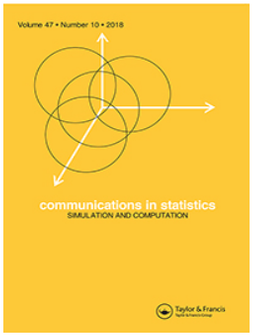

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

A robust sparse linear approach for contaminated data

Shirin Shahriari, Susana Faria & A. Manuela Gonçalves

To cite this article: Shirin Shahriari, Susana Faria & A. Manuela Gonçalves (2019): A robust sparse linear approach for contaminated data, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2019.1588304

To link to this article: https://doi.org/10.1080/03610918.2019.1588304

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2019.1588304

A robust sparse linear approach for contaminated data

Shirin Shahriari[a], Susana Faria[b], and A. Manuela Gonçalves[c]

[a]EPIUnit-Instituto de Saúde Pública, Universidade do Porto, Porto, Portugal; [b]DMA-Department of Mathematics and Applications, University of Minho, Guimarães, Portugal; [c]CMAT-Centre of Mathematics, DMA-Department of Mathematics and Applications, University of Minho, Guimarães, Portugal

ABSTRACT

A challenging problem in a linear regression model is to select a parsimonious model which is robust to the presence of contamination in the data. In this paper, we present a sparse linear approach which detects outliers by using a highly robust regression method. The model with optimal predictive ability as measured by the median absolute deviation of the prediction errors on JackKnife subsets is used to detect outliers. The performance of the proposed method is evaluated on a simulation study with a different type of outliers and high leverage points and also on a real data set.

ARTICLE HISTORY
Received 9 May 2017
Accepted 24 February 2019

KEYWORDS
JackKnife; Outlier detection; Robust variable selection; Sparsity

1. Introduction

Regression models are used in many different areas such as health, biology, environment, management, and finance-related fields. Motivated by various applications, there has been a dramatic growth in the automated means of collecting data, yielding data sets with huge number of observations and a large number of potentially relevant predictor variables. Usually, some predictor variables are correlated in large data sets, but entering all of them in a statistical model will not essentially improve the model's predictive ability. Also, interpreting models with reasonable and tractable amount of predictor variables is easier than for models with a large number of predictors. Therefore, a challenging problem is to sift the best predictor variables from all the candidate predictor variables.

A small proportion of outliers in the data may largely influence likelihood-type model selection methods such as AIC (Akaike 1970), Mallows' C_p (Mallows 1973), and BIC (Schwarz 1978). Under slight data contamination, the variance inflation factor (VIF) criterion (Lin, Foster, and Ungar 2011) may lead to a completely different and improperly selected model (Dupuis and Victoria-Feser 2013). Hence, when there are contaminations in data, we need a robust variable selection method that is resistant to outliers in order to select variables consistently.

Recently, robust variable selection methods have received more attention in the literature. There are various robust variable selection approaches that are based on robustifying classical selection criteria, namely robust AIC (Ronchetti 1985), robust C_p

CONTACT Shirin Shahriari shirin.shahriari22@gmail.com EPIUnit-Instituto de Saúde Pública, Universidade do Porto, Porto 4050-091, Portugal.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

(Ronchetti and Staudte 1994), robust final prediction error (Maronna, Martin, and Yohai 2006), and robust selection criteria based on stratified bootstrap (Müller and Welsh 2005). Another robust approach that is an added variable t-test in the context of regression based on the forward search procedure for variable selection has been proposed by Atkinson and Riani (2002). McCann and Welsch (2007) proposed to add a dummy variable identity matrix to the design matrix for performing robust variable selection using elemental set sampling. Also, for generalized linear models, a natural class of robust testing procedures based on robust deviances which are natural generalizations of quasi-likelihood functions has been proposed (Cantoni and Ronchetti 2001). Salibian-Barrera and Van Aelst (2008) used the fast and robust bootstrap to achieve a faster model selection method based on bootstrap, which makes it feasible to consider larger numbers of predictors. Yao and Wang (2013) imposed L_1 penalty in robust Minimum Average Variance Estimation (MAVE) (Čížek and Härdle 2006) to achieve a robust variable selection method. They investigated their method only in the presence of outliers in the data.

Most of the robust model selection methods need to fit a large number of submodels robustly. When the number of predictors is large, then it is computationally more efficient to use variable selection methods that sequence the predictors according to their importance, such as forward selection and backward elimination (Weisberg 2005).

The Least Angle Regression (LARS) algorithm proposed by Efron, Hastie, Johnstone, and Tibshirani (2004) is a modified version of the forward stagewise procedure. It is a powerful and computationally efficient procedure to sequence the predictor variables for least squares regression. LARS is based on the pairwise correlation between the predictors and the response variable, and therefore it is not robust to the presence of a small amount of contamination in data. Arslan (2012) proposed a weighted version of LAD-LASSO method, which is made resistant to outliers by down weighting leverage points, to find the robust regression estimators and select the appropriate predictors. A robust variable selection procedure via a class of penalized robust regression estimators based on exponential squared loss has been proposed by Wang, Jiang, Huang, and Zhang (2013). Fan, Fan, and Barut (2014) introduced the penalized quantile regression with weighted L_1–penalty, which is called the weighted robust Lasso (WR–Lasso), for robust regularization and also proposed an adaptive robust Lasso (AR-Lasso) through a two-step procedure. Gijbels and Vrinssen (2015) presented three robust versions of the non-negative garrote for linear regression models which are based on the S, M, and Least Trimmed Squares (LTS)-estimators. The proposed robust methods are robust to vertical outliers and leverage points.

Variance Inflation Factor (VIF) regression proposed by Lin et al. (2011) also inherits the spirit of a variation of forward stagewise regression. VIF regression selects those predictor variables among other available predictor variables that can reduce a statistically sufficient part of the variance in the predictive model. VIF regression approximates the partial correlation of each candidate variable with response variable by correcting the marginal correlations. Khan, Van Aelst, and Zamar (2007) proposed Robust LARS (RLARS) which replaces the means, variances, and correlations of the variables inside LARS with their robust versions which are medians, Median Absolute Deviations (MAD), and robust pairwise correlation estimates, respectively.

Dupuis and Victoria-Feser (2013) proposed a Robust version of VIF (RobVIF) regression, which robustly sequences the predictor variables. They used a robust weighted slope parameter to calculate the robust VIF selection criterion.

In this paper, we propose an algorithm that combines RLARS with LTS regression (Rousseeuw (1984)), which is a highly robust regression method (Rousseeuw and Van Driessen (2006)), and perform it on JackKnife (JK) subsets (Efron (1982)) to detect outliers. The merit of using JK subsets is to find the regression model with optimal predictive ability as measured by the MAD of the prediction errors obtained by cross-validation. This optimal regression model is devoted to do outlier detection in data. Then, the detected outliers are removed and standard LARS is performed on the cleaned data to obtain a robust sequenced predictor variable in order of importance.

The paper is organized as follows: Sec. 2 provides a brief description of two robust variable selection methods, RLARS and RobVIF, as the two counterparts of our proposed method. In Sec. 3, our strategy for outlier detection and robust variable selection is explained, as well as our proposed algorithm, JackKnife Robust Least Angle Regression (JKRLARS). In Sec. 4, we conduct different simulation studies to evaluate and compare the performance of JKRLARS with LARS, RLARS, and RobVIF. In Sec. 5, we conduct a real data comparison. Finally, in Sec. 6, the conclusions are presented.

2. Robust variable selection methods

Consider p predictor variables $\mathbf{X} = [1\ \boldsymbol{x}_1 ... \boldsymbol{x}_p]$ and a response variable $\mathbf{y} = (y_1, ..., y_n)^T$. The classical linear regression model supposes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

with vector of slope parameters $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_p]^T$. The errors $\mathbf{e} = [e_1, e_2, ..., e_n]^T$ are assumed to have $E(\mathbf{e}) = 0$ and $var(\mathbf{e}) = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

The aim of this paper is to select the most important predictor variables to enter the regression model when the data contains contaminations. Thus, in this section we briefly explain two robust variable selection methods as the two counterparts to our proposed method.

2.1. Robust LARS

Robust LARS (Khan et al. 2007) replaces the mean, variances, and correlations of the data with robust counterparts. As robust measures for mean and variance Khan et al. proposed to use the computationally fast median and MAD, respectively.

By generalizing the univariate winsorization (Huber and Ronchetti 2009), they introduced bivariate winsorization, which is a fast robust pairwise correlation estimator. After robustly standardizing the data, bivariate winsorization is obtained on the basis of an initial robust bivariate correlation matrix R_0 and a corresponding tolerance ellipse. For instance, consider the Mahalonobis distance $D(\boldsymbol{x})$, with $\boldsymbol{x} = (x_1, x_2)^T \in \mathbb{R}^2$ based on initial bivariate correlation matrix R_0 and set the tuning constant equal to the 95% quantile of the χ^2_2 distribution, which is $c = 5.99$. By using the bivariate transformation $u = \min(\sqrt{c/D(\boldsymbol{x})}, 1)\boldsymbol{x}$ with $\boldsymbol{x} = (x_1, x_2)^T$, the outliers are

shrunken to the boundary of the 95% tolerance ellipse, and therefore the resulting correlation estimate will be less affected by the outliers. Thus, a more robust correlation estimate is given.

An essential part of the bivariate winsorization procedure is choosing a proper initial correlation matrix R_0. Khan et al. proposed the adjusted winsorization as initial estimator (see details in Khan et al. 2007).

2.2. Robust VIF regression

Dupuis and Victoria-Feser (2013) proposed the robust VIF (RobVIF) regression, which is a robust version of VIF regression proposed by Lin et al. (2011). They used a robust weighted slope parameter to calculate the robust VIF selection criterion.

Let $\mathbf{X}_S$ be the design matrix that contains the selected variables at stage S, and $\tilde{\mathbf{X}}_S = [\mathbf{X}_S\ \mathbf{z}_j]$ with $\mathbf{z}_j$ the new candidate predictor to be evaluated for inclusion. Consider the following models

$$\mathbf{y} = \mathbf{X}_S\boldsymbol{\beta}_S + \mathbf{z}_j\beta_j + \mathbf{e}_{step}, \quad \mathbf{e}_{step} \sim N\left(0, \sigma^2_{step}\mathbf{I}\right), \tag{1}$$

$$\mathbf{r}_S = \mathbf{z}_j\gamma_j + \mathbf{e}_{stage}, \quad \mathbf{e}_{stage} \sim N\left(0, \sigma^2_{stage}\mathbf{I}\right), \tag{2}$$

where $\mathbf{r}_S$ are the residuals of the fit of $\mathbf{y}$ on $\mathbf{X}_S$, $\boldsymbol{\beta}_S$ and β_j are slope parameters, γ_j is the slope parameter of the fit of $\mathbf{z}_j$ on the residuals $\mathbf{r}_S$, and $\mathbf{e}_{step}$ and $\mathbf{e}_{stage}$ are the vector of errors. Lin et al. (2011) showed that when least squares are used to estimate, $\hat{\gamma} = \rho\hat{\beta}_j$ where $\rho = \mathbf{z}_j^T(\mathbf{I} - \mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T)\mathbf{z}_j$.

Dupuis and Victoria-Feser (2013) calculated the robust weighted slope estimators $\hat{\beta}_j^w$ using Tukey's redescending biweight weights (Huber and Ronchetti 2009), and then they computed an approximate robust test statistic in order to compare it with an adapted quantile to decide whether or not to add $\mathbf{z}_j$ to the current set of predictors. Let $\mathbf{X}_S^w = diag(\sqrt{w_{iS}^0})\mathbf{X}_S$ be the weighted design matrix with v columns (v – 1 predictors), $\mathbf{y}^w = diag(\sqrt{w_{iS}^0})\mathbf{y}$ be the weighted response variable at stage S, and $\mathbf{z}_j^w = diag(\sqrt{w_{ij}})\mathbf{z}_j$ be the new candidate predictor to be considered to enter the current set at stage $S+1$ (see details in Dupuis and Victoria-Feser (2013) for calculation of weights w_i and w_{ij}). Then $\hat{\beta}_j^w$, the robust weighted estimator of β_j, given in (1) is obtained. Let

$$\rho^w = \left(\mathbf{z}_j^{wT}\mathbf{z}_j^w\right)^{-1}\left(\mathbf{z}_j^{wT}\mathbf{z}_j^w - \mathbf{z}_j^{wT}\mathbf{H}_S^w\mathbf{z}_j^w\right),$$

then

$$\hat{\beta}_j^w = (\rho^w)^{-1}\hat{\gamma}_j^w$$

with $\hat{\gamma}_j^w = (\mathbf{z}_j^{wT}\mathbf{z}_j^w)^{-1}\mathbf{z}_j^{wT}\mathbf{r}_S^w$ the weighted estimator of the fit of $\mathbf{z}_j^w$ on the weighted residuals $\mathbf{r}_S^w$ (model 2). Since computing ρ using all the data is computationally expensive, Dupuis and Victoria-Feser (2013) used a subsampling approach followed by Lin et al. (2011) to estimate ρ^w on a randomly chosen subsample of size g. Then the approximate robust test statistic T_w based on $\hat{\gamma}_j^w$ by comparing the expected value of the estimated

variance of $\hat{\beta}_j^w$ and $\hat{\gamma}_j^w$ is given by

$$T_w = (\rho^w)^{-1/2} \frac{\hat{\gamma}_j^w}{\sqrt{\hat{\sigma}^2/n\left(1/n\sum_i z_{ij}^{w^2}\right)^{-1} e_c^{-1}}}, \tag{3}$$

with $\hat{\sigma}^2$ a robust mean squared error for the model with $\mathbf{r}_S^w$ as response and $\mathbf{z}_j^w$ as predictor variable where z_{ij}^w denotes the element of $\mathbf{z}_j^w$ (that is, model 2). Then T_w is compared with an adapted quantile as part of the decision rule to decide whether or not to add the new predictor variable (To calculate e_c and for more details, see Dupuis and Victoria-Feser 2013).

3. Proposed method

In this section we explain our strategy of outlier detection and of robust variable selection method. We seek a method to detect outliers in the data set and at the same time specify a robust sequence of the predictor variables in order of their importance. Therefore, we propose an approach that can perform outlier detection and robust variable selection simultaneously. To generate JK subsets, first the data is randomly partitioned into l non-overlapping equal size subsets. Then, each subset is retained once and RLARS is applied to the remaining subsets to find l sequence predictor variables in their order of importance. The most appropriate predictor variables relevant to these l RLARS sequences are used to fit a robust regression model. We use the LTS regression, which is a highly robust and computationally efficient robust regression method (Rousseeuw and Van Driessen 2006).[1] Denote the vector of squared residuals by $\mathbf{r}^2(\boldsymbol{\beta}) = (r_1^2, ..., r_n^2)^T$ with $r_i^2 = (y_i - \mathbf{x}_i\boldsymbol{\beta})^2, i = 1, ..., n$. Then the LTS estimator is defined as

$$\hat{\boldsymbol{\beta}}_{LTS} = \arg\min_{\beta} \sum_{i=1}^{h} \left(\mathbf{r}^2(\boldsymbol{\beta})\right)_{i:n}, \tag{4}$$

where $\left(\mathbf{r}^2(\boldsymbol{\beta})\right)_{1:n} \leq ... \leq \left(\mathbf{r}^2(\boldsymbol{\beta})\right)_{n:n}$ are the order statistics of the squared residuals and $h \leq n$. For $h = [(n + p + 1)/2]$ (here, $[a]$ denotes the integer part of a) the LTS breakdown point equals 50%, whereas for greater h its break down point is $(n-h)/n$. The usual choice $h \approx 0.75n$ yields the LTS breakdown point of 25% (Hubert, Rousseeuw, and Van Aelst 2008). Hence, LTS regression seeks to find the subset of h observations whose least squares fit gives the smallest sum of squared residuals. Based on the LTS regression model, predicted values of the observations are calculated by cross-validation. The MAD values of the prediction errors for each LTS regression model are calculated. Then, the optimal LTS regression model that yields minimum MAD value is selected. To identify outliers, the standardized prediction errors of this optimal model are computed. The detected outliers are left out, and LARS is applied to the cleaned data to find an improved sequence of predictor variables. The two goals of identifying outliers

[1]FAST-LTS algorithm (Rousseeuw and Van Driessen (2006)) was used inside the implementation of the proposed method.

in data and robustly sequencing predictor variables simultaneously reflect the specifications of our approach. (For code information, please refer to the authors.)

3.1. Description of the JackKnife Robust Lars (JKRLARS) algorithm

Consider a data set $(y_i, \mathbf{x}_i), i = 1, ..., n$. Let $J = \{1, ..., p\}$ and $I = \{1, ..., n\}$ be the set of indices for the candidate predictors and observations respectively, and $q \ll p$ the length of the most important predictor variables returned by RLARS. Then the JKRLARS algorithm proceeds as follows:

Step 1. The observations $(y_i, \mathbf{x}_i), i = 1, ..., n$ are partitioned into l randomized non-overlapping equally-sized JK subsets $I_f \subset I = \{1, ..., n\}$, with $f = 1, ..., l$. Clearly, I_f contains the indices of the observations in f-th subset, with $|I_f| \approx \frac{n}{l}$.

Step 2. With the f-th subset left out, RLARS is applied to the set $(y_i, \mathbf{x}_i), i \notin I_f$, of l – 1 subsets to find a sequence of predictor variables $(\boldsymbol{x}_j^{(f)})_{j \in J_f}$ with $J_f \subset J = \{1, ..., p\}$ such that $|J_f| \ll p$, where p is the number of predictor variables. Clearly, J_f contains the indices of the $|J_f| = q$ most important predictor variables returned by RLARS.

Step 3. The predictors $\boldsymbol{x}_j^{(f)}, j \in J_f$, are used as predictor variables for LTS regression. We thus consider the regression model

$$y_i = \mathbf{x}_i^{(f)} \boldsymbol{\beta}^{(f)} + e_i \tag{5}$$

where $\mathbf{x}_i^{(f)}$ denotes the i-th observation of the predictors $\boldsymbol{x}_j^{(f)}, j \in J_f$ for the ith observation and $\boldsymbol{\beta}^{(f)}$ is estimated by using LTS.

Step 4. To evaluate the prediction performance of each LTS regression fit, we perform l-fold cross-validation. In order to detect outliers, we compute the predicted values $\hat{y}_i$, which are defined as

$$\hat{y}_i = \mathbf{x}_i^{(f)} \hat{\boldsymbol{\beta}}^{(f)}, i \in I_f, \tag{6}$$

where $\hat{\boldsymbol{\beta}}^{(f)}$ denotes the LTS estimates of the regression coefficients with the f-th subset left out as obtained in the previous step. Thus, predicted values for all the observations are obtained. For each LTS regression the prediction errors $\mathrm{PE}_i = y_i - \hat{y}_i$, and the corresponding MAD value of the prediction errors as a measure of the model's predictive ability are calculated. The minimum MAD value over all LTS models is obtained to find the LTS model with optimal predictive ability. Then, the corresponding standardized prediction errors of this optimal model are used to detect outliers. The standardized prediction errors are defined by $\frac{\mathrm{PE}_i}{\hat{\sigma}}, i = 1, ..., n$ where $\hat{\sigma} = c_{h,n} \sqrt{\frac{1}{h} \sum_{i=1}^{h} (\mathbf{r}^2(\boldsymbol{\beta}))_{i:n}}$, and $c_{h,n}$ makes $\hat{\sigma}$ consistent and unbiased at Gaussian error distribution (Pison, Van Aelst, and Willems (2002)). It should be mentioned that the LTS scale estimate $\hat{\sigma}$ is itself highly robust, and therefore can be used to identify outliers by $\frac{\mathrm{PE}_i}{\hat{\sigma}}$. As in Hubert et al. (2008) we define the set of the indices of outlying observations as

$$I_{Out} = \left\{ i \in I : \left| \frac{\mathrm{PE}_i}{\hat{\sigma}} \right| > \sqrt{\chi^2_{1,0.975}} \right\}. \tag{7}$$

Step 5. The detected outliers $(y_i, \mathbf{x}_i), i \in I_{Out}$, are removed (or given weight zero) and LARS is applied on the cleaned data $(y_i, \mathbf{x}_i), i \in I_{Out}^c$, with I_{Out}^c the complement of I_{Out}.

The predictors $\boldsymbol{x}_j, j \in J_{final}$ with $J_{final} \subset J$ from the candidate predictors are obtained as the robust version of LARS sequenced predictor variables.

4. Simulation study

Here we conduct a simulation study to investigate and compare the performance of JKRLARS with its counterparts. We perform all the methods in R Core Team (2014). We use package *lars* (Hastie and Efron 2013) to perform LARS, and we perform RLARS (Khan et al. 2007) and RobVIF (Dupuis and Victoria-Feser 2013) using the codes available on their authors' website. We consider $h \approx 0.75n$ to have a breakdown point of 25% for LTS inside JKRLARS (Hubert et al. 2008). We use $l = 5$ subsets and $l = 10$ for the JKRLARS algorithm. For RobVIF, we consider the same subsample size of 200 with the same initial values for wealth and pay-out equal to 0.5 and 0.05 similar to Dupuis and Victoria-Feser (2013), respectively.

We consider a simulation setting similar to Khan et al. (2007) and Shahriari, Faria, Gonçalves, and Van Aelst (2014), which is based on the design of Frank and Friedman (1993). The linear model is created as

$$\mathbf{y} = \mathbf{l}_1 + \ldots + \mathbf{l}_k + \sigma\mathbf{e}, \tag{8}$$

with $k = 6$ latent independent standard normal variables $\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_k$ and an independent standard normal variable $\mathbf{e}$. The value of σ is chosen such that the signal to noise ratio is equal to 3. Let $\mathbf{e}_1, \ldots, \mathbf{e}_p$ be independent standard normal variables, then the set of p predictors is created as

$$\begin{aligned}
\boldsymbol{x}_j &= \mathbf{l}_j + \tau\mathbf{e}_j, j = 1, \ldots, k \\
\boldsymbol{x}_{k+1} &= \mathbf{l}_1 + \delta\mathbf{e}_{k+1} \\
\boldsymbol{x}_{k+2} &= \mathbf{l}_1 + \delta\mathbf{e}_{k+2} \\
\boldsymbol{x}_{k+3} &= \mathbf{l}_2 + \delta\mathbf{e}_{k+3} \\
\boldsymbol{x}_{k+4} &= \mathbf{l}_2 + \delta\mathbf{e}_{k+4} \\
&\vdots \\
\boldsymbol{x}_{3k-1} &= \mathbf{l}_k + \delta\mathbf{e}_{3k-1} \\
\boldsymbol{x}_{3k} &= \mathbf{l}_k + \delta\mathbf{e}_{3k} \\
\text{and } \boldsymbol{x}_j &= \mathbf{e}_j, \quad j = 3k+1, \ldots, p
\end{aligned}$$

where $\delta = 5$ and $\tau = 0.4$ so that target predictor variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are formed by low noise perturbations of the latent variables. Variables $\boldsymbol{x}_{k+1}, \ldots, \boldsymbol{x}_{3k}$ are noise predictor variables that are correlated with the latent variables, and variables $\boldsymbol{x}_{3k+1}, \ldots, \boldsymbol{x}_p$ are independent noise predictor variables.

We consider five different sampling distributions:

a. $e \sim (1-a)N(0,1) + aN(0,1)/U(0,1)$, symmetric, Slash contamination;
b. $e \sim \text{Cauchy}(0,1)$, heavy-tailed Cauchy contamination;
c. $e \sim (1-a)N(0,1) + aN(20,1)$, asymmetric, shifted Normal contamination;
d. same as (a), with high leverage X values, $X \sim N(50,1)$;
e. same as (b), with high leverage X values, $X \sim N(50,1)$;

Table 1. The True Positive Rate (TPR) and the Positive Predictive Value (PPV) with 5%, 10%, and 20% contamination, averaged over 200 simulation runs are reported for JKRLARS.

Case	a	b	c	d	e
5% contamination					
TPR	0.33	0.28	1	0.99	0.99
PPV	0.5	0.5	0.91	0.89	0.88
10% contamination					
TPR	0.26	0.20	1	0.96	0.97
PPV	0.52	0.50	0.91	0.88	0.87
20% contamination					
TPR	0.24	0.18	1	0.93	0.92
PPV	0.74	0.65	0.99	0.86	0.85

where a denotes the fraction of contamination in the data.

With different fraction of contamination, $a = 5\%, 10\%$, and 20%, from the above five simulation scenarios with $p = 50$ predictors, we generate 200 independent data sets of size $n = 150$, and each time we perform all the aforementioned methods on the same data set.

4.1. Performance measures

We evaluate the performance of JKRLARS concerning outlier detection by the True Positive Rate (TPR) and the Positive Predictive Value (PPV). A true positive is an observation that is contaminated in the data and is also detected as an outlier. PPV shows among those observations that are detected as outliers by the method which ones are really contaminated in the data. Thus, the sensitivity and the precision of the method concerning outlier detection can be measured by TPR and PPV, respectively.

Denote the set of the indices of the regular observations in the data by $I_R \subset I = \{1, \ldots, n\}$ and the set of the indices of the contaminated observations in the data by I_R^c. In mathematical terms, the TPR and PPV can then be defined as

$$TPR = \frac{|\{i : i \in I_{Out} \wedge i \in I_R^c\}|}{|I_R^c|}; \tag{9}$$

$$PPV = \frac{|\{i : i \in I_{Out} \wedge i \in I_R^c\}|}{|I_{Out}|}, \tag{10}$$

with I_{Out} given in (7). Higher TPR and higher PPV are desired which show that JKRLARS performs well concerning outlier detection.

In order to compare the JKRLARS performance with its counterparts, we determine for each sequence the number of target variables t_m included in the first m sequenced predictor variables entering the model as a function of varying m. With k number of target variables, good performance is achieved when the method can find the k target variables in the first t_k sequenced predictor variables, with t_k equal or close to k.

4.2. Results of the simulation study

In this section we present and discuss the simulation results for the five presented simulation scenarios. We perform LARS, RLARS, RobVIF, and JKRLARS on each of the 200 independent data sets.

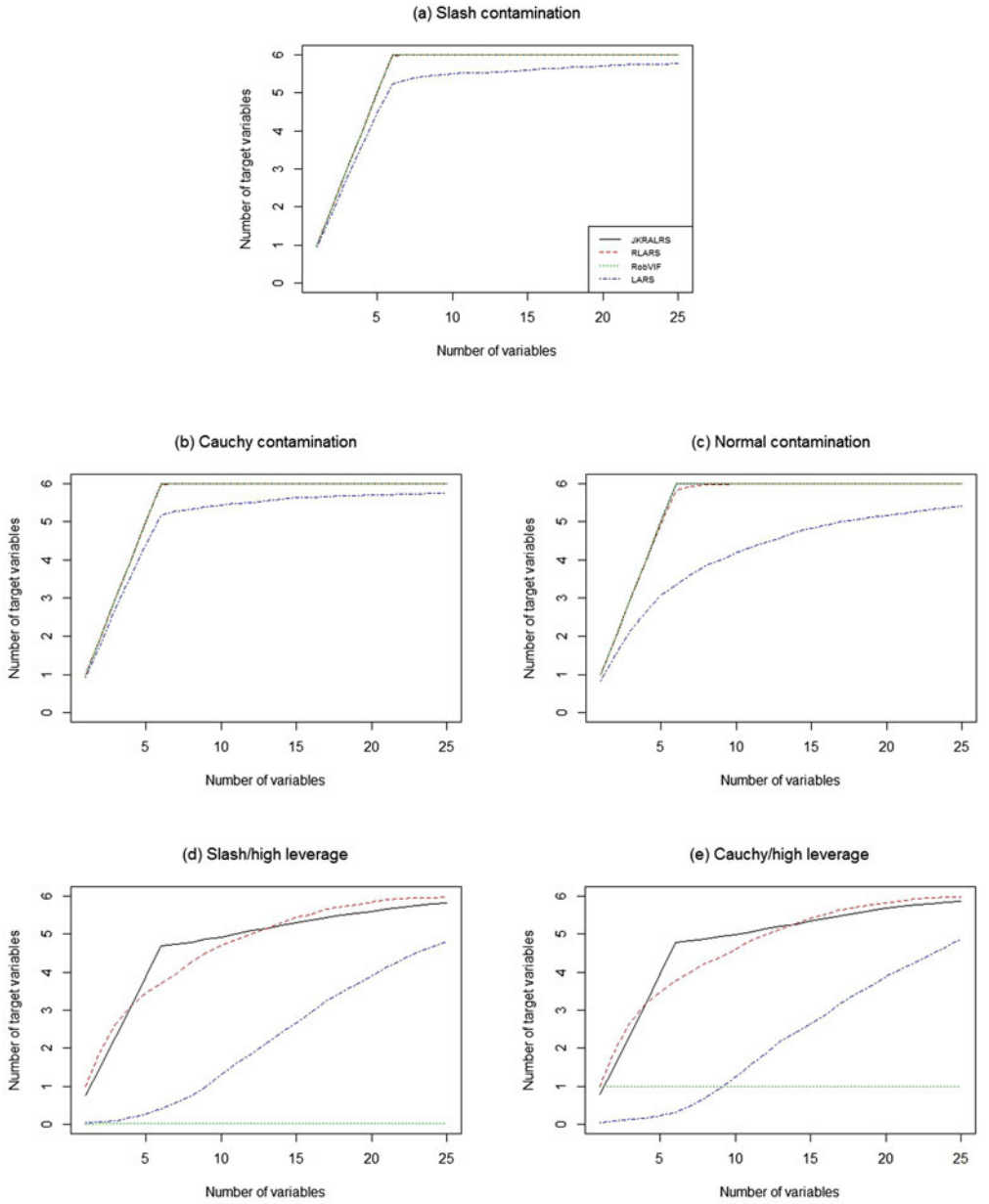

Figure 1. Average number of target variables t_m versus m for LARS, RLARS, RobVIF, and JKRLARS for scenarios (a)–(e) with 5% fraction of contamination. The lines shown in all plots follow the legend of figure (a).

First, we start by examining how JKRLARS performs to detect the outliers in the data sets. The results for TPR and PPV averaged over the 200 data sets for the 5 types of contamination considered with 5%, 10%, and 20% fraction of contamination and the number of $l = 10$ subsets for the JKRLARS algorithm are shown in Table 1. Similar results were obtained based on $l = 5$ subsets for the JKRLARS algorithm. As it can be seen in this table, TPR and PPV of the outlier detection procedure is almost perfect in

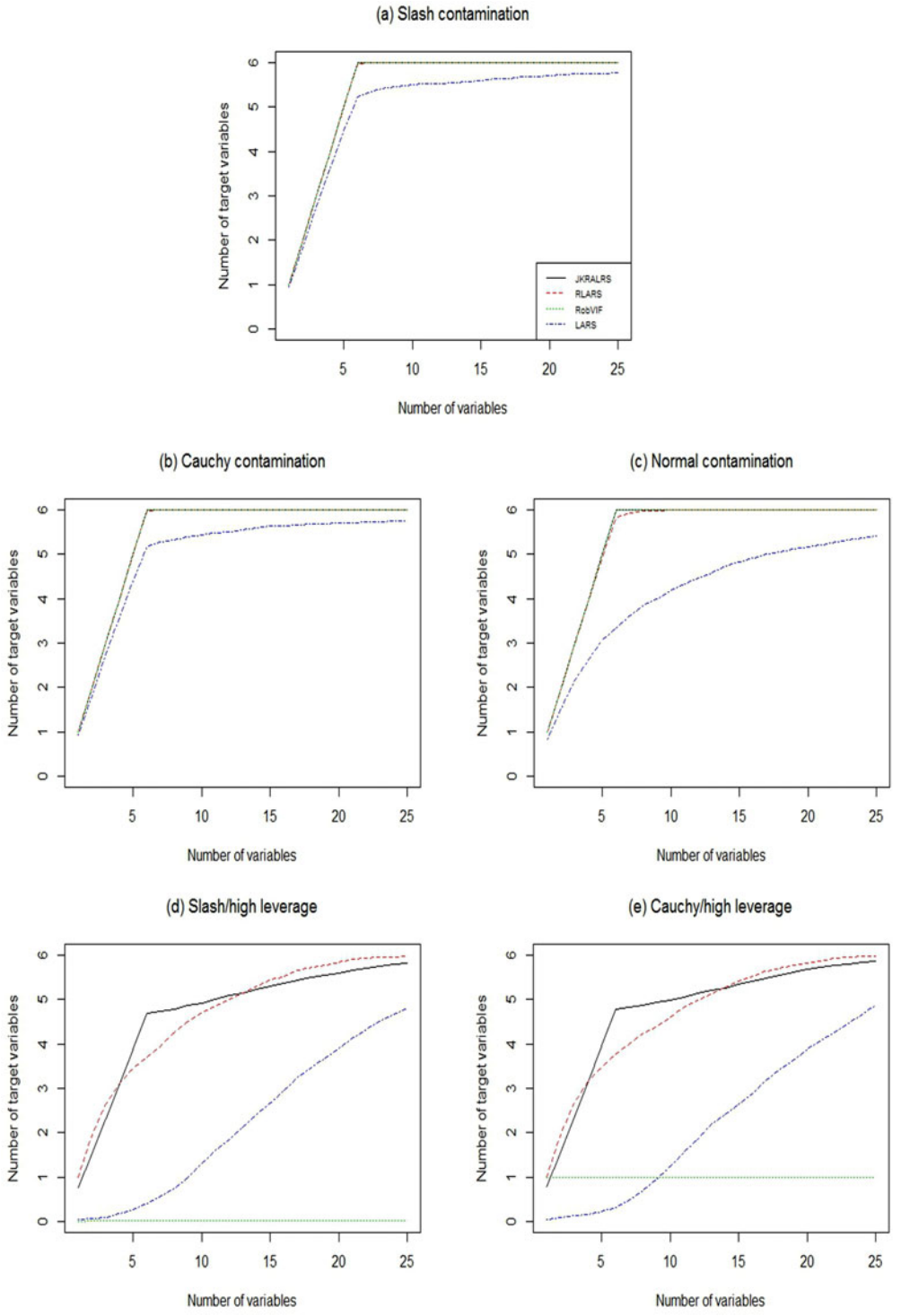

Figure 2. Average number of target variables t_m versus m for LARS, RLARS, RobVIF, and JKRLARS for scenarios (a)–(e) with 10% fraction of contamination. The lines shown in all plots follow the legend of figure (a).

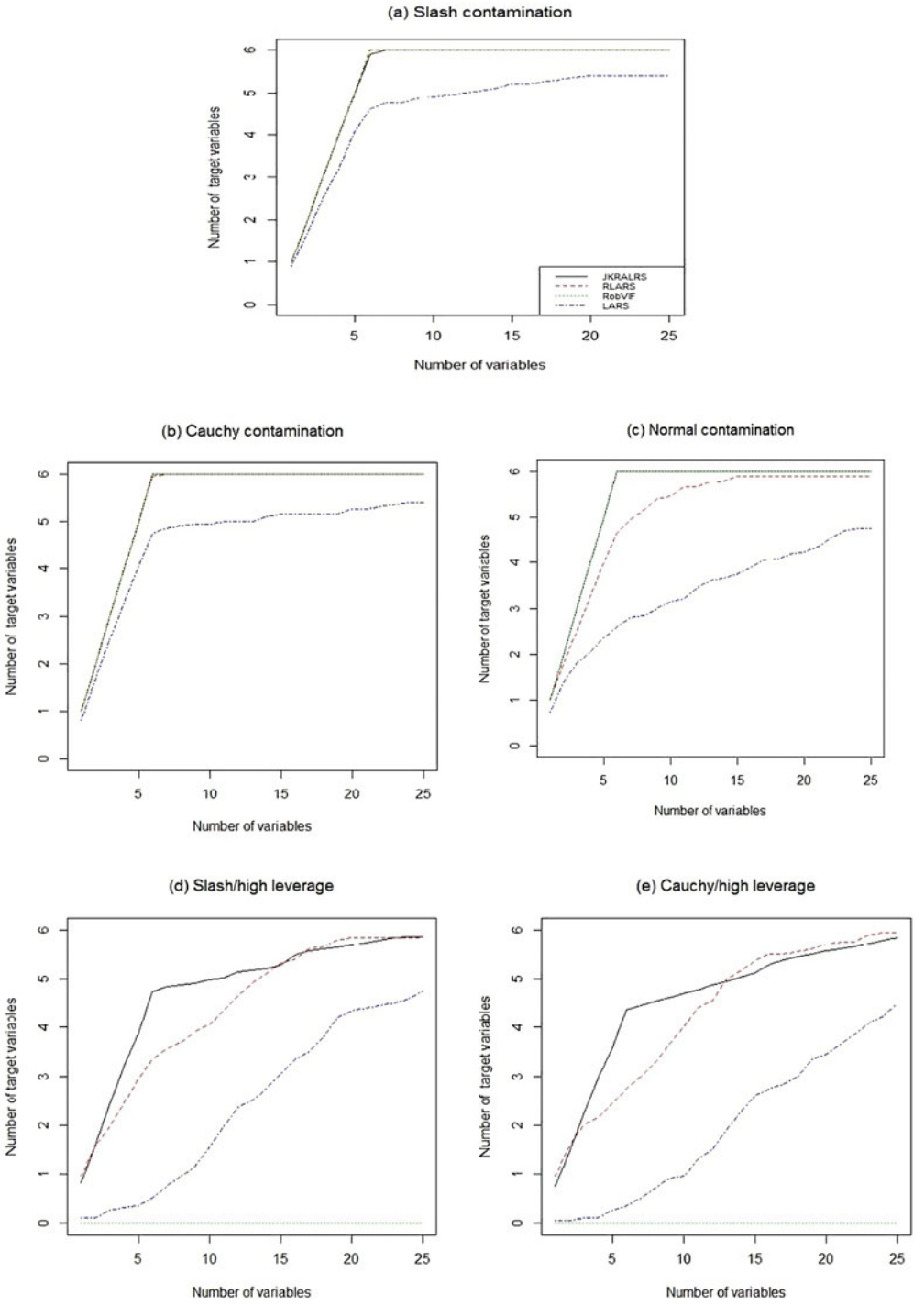

Figure 3. Average number of target variables t_m versus m for LARS, RLARS, RobVIF, and JKRLARS for scenarios (a)–(e) with 20% fraction of contamination. The lines shown in all plots follow the legend of figure (a).

scenarios (c), (d), and (e). High leverage points and clear vertical outliers can thus be detected with high recall and precision.

Although TPR and PPV for scenarios (a) and (b) seem much worse than the other scenarios, but we should bear in mind that in these scenarios the observations are contaminated by producing errors from the long-tailed Slash and Cauchy distributions, respectively. Not all of the errors produced from these distributions will lie in the tails of the distributions and thus not all of them will be actual outliers. Considering the cut-off value $\sqrt{\chi^2_{1,0.975}}$ for identifying outliers, it can easily be checked that only 35.2% of the produced Slash errors and 26.7% of the produced Cauchy errors are expected to produce outlying observations. Comparing TPR with these fractions, we see that the outlier detection procedure still performs reasonably well in these difficult scenarios. Therefore, JKRLARS does a good job of outlier detection considering both TPR and PPV (or recall and precision) in all scenarios.

The performances of LARS, RLARS, RobVIF and JKRLARS in terms of sequencing the predictor variables for each simulation scenario can be compared using Figures 1–3 for different fractions of contamination.

For each sequence of predictor variables we determine the number t_m of target variables included in the first m sequenced variables entering the model with m ranging from 1 to 25. Figures 1–3 show the average of t_m (over the 200 data sets) for all the methods and simulation scenarios with 5%, 10%, and 20% contamination, respectively.

All the methods perform similarly well when there is no contamination in the data, and they can select all the $k = 6$ target variables at the top of the sequence. In scenarios (a), (b) and (c), JKRLARS shows the same excellent performance as RLARS and RobVIF in sequencing the $k = 6$ target variables at the top of the sequence (Figures 1–3 (a–c)) for all 5, 10, and 20% contamination. In the high leverage scenarios (d) and (e), RobVIF fails to select the target variables properly for all 5%, 10, and 20 contamination (Figures 1–3 (d and e)). In the Slash and Cauchy scenarios with high leverage with 5% and 20% contamination, RobVIF can select none of the target variables (Figures 1 and 3 (d)). In the Slash with high leverage values with 10% contamination, RobVIF again cannot select any of the target variables (Figure 2 (d)), and in the Cauchy with high leverage scenario it can select only one of the target variables (Figure 2 (e)). Figures 1–3 (d) and (e) show that in the high leverage scenarios RLARS has much more problems to pick up the target variables in the beginning, while JKRLARS succeeds much better to pick up most of the important variables in the beginning of the sequence. In particular, in models with (less than) 10 predictors JKRLARS captures 5 of the 6 important predictor variables, while RLARS needs models with up to 15 predictors to include at least 5 of the 6 important predictor variables.

5. Application to real data

Here we evaluate and compare the performance of LARS, RLARS, RobVIF and JKRLARS on the 1990 US Census data. In order to investigate the selected variables obtained by each method, we measure their Median Absolute Prediction Error (MAPE) for the optimal number of selected variables as measured by 10 test subsets, i.e., the

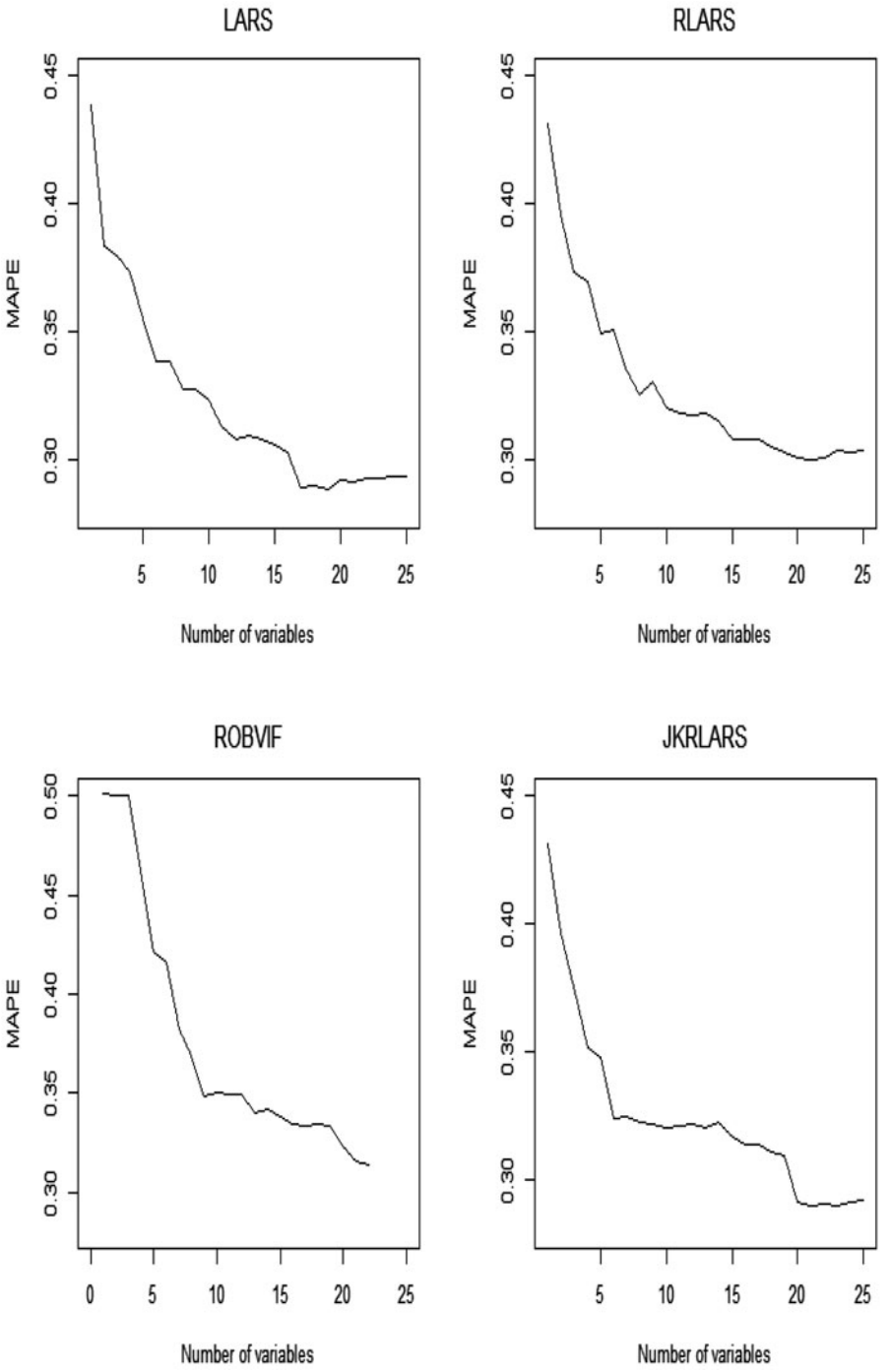

Figure 4. Learning curves for LARS, RLARS, RobVIF, and JKRLARS on Census data. Each learning curve suggests the optimal number of predictors for each method.

Table 2. Mean MAPE (standard deviation) over 10 test subsets using optimal number of selected variables for LARS, RLARS, RobVIF, and JKRLARS on Census data.

Method	LARS	RLARS	RobVIF	JKRLARS
MAPE	0.291 (0.012)	0.325 (0.015)	0.350 (0.017)	0.321 (0.014)
# Variables	17	8	9	6

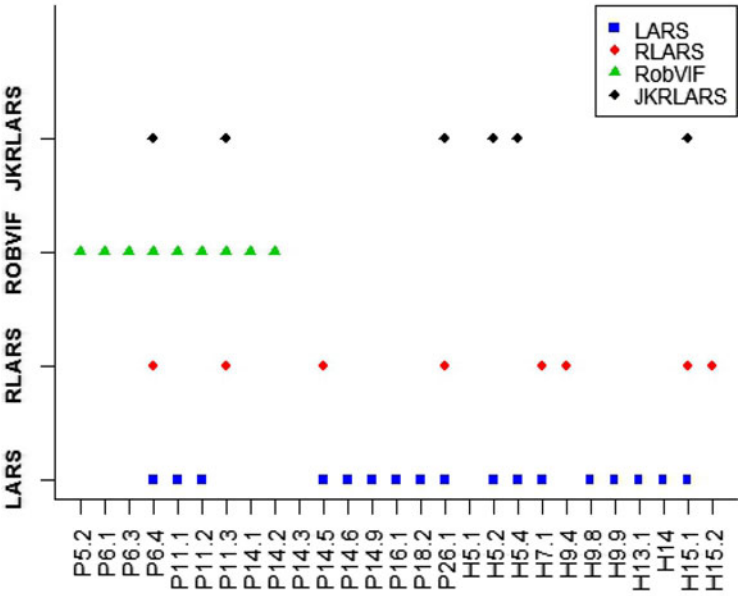

Figure 5. Selected variables by LARS, RLARS, RobVIF, and JKRLARS on Census data.

data is partitioned into 10 roughly equal-sized subsets and the MAPE of the methods is calculated on each test subset. We repeat this process 10 times and each time we use the same test subsets for all the methods. The optimal number of variables for each method is obtained by considering its learning curve. For this purpose, we plot the model size versus the MAPE as measured by 10 test subsets. The optimal model size can be chosen as the point where the learning curve does not show a considerable slope anymore.

The original Census data contains $n = 22{,}784$ observations and 139 variables and can be downloaded at http://www.cs.toronto.edu/~delve/data/census-house/desc.html. More details on how the data were obtained can be found in Dupuis and Victoria-Feser (2011). After removing the collinear predictors, and those that are all zeros or almost zeros, the data contains $n = 13{,}970$ observations with 50 predictors and the response variable, which is the average price asked for the housing unit. With these data, we do not seek a "true" model, but our purpose is to see which model can best explain the average asking price in a sector with few predictor variables as measured by MAPE over 10 test subsets.

As mentioned in Sec. 4, to have a breakdown point of 25% for LTS inside JKRLARS we consider $h \approx 0.75n$. The number of $l = 10$ subsets is considered for JKRLARS. For RobVIF, we consider the same subsample size of 200 with the same initial value for

Table 3. Descriptions of the selected predictors by LARS, RLARS, RobVIF, and JKRLARS in Figure 5.

Name	Description
P5.2	%-tage female
P6.1	%-tage white
P6.3	%-tage American Indian, Eskimo or Aleut (Indian)
P6.4	%-tage Asian or Pacific Islander (Asian)
P11.1	%-tage (0:11] years old
P11.2	%-tage [12:24] years old
P11.3	%-tage [25:64] years old
P14.1	%-tage males never married
P14.2	%-tage males married, not separated
P14.3	%-tage males separated
P14.5	%-tage males divorced
P14.6	%-tage females never married
P14.9	%-tage females widowed
P16.1	%-tage of HH-lds with 1 person
P18.2	%-tage of HH-lds with 1+ persons under 18 which are family HH-lds
P26.1	%-tage of HH-lds with 1+ non-relatives
H5.1	%-tage of vacant HU for rent
H5.2	%-tage of vacant HU for sale only
H5.4	%-tage of vacant HU for seasonal, recreational or occasional use
H7.1	%-tage of vacant HU with usual home elsewhere
H9.4	%-tage of ownOcc HU with Asian HH-lder
H9.8	%-tage of rentOcc HU with Indian HH-lder
H9.9	%-tage of rentOcc HU with Asian HH-lder
H13.1	%-tage of HU with 1-4 rooms
H14	Average number of rooms in a HU
H15.1	Average number of rooms in a ownOcc HU
H15.2	Average number of rooms in a rentOcc HU

wealth and pay-out equal to 0.5 and 0.05 similar to Dupuis and Victoria-Feser (2013), respectively.

As it can be seen from the learning curves in Figure 4, the optimal number of selected variables for LARS, RLARS, RobVIF, and JKRLARS is 17, 8, 9, and 6, respectively. Table 2 shows the mean MAPE (with standard deviation) values obtained using the optimal number of selected variables over 10 test subsets with 10 times replication for each method. As we can see in this table, JKRLARS outperfoms all the other considered methods by selecting the lowest number of variables, while its MAPE is slightly more than MAPE for LARS and almost equal to MAPE for RLARS, while RobVIF underperforms in comparison with other considered methods in both MAPE and the number of selected predictors. The decrease in selecting the number of variables by JKRLARS was 65%, 25%, and 33% relative to LARS, RLARS, and ROBVIF, respectively. Figure 5 shows the selected variables by the considered methods on Census data. Descriptions of the selected variables by the methods in Figure 5 can be found in Table 3 and descriptions of other variables are available in Dupuis and Victoria-Feser (2011).

6. Conclusion

Since outlier detection and variable selection in the presence of contaminations such as outliers and/or leverage points in the data are inseparable problems, we need a robust method capable of outlier detection and variable selection simultaneously. We proposed an outlier detection and robust variable selection method by combining RLARS with LTS regression as a highly robust regression method on JK subsets. The merit of using

JK subsets is to find the regression model with optimal predictive ability as measured by the MAD of the prediction errors obtained by cross-validation. This optimal regression model is devoted to the job of outlier detection. After removing the detected outliers, standard LARS is performed on the cleaned data to obtain a robust sequenced predictor variable in order of their importance.

The results of performing the proposed method on contaminated simulation data sets showed that it does a good job of outlier detection. Concerning robust variable selection, it does a good job of sequencing the predictor variables robustly for the different data configurations containing outliers and leverage points. Thus, a robust version of LARS sequenced predictor variables is yielded by JKRLARS. JKRLARS not only performs as well as its counterparts, RLARS and RobVIF, in robustly sequencing the predictor variables in simulation scenarios containing outliers, but it outperforms RLARS in situations with high leverage points, while RobVIF fails to robustly sequence the predictor variables in these situations.

Finally, through a real data set, we confirm that JKRLARS outperforms the other considered methods.

Acknowledgments

The authors would like to thank to the Associate Editor and the reviewers for their useful comments which led to a considerable improvement of the manuscript. This work was supported by FEDER Funds through "Programa Operacional Factores de Competitividade-COMPETE" and by Portuguese Funds through FCT "Fundação para a Ciência e a Tecnologia", within the SFRH/BD/51164/2010 and PEst-OE/MAT/UI0013/2017.

References

Akaike, H. 1970. Statistical predictor identification. *Annals of the Institute of Statistical Mathematics* 22 (1):203–17. doi:10.1007/BF02506337.

Arslan, O. 2012. Weighted lad-lasso method for robust parameter estimation and variable selection in regression. *Computational Statistics & Data Analysis* 56 (6):1952–65. doi:10.1016/j.csda.2011.11.022.

Atkinson, A. C., and M. Riani. 2002. Forward search added-variable t-tests and the effect of masked outliers on model selection. *Biometrika* 89 (4):939–46. doi:10.1093/biomet/89.4.939.

Cantoni, E., and E. Ronchetti. 2001. Robust inference for generalized linear models. *Journal of the American Statistical Association* 96 (455):1022–30. doi:10.1198/016214501753209004.

Čížek, P., and W. Härdle. 2006. Robust estimation of dimension reduction space. *Computational Statistics and Data Analysis* 51 (2):545–55. doi:10.1016/j.csda.2005.11.001.

Dupuis, D. J., and M.-P. Victoria-Feser. 2011. Fast robust model selection in large datasets. *Journal of the American Statistical Association* 106 (493):203–12. doi:10.1198/jasa.2011.tm09650.

Dupuis, D. J., and M.-P. Victoria-Feser. 2013. Robust VIF regression with application to variable selection in large data sets. *The Annals of Applied Statistics* 7 (1):319–41. doi:10.1214/12-AOAS584.

Efron, B. 1982. *The jackknife, the bootstrap and other resampling plans*. (Vol. 38). Philadelphia: SIAM NSF-CBMS.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. Least angle regression. *The Annals of Statistics* 32 (2):407–99. doi:10.1214/009053604000000067.

Fan, J., Y. Fan, and E. Barut. 2014. Adaptive robust variable selection. *Annals of Statistics* 42 (1): 324. doi:10.1214/13-AOS1191.

Frank, L. E., and J. H. Friedman. 1993. A statistical view of some chemometrics regression tools. *Technometrics* 35 (2):109–35. doi:10.1080/00401706.1993.10485033.

Gijbels, I., and I. Vrinssen. 2015. Robust nonnegative garrote variable selection in linear regression. *Computational Statistics & Data Analysis* 85:1–22. doi:10.1016/j.csda.2014.11.009.

Hastie, T., and B. Efron. 2013. lars: Least angle regression, lasso and forward stagewise [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=lars (R package version 1.2)

Huber, P. J., and E. M. Ronchetti. 2009. *Robust statistics*. New York: Wiley.

Hubert, M., P. J. Rousseeuw, and S. Van Aelst. 2008. High-breakdown robust multivariate methods. *Statistical Science* 23 (1):92–119. doi:10.1214/088342307000000087.

Khan, J. A., S. Van Aelst, and R. H. Zamar. 2007. Robust linear model selection based on least angle regression. *Journal of the American Statistical Association* 102 (480):1289–99. doi:10.1198/016214507000000950.

Lin, D., D. P. Foster, and L. H. Ungar. 2011. VIF regression: a fast regression algorithm for large data. *Journal of the American Statistical Association* 106 (493):232–47. doi:10.1198/jasa.2011.tm10113.

Mallows, C. L. 1973. Some comments on *Cp*. *Technometrics* 15 (4):661–75. doi:10.1080/00401706.1973.10489103.

Maronna, R. A., R. D. Martin, and V. J. Yohai. 2006. *Robust statistics: Theory and methods*. New York: J. Wiley & Sons.

McCann, L., and R. E. Welsch. 2007. Robust variable selection using least angle regression and elemental set sampling. *Computational Statistics and Data Analysis* 52 (1):249–57. doi:10.1016/j.csda.2007.01.012.

Müller, S., and A. Welsh. 2005. Outlier robust model selection in linear regression. *Journal of the American Statistical Association* 100 (472):1297–310. doi:10.1198/016214505000000529.

Pison, G., S. Van Aelst, and G. Willems. 2002. Small sample corrections for LTS and MCD. *Metrika* 55 (1–2):111–23. doi:10.1007/s001840200191.

R Core Team. 2014. R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Ronchetti, E. 1985. Robust model selection in regression. *Statistics and Probability Letters* 3 (1): 21–3. doi:10.1016/0167-7152(85)90006-9.

Ronchetti, E., and R. G. Staudte. 1994. A robust version of mallows' *cp*. *Journal of the American Statistical Association* 89 (426):550–9. doi:10.1080/01621459.1994.10476780.

Rousseeuw, P. J. 1984. Least median of squares regression. *Journal of the American Statistical Association* 79 (388):871–80. doi:10.1080/01621459.1984.10477105.

Rousseeuw, P. J., and K. Van Driessen. 2006. Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery* 12 (1):29–45. doi:10.1007/s10618-005-0024-4.

Salibian-Barrera, M., and S. Van Aelst. 2008. Robust model selection using fast and robust bootstrap. *Computational Statistics and Data Analysis* 52 (12):5121–35. doi:10.1016/j.csda.2008.05.007.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6 (2):461–4. doi:10.1214/aos/1176344136.

Shahriari, S., S. Faria, A. M. Gonçalves, and S. Van Aelst. 2014. Outlier detection and robust variable selection for least angle regression. In *International Conference on Computational Science and Its Applications* (pp. 512–522). Switzerland: Springer, Cham.

Wang, X., Y. Jiang, M. Huang, and H. Zhang. 2013. Robust variable selection with exponential squared loss. *Journal of the American Statistical Association* 108 (502):632–43. doi:10.1080/01621459.2013.766613.

Weisberg, S. 2005. *Applied linear regression*. New York: J. Wiley & Sons.

Yao, W., and Q. Wang. 2013. Robust variable selection through MAVE. *Computational Statistics and Data Analysis* 63:42–9. doi:10.1016/j.csda.2013.01.021.