Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 289 – 303
DOI : 10.15672/hujms.734212

Research Article

Robust regression estimation and variable selection when cellwise and casewise outliers are present

Onur Toka*[1], Meral Çetin[1], Olcay Arslan[2]

[1] *Hacettepe University, Faculty of Science, Department of Statistics, Ankara, Turkey*
[2] *Ankara University, Faculty of Science, Department of Statistics, Ankara, Turkey*

Abstract

Two main issues regarding a regression analysis are estimation and variable selection in presence of outliers. Popular robust regression estimation methods are combined with variable selection methods to simultaneously achieve robust estimation and variable selection. However, recent works showed that the robust estimation methods used in those estimation and variable selection procedures are only resistant to the casewise (rowwise) outliers in the data. Therefore, since these robust variable selection methods may not be able to cope with cellwise outliers in the data, some extra care should be taken when cellwise outliers are present along with the casewise outliers. In this study, we proposed a robust estimation and variable selection method to deal with both cellwise and casewise outliers in the data. The proposed method has three steps. In the first step, cellwise outliers were identified, deleted and marked with NA sign in each explanatory variable. In the second step, the cells with NA signs were imputed using a robust imputation method. In the last step, robust regression estimation methods were combined with the variable selection method LASSO (Least Angle Solution and Selection Operator) to estimate the regression parameters and to select remarkable explanatory variables. The simulation results and real data example revealed that the proposed estimation and variable selection procedure perform well in the presence of cellwise and casewise outliers.

Mathematics Subject Classification (2020). 62F35 , 62F07, 62J07

Keywords. Robust variable selection, outliers, cellwise outlier, LASSO

1. Introduction

One of the challenging problems in a regression analysis is to obtain estimators for the regression parameters that are robust against outliers in data sets. Until recently, outliers are defined as the observations that are not follow the model of the majority of the data. In a regression analysis, there are two types of outliers. One type is the outliers that may occur in the response variable and the other type of outliers occur in exploratory variables, which are usually called leverage points. Compared to the outliers in response variable,

*Corresponding Author.
Email addresses: onur.toka@hacettepe.edu.tr (O. Toka), meral@hacettepe.edu.tr (M. Çetin), oarslan@ankara.edu.tr (O. Arslan)
Received: 08.05.2020; Accepted: 23.11.2020

outliers in exploratory variables have a much greater influence on classical estimation procedures. If $X_{n\times p}$ is the data matrix formed by using the observations on the explanatory variables (rows as cases and columns as variables) the outliers in explanatory variables are used to be considered as the entire cases that correspond to the entire rows of $X_{n\times p}$. These outliers are called as casewise or rowwise outliers. Most of the robust regression methods, which are proposed against Huber-Tukey contaminated model, proceed by downweighting the entire rows that are considered as outliers (in response and/or casewise). Note that, in practice, the Huber-Tukey contaminated model corresponds to the casewise outliers [2]. However, in recent years, it has been realized that the observations considered as casewise outliers may not be completely contaminated. These observations may only have few contaminated cells and the rest of the cells may contain important information. These type of outliers are called as cellwise outliers [20]. That is, the cellwise outlier is a cell-deviated observation, so only outlier in one observation and one variable at the same time. The cellwise outliers may be the result of an independent contaminated model (ICM) [2]. In the presence of cellwise outliers, using ordinary robust regression estimation methods (for example using high breakdown point regression estimation methods) may be caused some loss of information since those methods try to downweight the entire row without considering non-contaminated cells in the outlying observations. Therefore, in recent papers new robust regression estimation methods have been proposed to take some extra care if cellwise and casewise outliers are present [1,6,17]. Debruyne *et al.* [7] argued that these outliers identification tools can be a thrilling topics. In order to compare outlier detection methods in the presence of cellwise and casewise outliers, Unwin [25] plotted the O3 graph, new visualization technique which is coded in a new R package called "cellWise" [19].

Another challenging problem in a regression analysis is to select a group of remarkable explanatory variables. To this extend, many variable selection methods have been proposed [11,24,31]. However, the popular ones are the methods that combine estimation and selection procedures together. These combined methods are also very effective for the high dimensional data sets. In particular, these methods are used for the regression problems involving data sets that have number of dimensions greater than the number of observes. The LASSO proposed by [24] is the first method in this direction. After the definition of LASSO, many other methods such as SCAD and bridge have been proposed to carry on simultaneous estimation and variable selection in a regression problem. Since LASSO and the other variable selection methods are based on the classical methods the researchers have been developed robust versions of these methods by using robust regression methods instead of the classical ones [3,4,8,15,28]. Since, the popular robust methods are designed to deal with the casewise outliers the combined robust estimation and variable selection methods, such as robust LASSO and robust SCAD, can only deal with the casewise outliers. However, recent works [1,9] show that the popular robust estimation methods may not be very successful when cellwise outliers are present. Especially, if we have high dimensional data and if the number of observations is rather small relative to the dimension of the data downweighting entire rows as casewise outliers may cause loss of information. Instead of doing so, monitoring those outliers and taking care only the outlying cells may reduce loss of information and improve estimation procedure.

Therefore, in recent papers, researchers have started concerning cellwise outliers and have proposed robust methods to deal with the cellwise outliers along with the casewise outliers. Some of these works are as follows. Raymaekers and Rousseeuw [18] proposed new identification technique which is based on LASSO regression with a stepwise application of constructed cutoff values for cellwise outliers. Leung *et al.* [12] proposed robust regression estimation methods under cellwise and casewise outliers contamination. However, there are few proposals for the robust estimation and variable selection in the presence of cellwise and casewise outliers [14]. In this paper, we will consider the robust estimation and the

variable selection in linear regression models when cellwise and casewise outliers are present in the data. Our proposal will have three steps. In the first step, we will try to identify the cellwise outliers in each explanatory variable. This will be done by independently monitoring each explanatory variable using outlier detection methods. After identifying cellwise outliers in each explanatory variable these outliers will be removed from the data and those cells will be marked by NA sign as it is done in [1, 13]. Then, in the second step, these cells will be regarded as missing observations and will be imputed by using the robust imputation method proposed by [5]. These two steps will make our explanatory data matrix as cellwise outliers free, but we may still have casewise outliers in the data. Finally, in the third step, we will combine robust regression estimation methods with LASSO, the variable selection method, to estimate the regression parameters and to select the remarkable explanatory variables without suffering from the casewise outliers. Our simulation results and real data example showed that the proposed estimation and selection method work well when casewise and cellwise outliers are possible in the data sets.

The rest of the paper is organized as follows. In Section 2 we will provide the details of the proposed method. In Section 3 the simulation and the real data examples will be given. The paper will be finalized with a conclusion section.

2. Three step robust regression estimation and variable selection in the presence of cellwise outliers

Consider the linear regression model

$$y_i = \alpha + \mathbf{x}_i^T \beta + \varepsilon_i, \quad i = 1, 2, 3, ..., n \tag{2.1}$$

where $y_i \in R$ is the response variable; $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ is the $p-$dimensional vector of the explanatory variables; $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ is the vector of regression parameters in R^p; and ε_i's are the iid random errors with zero mean, σ^2 variance and the distribution function F. Note that, distribution function F is symmetric distributions. Without loss of generality, we assume that $\alpha = 0$ and consider the model

$$y_i = \mathbf{x}_i^T \beta + \varepsilon_i, \quad i = 1, 2, 3, ..., n \tag{2.2}$$

The regression equation given in Equation (2.2) can also be written in matrix notation as

$$Y = X\beta + \varepsilon \tag{2.3}$$

where $X_{n\times p}$ is the design matrix, Y is the response vector, and ε is the vector of ε_i. Throughout this study, $\beta_0 = (\beta_{01}, \beta_{02}, ..., \beta_{0p})^T$ denotes the true parameter vector and $\Omega \subset R^p$ will denote the parameter space.

In this paper, our main aim is to estimate the regression parameters and select the important regressors under cellwise and casewise contaminations. As we have already mentioned, the casewise outliers can be identified using robust methods [16, 21] and are easily dealt with using robust variable selection methods if the variable selection is a concern. All of these can be done using combined robust estimation and variable selection methods. However, extra care should be taken to detect the cellwise outliers since they are not identified by examining the whole data matrix X. Each explanatory variable, that is; each column of X should be monitored to detect the cellwise outliers. Thus, before preforming estimation and variable selection each variable should be scanned in terms of cellwise outliers. As it is proposed by [1] and [13] after detecting the cellwise outlier, those cells should be imputed using robust imputation methods. Then, robust methods related to the problem of interests can be used to handle the casewise outliers. In the following subsections, starting from the identification of the cellwise outliers, we will describe the three steps of the proposed robust estimation and variable selection method when cellwise and casewise outliers are present.

2.1. Identifying cellwise outlier

Cellwise outlier (introduced in [2]) is not a big problem when the proportion of outliers compared to the sample size is not high. However, Alqallaf *et al.* [2] observed that even if there is a very small percent of outliers in every variables, but if the dimension of the data is large, popular robust estimators with high breakdown point will easily reach their possible breakdown point. In recent years, researchers have become aware of cellwise outliers and they have proposed several methods to deal with this problem. Most of the proposed methods first identifies the cellwise outliers and regard them as missing observations by changing them with NA sign [1, 9, 13]. That is, the outlier problem is transferred to a missing data problem. In order to obtain cellwise outliers, there is a new methodology which combines LASSO regression with a stepwise application of constructed cutoff values [18]. In this paper, following the same strategy, we will try to identify the cellwise outliers by using the outlier detection method described in [20]. First, we have to obtain robust estimates for the location and scale of each column. In this paper, we will use the sample median for location and MAD for scale. These estimates will be used as initial robust estimates to obtained the one-step M estimates for location and scale computed as

$$\begin{aligned} \hat{\mu}_M &= \frac{\sum w_i x_i}{\sum w_i} \\ \hat{\sigma}^2_M &= \frac{1}{n}\sum w_i \left(x_i - \hat{\mu}_M\right)^2 \end{aligned} \tag{2.4}$$

where weights $w\left(t\right) = \frac{\rho'(t)}{t}$ are computed using Tukey biweight ρ function. Note that, w_i is weights for i_{th} observation and W is a diagonal weight matrix. After robust location and scale estimates are computed, each column will be standardized using these robust estimates. Let z_i denote these standardized columns. Then, the observations x_i will be considered as outliers if

$$|z_i| \geq \sqrt{\chi^2_{1,q}} \tag{2.5}$$

where q is $q-th$ quantile of the chi-squared distribution. After screening all the columns and identifying all the cellwise outliers those cells will be replaced by NA signs, and hence the cellwise outlier problem will be transfered into the missing observation problem. This will be the first step of our proposed robust variable selection method. In next subsection we will describe the robust imputation algorithm to impute the observations that are flagged as NA.

2.2. Bypassing cellwise outlier: Robust imputation

After identifying cellwise outliers and replace them with NA, we have created a missing value problem. Thus, these missing values have to be imputed using some imputation methods. There are several procedures to deal with missing observations in the data. These procedures are classified according to the missingness patterns in the data. Cellwise outliers are considered as randomly occurred outliers. Therefore, deleting the cellwise outliers in the data causes the missingness case called as missing completely at random (MCAR). This type of missing data can be easily imputed using mean or median imputation method. In this paper we will use the robust imputation (ROBimpute) method proposed by [5]. Actually, the robust imputation method is a robust alternative to the sequential imputation (SEQimpute) method proposed by [26] and it can be summarized as follows. Let X_c be the completely observed part and X_m be the missing part of our explanatory data matrix X which contains missing observations. x^* be a row in X_m defined as $x^* = \left[(x^*_m)^T (x^*_o)^T\right]^T$, where x^*_m and x^*_o are the missing and observed part of that

row, respectively. As described in [26], let the matrix C defined as in Equation (2.6) be the inverse of the covariance matrix of X_c and let X^* be $\left[X_c^T, x^*\right]^T$. Further, let $\bar{x}_c$ be the rowwise sample mean of the complete data. Now minimizing the equation given in Equation (2.7), which can be also written as in Equation (2.8), will be an estimate for x^*. After finding x_m^* in X^*, it will be used instead of x^* in X^* to form new completed data. Then we have to take care the next missing observations. This procedure should be continued after all the missingness are imputed. The detailed information about SEQimpute can be found in [26].

$$C = \begin{bmatrix} C_{m,m} & C_{m,o} \\ C_{m,o}^T & C_{m,m} \end{bmatrix} \tag{2.6}$$

$$D\left(x^*\right) = \left(x^* - \bar{x}_c\right)^T \left(cov\left(X_c\right)\right)^{-1} \left(x^* - \bar{x}_c\right) \tag{2.7}$$

$$x_m^* = \left(\bar{x}_c\right)_m - \left(C_{m,m}\right)^{-1} C_{m,o}\left(x_o^* - \left(\bar{x}_c\right)_o\right) \tag{2.8}$$

However, since this SEQimpute algorithm is based on sample mean and sample covariance, it is not robust against the outliers in the whole dataset. Therefore, even a single outlier can badly ruin the algorithm and the imputed value for the missing observations will be far from the expected value. For this reason, robust alternative to the SEQimpute has been proposed in [5]. They use robust covariance estimator and the robust location estimator instead of sample mean and the sample covariance matrix. In particular, they use minimum covariance determinant (MCD) estimator as the covariance estimator and the sample median for the mean estimator. The rest of the imputation will be same as in the classical one described above. This imputation is called ROBimpute and the detail of the algorithm is found in [5]. In this paper we will use the ROBimpute to impute the missing cells that are created deleting the cellwise outliers.

2.3. Variable selection with robust LASSO

In this section, we will describe the third step of our proposal. Namely, we will explore the variable selection for the regression model using refined data. Variable selection methods are one of the most important part of modeling aspect. In particular, in regression methods, we are interested in the most important variables and the subsets of full model. Robust variable selection, such as LASSO, is the robust versions of the classical ones in the presence of outliers. In this paper, we used LASSO to carry on our variable selection. LASSO is a well known method which minimizes OLS loss function $(\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta})$ under the restriction $\sum_{j=1}^{p} |\boldsymbol{\beta_j}| \leq t$. Hence, this minimization problem with respect to β can be carried on using lagrange multiplier method. That is we have to minimize the following objective function,

$$Q_N = (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\boldsymbol{\beta_j}| \tag{2.9}$$

where λ is regularization parameter.

Using LASSO, parameter estimation and variable selection can be simultaneously obtained. Since the classical LASSO is based on OLS criterion, the resulting estimators will be sensitive to the outliers, the robust version of LASSO have been proposed in literature [3, 27]. In robust versions, OLS loss functions have been replaced with robust version of loss functions such as Huber or Tukey ρ functions.

Several algorithms have been proposed to obtain LASSO estimators. One of these algorithms to solve the robust LASSO problem is proposed by [28] and it is called using semi-smooth Newton coordinate descent (SNCD) algorithm. In this paper, we will use this algorithm to obtain robust LASSO estimates when we have outlier in y direction or we have heavy-tailed error distribution. The algorithm is provided in the same paper and it is available as R packages named "hqreg".

By using robust LASSO, we will get estimators that are resistant to the outliers in y direction. However, if we have casewise outliers in x direction, the robust LASSO obtained using Huber or Tukey ρ functions will be badly affected from the casewise outliers in x direction. Therefore, we have to modified the robust LASSO method to deal with the casewise outliers.

Concerning the casewise outliers, we will use the MM regression estimation method proposed by [29]. The MM estimation method will be used as follows. We will first obtain the MM estimators for the regression parameters. Then, using these MM estimators, we will compute the weights w_i for $i = 1, 2, ..., n$ for each observations using the weight function obtained from the Tukey ρ function (see e.g. [16], page 30). Then, we will form $W = diag(w_1, ..., w_n)$ matrix and transform our X and Y using W matrix as $X^* = W^{1/2}X$ and $Y^* = W^{1/2}Y$, respectively. Now we can apply classical LASSO to transform data to do variable selection.

Finally, these three steps can be combined to obtain robust parameter estimation and variable selection in the presence of cellwise and casewise outlier. The following algorithm will be used to carry on all of these procedures. In our simulation and real-data example, this algorithm will be implemented to demonstrate the performance of the proposed method. If it is followed from the algorithm, it will see that robust methods with robust imputation are preferred when there are both cellwise and casewise outliers. If there are only casewise outliers robust LASSO methods are preferred. If there are no outliers in dataset, classical LASSO method is preferred.

Algorithm 1: Variable Selection in the presence of cellwise and casewise outliers

---

Starting of Algorithm.
Data Obtain data (Generating data in simulation or use data from real world example)
If you suspect any Cellwise outliers, then Run
STEP 1: Identification of Cellwise Outliers
Loop 1. $i = 1, 2, \cdots, p$ (For each regressors)
Identify cellwise outliers using the procedure described in Section 2.1 and change them with NA
End Loop 1
STEP 2: Robust Imputation of NA
Loop 2. $m = 1, 2, \cdots, M$ (For each NA)
Impute the NA's by using robust imputation methods described in Section 2.2
End Loop 2
ElseIf Any Casewise Outlier
STEP 3: Robust Estimation and Variable Selection
Apply Robust LASSO described in Section 2.3
ElseIf No Outlier
Apply LASSO
End If
End of Algorithm.

---

3. Numerical studies

In the application part, we considered simulation study in R to compare the performance of variable selection methods in the presence of cellwise outliers. We considered the regression model given in Section 2. The explanatory variables were independently generated from the normal distribution $N(m, 1)$ with m coming from discrete uniform distribution randomly between zero and five. In the simulation study, the dimension of the parameter vector was taken as 7,15 and 30 and the sample sizes were taken as 50, 100

and 250. For the regression model, we took the regression parameters as $[1,0,1,0,1,1,0]'$ for dimension 7. For the dimensions 15, we formed β as follows: first five entries were taken as one and the others are zero. Similarly, for the dimension 30, the first 10 entries of β were one and the rest of the entries are as zero. In the regression model, we used three different error distributions. We first took the standard normal distribution ($N(0,1)$) to explore the case without outliers in y-direction. The other two error distributions were $0.9N(0,1) + 0.1N(3,1)$ and t_3. With these distributions we guaranteed the outliers in y-direction. For the outliers in x direction, we generated randomly observations from $N(50,1)$ and combined these observations with the major part of the data.

In this simulation study, cellwise outliers were generated as follows. We first generated explanatory variables and form our X matrix. Using missingmat() function in ForIMP R package (see [10]), we created missing observations which were completely at random and replaced the missing observations with NA signs. Now, we would apply three different imputation procedures to the X matrix. First, we used ROBimpute method to robustly impute this missing observations. Second, we used SEQimpute method to impute the missing observation in classical way (for the functions for imputation given in [5, 26] are used). Finally, to have data with cellwise outliers, we imputed the NAs with the values calculated by $\max(x_i)+2\sigma_{x_i}$. In this ad-hoc method, we easily obtained cellwise outliers in simulated data. To sum up, we had three different X matrices. One had cellwise outliers, the other ones had missingness wich were imputed by robust and the classical imputation methods. The proportion of cellwise outliers were 1%, 5% and 10%. Note that, when cellwise outliers were constructed, the proportion was calculated using $n \times p$, not just n. After we designed our data, we applied three different combination of LASSO methods using the glmnet [22] and hqregraw functions [28] in R. Note that, for the casewise outlier in x-direction we used glmnet function for the modified dataset described in previous section.

In the simulation results the methods were compared in three different ways. We randomly divided data in two subsections. We used one part for estimation and variable selection (training ; 80% of dataset) and the other part is testing (20 % of dataset). After we did estimation and variable selection, we counted the number of true zero- beta selection and we also calculated proportion of true model selection. Then, using the testing part of data, we computed the prediction error $\frac{1}{T}\sum_{i=1}^{T}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2/n$ where n is the number of observation and T is the number of iteration in testing data. We also provided some boxplot illustrations for estimated betas.

The simulation results were summarized in Tables 1-5. Tables 1-3 contained prediction errors. In Table 1, we displayed the results for the case normally distributed errors with cellwise and casewise outliers for the sample size $n = 50$. If we only had cellwise outlier, we observed the smallest prediction error for the case robust imputed data using classical LASSO (ROB-LASSO) and sequentially imputed data using classical LASSO (SEQ-LASSO). Therefore, we could say that robust imputation gave a better estimation for cellwise outliers. We also observed that when the number of cellwise outliers increased, the prediction errors for LASSO and robust imputed LASSO also increased. Overall, ROB-LASSO and SEQ-LASSO had superiority over the other methods for this case. When casewise outliers were introduced to the data, we observed that robust imputed robust LASSO (ROB-RLASSO) seems better performance for most of the cases compared to the other methods.

In Table 2, we gave the simulation results for the contaminated error distribution and we observed similar behavior for ROB-RLASSO. That is, the results for the ROB-RLASSO was superior to the other methods. In Table 3, simulation results for t_3 distributed error case were summarized. Concerning this case, without casewise outlier ROB-RLASSO gave smaller prediction errors for almost all the cases. However, when the casewise outliers were

introduced in the data, the performance of the ROB-RLASSO was getting worse compare to the robust LASSO (RLASSO).

Table 1. Prediction error for $n = 50$ and $\varepsilon \sim N(0,1)$

pr-casew	p	pr-cellw	LASSO	RLASSO	ROB-LASSO	ROB-RLASSO	SEQ-LASSO	SEQ-RLASSO
0	7	0.01	5.902	5.154	1.793	1.898	1.799	1.900
0	7	0.05	18.289	19.410	2.160	2.273	2.153	2.249
0	7	0.10	21.716	25.832	2.427	2.521	2.409	2.519
0	15	0.01	3.808	3.345	1.013	1.085	1.017	1.079
0	15	0.05	11.166	10.145	1.438	1.510	1.449	1.491
0	15	0.10	12.549	12.068	5.937	6.017	3.951	3.712
0	30	0.01	4.146	3.808	1.002	1.087	1.004	1.098
0	30	0.05	14.497	11.570	1.062	1.152	1.046	1.162
0	30	0.10	14.333	11.610	1.289	1.434	1.311	1.440
0.05	7	0.01	813.074	541.661	2.987	3.059	2.717	2.817
0.05	7	0.05	5346.473	3577.792	8.556	8.174	8.979	9.233
0.05	7	0.10	3616.270	9062.238	27.115	17.156	14.483	15.828
0.05	15	0.01	469.709	356.820	2.054	3.333	2.275	6.391
0.05	15	0.05	4307.260	2233.865	17.964	23.400	6.736	11.948
0.05	15	0.10	4108.178	4861.238	218.290	217.778	431.289	297.443
0.05	30	0.01	493.541	982.032	2.069	9.900	1.271	10.643
0.05	30	0.05	5886.633	4772.21	7.783	19.457	6.780	23.559
0.05	30	0.10	10434.33	8941.705	98.562	123.690	128.721	221.843

p: Number of parameters; pr-cellw: Cellwise outlier proportion; pr-casew: x direction outlier proportion; LASSO: Classical LASSO; RLASSO: Robust LASSO; ROB-LASSO: Robust imputed LASSO; ROB-RLASSO: Robust imputed Robust LASSO; SEQ-LASSO:Sequential imputed LASSO; SEQ-RLASSO: Sequential imputed Robust LASSO.

Table 2. MSE of beta for $n = 50$ and $\varepsilon \sim N(0,1) + N(3,1)$

pr-casew	p	pr-cellw	LASSO	RLASSO	ROB-LASSO	ROB-RLASSO	SEQ-LASSO	SEQ-RLASSO
0	7	0.01	7.022	5.963	3.050	3.115	3.072	3.113
0	7	0.05	20.945	20.988	3.579	3.586	3.582	3.595
0	7	0.10	25.313	30.345	4.063	4.052	4.056	4.059
0	15	0.01	4.397	4.244	1.716	1.766	1.700	1.777
0	15	0.05	12.502	10.808	2.088	2.139	2.073	2.105
0	15	0.10	12.339	11.738	7.121	6.931	4.446	4.483
0	30	0.01	4.733	4.435	1.517	1.596	1.517	1.597
0	30	0.05	15.298	12.179	1.632	1.671	1.706	1.666
0	30	0.10	14.507	12.227	1.850	1.972	1.876	1.956
0.05	7	0.01	38.653	661.211	3.318	3.841	3.513	5.331
0.05	7	0.05	6562.070	3857.103	3.816	5.648	3.672	5.011
0.05	7	0.10	3933.421	7910.298	77.336	37.420	18.466	28.275
0.05	15	0.01	450.921	367.050	1.660	2.110	1.662	3.253
0.05	15	0.05	3807.158	2180.618	23.511	23.329	3.181	8.353
0.05	15	0.10	3984.041	5059.752	260.543	257.044	83.259	384.752
0.05	30	0.01	606.122	841.965	9.369	41.843	11.931	64.211
0.05	30	0.05	7971.026	4716.980	32.025	50.330	46.235	103.099
0.05	30	0.10	9336.502	8642.231	87.116	141.860	141.060	235.443

p: Number of parameters; pr-cellw: Cellwise outlier proportion; pr-casew: x direction outlier proportion; LASSO: Classical LASSO; RLASSO: Robust LASSO; ROB-LASSO: Robust imputed LASSO; ROB-RLASSO: Robust imputed Robust LASSO; SEQ-LASSO:Sequential imputed LASSO; SEQ-RLASSO: Sequential imputed Robust LASSO.

Table 3. MSE of beta for $n = 50$ and $\varepsilon \sim t_3$

pr-casew	p	pr-cellw	LASSO	RLASSO	ROB-LASSO	ROB-RLASSO	SEQ-LASSO	SEQ-RLASSO
0	7	0.01	9.058	8.095	4.960	4.736	4.954	4.746
0	7	0.05	19.349	19.948	4.663	4.515	4.667	4.494
0	7	0.10	24.775	30.729	5.549	5.301	5.537	5.279
0	15	0.01	5.228	4.868	2.843	2.701	2.818	2.709
0	15	0.05	11.125	9.156	2.599	2.512	2.589	2.542
0	15	0.10	13.388	12.450	7.489	7.695	5.626	5.877
0	30	0.01	5.056	4.747	2.061	1.999	2.064	1.987
0	30	0.05	18.032	12.510	2.175	2.086	2.152	2.099
0	30	0.10	15.820	12.237	2.181	2.206	2.247	2.203
0.05	7	0.01	898.064	523.740	6.539	7.146	6.532	8.383
0.05	7	0.05	7288.761	4026.641	7.124	8.047	6.625	7.848
0.05	7	0.10	4171.419	8775.199	18.291	29.342	10.470	17.674
0.05	15	0.01	428.620	342.041	4.010	4.376	3.979	4.375
0.05	15	0.05	4510.567	2382.344	16.507	18.087	6.570	10.137
0.05	15	0.10	3913.375	5047.254	254.423	267.726	116.708	145.420
0.05	30	0.01	634.455	866.216	2.777	32.245	2.731	40.746
0.05	30	0.05	6331.625	4951.489	39.471	105.542	47.720	114.237
0.05	30	0.10	8828.778	8993.255	97.325	179.917	136.942	251.600

p: Number of parameters; pr-cellw: Cellwise outlier proportion; pr-casew: x direction outlier proportion; LASSO: Classical LASSO; RLASSO: Robust LASSO; ROB-LASSO: Robust imputed LASSO; ROB-RLASSO: Robust imputed Robust LASSO; SEQ-LASSO:Sequential imputed LASSO; SEQ-RLASSO: Sequential imputed Robust LASSO.

Table 4. Percents of true model selection - I

	cellwise pr	True Choice Pr.	LASSO	RLASSO	ROB-LASSO	ROB-RLASSO	SEQ-LASSO	SEQ-RLASSO
$\varepsilon \sim N(0,1)$	0.01	$\beta_2 = 0$	57.2	70	69.6	79.6	63.2	65.2
		$\beta_4 = 0$	53.6	70	70	80.8	54.8	54.8
		$\beta_7 = 0$	58.4	69.6	70.8	84	60.4	61.2
		True Model	16.4	36.4	38.4	51.2	24.8	27.2
$\varepsilon \sim N(0,1)$	0.05	$\beta_2 = 0$	29.6	66.4	63.6	82.4	56.8	57.6
		$\beta_4 = 0$	31.6	62.8	65.2	80.8	55.6	54.8
		$\beta_7 = 0$	30	61.6	62.4	74.4	53.2	52.4
		True Model	2.4	31.2	31.2	32.4	20.4	19.6
$\varepsilon \sim N(0,1)$	0.01	$\beta_2 = 0$	51.6	65.6	65.2	81.6	58.8	60.4
$+N(3,1)$		$\beta_4 = 0$	50.4	62.4	63.2	79.2	56.4	56.8
		$\beta_7 = 0$	50.8	62	61.6	76.8	53.2	56
		True Model	13.2	26.8	27.6	49.2	20.8	22
$\varepsilon \sim N(0,1)$	0.05	$\beta_2 = 0$	30.8	64	65.6	74.4	56.8	56.8
$+N(3,1)$		$\beta_4 = 0$	30	65.2	63.2	78.8	53.2	53.6
		$\beta_7 = 0$	28.4	62	63.6	76.8	52.4	51.2
		True Model	1.6	31.2	32	27.2	18.4	17.6
$\varepsilon \sim t_3$	0.01	$\beta_2 = 0$	56.8	62	64	77.6	60	62
		$\beta_4 = 0$	57.6	57.6	62	76	59.2	58.4
		$\beta_7 = 0$	50.8	59.2	60.8	74.8	53.2	52
		True Model	18	26.4	28	46.4	22.8	24
$\varepsilon \sim t_3$	0.05	$\beta_2 = 0$	30.8	58	57.6	70	55.6	54.4
		$\beta_4 = 0$	31.2	60.8	63.2	77.2	57.2	58
		$\beta_7 = 0$	30.8	61.6	62.8	77.2	53.2	54.8
		True Model	2	24.8	27.2	26.8	19.6	21.6
$\varepsilon \sim N(0,1)$	0.01	$\beta_2 = 0$	65.2	82.8	82.4	90.0	80.0	81.2
$+5\%$ *casewise*		$\beta_4 = 0$	62.4	81.6	82.0	90.4	77.6	75.6
		$\beta_7 = 0$	66.4	83.2	84.0	88.4	83.2	82.0
		True Model	2.7	21.4	21.8	28.0	20.6	20.5
$\varepsilon \sim N(0,1)$	0.05	$\beta_2 = 0$	87.2	79.6	80	98.8	79.6	79.6
$+$ 5% casewise		$\beta_4 = 0$	86.8	77.6	78.8	96.4	79.6	80.8
		$\beta_7 = 0$	89.2	77.6	78.8	96.8	74.8	77.6
		True Model	0	48	48	71.6	46.8	50.8

LASSO: Classical LASSO; RLASSO: Robust LASSO; ROB-LASSO: Robust imputed LASSO; ROB-RLASSO: Robust imputed Robust LASSO; SEQ-LASSO:Sequential imputed LASSO; SEQ-RLASSO: Sequential imputed Robust LASSO.

In Tables 4-5, we displayed the correctly selected number of zero betas and the correctly selected true models for $p = 7$ and $n = 50$ (Table 4) and $n = 250$ (Table 5). We observed that robust imputed robust LASSO performed the best correctly choosing zero betas and the correctly choosing true model. Robust LASSO seemed the second best among the others for identifying zero betas and the correct model. We observed that the other methods were broke-down for correctly choosing zero betas and correct model in the presence of cellwise and casewise outliers.

Table 5. Percents of true model selection - II

		True Choice Pr.	LASSO	RLASSO	ROB-LASSO	ROB-RLASSO	SEQ-LASSO	SEQ-RLASSO
$\varepsilon \sim N(0,1)$	0.01	$\beta_2 = 0$	26	93.2	92.8	96.4	91.6	92
		$\beta_4 = 0$	32	90.8	90.8	95.6	90	89.6
		$\beta_7 = 0$	26.4	92	93.2	97.2	89.2	90
		True Model	4.8	78	79.2	89.2	73.2	74.4
$\varepsilon \sim N(0,1)$	0.05	$\beta_2 = 0$	2.4	87.6	88	90.8	86	87.2
		$\beta_4 = 0$	7.2	88.8	89.2	92.8	88.8	89.2
		$\beta_7 = 0$	4.4	90.4	90	92.8	90	89.6
		True Model	0	71.6	71.6	77.2	68.4	69.2
$\varepsilon \sim N(0,1)$	0.01	$\beta_2 = 0$	26.8	82	81.6	92.4	82	82.4
$+N(3,1)$		$\beta_4 = 0$	28.4	79.2	76	89.6	83.2	82.8
		$\beta_7 = 0$	26.4	80.8	78.8	92	84.8	84.8
		True Model	2	55.2	52.4	76	58.8	58
$\varepsilon \sim N(0,1)$	0.05	$\beta_2 = 0$	5.6	79.2	78.4	84.4	85.6	86.4
$+N(3,1)$		$\beta_4 = 0$	8	74.8	76	80.8	77.2	77.2
		$\beta_7 = 0$	5.2	80.8	80	85.6	84.4	84.4
		True Model	0	50	48.4	58.4	55.6	56.4
$\varepsilon \sim t_3$	0.01	$\beta_2 = 0$	30.8	81.2	81.6	93.2	87.2	87.6
		$\beta_4 = 0$	32.8	79.6	79.2	89.6	84.8	84.4
		$\beta_7 = 0$	29.6	79.2	79.6	89.2	85.6	86.4
		True Model	3.2	54.8	56.4	76	63.2	64
$\varepsilon \sim t_3$	0.05	$\beta_2 = 0$	6	74.8	72.4	79.6	86	84.8
		$\beta_4 = 0$	4.4	78.4	76.4	82.8	84	84.4
		$\beta_7 = 0$	6.4	74.8	74.4	80.8	83.2	83.2
		True Model	0	48.4	45.2	53.2	59.6	59.2
$\varepsilon \sim N(0,1)$	0.01	$\beta_2 = 0$	64	91.6	91.6	97.6	91.6	91.6
$+5\%\, casewise$		$\beta_4 = 0$	65.2	93.6	93.6	98.8	92.8	92.8
		$\beta_7 = 0$	66.4	95.6	95.6	98.4	92.8	92.8
		True Model	0.8	81.2	81.2	94.8	78.0	78.0
$\varepsilon \sim N(0,1)$	0.05	$\beta_2 = 0$	1.6	94.8	94.8	100.0	93.6	92.4
$+5\%\, casewise$		$\beta_4 = 0$	2.0	94.0	94.0	100.0	94.0	93.6
		$\beta_7 = 0$	2.8	93.2	93.2	100.0	91.2	91.6
		True Model	0.0	82.8	82.8	98.8	79.2	78.8

LASSO: Classical LASSO; RLASSO: Robust LASSO; ROB-LASSO: Robust imputed LASSO; ROB-RLASSO: Robust imputed Robust LASSO; SEQ-LASSO:Sequential imputed LASSO; SEQ-RLASSO: Sequential imputed Robust LASSO.

Concerning the results given in Table 5, we observed exactly the similar performance of the methods. Robust imputed Robust LASSO had the excellent behavior for correctly choosing zero betas and for identifying the correct models. Comparing to the results given in Table 4, we noticed that the performances were getting better. For example, when the sample size was small for normally distributed error with 5% cellwise and 5% casewise outliers (see the 8th case in Table 4 and Table 5), the ratio choosing the corrected model is 71.6% . However, that ratio was 98.8% in Table 5. Therefore, increasing sample size affected for choosing correct model and correct zero betas.

Further to illustrate performance of the methods for higher dimensional cases, we gave boxplots of the some of the estimated zero betas (Mainly, we took last three zeros for simplicity). These boxplots were given in Figures 1-3. In these figures, dimension of the regression parameter is 15. We considered different outliers configurations in these figures. In Figures 1 and 2, heavy-tailed error distribution with cellwise outliers. On the other hand, in Figure 3, we had cellwise outlier and casewise outlier with normally distributed errors. We observed that robust imputed robust LASSO superior to the other methods

in terms of correctly choosing zero betas almost all the cases. Compare to the others, variability seemed smaller.

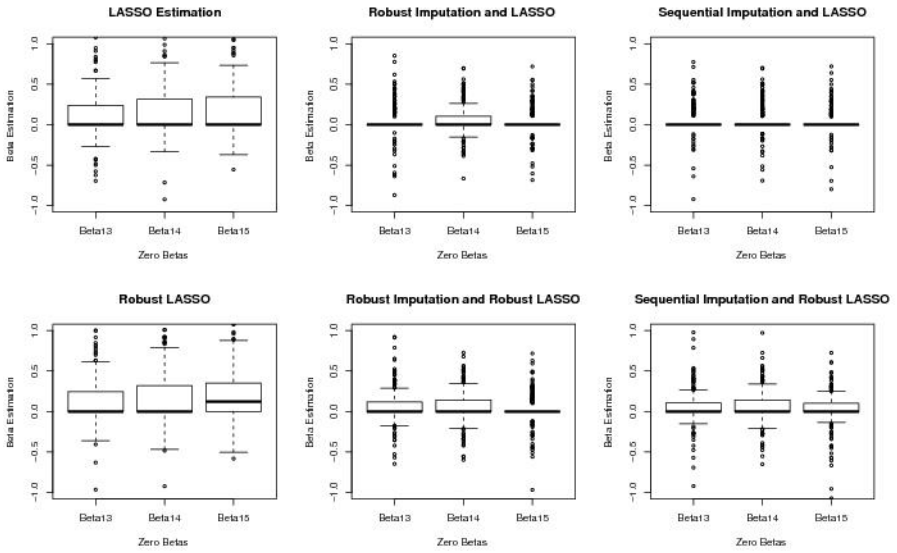

Figure 1. Results for $p = 15, n = 50, cellwise - pr = 0.05$, and $\varepsilon \sim N(0,1) + N(3,1)$

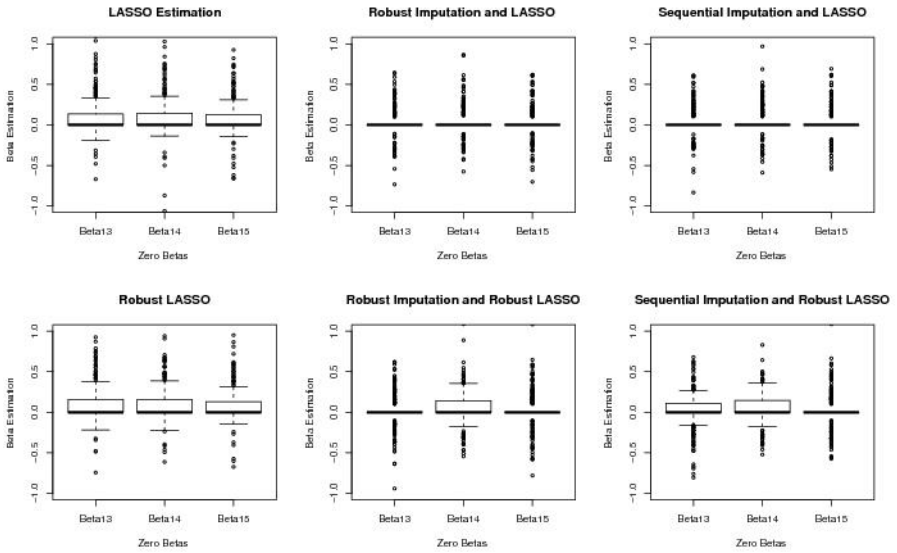

Figure 2. Results for $p = 15, n = 50, cellwise - pr = 0.01$ and $\varepsilon \sim t_3$

4. Real data example

To compare the methods in real data example, the most known model selection data, the prostate cancer data in [23] was examined. There are 97 observations collected from men who were about to receive a radical prostatectomy. The response variable was log(prostate specific antigen) (lpsa). The explanatory variables were log (cancer volume) x_1 : lcavol), log(prostate weight) (x_2 : lweight), age(x_3), log(benign prostatic hyperplasia amount) (x_4 : lbph), seminal vesicle invasion (x_5: svi), log(capsular penetration) (x_6: lcp), Gleason

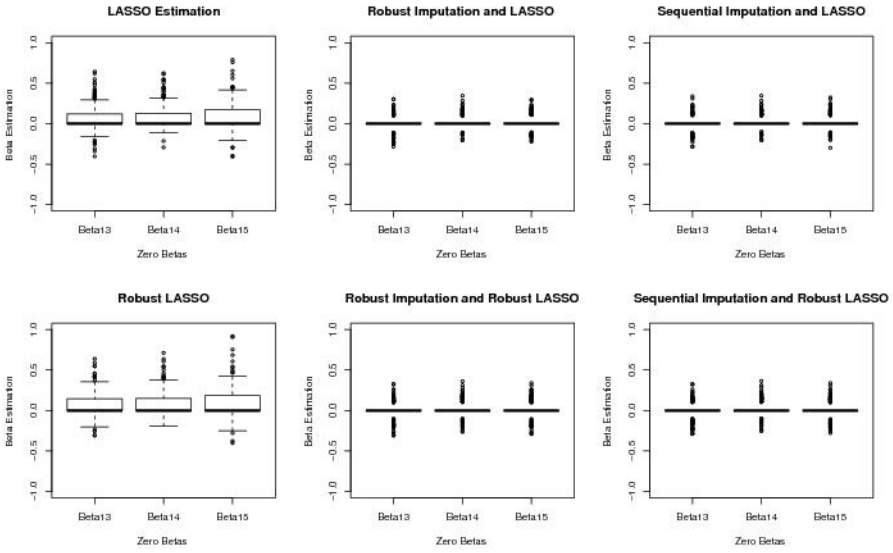

Figure 3. Results for $p = 15, n = 100, cellwise - pr = 0.05$, $casewise - pr = 0.05$ and $\varepsilon \sim N(0, 1)$

score (x_7: gleason) and percentage Gleason scores 4 or 5 (x_8 : pgg45). In literature, this dataset has been extensively used to access the performance of the model selection methods [24, 30]. In those papers, the variables x_1, x_4, x_5 were found the most important variables. In the applications of [24,30], explanatory variable x_3 was also found significant. In our paper, we compared the methods in terms of correctly selected non-significant betas (zero betas) and true model selection. We also checked the prediction errors for testing dataset which was randomly chosen 20% of the real dataset in each iteration. The results were given in Table 6 and Figure 4. All of these results confirmed that robust imputed robust LASSO was the best according to the criteria we were using. We also noticed that sequential imputed Robust LASSO had the similar behavior to the robust imputed robust LASSO.

5. Conclusion

After introducing cellwise outlier or independent contamination model, some problems occured in estimation even robust ones. Especially in high dimension, breakdown points of estimation will be exceeded even though there is very small proportion cellwise outliers. In this paper, we considered cellwise and the casewise outlier problem in a regression analysis when parameter estimation and variable selection is a concern. We used robust imputation method to deal with the cellwise outlier and we combined the robust regression estimation method with LASSO to deal with the variable selection in the presence of cellwise and casewise outliers. We did this procedure in three steps. In the first step, we had identified the cellwise outliers and in the second step, we had dealt with the cellwise outliers and use robust imputation to get rid of the cellwise outliers. Finally, in the last step, we combined robust estimation with LASSO to dealt with casewise outliers if they are in present. We provided an extensive simulation study to illustrate the performance of proposed method and observed that the proposed method has comparable results among the methods that have similar proposal. We had also explored the real data example using prostate cancer data which have been extensively used in literature to show the performance of the model selection methods. The result of the real data example have also confirm the simulation results in terms of the proposed method.

Table 6. Real data examples: prostate cancer data results

MSE of Beta for Prostate Cancer Data							
	pr-cellw	LASSO	RLASSO	ROB-LASSO	ROB-RLASSO	SEQ-LASSO	SEQ-RLASSO
	0.01	4.477	5.199	1.308	1.200	1.303	1.199
	0.05	14.845	14.047	1.272	1.192	1.272	1.195
	0.10	13.520	12.898	0.970	0.906	0.952	0.896
Zero Beta Selection for Prostate Cancer Data							
	pr-cellw	LASSO	RLASSO	ROB-LASSO	ROB-RLASSO	SEQ-LASSO	SEQ-RLASSO
		100.00	100.00	100.00	100.00	100.00	100.00
		33.60	87.60	85.60	90.00	38.80	38.00
0.01		63.60	94.40	93.20	99.60	80.40	81.20
		98.00	96.80	97.60	100.00	100.00	100.00
		100.00	100.00	100.00	100.00	100.00	100.00
		100.00	100.00	100.00	100.00	100.00	100.00
		66.80	95.20	95.20	100.00	86.00	85.60
0.05		98.40	96.00	95.60	100.00	76.80	78.80
		73.60	90.80	90.80	100.00	96.00	95.60
		100.00	100.00	100.00	100.00	100.00	100.00
		100.00	100.00	100.00	100.00	100.00	100.00
		84.40	72.00	70.80	87.20	49.20	55.60
0.10		99.60	36.00	21.60	99.20	26.00	19.60
		5.60	11.60	4.40	98.00	31.20	17.60
		100.00	100.00	100.00	100.00	100.00	100.00
True Model Selection for Prostate Cancer Data							
	pr-cellw	LASSO	RLASSO	ROB-LASSO	ROB-RLASSO	SEQ-LASSO	SEQ-RLASSO
	0.01	8.80	81.20	79.20	88.80	31.20	31.20
	0.05	0.80	88.40	88.00	98.40	65.20	66.4
	0.10	0.00	6.40	1.60	78.40	6.40	5.60

LASSO: Classical LASSO; RLASSO: Robust LASSO; ROB-LASSO: Robust imputed LASSO; ROB-RLASSO: Robust imputed Robust LASSO; SEQ-LASSO:Sequential imputed LASSO; SEQ-RLASSO: Sequential imputed Robust LASSO.

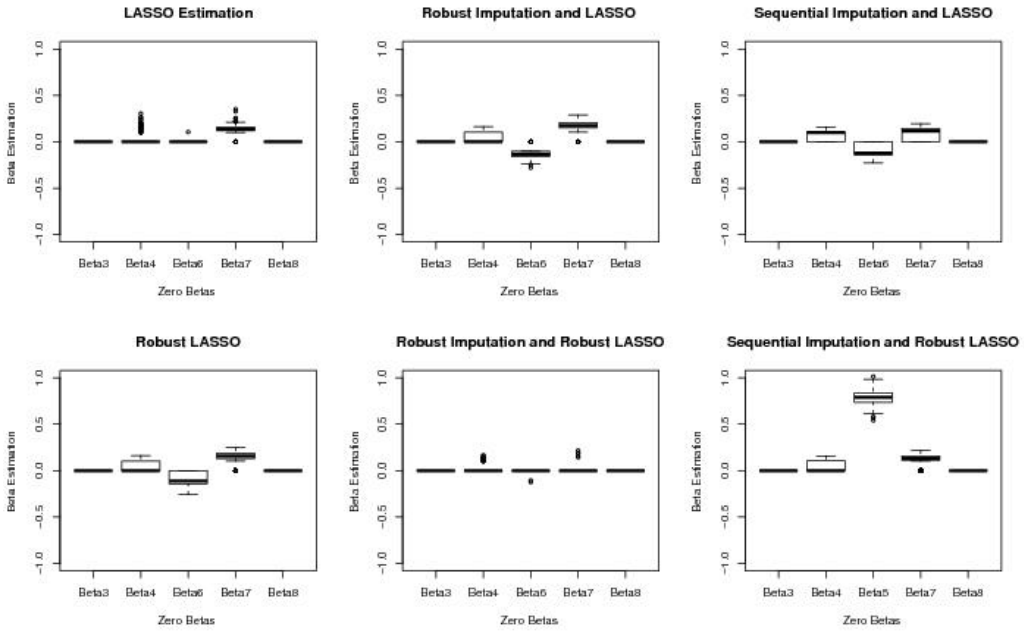

Figure 4. Results for prostate cancer data

References

[1] C. Agostinelli, A. Leung, V.J. Yohai and R.H. Zamar, *Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination*, Test, **24** (3), 441-461, 2015.

[2] F. Alqallaf, S. Van Aelst, V.J. Yohai and R.H. Zamar, *Propagation of Outliers in Multivariate Data*, Ann. Statist. **37** (1), 311-331, 2009.

[3] O. Arslan, *Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression*, Comput. Statist. Data Anal. **56** (6), 1952-1965, 2012.

[4] O. Arslan, *Penalized MM regression estimation with L γ penalty: a robust version of bridge regression*, Statistics **50** (6), 1236-1260, 2016.

[5] K.V. Branden and S. Verboven, *Robust data imputation*, Comput. Biol. Chem. **33** (1), 7-13, 2009.

[6] M. Danilov, *Robust estimation of multivariate scatter in non-affine equivariant scenarios*, University of British Columbia, 2010.

[7] M. Debruyne, S. Höppner, S. Serneels and T. Verdonck, *Outlyingness: Which variables contribute most?*, Stat. Comput. **29** (4), 707-723, 2019.

[8] J. Fan, Y. Fan and E. Barut, *Adaptive robust variable selection*, Ann. Statist. **42** (1), 324-351, 2014.

[9] A. Farcomeni, *Snipping for robust k-means clustering under component-wise contamination*, Stat. Comput. **24** (6), 907-919, 2014.

[10] P.A. Ferrari, P. Annoni, A. Barbiero and G. Manzi, *An imputation method for categorical variables with application to nonlinear principal component analysis*, Comput. Statist. Data Anal. **55** (7), 2410-2420, 2011.

[11] A.E. Hoerl and R.W. Kennard, *Ridge regression Biased estimation for nonorthogonal problems*, Technometrics **12** (1), 55-67, 1970.

[12] A. Leung, H. Zhang and R. Zamar, *Robust regression estimation and inference in the presence of cellwise and casewise contamination*, Comput. Statist. Data Anal. **99**, 1-11, 2016.

[13] A. Leung, V. Yohai and R. Zamar, *Multivariate location and scatter matrix estimation under cellwise and casewise contamination*, Comput. Statist. Data Anal. **111**, 59-76, 2017.

[14] J. Machkour, B. Alt, M. Muma and A.M. Zoubir, *The outlier-corrected-data-adaptive Lasso: A new robust estimator for the independent contamination model*, 25th European Signal Processing Conference (EUSIPCO), IEEE, 1649-1653, 2017.

[15] R.A. Maronna, *Robust ridge regression for high-dimensional data*, Technometrics **53** (1), 44-53, 2011.

[16] R.A. Maronna, R.D. Martin, V.J. Yohai and S.B. Matías, *Robust statistics: theory and methods (with R)*, John Wiley & Sons, 2019.

[17] V. Ollerer, A. Andreas and C. Croux, *The shooting S-estimator for robust regression*, Comput. Statist. **31** (3), 829-844, 2016.

[18] J. Raymaekers and P.J. Rousseeuw, *Flagging and handling cellwise outliers by robust estimation of a covariance matrix*, arXiv preprint arXiv:1912.12446, 2019.

[19] J. Raymaekers, P.J. Rousseeuw, W. Van den Bossche and M. Hubert, *cellWise: Analyzing Data with Cellwise Outliers*, CRAN, R package version: 2.0.9, 2019.

[20] P.J. Rousseeuw and W. Van den Bossche, *Detecting deviating data cells*, Technometrics **60** (2), 135-145, 2018.

[21] P.J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*, John Wiley & Sons, 2005.

[22] N. Simon, J. Friedman, T. Hastie and R. Tibshirani, *Regularization paths for Coxs proportional hazards model via coordinate descent*, J. Stat. Softw. **39** (5), 1-13, 2011.

[23] T.A. Stamey, J.N. Kabalin, J.E. McNeal, I. Johnstone, M. Iain, F. Freiha, E.A. Redwine and N. Yang, *Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients*, J. Urol. **141** (5), 1076-1083, 1989.

[24] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **58** (1), 267-288, 1996.

[25] A. Unwin, *Multivariate outliers and the O3 Plot*, J. Comput. Graph. Statist. **28** (3), 635-643, 2019.

[26] S. Verboven, K.V. Branden and P. Goos, *Sequential imputation for missing values*, Comput. Biol. Chem. **33** (5-6), 320-327, 2007.

[27] H. Xu, C. Caramanis and S. Mannor, *Robust regression and LASSO*, Adv Neural Inf Process Syst, 1801-1808, 2009.

[28] C. Yi and J. Huang, *Semismooth newton coordinate descent algorithm for elastic-net penalized huber loss regression and quantile regression*, J. Comput. Graph. Statist. **26** (3), 547-557, 2017.

[29] J.V. Yohai, *High breakdown-point and high efficiency robust estimates for regression*, Ann. Statist. **15** (2), 642-656, 1987.

[30] L. Zeng and J. Xie, *Regularization and variable selection for data with interdependent structures*, 2008.

[31] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **67** (2), 301-320, 2005.