Statistical Papers (2024) 65:2489–2525
https://doi.org/10.1007/s00362-023-01492-3

REGULAR ARTICLE

ROBOUT: a conditional outlier detection methodology for high-dimensional data

Matteo Farnè[1] · Angelos Vouldis[2]

Received: 10 August 2022 / Revised: 24 August 2023 / Published online: 29 September 2023

Abstract

This paper presents a methodology, called ROBOUT, to identify outliers conditional on a high-dimensional noisy information set. In particular, ROBOUT is able to identify observations with outlying conditional mean or variance when the dataset contains multivariate outliers in or besides the predictors, multi-collinearity, and a large variable dimension compared to the sample size. ROBOUT entails a pre-processing step, a preliminary robust imputation procedure that prevents anomalous instances from corrupting predictor recovery, a selection stage of the statistically relevant predictors (through cross-validated LASSO-penalized Huber loss regression), the estimation of a robust regression model based on the selected predictors (via MM regression), and a criterion to identify conditional outliers. We conduct a comprehensive simulation study in which the proposed algorithm is tested under a wide range of perturbation scenarios. The combination formed by LASSO-penalized Huber loss and MM regression turns out to be the best in terms of conditional outlier detection under the above described perturbed conditions, also compared to existing integrated methodologies like Sparse Least Trimmed Squares and Robust Least Angle Regression. Furthermore, the proposed methodology is applied to a granular supervisory banking dataset collected by the European Central Bank, in order to model the total assets of euro area banks.

Keywords Conditional outlier · Multivariate outlier · Robust regression · Variable selection · High dimension

Mathematics Subject Classification 62J07 · 62F07 · 65Y20 · 65F50

✉ Matteo Farnè
matteo.farne@unibo.it

Angelos Vouldis
angelos.vouldis@ecb.europa.eu

[1] University of Bologna, Via delle Belle Arti 41, Bologna, Italy

[2] European Central Bank, Sonnemannstrasse 20, Frankfurt am Main, Germany

1 Introduction

Data quality is a fundamental prerequisite for any kind of quantitative analysis and the large datasets which are becoming increasingly available present specific challenges to the task of monitoring and ensuring data quality. One critical aspect of the data quality monitoring is outlier detection, i.e. the identification of values which are either obviously mistaken or seem to be unjustified from an empirical perspective. An outlier is defined by Hawkins (1980) as 'an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism'. Similarly, Barnett and Lewis (1994) defines an outlier as 'an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data'. Classical statistical methods, which underpin analytical tools supporting analysis of large datasets, are sensitive to the presence of outliers and, consequently, could reach distorted conclusions (Rousseeuw and Hubert 2018).

In this paper, we focus on outlier detection in high-dimensional datasets, i.e., where the number of variables p (i.e. the dimension of the data space) is large, possibly larger than the number of observations n (i.e. the sample size). In this way, we also include so-called *fat* datasets, featuring $p > n$. Such datasets arise in diverse fields, such as bioinformatics, economics, neuroscience, signal processing and others.

Our aim is to retrieve from such datasets any anomaly in a target variable y with respect to a set of K related variables (the predictors of y) that constitute a subset of the $p \gg K$ variables of the dataset (the candidate predictors of y). The K predictors of y are ex-ante unknown and need to be identified from the p variables, which are also usually affected by multicollinearity effects. In addition, we tolerate the presence in the K predictors of leverage outliers, a feature that typically inflates the predictive power of irrelevant predictors, as well as the presence of multivariate outliers in the remaining variables of the dataset.

If the conditional outliers in y are present in the same observation with anomalous instances of the related predictors, it has been shown that predictors are typically not identified correctly (see Khan et al. 2007). That is the reason why we develop a robust preliminary imputation procedure, that prevents such points (i.e. leverage points) from corrupting predictor recovery, thus restoring the identification of true predictors. Ensuring the recovery of true predictors and limiting as much as possible the inclusion of irrelevant predictors is the key to successfully apply a subsequent robust regression in a consistent way and to identify conditional outliers by means of the robustly estimated residual scale. Other confounding factors for predictor recovery are multicollinearity, a high number of variables p compared to the sample size n, and a small overall signal-to-noise ratio.

The practical relevance of the problem as formulated above can be clarified by referring to the banking supervisory dataset that is used in this paper to show the behaviour of the proposed outlier detection method. A bank may present a particularly high value of e.g. total assets (the variable y in the above formulation), which may be spotted as an anomalous instance compared to the rest of banks. However, considering other related bank indicators, such as debt securities and derivatives, we may realize that the total asset value for the bank in question is perfectly in line with expectations. On the other hand, a bank with an average value of total assets may be judged as anomalous with respect to the predictors. This could be for example the case when for a specific observation (i.e. bank) a relatively moderate amount of assets corresponds to a disproportionally high amount of derivatives, a financial instrument that one would normally expect to be used extensively by the largest banks.

In general, whether an observation of variable y is considered to be an outlier does not depend solely on the distribution of y but also on the corresponding values of the most relevant predictors of y. This kind of outlier detection is referred to as conditional outlier detection (Hong and Hauskrecht 2015), because it focuses on the detection of improbable values in a target variable given (i.e. conditionally on) the observed values of a set of related variables.

We propose a new method, called ROBOUT, to solve the problem of conditional outlier detection in high dimensions. ROBOUT can accommodate diverse statistical properties of the examined dataset (such as multicollinearity, multivariate outliers in or out of the predictors and high-dimensional information sets) while at the same time being computationally efficient. Our specific contributions comprise the separation of predictor recovery and final model estimation, which renders the identification of true predictors possible under above mentioned perturbations, the study and recovery of conditional outliers with perturbed residual variance (beyond the usual ones with perturbed mean), and the adoption of a preliminary imputation method, which is effective in preserving model recovery from anomalies in the potential predictors.

The paper is structured as follows. Section 2 formulates the problem and motivates its solution. Section 3 presents a literature review on high-dimensional variable selection and robust regression methods. ROBOUT methodology is formally defined in Sect. 4, entailing a pre-processing, a robust preliminary imputation, a variable selection, a low-dimensional regression and a conditional outlier detection step. The wide simulation exercise that we undertake in Sect. 5 aims to test the performance of ROBOUT, placing emphasis on the versatility of the ensuing conditional outlier detection under several perturbed scenarios for different p/n ratios. Subsequently, we apply ROBOUT to a banking dataset collected by the European Central Bank in Sect. 6, and we provide concluding remarks in Sect. 7.

2 Motivating examples and first formalization

Figure 1 illustrates visually the problem that we aim to solve and the idea of the method that we propose. We consider the target variable y and two variables of the dataset x_1 and x_2 (in principle, the variables of the dataset could be many more than just two). x_1 is a predictor of y while x_2 is not, however this is not known ex-ante. We assume that there is an outlier in variable y, namely point 4 which is an outlier of y conditional on predictor x_1. Point $4'$ shows the 'correct' position that 4 would have occupied if it followed the same data generating process as the other three points. The perturbation of point 4 affects the linear relationship between y and x_1, as line l_0 is shifted below to line l_1, but also affects the relationship between y and x_2, potentially leading to a spuriously statistically significant relationship.

Identifying the perturbation in point 4 requires to solve two tasks: first to identify that x_1 is the relevant conditional variable and second to identify that 4 is an outlier of

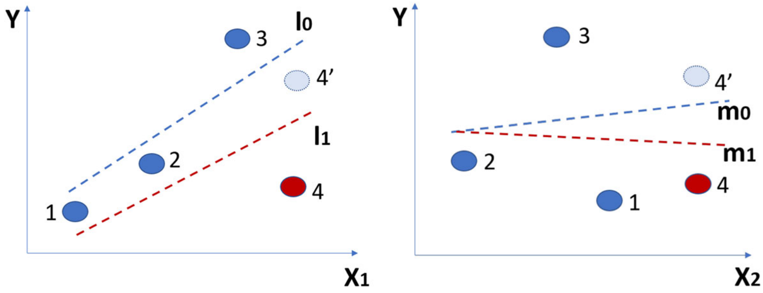


Fig. 1 Identification of conditional outliers when the predictors are not known ex-ante. The left panel depicts the scatter plot of the target variable y and its predictor variable x_1, while the right panel depicts the scatter plot of the target variable y and a non-predictor variable x_2. Point 4 is a conditional outlier of y on x_1. The blue line represents the OLS regression line with point 4 in the *correct* (theoretical) position, that is 4', while the red line represents the OLS regression line with point 4 in the *perturbed* (actual) position

y conditional on x_1. Since the presence of outliers in y may disrupt the identification of the statistical relationship between two related variables (in this case y and x_1) and may generate a misleading statistical relationship between unrelated variables (in this case y and x_2), in this paper we separate the two problems: we first run a robust variable selection step, with the aim to identify the relevant predictors, and we subsequently run a low-dimensional robust regression to identify conditional outliers. As will be argued in this paper, performing predictor recovery and the model estimation in consecutive steps has substantial advantages in terms of computational efficiency and effectiveness in detecting conditional outliers.

Let us formalize the described problem in probabilistic terms. In (Hong and Hauskrecht 2015), unconditional outliers are defined as instances which fall into a low probability density region of $\mathrm{f}(y)$, where $\mathrm{f}(y)$ is the unconditional probability density function of the scalar target variable y. Instead, conditional outliers are defined as instances of y which fall into a low probability density region of $\mathrm{f}(y|\mathbf{x}) = \mathrm{f}(y, \mathbf{x})/\mathrm{f}(\mathbf{x})$, where $\mathrm{f}(y|\mathbf{x})$ is the conditional probability density function of the scalar target variable y given a vector of predictors $\mathbf{x}$, and the set of predictors $\mathbf{x}$ is a subset of a wider noisy information set Ω, which is given at the outset of the problem. In this paper, we assume that the expected value of $\mathrm{f}(y|\mathbf{x})$ is a linear function of the variables in $\mathbf{x}$. Note that this does not constrain the nature of the prescribed relationships between y and the initial information set Ω from which $\mathbf{x}$ is identified, because we can always include e.g. quadratic and exponential functions of specific variables in the vector $\mathbf{x}$.

For the sake of simplicity, let us suppose that for the single observation $i \in \{1, \dots, n\}$ it holds that $y_i|\mathbf{x}_{D,i} \sim N(a_i + \mathbf{x}'_{D,i}\beta, \sigma_i^2)$, where, for observation i, y_i is the specific value of y, a_i is the intercept term, $\mathbf{x}_{D,i}$ is the $K \times 1$ vector of predictors, β is the $K \times 1$ vector of regression coefficients, and σ_i^2 is the variance of y_i conditional on $\mathbf{x}_{D,i}$. We denote by D the set of predictor indices, such that $|D| = K$ and $D \subseteq \{1, \dots, p\}$, and by $\bar{D}$ the complementary set of D with respect to the set $\{1, \dots, p\}$. The $n \times K$ matrix $\mathbf{X}_D$ contains as columns the predictor variables, indexed by D, the $n \times (p - K)$ matrix $\mathbf{X}_{\bar{D}}$ contains as columns the non-predictor variables, indexed by $\bar{D}$, and the $n \times p$ complete matrix $\mathbf{X} = [\mathbf{X}_D|\mathbf{X}_{\bar{D}}]$ represents the available information set Ω.

We define a set of outlier indices O with $O \subseteq \{1, \dots, n\}$ such that $|O| = [\alpha n]$. The parameter α represents the contamination rate, with $\alpha \in [0, 0.5]$. We assume, without loss of generality, that for any $i \in \bar{O}$ it holds $a_i = a$ and $\sigma_i^2 = \sigma^2$, i.e. that all non-outlying observations feature a constant conditional mean and variance.

Definition 1 A conditional outlier in mean is defined as any observation i' such that $y_{i'}$ falls in a low probability density region of $N(a + \mathbf{x}'_{D,i'}\beta, \sigma^2)$ because $a_{i'} \gg a$ or $a_{i'} \ll a$.

Definition 2 A conditional outlier in variance is defined as any observation i' such that $y_{i'}$ falls in a low probability density region of $N(a + \mathbf{x}'_{D,i'}\beta, \sigma^2)$ because $\sigma_{i'}^2 \gg \sigma^2$.

Definition 3 A leverage outlier is defined as any observation i' such that its vector of predictors $\mathbf{x}_{D,i'}$ is perturbed by replacing $\mathbf{x}_{D,i'k}$ with $\mathbf{x}_{D,i'k} + \mathbf{m}$, $\mathbf{m} \in \mathbb{R}^K$, $|m_k| \gg 0$, $k = 1, \dots, K$.

Definition 4 A rowwise outlier is defined as any observation i' such that a subset of its vector of non-predictors, say, $\mathbf{x}_{\tilde{D},i'}$, with $\tilde{D} \subset \bar{D}$, is perturbed by replacing $\mathbf{x}_{\tilde{D},i'k'}$ with $\mathbf{x}_{\tilde{D},i'k'} + \mathbf{m}$, $\mathbf{m} \in \mathbb{R}^{|\tilde{D}|}$, $|m_{k'}| \gg 0$, $k' \in \tilde{D}$.

In practice, Definitions 1 and 2 may reflect that the target variable y is affected by measurement errors, idiosyncratic events, unpredictable shocks etc., as well as structural differences in the data generating mechanism. Definitions 3 and 4 represent the same situations in the matrix of predictors $\mathbf{X}_D$ and in the matrix of non-predictors $\mathbf{X}_{\bar{D}}$, respectively. Conditional outlier recovery is critical to spot hidden inconsistencies or frauds in y, and cannot be performed properly if the unknown predictors of y, contained in $\mathbf{X}_D$, are not identified.

Few integrated methods have been presented in the literature that recover simultaneously both the K predictors of the response variable y (out of the p candidate predictors) and the conditional outliers in y, such as Sparse Least Trimmed Squares (SPARSE-LTS, Alfons et al. 2013) and Robust Least Angle Regression (RLARS, Khan et al. 2007). To the best of our knowledge, their performance as conditional outlier detection methods in the presence of conditional outliers in variance in the sense of Definition 2 has not been explored yet. Another existing method, called SNCD (Semismooth Newton Coordinate Descent) algorithm (Yi and Huang 2017), provides a fast and reliable solution to the recovery of predictors, by minimizing a Huber or a Least Absolute Deviation (LAD) loss of the residuals penalized by an elastic net (Zou and Hastie 2005). However, SNCD has a significant drawback: it is not robust to leverage outliers, as shown in Alfons et al. (2013).

For this reason, we propose in Sect. 4.2.2 a preliminary imputation procedure which is robust to potentially disruptive outliers in $\mathbf{X}$ while being computationally efficient. In the following, we apply SNCD to the clean imputed dataset to identify the *right* predictors, on which a robust regression model is finally calculated to spot conditional outliers. The described method, which we call ROBOUT, is very efficient as regards computational cost, which is a direct function of the degree of perturbation in the dataset. ROBOUT improves existing methods in the literature by enhancing both predictor recovery and conditional outlier detection performance in high dimensions also under challenging conditions, when other methods may fail. This result is achieved by exploiting the large cross section to setup the preliminary imputation procedure, whose accuracy benefits from the presence of many pairs of highly correlated variables, thus turning the curse of dimensionality to a blessing.

3 State of the art

Large and high-dimensional datasets appear very suitable for conditional outlier detection in the regression context (Rousseeuw and Leroy 1987). As Varian (2014) notes, the large size of the data requires automated techniques for the identification of subsets of variables which are statistically linked. In such datasets, the challenge is to identify the critical predictors and then to perform a conditional outlier detection for the variables of interest. To this end, we need to define both a variable selection and a robust regression procedure, in order to recover consistently both the true set of predictors and the regression coefficients. This type of outlier detection may spot outliers which could remain unnoticed if single variables are considered separately, thus taking advantage of the information content present in a large cross section.

The original idea of robust regression is based on the application of a weighting scheme to observations, with the aim to dampen the influence of outlying points on regression estimates, both in $\mathbf{y}$ (conditional outliers) and in $\mathbf{X}$ (leverage outliers). Two established methods for low-dimensional robust regression are:

- least trimmed squares (LTS) estimation (Rousseeuw 1984), which identifies the $100 \times (1 - \alpha)\%$ most concentrated observations and estimates the regression coefficients on those via ordinary least squares;
- MM-estimation (Yohai 1987), which is a three-stage procedure that minimizes a Tukey's bisquare function of the residuals using a robust initialization of the coefficients in β (obtained by S-regression) and of the residual scale σ (obtained by M-estimation).

In the $p > n$ case, the mentioned traditional robust regression techniques no longer work, because they simply are weighted versions of the least squares method. Filzmoser and Nordhausen (2021) provides an exhaustive literature review on robust linear regression for high-dimensional data and describes the different strategies that have been consequently proposed to perform robust regression in high dimensions.

There are few methods proposed in the literature that perform simultaneously variable selection and conditional outlier detection. A first group is based on LASSO regression (Tibshirani 1996). A robust LASSO regression method performing both variable selection and conditional outlier detection is SPARSE-LTS (Alfons et al. 2013), that is based on the simultaneous optimization of a trimmed least squares loss and a LASSO penalty. The coefficient estimates are derived by performing at each iteration a LASSO estimation on the $100 \times (1 - \alpha)\%$ observations with the smallest squared residuals. Another robust LASSO regression method is MM-LASSO (Smucler and Yohai 2017). Building on the S-ridge approach to high-dimensional regression model estimation in Maronna (2011), MM-LASSO combines (adaptive) MM-regression and

LASSO penalization to provide consistent identification of predictors and estimation of coefficients under fat tailed distributions.

A previous relevant solution of different nature is plug-in RLARS (Khan et al. 2007), that is the robustified version of LARS (Efron et al. 2004) where the covariance matrix of the features is robustly estimated via an adjusted Winsorization step. RLARS manages to deal with outliers by rendering them uninfluential in the iterative computation of the covariances that are needed to retrieve predictors. Predictor selection is then performed by iteratively selecting the best angle to update coefficients, as in LARS, and the MM regression is used as final step.

A further approach to identify predictors and spot conditional outliers in high dimensions is by robustifying the elastic net regression. In this respect, a robust solution is ENET-LTS (Kurnaz et al. 2018), that provides the trimmed version of the elastic net. In Freue et al. (2019), the penalized elastic net S-estimator PENSE and its refinement, the penalized elastic net M-estimator PENSEM, provide a consistent solution to conditional outlier detection, with the relevant advantage for PENSEM to provide the most precise τ-estimate of the residual scale.

Forward Search (see Atkinson et al. 2004) is a fascinating method to iteratively identify the subset of the least anomalous observations from a multivariate dataset, which are ranked according to their degree of outlyingness. This method has been scaled to very large samples (Riani et al. 2015) via a fast calculation of robust Mahalanobis distances. A Bayesian version of it has been presented in Atkinson et al. (2017).

The SNCD algorithm (Yi and Huang 2017), that minimizes simultaneously a Huber loss or a LAD loss of the residuals and the elastic net penalty by explicitly deriving the Karush-Kuhn-Tucker (KKT) conditions of the objective function, is instead not robust to leverage outliers, even if it is scalable to both large sample sizes and high dimensions, as its computational cost is $O(pn)$.

Another related strand of the literature concerns the cellwise contaminated outliers: Rousseeuw and Bossche (2018) identifies deviating data cells, Bottmer et al. (2022) adapts the sparse shooting S-estimator of Öllerer et al. (2016) in the presence of cellwise outliers, Filzmoser et al. (2020) proposes cellwise robust M regression. Even if the performed task is similar with that of ROBOUT (sparse regression model identification and estimation in a perturbed context), the perturbations which we consider in the data matrix are somewhat different, as we consider the presence of multivariate outliers in the true predictors (similarly to SPARSE-LTS and RLARS) as well as beyond the true predictors, with the aim to extend as much as possible robustness to multivariate outliers. In addition, our method avoids running pairwise robust regressions in order to keep computational cost low and aiming to be fit for high-dimensional data. We leave to future studies a formal comparison with the sparse shooting S-estimator and a modification of ROBOUT able to deal with randomly scattered individual outliers.[1]

[1] Our approach distinguishes between a response variable and the conditional information set while (Rousseeuw and Bossche 2018) scans the whole original dataset for cell-wise outliers. This difference however can be reconciled by running the proposed ROBOUT method consecutively defining each column of the dataset as the response variable.

4 ROBOUT: a comprehensive approach for conditional outlier detection

In this section, we describe ROBOUT procedure. We first present the data model behind its usage in Sect. 4.1 and we elaborate on its five steps in Sect. 4.2.

4.1 Model

Let us consider n numerical observations of one response variable y and p additional variables. We call the unknown set of conditional outlier indices O with $|O| = [\alpha n]$ and $\alpha \in [0, 0.5]$, and $\bar{O}$ the index set of non-outlying points. The response variable vector $\mathbf{y}$ may be expressed in terms of the following regression model

$$\mathbf{y} = \mathbf{a} + \mathbf{X}_D \beta + \epsilon, \tag{1}$$

where $\mathbf{y}$ is the $n \times 1$ vector of the response variable, $\mathbf{a}$ is the $n \times 1$ vector of intercepts, $\mathbf{X}_D$ is the $n \times K$ matrix of predictors (whose columns are indexed by D), β is the $K \times 1$ vector of regression coefficients and ϵ is the $n \times 1$ vector of residuals. The same regression model for the single observation $i \in \{1, \ldots, n\}$ can be written as

$$y_i = a_i + \mathbf{x}'_{D,i} \beta + \epsilon_i, \tag{2}$$

where a_i denotes the intercept and $\mathbf{x}_{D,i}$ denotes the $K \times 1$ vector of predictors. We name $\mathbf{R}_\mathbf{X}$ the $p \times p$ covariance matrix of $\mathbf{X}$, and we identify the covariance matrix of the K predictors as $\mathbf{R}_D$, of the $p - K$ non-predictors as $\mathbf{R}_{\bar{D}}$, of the variables in $\tilde{D}$ as $\mathbf{R}_{\tilde{D}}$.

For each non-outlier index $i \in \bar{O}$, we assume that $a_i = a$, $\epsilon_i \sim N(0, \sigma^2)$, $\mathbf{x}_{D,i} \sim MVN(\mathbf{0}_K, \mathbf{R}_D)$ and $\mathbf{x}_{\bar{D},i} \sim MVN(\mathbf{0}_{p-K}, \mathbf{R}_{\bar{D}})$, where $\mathbf{x}_{\bar{D},i}$ is the vector of non-predictors ($\bar{D}$ stores the indices of non-predictor variables). We distinguish in the outlier set O: the index sets of conditional outliers in y, O_y, with size $[\alpha_y n]$; of leverage outliers in the predictors, O_{lev}, with size $[\alpha_{lev} n]$; of rowwise outliers out of the predictors, O_{row}, with size $[\alpha_{row} n]$. Note that the proportions of index sets α_y, α_{lev}, α_{row} are allowed to vary, and we allow O_y, O_{lev}, and O_{row} to overlap.

Let us define for any $x \in \mathbb{R}$ the sign operator sgn, such that $\text{sgn}(x) = 1$, if $x > 0$, $\text{sgn}(x) = 0$, if $x = 0$, $\text{sgn}(x) = -1$, if $x < 0$. Then, for each outlier index $i' \in O$, we fix $a_{i'} = a$, and we generate the four types of outliers (Definitions 1-4) in the following way:

1) conditional outliers in mean are generated consistently with Definition 1 as $\epsilon_{i'} \sim N((m-1)a, \sigma^2)$, $i' \in O_y$, $m \in \mathbb{R}$, $m > 1$;
2) conditional outliers in variance are generated consistently with Definition 2 as $\epsilon_{i'} \sim N(0, m^2\sigma^2)$, $i' \in O_y$, $m \in \mathbb{R}$, $m > 1$;
3) the vector of predictors is generated consistently with Definition 3 as $\mathbf{x}_{D,i'} \sim MVN(\mathbf{0}_K, \mathbf{R}_D)$, and leverage outliers are generated by replacing *a posteriori* $x_{D,i'k}$ with the perturbed values $x_{D,i'k} + (m+2) \times \text{sgn}(\text{Unif}[-1, 1])$, $i' \in O_{lev}$, $k = 1, \ldots, K$, $m \in \mathbb{R}$, $m > 1$;

4) the vector $\mathbf{x}_{\tilde{D},i'}$ is generated consistently with Definition 4 as $\mathbf{x}_{\tilde{D},i'} \sim MVN(\mathbf{0}_K, \mathbf{R}_{\tilde{D}})$, and rowwise outliers are generated by replacing *a posteriori* $x_{\tilde{D},i'k'}$ with the perturbed values $x_{\tilde{D},i'k'} + (m+2) \times \text{sgn}(\text{Unif}[-1,1])$, $i' \in O_{row}$, $k' \in \tilde{D}$, $\tilde{D} \subset \bar{D}$, $m \in \mathbb{R}$, $m > 1$.

Normality is assumed to ensure the validity of residual diagnostics and model inference (also see Sect. 4.2.1). The multiplier m is the outlyingness parameter, and the value $m+2$ is chosen in 3) and 4) to ensure that non-robust correlation estimates are impacted by those outliers, as well as the randomized sign on individual perturbations obtained by the uniform distribution $\text{Unif}[-1, 1]$ contributes to ensure.

In all the above cases of outliers, the parameter $m > 1$ represents the degree of perturbation. Because more than one type of outliers can co-exist within the given observation (row), there are in total nine outlier schemes for each $i' \in O$:

(1–2) “Only mean" or “Only variance" cases: only response variable outliers (in mean or variance) are present. The outliers i' are only generated in $y_{i'}$ according to Definition 1 or 2.

(3–4) “Mean-leverage" and “Variance-leverage" cases: simultaneous presence of conditional outliers (in mean or in variance) and leverage outliers, where the outliers i' are generated in $y_{i'}$ according to Definition 1 or 2, and in $x_{i'k}$ according to Definition 3, $k \in D$.

(5–6) “Mean-row" and “Variance-row" cases: simultaneous presence of conditional outliers (in mean or in variance) and rowwise outliers, where the outliers i' are generated in $y_{i'}$ according to Definition 1 or 2, and in $x_{i'k'}$ according to Definition 4, $k' \in \tilde{D}$.

(7–8) “Mean-leverage-row" and “Variance-leverage-row" cases: simultaneous presence of conditional outliers (in mean or in variance), leverage outliers and rowwise outliers, where the outliers i' are generated in $y_{i'}$ according to Definition 1 or 2, in $x_{i'k}$ according to Definition 3, $k \in D$, and in $x_{i'k'}$ according to Definition 4, $k' \in \tilde{D}$.

(9) "Leverage-row" case: simultaneous presence of leverage outliers and rowwise outliers, where the outliers i' are generated in $x_{i'k}$ according to Definition 3, $k \in D$, and in $x_{i'k'}$ according to Definition 4, $k' \in \tilde{D}$.

The technical derivations of expected perturbations in the expected value and in the variance of ϵ_i are reported in supplement Sect. 1 for each outlier scheme. They allow us to define $O_{all} = O_y$ for the cases where outliers in the response variable coexist with only one other type of outlier in the remaining variables (i.e., scenarios from 1 to 6), $O_{all} = O_y + O_{lev}$ for cases where all three types of outliers co-exist (scenarios 7–8), $O_{all} = O_{lev}$ for the case where rowwise and leverage outliers occur without any outlier in the response variable (scenario 9). We then define $\alpha = |O_{all}|$, where O_{all} is the set of actual conditional outliers to be recovered. Note that leverage outliers become conditional outliers under cases 7–9 (i.e., all “...leverage-row" cases), due to the contemporaneous presence of rowwise outliers and outliers in the predictors.

In the end, we can derive the expected overall Signal-to-Noise Ratio $SNR = \sqrt{ESS/RSS}$ of model (2) under the different cases, where ESS, the expected Explained Sum of Squares, is constant across cases and equal to $ESS = n\beta'\mathbf{R}_D\beta =$

$n\sum_{k=1}^{K}\beta_k[\beta_k+\mathbf{R}_{kk'}\sum_{k'\neq k}\beta_{k'}]$, and the formula for RSS, the expected Residual Sum of Squares, varies under each case (see Supplement Sect. 1). We highlight that ESS depends on the coefficient vector β (signs and magnitudes), the covariance matrix of predictors $\mathbf{R}_X$, and the number of predictors K, while RSS may depend on the outlier proportion α, the outlyingness parameter m, the intercept a, the residual variance σ^2, the number of predictors K, the perturbations occurred in the predictors, and the coefficient vector β, according to the underlying outlier scheme.

4.2 Methodology

We present here our proposed conditional outlier detection method, ROBOUT, consisting of five robust steps, namely, preprocessing (Sect. 4.2.1), preliminary imputation (Sect. 4.2.2), variable selection (Sect. 4.2.3), low-dimensional regression (Sect. 4.2.4) and conditional outlier detection (Sect. 4.2.5). Approximate normality is needed both in the target variable and in potential predictors, to make a linear model meaningful and to rely on residual diagnostics tools. The preliminary imputation procedure is needed to prevent anomalous points in potential predictors from corrupting predictor recovery, which is then performed via SNCD, applied to the robustly imputed dataset. In the end, a low-dimensional regression method like MM, which is doubly robust with respect to y and $\mathbf{X}_D$, is applied to spot conditional outliers. Note that ROBOUT, unlike SPARSE-LTS and RLARS, separates the steps of variable selection and conditional outlier detection, because this leads to a better performance for predictor selection and conditional outlier identification, as it is shown by our simulation study in Sect. 5.

4.2.1 Pre-processing

A pre-screening based on a robust correlation measure aimed at excluding one of two very correlated variables (say, more than 0.8 in absolute value) is first performed to detect nearly identities (see also Sect. 6) and to annihilate the impact of possibly masked bivariate outliers (see Sect. 4.2.2). For this purpose, we calculate robust pairwise correlations $\hat{\varrho}_{j'j''}$ for each pair $j'j''$ of variables, $j', j'' = 1, \ldots, p$, $j' \neq j''$. One can consider Spearman's rho (Spearman 1904), Kendall's tau (Kendall 1938), or Huber correlation (as employed in Khan et al. 2007), or other robust measures (see Raymaekers and Rousseeuw 2021).

If the model is intrinsically non-linear, polynomial effects could be included in the data matrix if known. Otherwise, as suggested by Rousseeuw and Bossche (2018), one can operate a pre-processing of the single variables in $\mathbf{X}$ by applying a non-linear transformation (like Box-Cox, Yeo-Johnson, etc.) able to ensure at least the symmetry of predictors, in order to avoid the insurgence of non-linear effects. The same holds *a fortiori* for the target variable $\mathbf{y}$, for which residual normality must hold in order to justify residual diagnostic analysis. In the empirical application of Sect. 6, applying the logarithmic transform is enough to obtain a normal y and a valid linear model. In general, the R function *bestNormalize* may be of great help.

If the variables in the dataset are intrinsically non-symmetric, then different specific procedures apply to determine the relevant outlyingness cut-offs (see for instance Rousseeuw and Hubert 2018). We do not include this case in the present paper.

4.2.2 Preliminary imputation

Let us consider the $n \times p$ matrix of potential predictors $\mathbf{X}$. First, we identify univariate outliers in that matrix. We calculate the median and the unbiased median absolute deviation of each variable. This means that, for each $j = 1, \dots, p$, we compute $x_{med,j} = \text{med}(x_j)$ and $x_{mad,j} = 1.4826\text{mad}(x_j)$.[2] Then, relying on $x_{med,j}$ and $x_{mad,j}$, we derive a robust z-score for each entry ij of $\mathbf{X}$: $z_{ij} = \frac{x_{ij} - x_{med,j}}{x_{mad,j}}$, $i = 1, \dots, n$, $j = 1, \dots, p$. In the end, we flag as outliers the entries that present an outlying z_{ij}, i.e. we set $w_{ij} = 0$ if $|z_{ij}| > \phi_z^{-1}(1 - \tilde{\delta}/2)$, where ϕ_z^{-1} is the inverse standard normal distribution function and $\tilde{\delta}$ is the significance level of the outlyingness test, usually set in a first run as 0.01 or 0.05, and $w_{ij} = 1$ otherwise.

Second, we impute values to the previously identified univariate outliers, utilising the information present in the whole matrix of potential predictors. Then, for each entry ij, $i = 1, \dots, n$, $j = 1, \dots, p$, we apply Algorithm 1.

Algorithm 1 Algorithm for preliminary imputation (depending on the outlyingness test level $\delta \in [0, 0.1]$, and the minimum neighbor set size $\zeta \in \mathbb{Z}^+ \cup 0$).

```
For each $ij$, $i = 1, \dots, n$, $j = 1, \dots, p$, such that $w_{ij} = 0$:
  1. derive the set of ordered variable indices $\widetilde{D}_j$ by sorting $|\hat{\varrho}_{j'j}|$, $j' \neq j$, in decreasing order;
  2. find the first index in the ordered set $\widetilde{D}_j$, $\tilde{j}$, such that $w_{i\tilde{j}} = 1$, if it exists;
  3. if $\tilde{j}$ does not exist, impute $x^*_{ij} = \text{med}(\mathbf{X}_{.j})$ and exit, otherwise, go to step 4;
  4. identify among all the points $i' \neq i$ the set $I_{ij}^{MAD}$ of all the neighbours of $x_{i\tilde{j}}$ with $w_{i'\tilde{j}} = 1$ such that
     $|x_{i'\tilde{j}} - x_{i\tilde{j}}| \leq x_{mad,\tilde{j}}$;
  5. calculate $\chi_{i',C}^{j\tilde{j}} = \mathbf{z}'_{i',j\tilde{j}} \widehat{\mathbf{R}}_{j\tilde{j}}^{-1} \mathbf{z}_{i',j\tilde{j}}$ for all points $i' \in I_{ij}^{MAD}$;
  6. derive the set $I_{ij}^{NB}$ of points $i' \in I_{ij}^{MAD}$ such that $\chi_{i',C}^{j\tilde{j}} \leq \phi_\chi^{-1}(1 - \delta, 2)$ and $w_{i'j} = 1$;
  7. if $|I_{ij}^{NB}| \geq \zeta$, go to step 8, otherwise, increase $\tilde{j}$ till $w_{i\tilde{j}} = 1$, and go to step 3;
  8. impute $x^*_{ij} = \text{med}(\mathbf{X}_{I_{ij}^{NB}j})$ and exit.
For each $ij$, $i = 1, \dots, n$, $j = 1, \dots, p$, such that $w_{ij} = 1$, set $x^*_{ij} = x_{ij}$.
```

Let us explain Algorithm 1 in more detail. We first identify univariate outliers in each variable by robustified z-scores with median and MAD. For any outlying cell (i, j), we use the robust correlation measure $\hat{\varrho}_{jj'}$, $j' \neq j$, to obtain a ranking of the most correlated variables with x_j. We consider the first one in the ranking such that $w_{ij'} = 1$, and we call it $\tilde{j}$. We derive the set of nearest neighbors $I_{ij,MAD}$ to be used for imputation by including all the points i' presenting $w_{i'\tilde{j}} = 1$ with $x_{i'\tilde{j}}$ within one MAD from $x_{i\tilde{j}}$. At this stage, since bivariate outliers may impact on the predictor set choice, we calculate the Mahalanobis distances of any point $i' \in I_{ij}^{MAD}$ by means of the vector of robustified scores $\mathbf{z}_{i',jj'}$ and the estimated covariance matrix $\widehat{\mathbf{R}}_{j\tilde{j}} = \begin{bmatrix} 1 & \hat{\varrho}_{jj'} \\ \hat{\varrho}_{jj'} & 1 \end{bmatrix}$ and we identify the set of bivariate outliers I_{ij}^{BIV} with a significance level equal to δ,

[2] In R, the function *mad* automatically calculates the re-scaled MAD $x_{mad,j}$.

typically fixed to 0.01 or 0.05. In the end, we derive the set $I_{ij}^{NB} = I_{ij}^{MAD} \setminus I_{ij}^{BIV}$ and we obtain the imputed value x_{ij}^{*} as the median value $\mathrm{med}(\mathbf{X}_{I_{ij}^{NB}j})$.

Three relevant features of Algorithm 1 need to be stressed. First, the computational cost is a direct function of $\sum_{i=1}^{n} \sum_{j=1}^{p} \mathbb{1}(w_{ij} = 0)$, that represents the degree of dataset perturbation. Second, Algorithm 1 exploits the multicollinearity structure of a large cross section, such that a large p and a rich multicollinearity structure are actually improving the imputation procedure systematically. Third, the procedure adapts the number of neighbors to the distribution of the closest related variable, by allowing all the points within one MAD in the neighbor set. The overall computational cost of the imputation step is thus proportional to $O(pn \log n)$ as p and n diverge.

At this stage, it may still be objected that Algorithm 1 may miss a fraction of bivariate outliers, namely, all bivariate outliers $\mathbf{x}_{i,j'j''}$ such that both $w_{ij'} = 1$ and $w_{ij''} = 1$. Although this occurs very rarely in practice, it is indeed true, and the size of this set of masked bivariate outliers is actually enlarging if $\mathrm{sgn}(x_{ij'}) = \mathrm{sgn}(x_{ij''})$ and $\varrho_{j'j''}$ is strongly negative, or if $\mathrm{sgn}(x_{ij'}) \neq \mathrm{sgn}(x_{ij''})$ and $\varrho_{j'j''}$ is strongly positive.

The practical remedy for this potential drawback is to re-apply Algorithm 1 with larger levels $\tilde{\delta}$ and δ. In particular, it can be noted that setting $\tilde{\delta} = 0.1$ is enough to obtain a bivariate outlyingness test level $\delta = 0.01$ when $\mathrm{sgn}(x_{ij'}) = \mathrm{sgn}(x_{ij''})$ and $\varrho_{j'j''} = -0.4125$ (or $\mathrm{sgn}(x_{ij'}) \neq \mathrm{sgn}(x_{ij''})$ and $\varrho_{j'j''} = 0.4125$). This means that if the predictor set retrieved in Sect. 4.2.3 changes abruptly when $\tilde{\delta}$ is lower than a certain level, one should stop the descent of $\tilde{\delta}$ at the previous value. We suggest trying $\tilde{\delta} = \delta = 0.1, 0.05, 0.01$ and verifying the stability of the recovered predictor set, as we do in Sect. 6. In the supplement, at the end of Sect. 6, we have devoted a specific experiment to this limit case.

Concerning potential outliers in more than two dimensions, we can note that their impact on predictor selection by LASSO-penalized robust regression (see Sect. 4.2.3) is actually negligible, because Algorithm 1 preserves the correct estimation of pairwise correlations, which is crucial to solve iteratively reweighted least squares problems like (3). For this reason, Algorithm 1 is effective for the purpose of preserving the true predictor recovery in the presence of outliers such as those of Definitions 3 and 4, under which SPARSE-LTS and RLARS break down (see Sect. 5). In the robust regression step (see Sect. 4.2.4), a doubly robust method like MM is employed, so that those points can eventually be recovered as conditional outliers in y.

Given that we allow for the presence of bivariate outliers, we should wonder if the three considered robust correlation measures, Spearman, Kendall and Huber, are sensitive to bivariate outliers and how much. This additional simulation study is reported in Supplement Sect. 4. We discover that, among the three measures, Kendall correlation is the most biased, followed by Spearman and Huber. At the same time, once fixed $R_{ij} = \varrho^{|i-j|}$, even Kendall correlation correctly selects the most correlated variable almost always, unless ϱ is very small, like the other two measures. This means that, even using a correlation measure which is non-robust to bivariate outliers, the identification of the closest variable is not affected. This is why in the simulation study of Sect. 5 we employ Kendall correlation, thus proving that the most biased correlation measure among the three still produces excellent results for ROBOUT, although the calculation of χ-score at step 5 may be biased. What is more, a large ϱ enlarges the set of possibly masked bivariate outliers, but at the same time ensures the selection of the closest variable via any correlation measure (although not robust to bivariate outliers) under any p/n ratio.

Algorithm 1 is also equipped with systematic protections against mis-imputation. In particular, step 4 selects the neighbor set as the one containing all points within one MAD in the closest variable, in order to prevent overfitting. Then, the minimum neighbor set size ζ can further control at step 7 that the neighbor set is large enough to avoid overfitting. Throughout the paper we set $\zeta = 3$, but ζ may also be selected by cross-validation. A large p increases prediction informativeness (thus decreasing imputation bias), and a large n prevents overfitting (thus decreasing imputation variance). Increasing both p and n concurs to prevent any point to be imputed by the general median of x_j, which occurs at step 3 if no close variables are available for imputation. That situation is very unlikely to occur, unless p and n are really small. For these reasons, our method is a large p large n one, where multicollinearity actually improves imputation accuracy, because it increases the probability to select the closest variable in step 2. We stress that Algorithm 1 analogously functions when applied to missing data entries.

Alternative methods include employing a correlation measure proved to be robust to any bivariate outlier, like the one in Raymaekers and Rousseeuw (2021), or computing a fast MCD as proposed in Riani et al. (2015). These two alternatives would not help anyway overcome the current limitations of ROBOUT, which require to have less than 50% univariate outliers per variable, or bivariate outliers per variable pair, and require that all points have enough variable values to exploit for a safe imputation, which means that the maximum number of variables involved in outliers per each row must be limited, and enough close points for imputation must be present (that is, we need approximately symmetric distributions). This is the reason why randomly scattered outliers in the style of Rousseeuw and Bossche (2018) and Bottmer et al. (2022) are not well managed by Algorithm 1. We leave to future studies to modify Algorithm 1 to overcome these limitations, in order to compare theoretically and empirically its performance to the shooting S-estimator proposed by Bottmer et al. (2022).

4.2.3 Variable selection

Starting with our target response variable y_i, $i = 1, \dots, n$, we aim to consistently select the relevant set of its predictors from a large set of variables under the presence of conditional outliers in mean or variance. Consequently, we aim to identify a model such as (2) from the data. Importantly, the focus of this step is on the selection of predictors rather than on the estimated coefficients and residual scale. This is a core idea of ROBOUT, i.e. that a robust predictor selection step is distinguished from the robust regression step, as this renders the method more reliable in challenging situations compared to the existing methods that combine these two steps.

Predictor retrieval receives as input the dataset imputed by Algorithm 1 without further standardization or normalization.[3] Two options are considered for this step.

[3] The pre-processing operations of Sect. 4.2.1 prevent the variables to have tremendously different scale. If the variables are of different nature, a preliminary robust standardization is adviced.

The first is based on Yi and Huang (2017):

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} \rho_{H,\theta}(\epsilon_i) + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{3}$$

where $\epsilon_i = y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}^*$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, λ is a penalization parameter, and

$$\rho_{H,\theta}(t) = \begin{cases} \frac{t^2}{2\theta}, \ |t| \le \theta \\ |t| - \frac{\theta}{2}, \ |t| > \theta, \end{cases}$$

is the Huber weight function (Ronchetti and Huber 2009), where θ is a tuning parameter. Henceforth, we call (3) SNCD-H objective function.

SNCD-H estimates an elastic-net penalized Huber loss regression, optimized by using the semi-smooth Newton coordinate descent algorithm presented in Yi and Huang (2017). SNCD has a computational cost proportional to $O(pn)$, i.e. linear in both p and n. Weighting observations is precisely what renders the results robust in the face of conditional outliers in y, because it annihilates their influence.

The other robust alternative that we consider in the simulation study substitutes $\rho_{H,\theta}(\epsilon_i)$ in (3) with the absolute loss $\rho_L(\epsilon_i)$, where

$$\rho_L(t) = t\left(\frac{1}{2} - \mathrm{I}(t < 0)\right), \ t \in \mathbb{R}.$$

We call that version the SNCD-L objective function. SNCD-L estimates an elastic-net penalized absolute loss regression, optimized in the same way.

The minimization of (3) with $\rho_{H,\theta}(t)$ or $\rho_L(t)$ is practically performed for a decreasing sequence of values $\lambda = \lambda_t$, $t = 1, \ldots, T$, such that $\lambda_0 = \lambda_{max}$ and $\lambda_T = \lambda_{min}$, where λ_{max} returns no predictors, and λ_{min} returns K_{max} predictors (we set $K_{max} = [p/10]$). Predictor selection is performed by the adaptive version of the strong rule of Tibshirani et al. (2012), proposed in Yi and Huang (2017). At each value of λ_t, that rule exploits the coefficient estimates at λ_{t-1}. The optimal λ is then selected by cross-validation, using as Out-Of-Sample (OOS) metric the loss $\rho_{H,\theta}(t)$ for SNCD-H or the loss $\rho_L(t)$ for SNCD-L.

Following Friedman et al. (2010), two optimal values of λ can be selected. The value λ_{0SE} is the one returning the minimum OOS loss $\rho_{H,\theta}^{min}$ or ρ_L^{min} along the sequence λ_t. The value λ_{1SE} is the maximum value along the sequence λ_t such that the OOS metric $\rho_{H,\theta}$ or ρ_L stands within one standard error by $\rho_{H,\theta}^{min}$ or ρ_L^{min}. In this application, we use $\lambda_{min} = \lambda_{0SE}$, and then we apply a screening rule on the selected predictors by trimming the relative estimated coefficients $\widehat{\beta}_{min}$ as follows: $\widehat{\beta}_{min}^* = \widehat{\beta}_{min} \mathbb{1}(|\widehat{\beta}_{min}| > \phi_z^{-1}(1-\delta/2)\rho_{H,\theta}^{min})$ for SNCD-H, or $\widehat{\beta}_{min}^* = \widehat{\beta}_{min} \mathbb{1}(|\widehat{\beta}_{min}| > \phi_z^{-1}(1-\delta/2)\rho_L^{min})$ for SNCD-L (δ is typically set to 0.01 or 0.05). The predictor indices corresponding to the non-null entries of $\widehat{\beta}_{min}^*$ constitute the set $\widehat{D}$ of selected predictor indices, and the estimate of K is the size of $\widehat{D}$, $\widehat{K} = |\widehat{D}|$. This screening process is able to systematically avoid redundancy in the predictor set, which is typical of existing procedures such as SPARSE-LTS and RLARS, by killing irrelevant predictors while keeping the right predictors contained in D into $\widehat{D}$.

4.2.4 Robust regression

Once identified in the previous step the predictors of y in the matrix $\mathbf{X}_{\widehat{D}}$, we apply a robust low-dimensional regression method in order to robustly estimate coefficients, residuals, and residual scale. To this aim, we utilise MM regression (Yohai 1987). In the simulation study, we also test the performance of Least Trimmed Squares (LTS) (Rousseeuw 1984). Concerning computational cost, we know that both LTS and MM share a burden of $O(n)$ operations, due to the use of Fast-LTS (Rousseeuw and Van Driessen 2006) for LTS, and of Fast-S (Salibian-Barrera and Yohai 2006) for the initial scale in MM. Residual indipendence, homoscedasticity, and normality must be appropriately verified by residual diagnostics, and must hold in the recovered non-outlier set. At the end of this step, we get the estimated $\widehat{K} \times 1$ vector of coefficients $\widehat{\beta}$, the estimated $n \times 1$ vector of residuals $\widehat{\epsilon}$, and the residual scale estimate $\widehat{\sigma}$.

4.2.5 Outlier detection

As a last step, we recover the vector $\widehat{O}$ of conditional outlier indices as the set of all the points $i \in \{1, \ldots, n\}$ with robustly rescaled residuals $\widehat{\epsilon}_i/\widehat{\sigma}$ larger than $\phi_z^{-1}(1-\delta/2)$ in absolute value, where δ is typically set to 0.01 or 0.05 (see for instance Atkinson and Riani 2000; Rousseeuw and Van Driessen 2006; Rousseeuw and Bossche 2018; Alfons et al. 2013, or Khan et al. 2007).

5 A comparative simulation study

In this section, we conduct simulation experiments aiming both to compare the performance of ROBOUT to competitor methods but also to identify the optimal design of ROBOUT with respect to its constituent components. The competitor methods against which the ROBOUT versions are tested are SPARSE-LTS and RLARS.[4]

5.1 Parameter settings

We test the performance of the four variants of the ROBOUT methodology, distinguished on the basis of the variable selection and robust regression estimation options, as presented in Sect. 4.2. Specifically, in the first step, either the SNCD-H or the SNCD-L objective function can be used, while in the second step either LTS or MM

[4] We also experimented with additional competitor methods, namely MM-LASSO, ENET-LTS and PENSEM, on the same simulation scenarios, but the results were clearly worse compared to the rest of the methods we tried, therefore we do not present the results for simplicity. More, we did not implement (Bottmer et al. 2022) because of computational cost, difficult parameter tuning, and different underlying assumptions.

can be used to estimate the regression equation. We call the ensuing four variants of ROBOUT as H+LTS, L+LTS, H+MM and L+MM, where H and L refer to SNCD-H and SNCD-L, respectively. We provide detailed information about the parameterisation of the various methods in the Sect. 2 of the supplementary material.

Both competitors are run with the default preprocessing step for all the potential predictors: unit-norm normalization for SPARSE-LTS, robust standardization for RLARS. The intercept is included for all estimations, and the levels δ and $\tilde{\delta}$ are fixed to 0.01. Note that, according to the above settings, H+MM, H+LTS, and SPARSE-LTS present a breakdown point equal to 25%, L+MM, L+LTS, and RLARS present a breakdown point equal to 50%.

5.2 Data settings

To fully define the scenarios that we use in the simulation study, we complement the nine outlier schemes presented in Sect. 4 with specifications of the dataset dimensions and parameters.

The set of predictors D is randomly generated with $K = 3$. The intercept is $a_i = 10$ for all $i = 1, \dots, n$. The coefficients are generated in the following way: $\beta_1, \beta_3 \sim U(15, 20)$, $\beta_2 \sim U(-20, -15)$, with residual variance $\sigma^2 = 1$. According to the definitions provided in Sect. 4.1, we generate the data from model (1) for $m = 1, 5, 9, 13, 17, 23, 29, 37, 45, 55$, where m represents the outlyingness parameter (the larger it is, the more perturbed the data setting). In practice, we generate leverage outliers i' as $\mathbf{x}_{D,i'} \sim MVN(\mathbf{m}, \mathbf{R}_D/m)$, where $\mathbf{m} \in \mathbb{R}^K$, $|m_k| \gg 0$, $k = 1, \dots, K$, to obtain some (small) variability among outliers.

Each scenario is further defined by the relative size of the matrix (i.e., the p/n ratio). We examine four different settings for the dimensions and the ensuing p/n ratio of the dataset: “Very fat" (VF), with $p = 300$, $n = 60$ (i.e. $p/n = 5$, $p \gg n$); “Fat" (F), with $p = 200$, $n = 100$ (i.e. $p/n = 0.5$, $p < n$); “Tall" (T), with $p = 100$, $n = 200$ (i.e. $p/n = 2$, $p > n$); “Very tall" (VT), with $p = 60$, $n = 300$ (i.e. $p/n = 0.2$, $p \ll n$). As suggested by Alfons (2021), we set the $p \times p$ covariance matrix $\mathbf{R}_\mathbf{X}$ as $\mathbf{R}_{j'j''} = 0.5^{|j'-j''|}$, $j', j'' = 1, \dots, p$, $j' \neq j''$.

The rationale for differentiated scenarios with respect to the p/n ratio is that relative dimensions affect the performance of the various methods that are tested, for example by impacting on the effectiveness of the selection of predictors.

The contamination rates α_y, α_{lev}, and α_{row} and the number of variables in the set of perturbed predictors $\tilde{D}$, $|\tilde{D}|$, vary according to the outlier scheme and the dimension setting. The general idea behind the choice of α_y, α_{lev}, α_{row} and $|\tilde{D}|$ is that a larger contamination rate is expected as n increases because more reasons for outlying behaviour may materialise (e.g., presence of sub-populations generated by a different statistical process). More details about the contamination rates and the signal-to-noise ratios across scenarios are provided in the Supplementary material (Sect. 3).

Henceforth, we refer to scenarios by combining the name of the outlier scheme, e.g., “mean-row" and the dimensional set-up, e.g., “tall" dataset. Each scenario is run 100 times for each value of m and the various performance metrics, presented in the next section, are calculated by averaging across these 100 replications.

5.3 Performance metrics

Performance metrics, as stated in the introductory section, pertain to three different dimensions: outlier detection, predictor recovery, and predictors' coefficient estimation.

Let us define D_k as the k-th predictor in the set D and $\setminus$ as the set difference. To measure the performance with respect to predictor recovery, we calculate for each scenario, each value of parameter m, each outlier detection method, and each iteration $r = 1, \dots, 100$, the set of recovered predictors $\widehat{D}^{(r)}$. $\widehat{D}^{(r)}$ could be in general either a subset, superset or non-overlapping set with respect to D, ideally however the two sets should be identical. Consequently, we calculate the following metrics:

- masked predictor rate, $MPR_r = K^{-1}|D \setminus \widehat{D}^{(r)}|$, which is the false negative rate of recovered predictors;
- swamped predictor rate, $SPR_r = |\widehat{D}^{(r)}|^{-1}|\widehat{D}^{(r)} \setminus D|$, which is the false positive rate of recovered predictors;
- recovered predictor ratio, $RPR^{(r)} = K^{-1}|\widehat{D}^{(r)}|$, which is a measure of the propensity to recover irrelevant predictors;
- true predictor rate, defined as $TPR^{(r)} = \mathbb{1}(MPR^{(r)} = 0)$, which is the predictor set recovery rate;
- true predictor rate for each predictor D_k, $k = 1, \dots, K$, defined as $TPR_k^{(r)} = \mathbb{1}(D_k \in \widehat{D}^{(r)})$;
- adjacent predictor rate, defined as $APR^{(r)} = 1$ if $K^{-1}\sum_{k=1}^{K} \mathbb{1}(D_k \in \widehat{D}^{(r)} | (D_k - 1) \in \widehat{D}^{(r)} | (D_k + 1) \in \widehat{D}^{(r)}) = 1$, $APR^{(r)} = 0$ otherwise, which is a measure of the propensity to recover wrong models having strongly correlated predictors with the true ones.

To quantify the performance of coefficient estimation, we obtain estimated intercept $\widehat{\beta}_0^{(r)}$ and coefficient vector $\widehat{\beta}^{(r)}$, and we calculate:

- for each coefficient β_k, $k = 1, \dots, K$, the coefficient squared error, defined as $CSE_k^{(r)} = (\widehat{\beta}_k^{(r)} - \beta_k)^2$ if $TPR_k^{(r)} = 1$, $CSE_k^{(r)} = 0$ if $TPR_k^{(r)} = 0$;
- the intercept squared error, $ISE^{(r)} = (\widehat{\beta}_0^{(r)} - \beta_0)^2$.

The two indicators are then averaged as follows:

- the average root relative coefficient squared error, $RCSE = K^{-1}\sum_{k=1}^{K} |\beta_k|^{-1} \sqrt{(\sum_{r=1}^{R} TPR_k^{(r)})^{-1}\sum_{r=1}^{R} CSE_k^{(r)}}$;
- the average root relative intercept squared error, $RISE = |\beta_0|^{-1}\sqrt{R^{-1}\sum_{r=1}^{R} ISE^{(r)}}$.

To assess performance with regard to conditional outlier detection, we derive for each iteration $r = 1, \dots, 100$ the set of recovered outliers $\widehat{O}^{(r)}$. Then, we calculate the following performance metrics:

- outlier rate ratio, defined as the ratio $OR^{(r)} = [\alpha n]^{-1}|\widehat{O}^{(r)}|$;
- zero-outlier indicator, defined as $OZ^{(r)} = 1$ if $|\widehat{O}^{(r)}| = 0$, 0 otherwise;
- the masking rate $MR^{(r)}$, defined as the proportion of masked outliers (i.e. false negatives) over the true number of outliers: $MR^{(r)} = [\alpha n]^{-1}\sum_{i \in O} \mathbb{1}(i \notin \widehat{O}^{(r)})$;

- the swamping rate $SR^{(r)}$, defined as the proportion of swamped outliers (i.e. false positives) over the number of recovered outliers: $SR^{(r)} = |\widehat{O}^{(r)}|^{-1} \sum_{i \in \widehat{O}^{(r)}} \mathbb{1}(i \notin O)$;
- the F_1 score, defined as $2\frac{PREC^{(r)} \times REC^{(r)}}{PREC^{(r)}+REC^{(r)}}$, where the precision $PREC^{(r)}$ is equal to $PREC^{(r)} = 1 - MR^{(r)}$, and the recall $REC^{(r)}$ is equal to $REC^{(r)} = 1 - SR^{(r)}$, which is an overall performance measure incorporating both masking and swamping effects.

The above measures are averaged across the 100 replications of each scenario and value of m for each method, and their standard deviation is calculated where appropriate.

5.4 Simulation results

Our simulations, presented in this and the next Sect. 5.5 and complemented by the supplementary material, show that the ROBOUT approach compares favourably to the competitors, as it is not plagued by the systematic predictor redundancy of RLARS and the systematic sub-optimal estimation of regression coefficients of SPARSE-LTS. We also find that overall H+MM is the most reliable ROBOUT option because it is expected to recover predictors, estimate coefficients and identify outliers in a systematically better and more stable way than competitors. In other words, the optimal ROBOUT version utilises the MM regression as the final robust regression step, similarly to RLARS, while modifying RLARS predictor selection procedure to be robust to multivariate outliers and parsimonious.

In more detail, under all scenarios ROBOUT performs better than SPARSE-LTS and RLARS with respect to coefficient estimation (see Table 1). The competitor methods exhibit relatively high values of $RCSE$; SPARSE-LTS even exceeds 60%, especially as p/n decreases, i.e., as the dataset becomes “taller", and as the outlier scheme becomes more complex, e.g., in the case of the outlier scheme 7. RLARS faces problems in avoiding the inclusion of irrelevant predictors, presenting RPR much larger than 1 (see Fig. 2 and also Sects. 6 and 7 in the supplementary material). The above patterns hold both for scenarios with conditional outliers in mean and in variance.

Concerning outlier detection, the performance of ROBOUT is clearly superior in the more complex outlier scheme “mean/variance-leverage-row" (schemes 7 and 8), see, e.g., Fig. 3. SPARSE-LTS is doing systematically worse than competitors when conditional outliers in variance are present (see the results for the outlier schemes 2, 4 and 6 in Table 2). RLARS faces challenges in recovering the true predictors and the same holds for SPARSE-LTS when $p/n < 1$ for “tall" and “very tall" datasets (see the supplementary material for the respective TPR scores).

We also observe a difference between the ROBOUT versions SNCD-H and SNCD-L, in that the former is slightly more parsimonious than the latter in predictor recovery for the “very fat" dataset as reflected in the RPR metric (see Sects. 6 and 7 in the supplementary material). Concerning outlier detection, a difference between MM and LTS emerges, that is, the LTS options of ROBOUT are systematically more erratic when the “very fat" dataset is used (see Table 2).

Table 1 Average root relative coefficient squared error $RCSE$ across all scenarios and for all methods considered

Scenario	H+MM	H+LTS	L+MM	L+LTS	S-LTS	RLARS
“Very fat"						
Scheme						
1	0.024	0.035	0.041	0.047	0.081	0.027
2	0.009	0.010	0.009	0.010	0.205	0.009
3	0.028	0.050	0.046	0.040	0.074	0.040
4	0.007	0.008	0.007	0.008	0.065	0.008
5	0.016	0.018	0.036	0.038	0.080	0.010
6	0.020	0.019	0.024	0.024	0.212	0.009
7	0.056	0.060	0.074	0.098	0.070	0.171
8	0.010	0.010	0.011	0.012	0.103	0.140
9	0.004	0.005	0.004	0.005	0.048	0.005
“Fat"						
Scheme						
1	0.007	0.008	0.007	0.008	0.053	0.007
2	0.007	0.007	0.007	0.007	0.063	0.007
3	0.006	0.007	0.006	0.007	0.054	0.011
4	0.007	0.008	0.007	0.008	0.065	0.008
5	0.007	0.007	0.007	0.007	0.052	0.007
6	0.007	0.007	0.007	0.007	0.061	0.008
7	0.006	0.007	0.006	0.007	0.206	0.215
8	0.006	0.007	0.006	0.006	0.042	0.046
9	0.004	0.004	0.004	0.004	0.044	0.008
“Tall"						
Scheme						
1	0.004	0.005	0.004	0.005	0.049	0.007
2	0.006	0.006	0.006	0.006	0.060	0.006
3	0.004	0.004	0.004	0.004	0.048	0.005
4	0.006	0.006	0.006	0.006	0.080	0.006
5	0.004	0.005	0.004	0.005	0.048	0.005
6	0.005	0.005	0.005	0.005	0.056	0.005
7	0.005	0.005	0.005	0.005	0.643	0.336
8	0.005	0.005	0.005	0.005	0.031	0.052
9	0.004	0.004	0.004	0.004	0.044	0.006
“Very tall"						
Scheme						
1	0.004	0.004	0.004	0.004	0.044	0.008
2	0.004	0.004	0.004	0.004	0.055	0.004
3	0.004	0.004	0.004	0.004	0.044	0.006
4	0.004	0.004	0.004	0.004	0.055	0.004
5	0.004	0.004	0.004	0.004	0.045	0.009

Table 1 continued

Scenario	H+MM	H+LTS	L+MM	L+LTS	S-LTS	RLARS
6	0.005	0.005	0.005	0.005	0.055	0.004
7	0.005	0.005	0.005	0.005	0.630	0.291
8	0.005	0.005	0.005	0.005	0.044	0.230
9	0.004	0.004	0.004	0.004	0.045	0.009

The results are obtained when the perturbation factor is $m = 23$, i.e., in the middle of the considered range. Each scenario is characterised by the dimensions of the dataset ("Very fat", "Fat", "Tall" and "Very tall") and the outlier scheme numbered as in Sect. 4.1, e.g., Scheme 1 corresponds to the "Only mean" outlier scheme

Furthermore, under scenarios featuring the outlier scheme "leverage-row" (case 9), SPARSE-LTS and RLARS do not perform well under the "very fat" case and their performance improves as the dataset becomes "taller". Finally, all ROBOUT options are more effective than competitors also in the absence of outliers.

We now report in detail the results for scenarios featuring outliers in mean (schemes 1, 3, 5 and 7), which provide a representative overview of the relative performance of the various methods. We refer to supplementary material for the results of all other scenarios.

5.5 Scenarios with outliers in mean

We first examine the cases where a maximum two types of outliers are present, one of them in being outliers in mean for the response variable. Specifically, this applies for the outlier schemes 1, 3 and 5. A representative sample of these findings is shown in Fig. 4, for the scenario with the "very fat" dataset and the "mean-leverage" outlier scheme. In these scenarios and irrespective of the p/n dimensions, the average of TPR is around 1 for all methods, meaning that all methods perform satisfactorily with respect to predictor identification in these relatively simple scenarios. Concerning outlier detection, the F_1 score of all methods is very close to 1, however we note that H+LTS and L+LTS are slightly more erratic than the competitors, reflected in their somewhat elevated standard deviation across iterations.

In general and especially as the p/n ratio increases, the performance of ROBOUT stands out compared to the competitor methods, which face challenges regarding predictor recovery and coefficient estimation. This is shown clearly in Fig. 4 where the "very fat" dataset is used. These challenges are caused by the relatively high dimensions of the potential predictors set. For example, SPARSE-LTS faces severe challenges as regards the coefficient estimation, performing systematically worse than other methods with respect to the $RCSE$ and $RISE$ indicators. The SPARSE-LTS estimated coefficients are structurally worse: on average, $RISE$ is around 4% for SPARSE-LTS, while all the other methods stand below 1%. Furthermore, RLARS performs worse than the other methods with regard to predictor recovery: the RPR indicator for RLARS is well beyond 1. Finally, for high values of m RLARS often crashes, due to multicollinearity issues.

Table 2 F_1 score (mean and standard deviation), considering outlier schemes 1 to 6 as in Sect. 4.1, e.g., the number 1 corresponds to the "Only mean" outlier scheme

Scenario		H+MM	H+LTS	L+MM	L+LTS	S-LTS	RLARS
"Very fat"							
Scheme							
1	Mean	0.987	0.973	0.986	0.964	0.973	0.987
	Std	0.035	0.142	0.037	0.172	0.063	0.029
2	Mean	0.921	0.909	0.921	0.909	0.726	0.923
	Std	0.084	0.094	0.084	0.094	0.233	0.082
3	Mean	0.983	0.967	0.981	0.967	0.971	0.986
	Std	0.037	0.172	0.041	0.172	0.064	0.033
4	Mean	0.936	0.911	0.936	0.911	0.857	0.934
	Std	0.047	0.066	0.047	0.066	0.078	0.045
5	Mean	0.983	0.985	0.984	0.965	0.973	0.989
	Std	0.036	0.101	0.036	0.172	0.053	0.028
6	Mean	0.923	0.915	0.923	0.915	0.737	0.933
	Std	0.109	0.123	0.111	0.121	0.242	0.075
7	Mean	0.992	0.950	0.991	0.955	0.985	0.943
	Std	0.050	0.219	0.049	0.199	0.085	0.126
8	Mean	0.961	0.934	0.949	0.938	0.911	0.914
	Std	0.047	0.108	0.108	0.108	0.150	0.158
9	Mean	0.981	0.986	0.975	0.983	0.919	0.959
	Std	0.046	0.042	0.052	0.043	0.203	0.146
"Fat"							
Scheme							
1	Mean	0.995	1	0.995	1	0.997	0.995
	Std	0.013	0	0.013	0	0.011	0.014
2	Mean	0.939	0.919	0.939	0.919	0.845	0.939
	Std	0.048	0.052	0.048	0.052	0.082	0.049
3	Mean	0.997	1	0.997	1	0.997	0.998
	Std	0.009	0	0.009	0	0.012	0.008
4	Mean	0.936	0.911	0.936	0.911	0.857	0.934
	Std	0.047	0.066	0.047	0.066	0.078	0.045
5	Mean	0.997	1	0.997	1	0.997	0.996
	Std	0.009	0	0.009	0	0.009	0.011
6	Mean	0.934	0.913	0.934	0.913	0.864	0.932
	Std	0.053	0.059	0.053	0.059	0.075	0.052
7	Mean	0.999	1.000	0.998	0.998	0.955	0.926
	Std	0.007	0.000	0.006	0.010	0.118	0.125
8	Mean	0.967	0.952	0.969	0.953	0.941	0.944
	Std	0.026	0.036	0.026	0.036	0.041	0.083

Table 2 continued

Scenario		H+MM	H+LTS	L+MM	L+LTS	S-LTS	RLARS
9	Mean	0.984	0.993	0.982	0.996	0.989	0.979
	Std	0.031	0.019	0.031	0.013	0.022	0.033
“Tall"							
Scheme							
1	Mean	0.999	1	0.999	1	0.999	0.999
	Std	0.004	0	0.004	0	0.003	0.004
2	Mean	0.933	0.893	0.933	0.893	0.846	0.933
	Std	0.031	0.040	0.031	0.040	0.052	0.030
3	Mean	0.999	1	0.999	1	0.999	0.999
	Std	0.004	0	0.004	0	0.003	0.003
4	Mean	0.933	0.894	0.933	0.894	0.844	0.934
	Std	0.030	0.041	0.030	0.041	0.050	0.030
5	Mean	0.999	1	0.999	1	1	1
	Std	0.003	0	0.003	0	0	0
6	Mean	0.931	0.893	0.931	0.893	0.843	0.931
	Std	0.032	0.048	0.032	0.048	0.060	0.032
7	Mean	1.000	1.000	1.000	1.000	0.620	0.650
	Std	0.000	0.000	0.000	0.000	0.065	0.158
8	Mean	0.963	0.929	0.963	0.930	0.943	0.919
	Std	0.018	0.024	0.018	0.024	0.019	0.123
9	Mean	0.996	0.999	0.995	1.000	0.999	0.971
	Std	0.009	0.010	0.009	0.000	0.005	0.136
“Very tall"							
Scheme							
1	Mean	1	1	1	1	1	1
	Std	0	0	0	0	0	0
2	Mean	0.929	0.870	0.929	0.870	0.834	0.929
	Std	0.024	0.034	0.024	0.034	0.044	0.025
3	Mean	1	1	1	1	1	1
	Std	0	0	0	0	0	0
4	Mean	0.928	0.867	0.928	0.867	0.835	0.928
	Std	0.019	0.035	0.019	0.035	0.044	0.019
5	mean	1	1	1	1	1	1
	Std	0	0	0	0	0	0
6	Mean	0.927	0.867	0.927	0.867	0.835	0.928
	Std	0.023	0.033	0.023	0.033	0.044	0.019
7	Mean	1.000	1.000	1.000	1.000	0.724	0.751
	Std	0.000	0.000	0.000	0.000	0.012	0.075

Table 2 continued

Scenario		H+MM	H+LTS	L+MM	L+LTS	S-LTS	RLARS
8	Mean	0.951	0.899	0.952	0.899	0.874	0.688
	Std	0.014	0.023	0.014	0.023	0.014	0.406
9	Mean	0.999	1.000	0.999	1.000	1.000	0.830
	Std	0.003	0.000	0.003	0.000	0.001	0.363

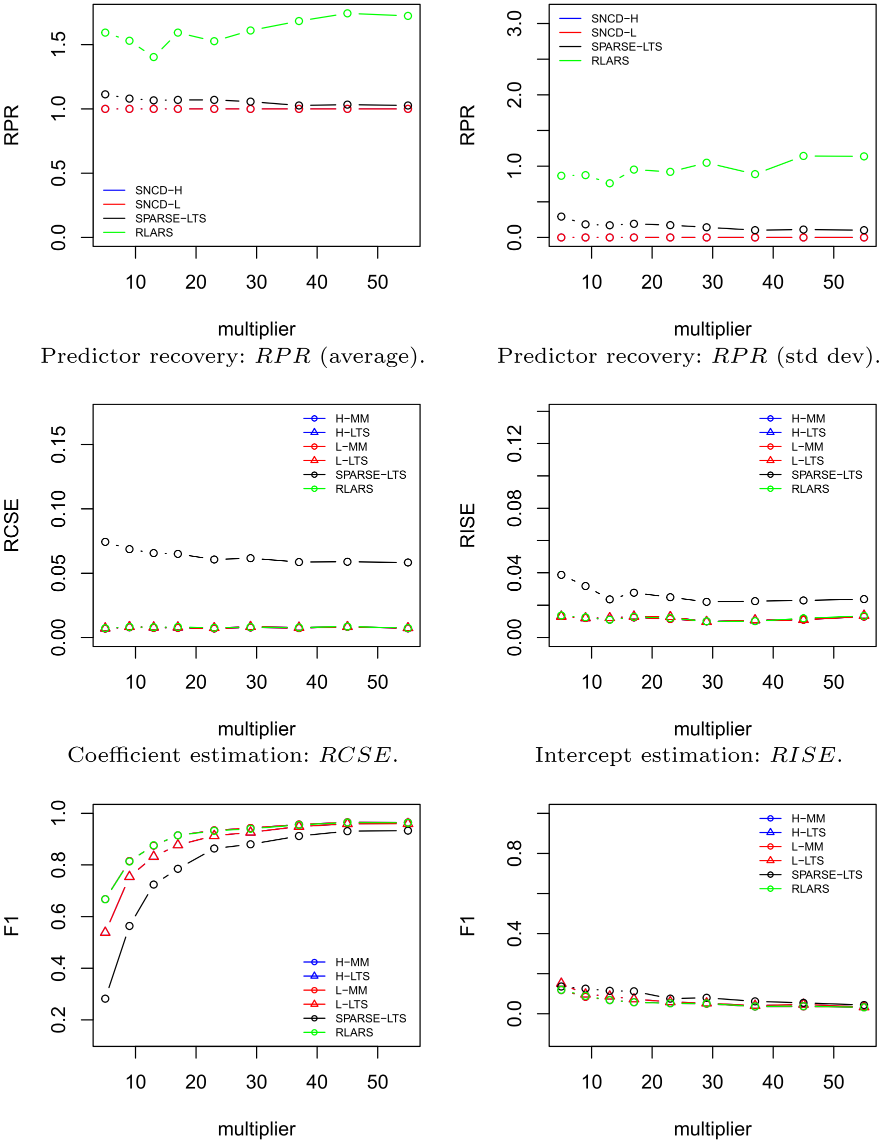
The results are obtained when the perturbation factor is $m = 23$, i.e., in the middle of the considered range. Results for all four datasets are presented

In the less extreme case of a “fat" dataset, these challenges to the competitors are attenuated, however ROBOUT remains overall the preferred choice. The metrics for the scenario with a “fat" dataset and a “mean-row" outlier scheme are shown in Fig. 5. On the one hand, the RPR of SPARSE-LTS converges to 1, the instability of F_1 score for H+LTS and L+LTS disappears, and the patterns of $RCSE$ and $RISE$ are much smoother. However, the RPR of RLARS keeps well beyond 1, and the relative standings of $RCSE$ and $RISE$ maintain the same pattern, with SPARSE-LTS standing around 4% and the other methods below 1%. For “tall" and “very tall" datasets, the patterns remain qualitatively similar.

We now examine the more challenging scenarios featuring all three types of outliers, i.e., the “mean-leverage-row" outlier scheme (case 7). In the “very fat" dataset case, predictor recovery is suffering more than under the simpler schemes 1, 3 and 5 examined previously (see Fig. 6). The ROBOUT versions SNCD-H and SNCD-L are consistently the best, in all aspects. First, both the two competitor methods underperform with respect to predictor recovery. For small values of m, SPARSE-LTS is not effective, exhibiting values of TPR below 0.8, even if the performance improves for larger m. RLARS performs even worse with values of the TPR persistently low, irrespective of the degree of perturbation.[5] Similarly, as regards the coefficient estimation, we observe that RLARS is standing around 20% for $RCSE$, while the performance of SPARSE-LTS improves as m increases, paralleling the Huber options of ROBOUT, which are the most stable. The competitor methods are also worse than ROBOUT with respect to the outlier detection, as can be seen clearly in the F_1 score. Based on this metric, RLARS exhibits by far the worst performance, due to a masking rate which stabilizes around 20%. SPARSE-LTS starts from a masking and swamping rate above 15% and its performance improves when m increases. The MM-based versions of ROBOUT are the best performing because the versions using the LTS option present a small but systematic masking rate (around 5%). Overall, it is found that the H+MM option of ROBOUT is the most stable and reliable method when predictor selection, coefficient estimation and outlier detection are jointly considered.

In the milder case where the outlier scheme “mean-leverage-row" (case 7) is combined with a “fat" dataset (see Fig. 7), the pattern of TPR does not qualitatively change compared to the “very fat" dataset case. Concerning predictor recovery, the

[5] When the APR metric is considered, which considers as successes also the cases where the recovered predictors are contiguous to the true ones, RLARS improves but does not attain values above the range 0.7-0.8.



Predictor recovery: RPR (average).

Predictor recovery: RPR (std dev).

Coefficient estimation: $RCSE$.

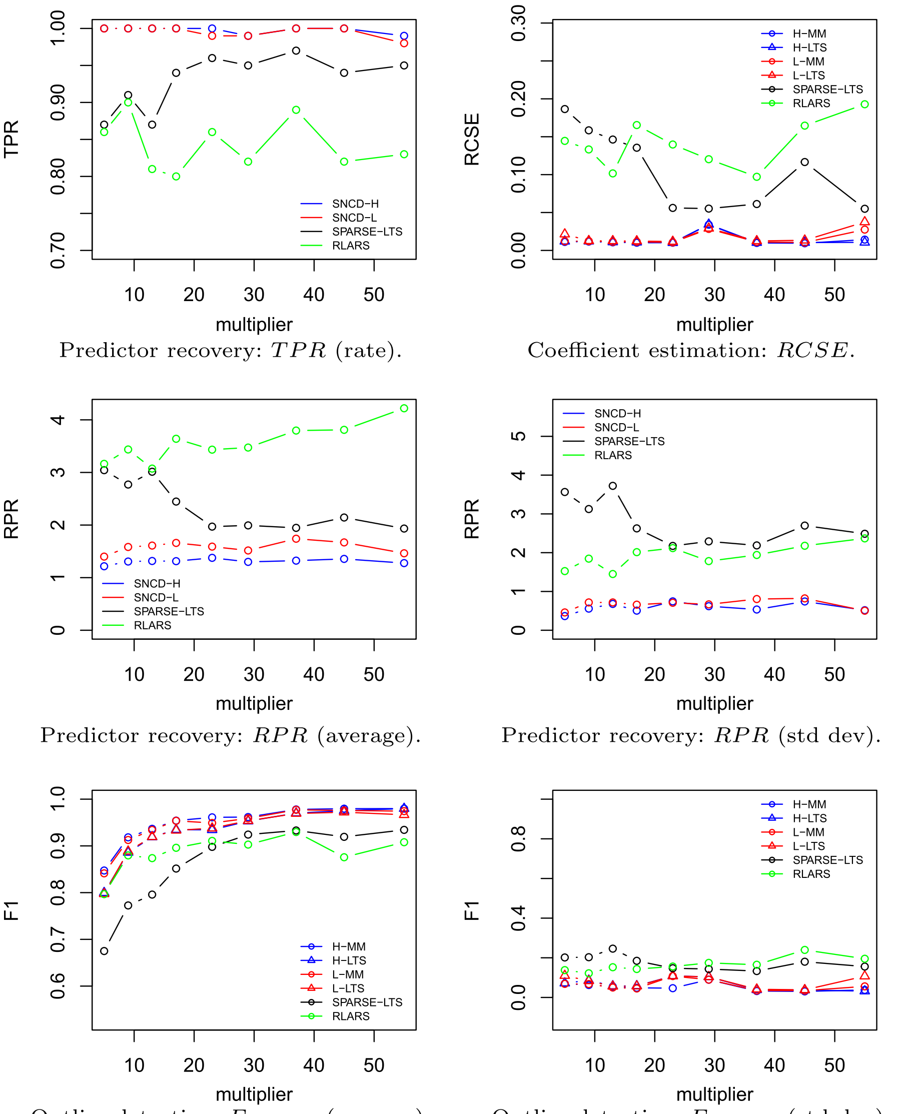
Intercept estimation: $RISE$.

Outlier detection: F_1 score (average).

Outlier detection: F_1 score (std dev).

Fig. 2 Scenario featuring the “fat" dataset, outliers in variance and row-wise outliers: predictor recovery, coefficient estimation and outlier detection performance

RPR of SPARSE-LTS converges to Huber and LAD, while RLARS is still clearly the worse performing. Coefficient estimation metrics $RCSE$ and $RISE$ show that coefficients are not recovered well by SPARSE-LTS and RLARS, while all ROBOUT options are doing almost perfectly. The F_1 score of all methods is now fine, apart from SPARSE-LTS and RLARS when m is small.



Predictor recovery: TPR (rate).

Coefficient estimation: $RCSE$.

Predictor recovery: RPR (average).

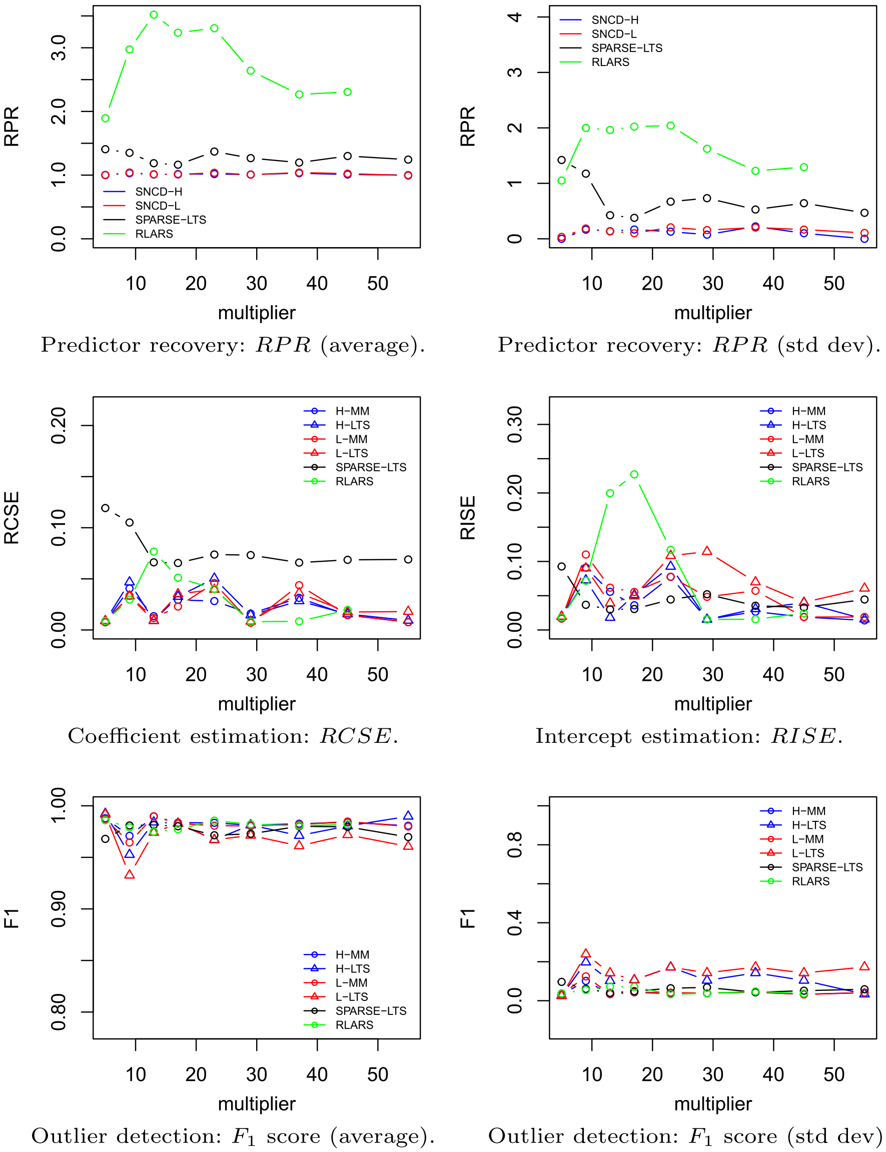
Predictor recovery: RPR (std dev).

Outlier detection: F_1 score (average).

Outlier detection: F_1 score (std dev).

Fig. 3 Scenario featuring the “very fat" dataset, outliers in variance, leverage outliers and row-wise outliers: predictor recovery, coefficient estimation and outlier detection performance

When the outlier scheme “mean-leverage-row" (case 7) is combined with a “tall" dataset (see Fig. 8), SPARSE-LTS breaks down. Its TPR metric drops toward zero at $m = 23$, because it systematically recovers a contiguous predictor. Its RPR attains values as high as 20. At $m = 23$, its $RCSE$ approaches 1, its masking rate 0.4, its swamping rate 0.25, and its F_1 score is as low as 0.7. The other methods do not behave differently compared to the “fat" dataset case.



Predictor recovery: RPR (average).

Predictor recovery: RPR (std dev).

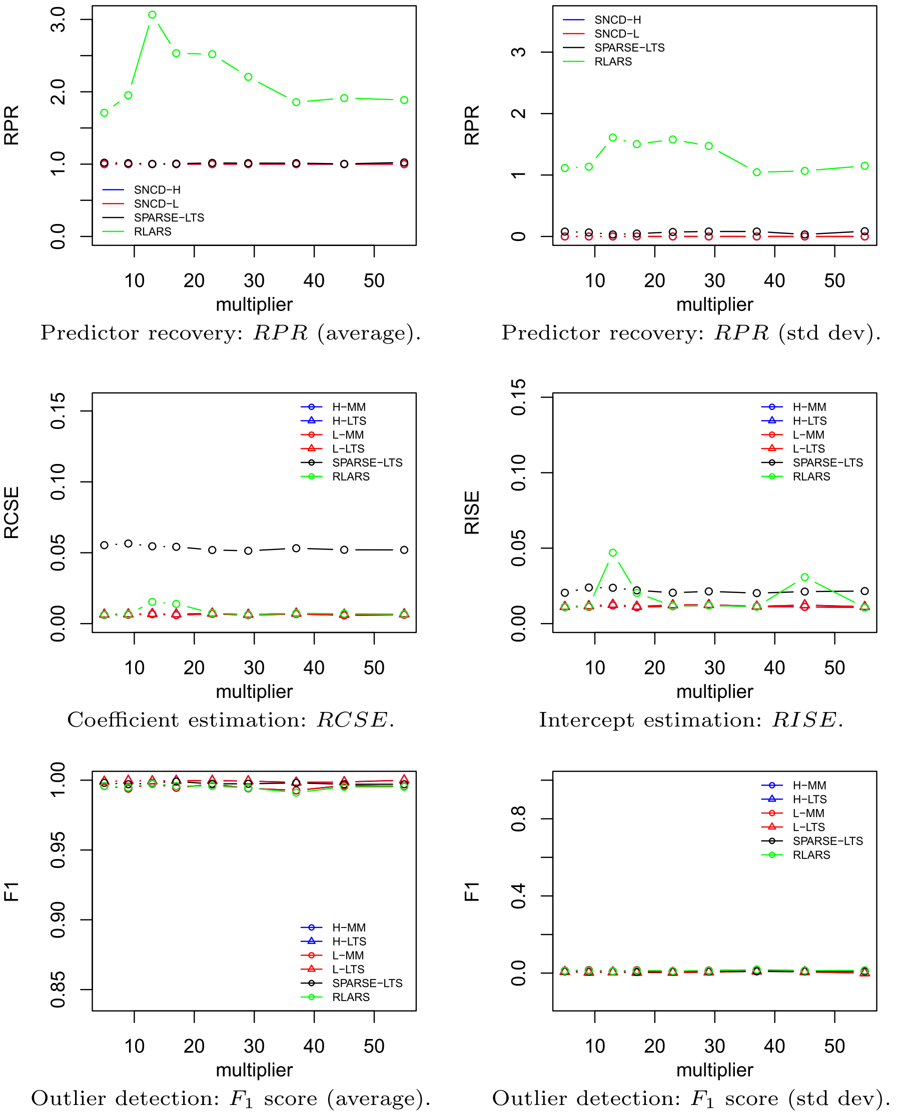
Coefficient estimation: $RCSE$.

Intercept estimation: $RISE$.

Outlier detection: F_1 score (average).

Outlier detection: F_1 score (std dev).

Fig. 4 Scenario “mean-leverage" outliers & “very fat" dataset. Performance metrics for predictor recovery (RPR mean and standard deviation), coefficient estimation (average coefficient and intercept error) and outlier detection performance (F_1 score mean and standard deviation). RLARS not shown at $m = 55$ because the method crashes



Predictor recovery: RPR (average).

Predictor recovery: RPR (std dev).

Coefficient estimation: $RCSE$.

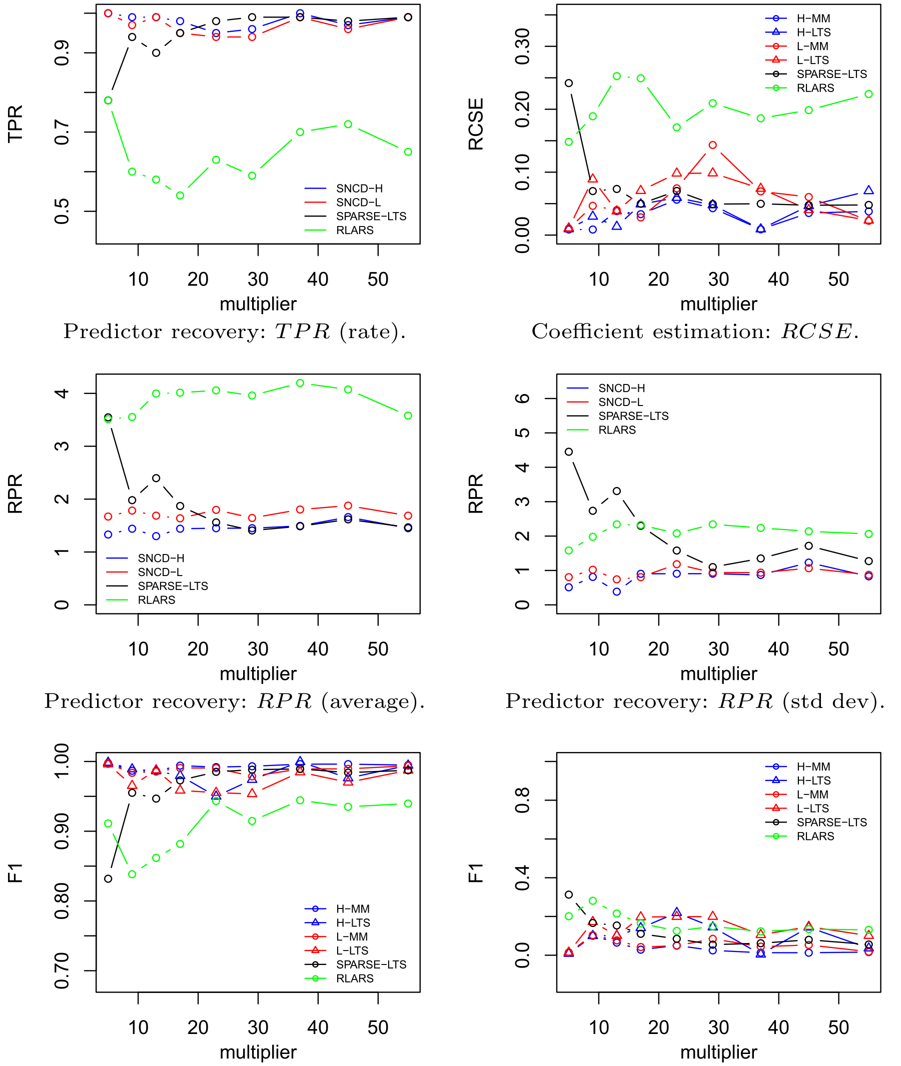
Intercept estimation: $RISE$.

Outlier detection: F_1 score (average).

Outlier detection: F_1 score (std dev).

Fig. 5 Scenario with a “fat" dataset and a “mean-row" outlier scheme. Performance metrics for predictor recovery, coefficient estimation and outlier detection are shown

Under the “very tall" dataset, also RLARS crashes (see Fig. 9). Its TPR becomes close to zero as m approaches 13, descending faster than SPARSE-LTS. RPR remains very high (above 10) for SPARSE-LTS, and high for RLARS (around 3). It follows that their estimated coefficients are completely inconsistent. Both methods also present a masking rate above 40%, which leads to a very poor F_1 score across m. We consider also no-outlier scenarios, that is when $m = 1$ is combined with the outlier scheme “only mean" (case 1). The results for this case are presented in the Supplementary



Predictor recovery: TPR (rate).

Coefficient estimation: $RCSE$.

Predictor recovery: RPR (average).

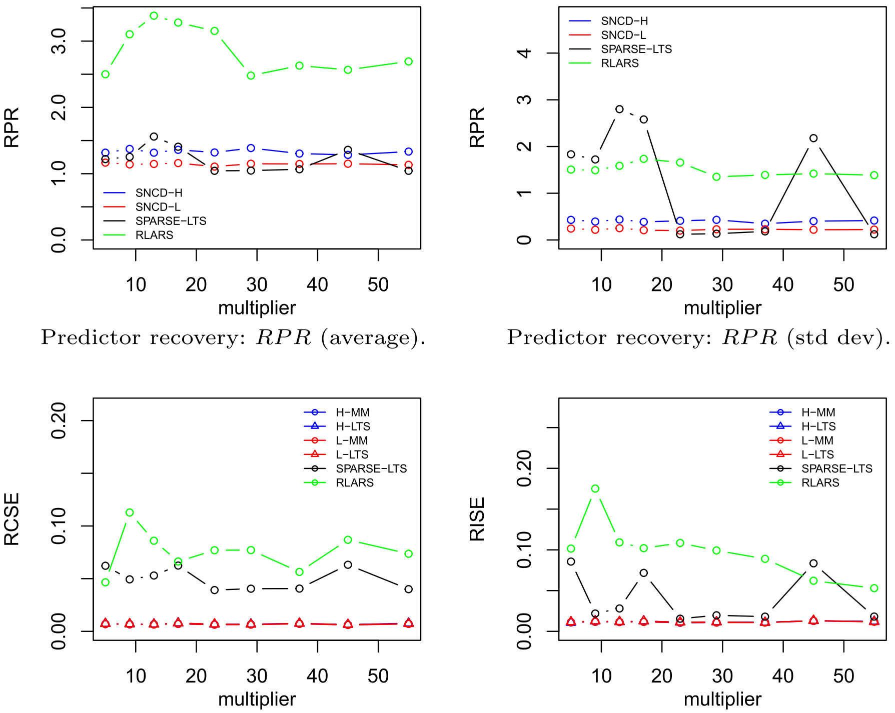
Predictor recovery: RPR (std dev).

Outlier detection: F_1 score (average).

Outlier detection: F_1 score (std dev).

Fig. 6 Scenario with a "very fat" dataset and a "mean-leverage-row" outlier scheme. Metrics for predictor recovery, coefficient estimation and outlier detection are shown

material Sect. 5. Finally, the results of scenarios with outliers in variance are also presented in the Supplementary material Sect. 3 and are consistent with the ranking of methods with regard to their performance as it has been derived based on the scenarios examined in this section.



Predictor recovery: RPR (average).

Predictor recovery: RPR (std dev).

Coefficient estimation: $RCSE$.

Intercept estimation: $RISE$.

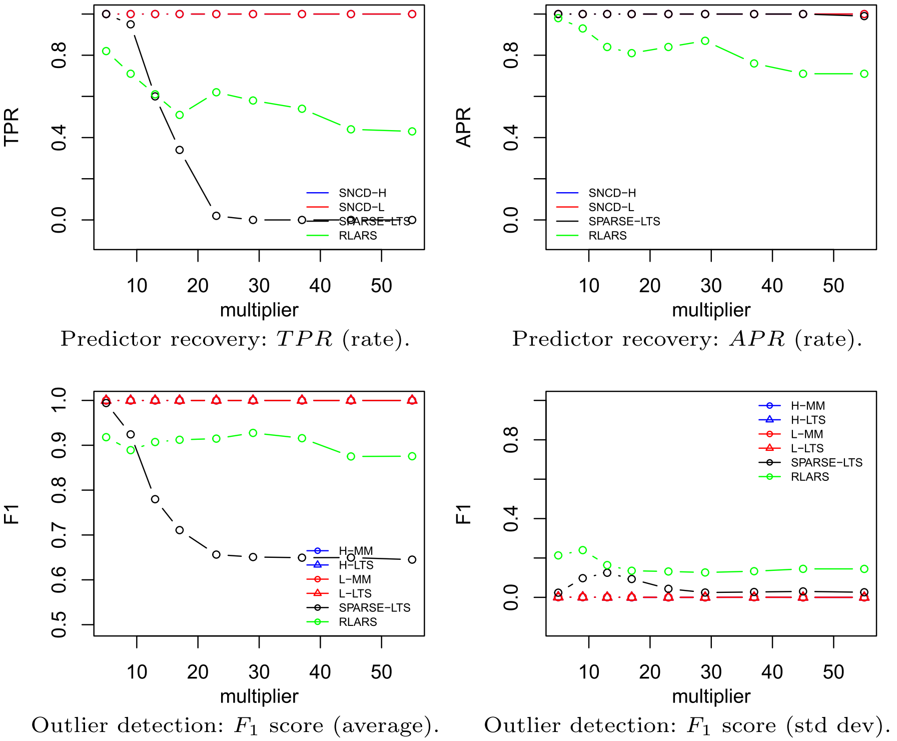
Fig. 7 Scenario with a “fat" dataset and a “mean-leverage-row" outlier scheme. Metrics for predictor recovery, coefficient estimation and outlier detection are shown

6 A real banking supervisory data example

In this section, we apply the ROBOUT conditional outlier detection procedure to a real dataset that contains granular data on the activities of the largest euro area banks, both on the asset and the liability side of their balance sheet. These data are submitted by the euro area banks to the European Central Bank in the context of their supervisory reporting requirements. The dataset in question is compositional, i.e. it includes some parent categories, like ‘Debt’, and their sub-parent categories like ‘Debt versus other banks’, ‘Debt versus central government’, etc. In addition, the dataset is very sparse, as not all banks are engaged in all the activities spanned by the granular set of variables. The reference date of the data is end-2014.

Since the original scale of the variables is in the order of billions (expressed in euros), we apply to each entry of the data matrix $\mathbf{X}$ a logarithmic transformation of the following kind:

$$t(\mathbf{X}_{i,j}) = \begin{cases} -\log(-\mathbf{X}_{i,j}), & \text{if} \quad \mathbf{X}_{i,j} < -1; \\ 0, & \text{if} \quad -1 \leq \mathbf{X}_{i,j} \leq 1; \\ \log(\mathbf{X}_{i,j}), & \text{if} \quad \mathbf{X}_{i,j} > 1; \end{cases}$$



Predictor recovery: TPR (rate).

Predictor recovery: APR (rate).

Outlier detection: F_1 score (average).

Outlier detection: F_1 score (std dev).

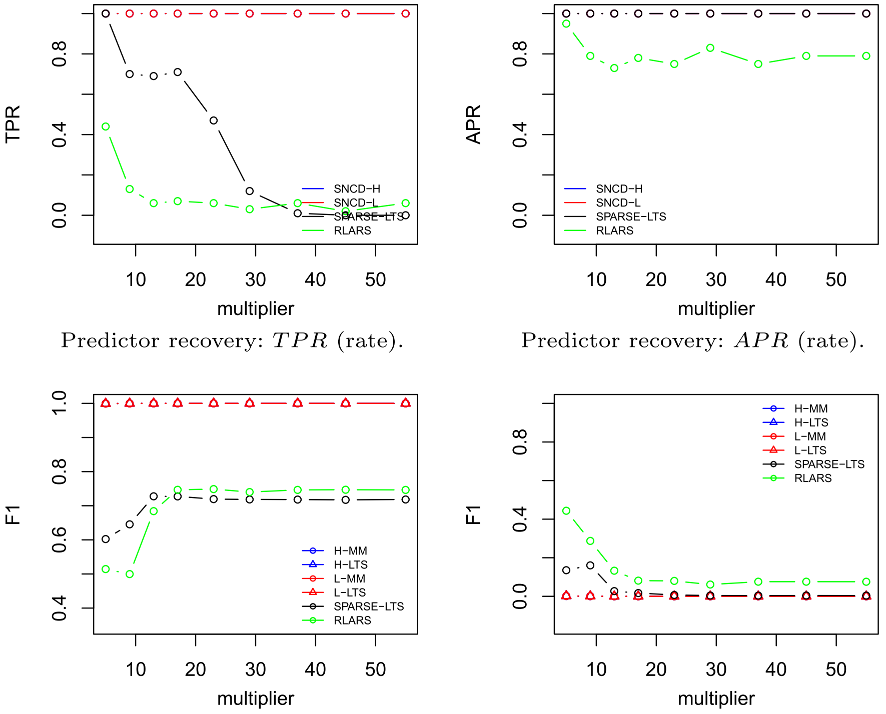
Fig. 8 Scenario with a "tall" dataset and a "mean-leverage-row" outlier scheme. Metrics for predictor recovery, coefficient estimation and outlier detection are shown

where $i = 1, \ldots, 364$ banks and $j = 1, \ldots, 771$ variables. This preliminary step is needed to render the distribution shapes symmetrical. In this way, the few variables with negative values, representing cost or loss items, are measured in the same scale of other variables, without neglecting the sign information. Then, we apply a variable screening based on robust Spearman correlation: we derive the list of all the variable pairs with Spearman correlation exceeding 0.8, and we exclude from the dataset the variable with smaller index. At the end of this procedure, 360 variables survive.

Our final data matrix presents 79.23% of zero entries, 18.86% of positive entries, and 2.08% of negative entries. The retained variables show a mean Spearman correlation of 0.089 and a mean absolute Spearman correlation of 0.171. The Spearman correlation matrix displays the presence of a very rich multicollinearity structure, whose pattern matches the order of the dataset variables. A rich negative correlation is especially present among the variables related to loans and receivables.

We set the log of bank's size as the target variable on which we would like to identify conditionally outlying observations. The distribution of the log of banks' size shows that the log-normality assumption on total assets cannot be rejected.[6] The fact that the size follows a log-normal distribution is intuitive, given the high variance, due

[6] The p-value of the Jarque-Bera test is 46.72%.



Predictor recovery: TPR (rate).

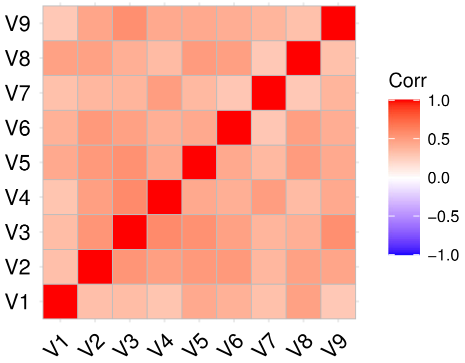
Predictor recovery: APR (rate).

Outlier detection: F_1 score (average).

Outlier detection: F_1 score (std dev).

Fig. 9 Scenario with a "very tall" dataset and a "mean-leverage-row" outlier scheme. Metrics for predictor recovery, coefficient estimation and outlier detection are shown

Fig. 10 Heat map of the Spearman correlation matrix estimated on the data matrix restricted to the nine predictors recovered by SNCD-H (see Table 3)



to the existence of both large and small banks, reinforced by the size dispersion of home countries in the sample e.g. with respect to their GDP, and the non-negativity of the size variable. The mean and the median of the log-size are almost equal (23.04 vs 23.14), while the standard deviation is slightly larger than the rescaled MAD (2.14 vs 1.73).

We apply the four variants of ROBOUT method plus the competitor methods SPARSE-LTS and RLARS on the final data matrix to identify conditional outliers. The application of the ROBOUT versions that utilise SNCD-H and SNCD-L is conducted as follows. First, we employ Spearman correlation in Algorithm 1 and we set $\tilde{\delta} = \delta = 0.01$. For the selection of the predictors we experiment with several values of $K_{\max}$: $\{12, 15, 18, 21, 24, 27, 30\}$. We note that the are two stability areas, one with $\widehat{K} = 6$ and the other with $\widehat{K} = 9$, and we observe small differences between the predictor sets returned by SNCD-H and SNCD-L. We select $K_{\max} = 24$ (with $\widehat{K} = 9$) because all the coefficients estimated by MM and LTS are statistically significant, considering both predictor sets as returned by SNCD-H and SNCD-L. The adjusted R^2 of the MM and the LTS models are both slightly larger when the SNCD-H predictor set is used, compared to the SNCD-L set. Consequently, we select as the preferred predictor set that estimated by H+MM and H+LTS, which is the same.[7] The nine predictors identified by the Huber method are reported in Table 3.

As a robustness test, we repeat the same exercise but we set $\tilde{\delta} = \delta = 0.05, 0.1$. Although a larger $\tilde{\delta}$ leads to a systematically more redundant model, due to the variable selection procedure of Sect. 4.2.3, we observe that the predictors recovered with $\delta = \tilde{\delta} = 0.01$ always appear when $\tilde{\delta} = \delta = 0.05, 0.1$, which is a confirmation that the predictor set recovered with $\tilde{\delta} = 0.01$ is valid.

In contrast, when RLARS is applied, a substantially larger number of predictors is returned under all values of $K_{\max}$. Similarly when SPARSE-LTS is used, also when setting $\alpha = 0.5$ or $\alpha = 0.9$ instead of $\alpha = 0.75$.[8] These findings seem to be consistent with what is observed in the simulation study under the scenarios with the outlier scheme “mean-leverage-row", i.e., where the predictor recovery rate for SPARSE-LTS and RLARS were found to be clearly lower than that of ROBOUT while simultaneously their swamping rates was higher.[9]

The predictor set that we use includes deposits (related to trading), loans and advances, debt securities and trading items such as interest rate derivatives. These are all fundamental elements of the banks’ activities, therefore it makes economic sense that they are chosen as predictors for banks’ total assets. In addition, these are variables that are expected to appear more frequently in the balance sheets of larger and more sophisticated institutions, therefore these variables can discriminate banks with respect to their size.

The application of the ROBOUT method starts with the step comprising the procedure described in Sect. 4.2.2, applied across the 360 potential predictors. It is found that approximately one third of the variables presents no flagged outlying cells, and 36 variables out of 360 present more than 6 flagged cells (out of 364).

[7] The LTS options of ROBOUT return multi-collinearity warnings when $K_{\max} = 12, 15, 18$, because the MCD covariance matrix is singular, due to the high number of predictors.

[8] When the two components of the vector $nsamp = c(500, 10)$, representing subset size and number, are lowered, SPARSE-LTS may not work, due to the rank deficiency of covariance matrix in the sub-samples.

[9] From a robustness perspective it can be noted that when we run the outlier methods inserting a duplicate row in the 364×361 data matrix the predictor sets retrieved by ROBOUT do not change, while SPARSE-LTS and RLARS still return a massive number of predictors. This suggests that ROBOUT is robust to possible duplications of specific records.

Table 3 Description of the predictor set identified, which is used in the model estimated in Table 4

V1	Other demand deposits.
V2	Debt securities. Carrying amount.
V3	Loans and advances. Households. Impaired assets.
V4	Loans and advances. Households. Specific allowances for financial assets, individually estimated.
V5	Deposits. Current accounts / overnight deposits. Held for trading.
V6	Deposits. Current accounts / overnight deposits.
	Designated at fair value through profit or loss.
V7	Interest rate. Negative value. Trading.
V8	Interest rate. OTC other. Trading.
V9	Equity. OTC options. Financial liabilities held for trading.

Table 4 Output of the MM regression applied to the predictors retrieved by SNCD-H

	Estimate	Std. error	t-value	Pr(> |t|)	
(Intercept)	21.0829	0.1158	182.041	0.0000	***
V1	0.0290	0.0073	3.954	0.0001	***
V2	0.0301	0.0078	3.877	0.0001	***
V3	0.0291	0.0098	2.976	0.0031	**
V4	0.0238	0.0075	3.185	0.0016	**
V5	0.0342	0.0079	4.351	0.0000	***
V6	0.0414	0.0063	6.575	0.0000	***
V7	0.0324	0.0067	4.805	0.0000	***
V8	0.0357	0.0069	5.134	0.0000	***
V9	0.0186	0.0075	2.488	0.0133	*

The response variable is the log-size of the banks. The variable legend is reported in Table 3. '*', '**' and '***' denote significance at 5%, 1% and 0.1% respectively
Robust residual standard error = 0.9908
Multiple R-squared = 0.7454
Adjusted R-squared = 0.7389

Table 4 reports the results of the MM regression estimated on the predictors reported in Table 3. All estimated coefficients are positive and strongly significant. We observe that extremely significant predictors are V5-V8, that reflect the extent of trading activities (on the liabilities side). The nine predictors are all positively correlated, with an average Spearman estimated correlation of 0.4252. The heat map of the Spearman predictor correlation matrix is reported in Fig. 10. It is remarkable that 63.83% of entries in the 364×9 predictor matrix is zero, with 64 of 364 rows being completely null.

Starting from the retrieved predictors displayed in Table 3, MM and LTS provide very similar estimated coefficients, and recover the same outliers, irrespective

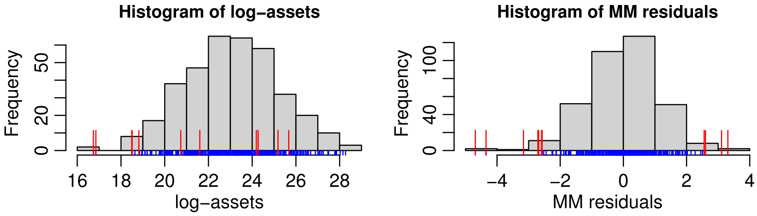


Fig. 11 Histogram of total log-assets (left) and estimated MM residuals (right). Recovered outliers are depicted in red, non-outliers are depicted in blue

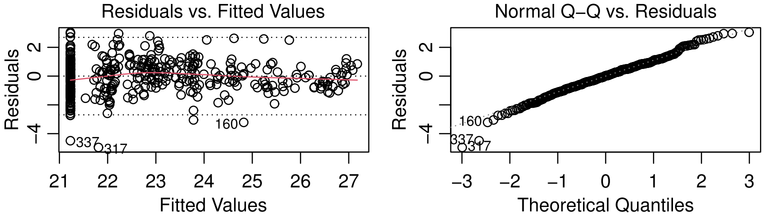


Fig. 12 Residual diagnostic plots for MM residuals: residuals VS fitted values plot (left panel) and normal Q-Q plot (right panel)

of whether $\delta = 0.01$ or $\delta = 0.05$. With a level $\delta = 0.01$, the recovered outliers are eleven, of which seven negative and four positive.

In Fig. 11, two recovered outliers on the right side of the distribution present a very low value of log-assets, while most of the remaining outliers lie within the range of intermediate values for the log-assets. While the former two clearly are reporting errors due to a wrong scale (expressed in 1000s of euros instead of simply euros), the other outliers correspond to banks with a disproportionately large or small value of log-assets compared to the values of the predictors, i.e., conditional outliers. As reported in Table 5, recovered outliers are values with a systematically larger presence of zeros in the predictors compared to non-outliers, that is, they are banks with a more sparsely populated balance sheet. The presence of more zero elements has a bearing on the ability of predictors to explain the observed amount of log-assets. Upon closer examination, the outliers are mainly idiosyncratic banks such as public financing institutions, branches of investment banks, or clearing houses, presenting an anomalous value of total assets with respect to the values recorded in the main predictors of total assets.

Figure 12 shows residual diagnostic plots for the MM residuals to check the validity of the estimated model. We can see that the estimated residuals respect the assumptions of independence, homoscedasticity and normality, with the only exception of the two aforementioned strongly negative outliers. Similar plots are observed for the LTS residuals, which are reported alongside the table of LTS coefficients in the Supplementary material (Sect. 8).

Table 5 Proportion of zero values on relevant predictors for the set of outliers and non-outliers that is recovered by H+MM. See Table 3 for the variable legend

	V1	V2	V3	V4	V5	V6	V7	V8	V9
NON-OUT	0.691	0.669	0.637	0.717	0.501	0.717	0.504	0.518	0.740
OUT	0.909	0.818	0.727	0.909	0.818	1.000	0.636	0.636	0.909

7 Conclusions

In this paper, we propose a new conditional outlier detection method, called ROBOUT. ROBOUT is versatile and flexible, as it is able to robustly spot conditional outliers under many different perturbed conditions and combinations of sample size n and dimension p. ROBOUT works efficiently on datasets with many observations and variables and it is able to robustly select the most relevant predictors of any target variable. Importantly, ROBOUT is effective when the ratio p/n is larger than 1, multicollinearity is present, and potential predictors are perturbed.

ROBOUT presents two options to select relevant predictors, based on LASSO-penalized Huber (SNCD-H) or the Least Absolute Deviation (SNCD-L) loss, and two options to estimate a robust regression, namely, the LTS and the MM methods. In a comprehensive simulation study, we have tested perturbation conditions including conditional outliers in mean or in variance in the response variable, multicollinearity and multivariate outliers in the potential predictors, and we have considered cases when the dimension is both (much) smaller and (much) larger than the sample size. The simulation study shows that when the ROBOUT performance is compared to that of the competitors such as RLARS and SPARSE-LTS, the option SNCD-H+MM is overall the most resilient with respect to predictor recovery and conditional outlier detection across all tested scenarios.

We also test ROBOUT in a large granular banking dataset containing several balance sheet indicators for euro area banks. We are able to robustly model the log-size of euro area banks through a set of predictors that includes loans, deposits, securities, and trading assets, and to identify the banks presenting anomalous values in the total assets conditional on the identified predictors. The recovered outliers constitute a set of idiosyncratic banks in comparison to the textbook prototype of traditional bank, comprising public financing institutions, local branches of investment banks, and clearing houses. ROBOUT may thus be a useful tool for bank supervisors, who need to spot hidden anomalies from raw balance sheet data.

Acknowledgements We gratefully thank the Supervisory Statistics Division of the European Central Bank for granting the permission to conduct research using the supervisory data of Sect. 5, and one associate editor and two anonymous reviewers, who relevantly improved the paper by their insightful and constructive comments.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement.

Data availibility The dataset used in Sect. 5 is confidential, subject to institutional constraints. For these reasons, it may not be made publicly available. However, the paper is accompanied by two R functions, cond_out_gen and robout, which can regenerate all the simulated contamination rates of this paper and compute all the possible variants of ROBOUT procedure. We have no conflict of interest to disclosure of any kind.

References

Atkinson AC, Riani M (2000) Robust diagnostic regression analysis, vol 2. Springer, New York

Atkinson AC, Riani M, Cerioli A (2004) Exploring multivariate data with the forward search, vol 1. Springer, New York

Atkinson AC, Corbellini A, Riani M (2017) Robust Bayesian regression with the forward search: theory and data analysis. Test 26:869–886

Alfons A, Croux C, Gelper S (2013) Sparse least trimmed squares regression for analyzing high-dimensional large data sets. Ann Appl Stat 7(1):226–48

Alfons A (2021) RobustHD: an R package for robust regression with high-dimensional data. J Open Source Softw 6(67):3786

Barnett V, Lewis T (1994) Outliers in statistical data. Wiley series in probability and mathematical statistics. Wiley, New York

Bottmer L, Croux C, Wilms I (2022) Sparse regression for large data sets with outliers. Eur J Oper Res 297(2):782–794

Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Stat 32(2):407–99

Filzmoser P, Höppner S, Ortner I, Serneels S, Verdonck T (2020) Cellwise robust M regression. Comput Stat Data Anal 147:106944

Filzmoser P, Nordhausen K (2021) Robust linear regression for high-dimensional data: an overview. Wiley Interdiscip Rev 13(4):e1524

Freue GVC, Kepplinger D, Salibián-Barrera M, Smucler E (2019) Robust elastic net estimators for variable selection and identification of proteomic biomarkers. Ann Appl Stat 13(4):2065–2090

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1–22

Hawkins DM (1980) Identification of outliers. Chapman and Hall, London

Hong C, Hauskrecht M (2015) Multivariate conditional anomaly detection and its clinical application. Proc AAAI Conf Artif Intell 29(1):4239–4240

Ronchetti EM, Huber PJ (2009) Robust statistics. Wiley, Hoboken

Kendall MG (1938) A new measure of rank correlation. Biometrika 30(1/2):81–93

Khan JA, Van Aelst S, Zamar RH (2007) Robust linear model selection based on least angle regression. J Am Stat Assoc 102(480):1289–99

Kurnaz FS, Hoffmann I, Filzmoser P (2018) Robust and sparse estimation methods for high-dimensional linear and logistic regression. Chemomet Intell Lab Syst 172:211–222

Maronna RA (2011) Robust ridge regression for high-dimensional data. Technometrics 53(1):44–53

Öllerer V, Alfons A, Croux C (2016) The shooting S-estimator for robust regression. Comput Stat 31:829–844

Raymaekers J, Rousseeuw PJ (2021) Fast robust correlation for high-dimensional data. Technometrics 63(2):184–198

Riani M, Perrotta D, Cerioli A (2015) The forward search for very large datasets. J Stat Softw 67:1–20

Rousseeuw PJ (1984) Least median of squares regression. J Am Stat Assoc 79(388):871–80

Rousseeuw PJ, Van Driessen K (2006) Computing LTS regression for large data sets. Data Min Know Discov 12(1):29–45

Rousseeuw PJ, Hubert M (2018) Anomaly detection by robust statistics. Wiley Interdiscip Rev 8(2):e1236

Rousseeuw PJ, Leroy AM (1987) Robust regression and outlier detection. Wiley-Interscience, New York

Rousseeuw PJ, Bossche WVD (2018) Detecting deviating data cells. Technometrics 60(2):135–145

Salibian-Barrera M, Yohai VJ (2006) A fast algorithm for S-regression estimates. J Comput Graph Stat 15(2):414–27

Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15(1):72–101

Smucler E, Yohai VJ (2017) Robust and sparse estimators for linear regression models. Comput Stat Data Anal 111:116–130

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B 58(1):267–88

Tibshirani R, Bien J, Friedman J, Hastie T, Simon N, Taylor J, Tibshirani RJ (2012) Strong rules for discarding predictors in lasso-type problems. J R Stat Soc B 74(2):245–266

Varian HR (2014) Big data: new tricks for econometrics. J Econ Perspect 28(2):3–28

Yohai VJ (1987) High breakdown-point and high efficiency robust estimates for regression. Ann Stat 15:642–656

Yi C, Huang J (2017) Semismooth newton coordinate descent algorithm for elastic-net penalized huber loss regression and quantile regression. J Comput Graph Stat 26(3):547–557

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc B 67(2):301–320

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.