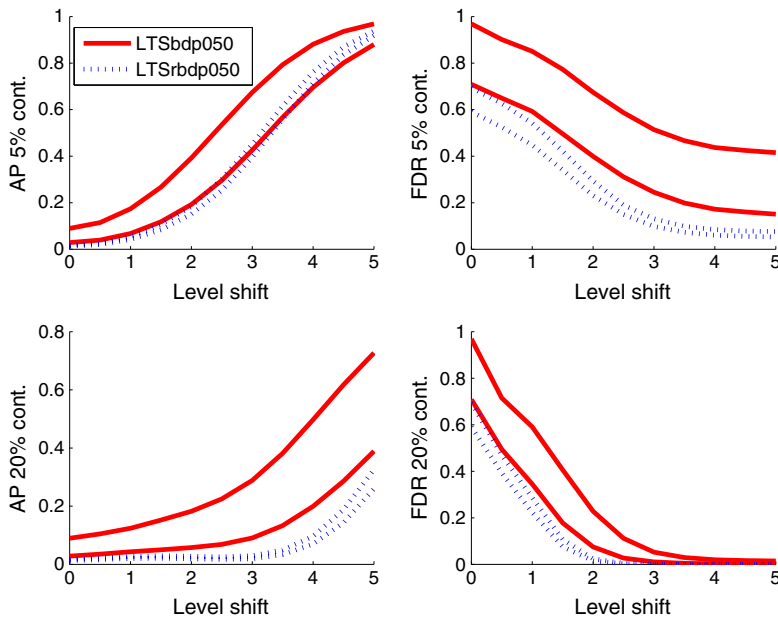


Figure 11. Plots of average power (AP) (left panels) and false discovery rate (FDR) (right panels) of the tests of the n individual hypotheses (1.4), at nominal size 1%, for the LTS estimator (2.8) with $\text{bdp} = 0.5$ and for its reweighted version with weights (2.10), in the case $n = 100$ and $p = 5$. Contamination model (5.1), with different values of the contamination rate τ and level shift λ . In each plot, the upper curve corresponds to the standard robust procedure, while the lower curve corresponds to the test based on our reliable robust diagnostics. 10 000 simulations for each simulation setting.

Table 10. Outcome in testing n observations for outlyingness.

Null hypotheses (1.4)			
	Not rejected	Rejected	Total
True	$N_{0 0}$	$N_{1 0}$	M_0
False	$N_{0 1}$	$N_{1 1}$	M_1
Total	$n - R$	R	n

$$\text{FDR} = \mathbb{E} \left[\frac{N_{1|0}}{R} | R > 0 \right] \Pr(R > 0),$$

where the relevant quantities in the context of outlier detection are defined in Table 10. The nominal test size is 1%. Figure 12 repeats the analysis for the S-estimator (2.2) with Tukey ρ function (2.3) and $\text{bdp} = 0.5$, and for the MM-estimator (2.7) with Tukey ρ function (2.3) and $\text{eff} = 0.85$. In all plots, we report the outlier detection performance of both the standard diagnostics (upper curve) and our corrected procedure (lower curve). Although the power in identifying contaminated observations is smaller when correction (4.6) is applied, as expected, we clearly see that in all cases, the shapes of our average power curves are similar to those of the uncorrected liberal diagnostics and converge to them when λ grows. Furthermore, we have obtained further evidence (not displayed here) that the gap reduces for larger values of n and/or smaller values of p . We can conclude that our reliable diagnostics do not suffer from masking, as the standard high-breakdown ones, and share reasonable detection properties under model (5.1). On the other hand, it is clear from the plots that the impact of false discoveries—now measured by the FDR—is consistently smaller for our approach than for the standard procedure. Our method is thus able to reduce the degree of swamping also when outliers are present, not only under the null model considered in Section 4.2. We could also think of more sophisticated strategies, such as those developed by Cerioli (2010), Cerioli & Farcomeni (2011) and Lourenco & Pires (2014), that can improve the ability of our tests to detect contaminated observations. However, these extensions are not addressed here and will be the subject of future work.

In our second contamination framework, we introduce ‘bad leverage points’ by contaminating the explanatory variables. This step is performed by replacing in each simulation setting a proportion τ of vectors x_l with $z_l = (1, z_{l1}, \dots, z_{l(p-1)})'$, where

$$z_{lj} \sim N(\lambda, 1) \quad l = 1, \dots, \lfloor n\tau \rfloor, \quad j = 1, \dots, p-1, \quad (5.3)$$

for a fixed level shift λ . Then, conditionally on the realised values of $x_1, \dots, x_n$, we simulate our Monte Carlo values of the response as in the postulated model (1.1) with

$$\beta_0 = \beta_1 = \dots = \beta_{p-1} = 0.7. \quad (5.4)$$

This contamination scheme produces observations that clearly deviate from the model describing the bulk of the data and that can strongly attract the parameter estimates. Figures 13 and 14 report the results, again in terms of average power and false discovery rate, for both the standard liberal diagnostics and our reliable robust approach. We see that the plots tell virtually the same story as in the case of vertical outliers. Furthermore, our experience is that different choices for the values of the regression parameters in (5.4) do not affect the comparison between the two approaches, although the detection rates might change.

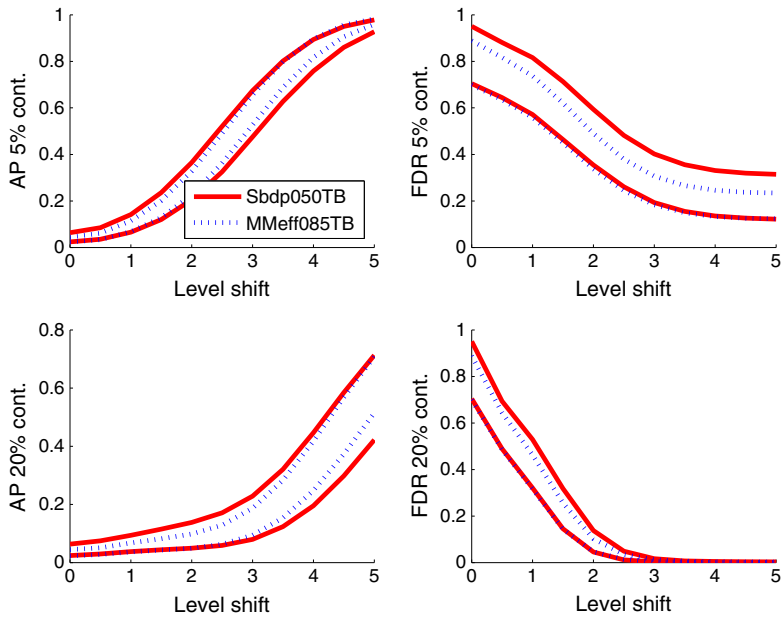


Figure 12. Plots of average power (AP) and false discovery rate (FDR) as in Figure 11 but now for the S -estimator (2.2) with the Tukey p function (2.3) and $b\hat{d}p = 0.5$, and for the MM -estimator (2.7) with the Tukey p function (2.3) and $e\hat{e}f = 0.85$.

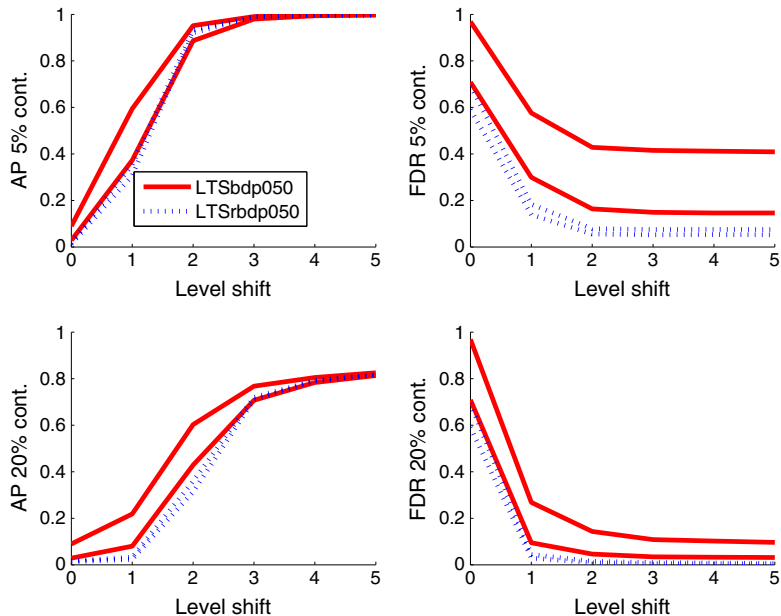


Figure 13. Plots of average power (AP) and false discovery rate (FDR) of the tests of the n individual hypotheses (1.4), at nominal size 1%, for the LTS estimator (2.8) with $b\hat{d}p = 0.5$ and for its reweighted version with weights (2.10), in the case $n = 100$ and $p = 5$. Contamination model (5.3) for the explanatory variables giving rise to ‘bad leverage points’, with different values of the contamination rate τ and level shift λ . In each plot, the upper curve corresponds to the standard robust procedure, while the lower curve corresponds to the test based on our reliable robust diagnostics. 10 000 simulations for each simulation setting.

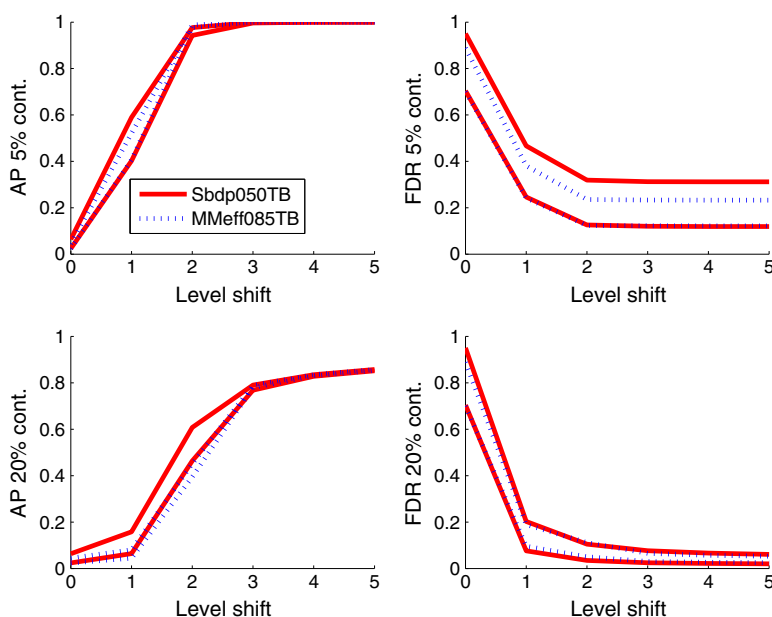


Figure 14. Plots of average power (AP) and false discovery rate (FDR) as in Figure 13 but now for the S -estimator (2.2) with the Tukey ρ function (2.3) and $\text{bdp} = 0.5$, and for the MM -estimator (2.7) with the Tukey ρ function (2.3) and $\text{eff} = 0.85$.

6 Discussion

In this work, we have studied the behaviour of diagnostic procedures obtained from popular high-breakdown regression estimators when no outlier is present in the data. We argue that controlling the number of false discoveries is an important issue that arises in several application fields. We have found that the empirical error rates for many of the available robust techniques are surprisingly far from the prescribed nominal level, thus considerably limiting their practical appeal. To overcome this drawback, we have proposed a simulation-based approach to correct the liberal diagnostics and reach reliable inferences. We have provided extensive evidence that our approach performs well in many settings of practical interest and for different robust regression techniques. Furthermore, we have shown that it has reasonably good diagnostic properties when outliers are present, under different contamination schemes. We thus claim that our correction method has a wide applicability. We also conjecture that it could be extended to other high-breakdown procedures not considered here.

The simulated thresholds $\hat{\gamma}_V^2$, defined in (4.4), are available from www.riani.it. We have also implemented a Matlab function, within the FSDA Toolbox (called `RobRegrSize`), which provides the interpolated correction factors of §§4.2–4.3 according to the selected options of robust estimators, ρ functions, breakdown point and efficiency.

A more complex research question than the one addressed in this work is to obtain accurate approximations for the distribution of the diagnostic quantities that we consider. To our knowledge, such approximations are only available for the deviance measures computed from the forward search (Riani *et al.*, 2009) and for some multivariate (adaptive) trimming estimators (Hardin & Rocke, 2005; Cerioli, 2010). How to extend these distributional approximations to other classes of high-breakdown regression techniques is the subject of ongoing research.

Acknowledgements

We are grateful to two anonymous referees, an Associate Editor and the Co-editor in Chief for many insightful comments that helped us to improve the manuscript, and, in particular, for the suggestion to develop the work described in Sections 3.4 and 5. We also thank Professor Anthony C. Atkinson for his comments on previous drafts. This research was partly supported by the project MIUR PRIN ‘MISURA—Multivariate models for risk assessment’.

References

- Alfons, A., Croux, C. & Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann. Appl. Stat.*, **7**, 226–248.
- Atkinson, A.C. & Riani, M. (2000). *Robust Diagnostic Regression Analysis*. New York: Springer-Verlag.
- Ceroli, A. (2010). Multivariate outlier detection with high-breakdown estimators. *J. Am. Stat. Assoc.*, **105**, 147–156.
- Ceroli, A. & Farcomeni, A. (2011). Error rates for multivariate outlier detection. *Comput. Stat. Data Anal.*, **55**, 544–553.
- Ceroli, A. & Perrotta, D. (2014). Robust clustering around regression lines with high density regions. *Adv. Data. Anal. Classi.*, **8**, 5–26.
- Ceroli, A., Riani, M. & Atkinson, A.C. (2009). Controlling the size of multivariate outlier tests with the MCD estimator of scatter. *Stat. Comput.*, **19**, 341–353.
- Ceroli, A., Farcomeni, A. & Riani, M. (2013a). Robust distances for outlier-free goodness-of-fit testing. *Comput. Stat. Data Anal.*, **65**, 29–45.
- Ceroli, A., Riani, M. & Torti, F. (2013b). Size and power of multivariate outlier detection rules. In *Algorithms from and for Nature and Life*, Eds. B. Lausen, D. Van den Poel & A. Ultsch, pp. 3–17. Berlin: Springer-Verlag.
- Ceroli, A., Farcomeni, A. & Riani, M. (2014). Strong consistency and robustness of the forward search estimator of multivariate location and scatter. *J. Multivariate Anal.*, **126**, 167–183.
- Cizek, P. (2013). Reweighted least trimmed squares: an alternative to one-step estimators. *Test*, **22**, 514–533.
- Cook, R.D. & Hawkins, D.M. (1990). Comment on Rousseeuw and van Zomeren (1990). *J. Am. Stat. Assoc.*, **85**, 640–644.
- Croux, C. & Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *J. Multivariate Anal.*, **71**, 161–190.
- Davies, L. (1990). The asymptotics of S-estimators in the linear regression model. *The Ann. Stat.*, **18**, 1651–1675.
- De Battisti, F. & Salini, S. (2013). Robust analysis of bibliometric data. *Stat. Method. Appl.*, **22**, 269–283.
- Fauconnier, C. & Haesbroeck, G. (2009). Outliers detection with the minimum covariance determinant estimator in practice. *Stat. Methodol.*, **6**, 363–379.
- Filzmoser, P., Maronna, R. & Werner, M. (2008). Outlier identification in high dimensions. *Comput. Stat. Data Anal.*, **52**, 1694–1711.
- García-Escudero, L. A. & Gordaliza, A. (2005). Generalized radius processes for elliptically contoured distributions. *J. Am. Stat. Assoc.*, **100**, 1036–1045.
- García-Escudero, L. A. & Gordaliza, A. (2006). The importance of the scales in heterogeneous robust clustering. *Comput. Stat. Data Anal.*, **51**, 4403–4412.
- Hampel, F.R., Rousseeuw, P.J. & Ronchetti, E. (1981). The change-of-variance curve and optimal redescending M-estimators. *J. Am. Stat. Assoc.*, **78**, 643–648.
- Hardin, J. & Rocke, D.M. (2005). The distribution of robust distances. *J. Comput. Graph. Stat.*, **14**, 910–927.
- Hawkins, D.M. & Olive, D.J. (2002). Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm (with discussion). *J. Am. Stat. Assoc.*, **97**, 136–159.
- Hubert, M., Rousseeuw P.J. & Van Aelst, S. (2008). High-breakdown robust multivariate methods. *Stat. Sci.*, **23**, 92–119.
- Johansen, S. & Nielsen, B. (2015). Analysis of the forward search using some new results for martingales and empirical processes. *Bernoulli* **21**. (In press).
- Johnson, N.L., Kotz, S. & Balakrishnan, N. (1994). *Continuous Univariate Distributions—I*, 2nd edition. New York: Wiley.
- Korosok, M.R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.
- Lourenco, V.M. & Pires, A. (2014). M-regression, false discovery rates and outlier detection with application to genetic association studies. *Comput. Stat. Data Anal.*, **78**, 33–42.
- Maronna, R.A. & Yohai, V.J. (2010). Correcting MM estimates for “fat” data sets. *Comput. Stat. Data Anal.*, **54**, 3168–3173.

- Maronna, R.A., Martin, R.D. & Yohai, V.J. (2006). *Robust Statistics*. Chichester: Wiley.
- Pison, G., Van Aelst, S. & Willems, G. (2002). Small sample corrections for LTS and MCD. *Metrika*, **55**, 111–123.
- Riani, M., Atkinson, A.C. & Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *J. R. Stat. Soc., Ser. B*, **71**, 447–466.
- Riani, M., Perrotta, D. & Torti, F. (2012). FSDA: a MATLAB toolbox for robust analysis and interactive data exploration. *Chemometr. Intell. Lab.*, **116**, 17–32.
- Riani, M., Atkinson, A.C. & Perrotta, D. (2014a). A parametric framework for the comparison of methods of very robust regression. *Stat. Sci.*, **29**, 128–143.
- Riani, M., Cerioli, A., Atkinson, A.C. & Perrotta, D. (2014b). Monitoring robust regression. *Electron. J. Stat.*, **8**, 646–677.
- Riani, M., Cerioli, A. & Torti, F. (2014c). On consistency factors and efficiency of robust S-estimators. *Test*, **23**, 356–387.
- Rousseeuw, P.J. & Hubert, M. (2011). Robust statistics for outlier detection. *WIREs Data Mining and Knowledge Discovery*, **1**, 73–79.
- Rousseeuw, P.J. & Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.
- Van Aelst, S. & Willems, G. (2011). Robust and efficient one-way MANOVA tests. *J. Am. Stat. Assoc.*, **106**, 706–718.
- Van Aelst, S., Willems, G. & Zamar, R.H. (2013). Robust and efficient estimation of the residual scale in linear regression. *J. Multivariate Anal.*, **116**, 278–296.

[Received July 2014, accepted March 2015]