Least Correntropic Loss Regression

Mujahid N. Syed(✉)

Department of Systems Engineering,
King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
smujahid@kfupm.edu.sa, snumujahid@gmail.com

Abstract. Robust linear regression is one of the well known areas in data analysis, and various methods to solve the robust regression problems are available in the literature. However, one of the key issues in these methods is the adaptability of the scale/tuning parameter to the data demographics. In this work, a correntropic loss based linear regression model is proposed. An approximation and simplification of the model reduces the model to the well known class of weighted linear regression models. Iterative solution methodology is proposed to solve the proposed formulation. Performance of the proposed approach is evaluated on simulated data. Results of the experiments highlight the usability and importance of the proposed approach.

Keywords: Robust linear regression · Correntropic loss · Weighted least square errors

1. Introduction

Least square error minimization is one of the earliest and commonly known form of linear regression. The method was coined in early 1800's by the individual seminal works of Legendre and Gauss. The survival of linear regression over the past 2 centuries can be attributed to its simplicity and applicability in multitude of pragmatic applications. The literal meaning of word 'regression' is 'return to a formal state'. The linear regression problem can be described as follows. Given $\varDelta$ independent and identical (iid) records $(\mathbf{x}_r, y_r)$, for $r = 1, \ldots, \varDelta$ collected from a system, where $\mathbf{x}_r \in \mathbb{R}^{1\times D}$ corresponds to the system's input parameters or regressors, and $y_r \in \mathbb{R}$ corresponds to the system's output or response for $r = 1, \ldots, \varDelta$; find $\boldsymbol{\beta}$ such that the following relation holds:

$$y_r = \mathbf{x}_r^T \boldsymbol{\beta}_\bullet + \beta_0 + \varepsilon_r \qquad \forall o = 1, \ldots, \varDelta, \tag{1}$$

where $\boldsymbol{\beta} = [\beta_0, \boldsymbol{\beta}_\bullet]$ is unknown $(D+1)\times 1$ vector, and ε_r's are iid errors that are independent of $\mathbf{x}_r$ with $E(\varepsilon_r|\mathbf{x}_r) = 0$. Equation (1) can be written in the compact form as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

I. S. Kotsireas and P. M. Pardalos (Eds.): LION 14 2020, LNCS 12096, pp. 402–413, 2020.
https://doi.org/10.1007/978-3-030-53552-0_36

where $\mathbf{X} \in \mathbb{R}^{\Delta \times (D+1)}$ and the r^{th} row of $\mathbf{X}$ is defined as $[1, \mathbf{x}_r]$, and $\mathbf{y}$, $\boldsymbol{\varepsilon} \in \mathbb{R}^{\Delta}$ are vectors containing responses and errors respectively. The mathematical formulation of least square linear regression, a.k.a, Ordinary Least Square (OLS) regression can be modeled as follows:

$$\begin{aligned} &minimize: \\ &\qquad || \ \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \ ||_2, \end{aligned} \tag{3}$$

where $|| \ ||_2$ is the second norm or the quadratic norm. Although Formulation (3) is nonlinear, it is a convex optimization problem. Furthermore, for a reasonable size of data and computing power, the formulation has a closed form solution. By reasonable size and computation power, we mean that the computer system is capable to inverse or handle $\mathbf{X}^T\mathbf{X}$. The closed form solution for Formulation (3) is an immediate result of the optimality conditions for unconstrained non-linear programs [1]. Since Formulation (3) is convex, the necessary and sufficient conditions for $\widehat{\boldsymbol{\beta}}$ to be optimal is, $\boldsymbol{\nabla} f(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$, where $f(\widehat{\boldsymbol{\beta}}) = || \ \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \ ||_2$. Upon further simplification, the optimality conditions can be stated as:

$$\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{0} \qquad or \qquad \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{y} \tag{4}$$

If $\mathbf{X}^T\mathbf{X}$ is non-singular, then the solution of Eq. (4) can be written as:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{5}$$

With the development of numerical methods and computing power, $\mathbf{X}^T\mathbf{X}$ of $10,000 \times 10,000$ can be easily inverted in a single go (See [14]). Furthermore, there are many iterative methods to solve Eq. (4), which can extend the usage of OLS to big data. For example, in [5], OLS estimates of 10^{11} regressors are estimated.

2. Relevant Work

One of the critical drawbacks of OLS is its sensitivity to outliers (data points that do not fit in with the majority of the data points). Even a single outlier can have huge impact on the OLS estimate, $\widehat{\boldsymbol{\beta}}$. For example, see Figs. 1 & 2. In Fig. 1, the data is free from outliers. Whereas, in Fig. 2 10% of the observations are replaced with outliers.

To overcome the above limitation for OLS, many approaches have been developed by data scientists (see [7,10,13,20,22,24,29] and the reference there in). Indeed robust regression approaches have been well studied in many research areas originating from various disciplines over the past five decades. The robust approaches typically vary in degree of robustness, type of robustness, and computational complexity. It is out of scope of this work to review or list all the robust regression approaches. Interested readers are directed to see [16,31] in addition to the above references. From the literature, two major ideas for robust linear

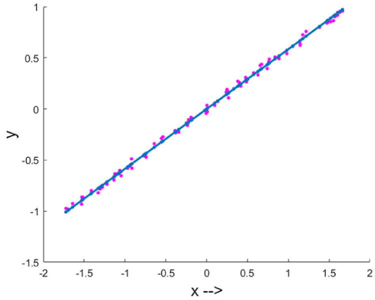

Fig. 1. OLS estimates without outliers

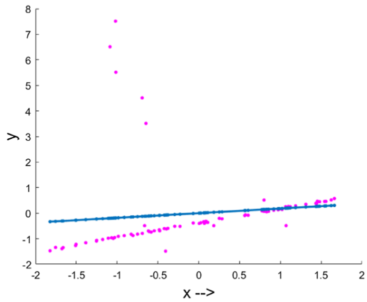

Fig. 2. OLS estimates with 10% outliers

regression can be grouped as: robust approach methods, and robust statistic methods.

Robust Approach: In robust approach methods, the key idea is to use the current OLS method with sampling mechanisms. For example, RANdom SAmple Consensus (RANSAC) is one of the robust approaches that withstood the test of time. In 1981, Fischler and Bolles proposed a generic framework called RANSAC that handles outliers in parameter estimation [4]. Usage of OLS with RANSAC strategy has since then became a popular approach to handle outliers in linear regression. The wide applicability of RANSAC can be attributed to its simple and generic characteristics. Many extensions of RANSAC are also available in the literature [17].

Robust Statistic: In robust statistic methods, the key idea is to use replace the squared error measure with a measures that is insensitive to the outliers. Among myriad of robust methods, some of the well known robust statistic methods used in robust linear regression are: Huber's M-estimates [9,10], MM-estimates [13,30], Generalized M-estimates [2,8], R-estimates [11,15], S-estimates [19], GS-estimates [3,18], LMS-estimate [24], LTS estimates [21], REWLSE estimates [6], and regularized estimates [12,23].

In this work, an adaptive weighted linear regression method that is robust to outliers is proposed. The proposed method uses a robust measure called correntropic loss. Although, weighted methods are available in the literature, the adaptive nature of the weights proposed in this paper improves the quality of the estimates. The rest of the paper is organized as follows: Sect. 3 presents the proposed model, followed by the proposed solution methodology. A numerical study involving simulated data is illustrated in Sect. 4. Some discussion and concluding remarks are depicted in Sect. 5.

3. Methodology

In this section, a mathematical model that is robust and/or insensitive to outliers is presented. An iterative solution methodology for the proposed formulation is developed in the latter part of this section.

3.1 Proposed Model

The following model is proposed for linear regression:

$$\begin{aligned} &minimize: \\ &\qquad \sum_{r=1}^{\Delta}\left(1 - e^{-\frac{(y_r - \mathbf{x}_r\beta)^2}{2\sigma^2}}\right), \end{aligned} \tag{6}$$

where $\sigma > 0$ is a scale parameter. The exponential objective function (also defined as correntropic loss) in Formulation (6) appears in many data analysis works including [25–28]. From the theory of optimality conditions for unconstrained non-linear programs [1], a local optimal solution to the above formulation should satisfy the following necessary condition:

$$\sum_{r=1}^{\Delta} e^{-\frac{(y_r - \mathbf{x}_r\beta)^2}{2\sigma^2}} \left(\frac{(y_r - \mathbf{x}_r\boldsymbol{\beta})(x_{rf})}{\sigma^2}\right) = 0 \qquad \forall\, f \in D \tag{7}$$

Let $\mathbf{w}(\boldsymbol{\beta}) : \mathbb{R}^D \mapsto \mathbb{R}^{\Delta}$ be defined as $w_r(\boldsymbol{\beta}) = e^{-\frac{(y_r - \mathbf{x}_r\beta)^2}{2\sigma^2}}$ for $r = 1, \ldots, \Delta$. The above conditions can be recast as:

$$\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \;=\; \mathbf{0} \tag{8}$$

where $\mathbf{W}(\boldsymbol{\beta})$ is a diagonal matrix containing $\mathbf{w}(\boldsymbol{\beta})$ as its diagonal.

3.2 Proposed Solution Approach

In order to find $\boldsymbol{\beta}$ that satisfies the above necessary conditions, an iterative procedure is proposed. The update rule for the procedure is described as follows:

$$\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta}_{old})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{new}) \;=\; \mathbf{0} \tag{9}$$

Since $\mathbf{W}(\boldsymbol{\beta})$ is a diagonal matrix with positive elements, and if $\mathbf{X}^T\mathbf{X}$ is non-singular, then we have the following closed form solution:

$$\boldsymbol{\beta}_{new} = (\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta}_{old})\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta}_{old})\mathbf{y}. \tag{10}$$

Notice that the above update rule is similar to the update rule obtained when solving the following weighted linear regression problem:

$$\begin{aligned} &minimize: \\ &\qquad \sum_{r=1}^{\Delta} \omega_r (y_r - \mathbf{x}_r\boldsymbol{\beta})^2 \end{aligned} \tag{11}$$

where ω_r is the non-negative weights assigned to the r^{th} record. The simplification depicted in Eq. (10) drastically reduces the complexity involved in obtaining the solution to Formulation (6). However, the key issue lies in obtaining $\boldsymbol{\beta}_{old}$ for any value of σ, such that the Hessian of the objective function in Formulation (6) is Positive Semi Definite (PSD) at $\boldsymbol{\beta}_{old}$. Obtaining such $\boldsymbol{\beta}_{old}$ ensures that the necessary conditions stated in Eq. (8) are also sufficient for local optimality. When σ is very large, the Hessian is PSD everywhere. Thus, the main difficulty is to obtain such $\boldsymbol{\beta}_{old}$ for smaller values of σ. The iterative procedure depicted in Algorithm-1 obtains such $\boldsymbol{\beta}_{old}$ at each iteration.

Algorithm 1: Proposed Algorithm

Input : $\mathbf{X} \in \mathbb{R}^{\Delta\times(D+1)}, \mathbf{y} \in \mathbb{R}^{\Delta}, \boldsymbol{\beta}^{OLS} \in \mathbb{R}^{D+1}$ and α
Output: $\boldsymbol{\beta} \in \mathbb{R}^{D+1}$

1 Set $\sigma \leftarrow \sqrt{\max_{1\leq r\leq\Delta}\{(y_r - \mathbf{x}_r\boldsymbol{\beta}^{OLS})^2\}}$;
2 Set $\boldsymbol{\beta}_{new} \leftarrow \boldsymbol{\beta}^{OLS}$;
3 **while** *termination criteria is False* **do**
4 Set $\boldsymbol{\beta}_{old} \leftarrow \boldsymbol{\beta}_{new}$;
5 Construct $\mathbf{W}(\boldsymbol{\beta}_{old})$;
6 Get $\boldsymbol{\beta}_{new} \leftarrow (\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta}_{old})\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta}_{old})\mathbf{y}$;
7 Update $\sigma \leftarrow \alpha\sigma$;
8 **end**
9 Set $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}_{new}$

In Algorithm-1, $\boldsymbol{\beta}^{OLS}$ are the OLS estimates for the given data, $0 < \alpha < 1$ is a tuning parameter. The termination criterion used in the current work is $\sigma < \epsilon$ for some prespecified threshold ϵ. Upon termination at a low value of σ, the algorithm gives a local minimum of Formulation (6). In addition to that, the proposed approach to solve Formulation (6) involves solving Formulation (11) at each iteration (Line-6 in Algorithm-1). Thus, the proposed algorithm may give global minimum of Formulation (6) when $\alpha \longrightarrow 1$.

4. Experimentation

In order to compare the performance of the proposed method, following existing weighted linear regression methods from the literature of robust linear regression are considered (see Table 1). The first column in Table 1 indicates the commonly known name of the method. The second column describes the mechanism to generate w_r's for each of the methods based on the value of error/residual e_r. Some of these methods are require a tuning parameter, and the third column displays the suggested parameter value.

Table 1. Some robust linear regression methods

Method	Description	Constant
'andrews'	$\omega_r = \begin{cases} sin(e_r)/e_r & \lvert e_r \rvert < pi, \\ 0 & o/w \end{cases}$	1.339
'bisquare'	$\omega_r = \begin{cases} (1 - e_r^2)^2 & \lvert e_r \rvert < 1, \\ 0 & o/w \end{cases}$	4.685
'cauchy'	$\omega_r = 1/(1 + e_0^2)$	2.385
'huber'	$\omega_r = 1/\max(1, \lvert e_r \rvert)$	1.345
'logistic'	$\omega_r = tanh(e_r)/r_r$	1.205
'talwar'	$\omega_r = \begin{cases} 1 & \lvert e_r \rvert < 1, \\ 0 & o/w \end{cases}$	2.795
'welsch'	$\omega_r = exp(-(e_r^2))$	2.985

Following sequence of experiments are conducted in this section. At first, simulated data containing no outliers is used for checking the validity of the proposed methodology. Next, simulated data containing outliers is used for demonstrating the capability of the proposed methodology to handle outliers. Finally, simulated data that contains outliers in a linear structure is considered.

4.1 Experiment-1

Setup: In this experiment, the data is simulated using the following equation:

$$y = \beta_1 x + \beta_0 + 0.1\varepsilon, \tag{12}$$

where the values of x are uniformly selected from 0 to 1, and ε is a Gaussian noise with zero mean and unit variance. This experiment consists of 30 scenarios, where each scenario contains 100 trials. At the beginning of each scenario, β_0 and β_1 are uniformly randomly selected from the following interval $[1, 10]$. The values for β_0 and β_1 will not be changed during the trials for a given scenario. However, these values will be updated at the start of every scenario. **Validation:** For each scenario, following measure is used for reporting the quality of the estimates:

$$\mu = \frac{1}{|T|} \sum_{t \in T} ||\boldsymbol{\beta}^{act} - \boldsymbol{\beta}_t||_2, \tag{13}$$

where T represents the set of all trials, $\boldsymbol{\beta}^{act}$ are the actual coefficients used for data generation in the scenario, and $\boldsymbol{\beta}_t$ represents the estimated coefficients. A two sided hypothesis sign test is utilized for concluding any differences between existing and the proposed method estimates. The null hypothesis (H_0) is that the mean of μ values of the existing method and proposed method are same. The alternate hypothesis (H_a) is that the proposed method's mean μ value is lower than the existing method's mean μ value. **Results:** The results of this experiment are displayed in Table 2. Column labeled Avg μ(Std μ) represents the average(standard deviation) μ value for the method over the 30 scenarios. Column labeled Avg Time(Std Time) represents the average(standard deviation) time in seconds used by the method per trial per scenario. Column labeled Ha contains either 0 or 1. A value of 1 in Ha implies that the sign test supports/favors the alternate hypothesis at 5% significance level. Similarly, a value of 0 in Ha indicates that, at 5% significance level, the test fails to reject the null hypothesis. For Logistic method, based on the sign test, at the 5% significance level, the test favors the alternate hypothesis.

4.2 Experiment-2

Setup: In this experiment, the data is simulated similar to Experiment-1. However, 10% of the response values are modified by updating the response values to $max\{y\} + 10$. **Validation:** The measure, null and alternate hypotheses are similar to Experiment-1. **Results:** The results of this experiment are displayed in Table 3. The columns have similar meaning as described in Experiment-1. For OLS, Logistic and Huber methods, based on the sign test, at the 5% significance level, the test favors the alternate hypothesis.

4.3 Experiment-3

Setup: In this experiment, the data is simulated similar to Experiment-2. However, the number of regressors in this experiment are 10, i.e., $\boldsymbol{\beta} \in \mathbb{R}^{11}$. In addition to that, three cases are considered in this experiment. In Case-1 10% of the data are outliers, in Case-2 20% of the data are outliers, and in Case-3, 30% of the data are outliers. **Validation:** The measure, null and alternate hypotheses are similar to Experiment-1. **Results:** The results of this experiment are displayed in Table 4. The columns have similar meaning as described in Experiment-1. From the results, it can be concluded that as the percentage of outliers increase, the number of existing methods favoring the alternate hypothesis increase (at 5% significance level).

Table 2. Experiment-1 results

Method	Avg μ	Std μ	Avg Time	Std Time	H_a
'OLS'	0.051	0.0015	0	0	0
'proposed'	0.0511	0.0018	0.0001	0.0002	
'andrews'	0.0512	0.0017	0.0008	0.0029	0
'bisquare'	0.0512	0.0017	0.0006	0.0005	0
'cauchy'	0.0512	0.0017	0.0007	0.0002	0
'huber'	0.0511	0.0016	0.0005	0.0002	0
'logistic'	0.0513	0.0018	0.0008	0.0008	1
'talwar'	0.051	0.0015	0.0003	0.0001	0
'welsch'	0.0512	0.0017	0.0006	0.0001	0

Table 3. Experiment-2 results

Method	Avg μ	Std μ	Avg Time	Std Time	H_a
'OLS'	1.9897	0.3105	0	0	1
'proposed'	0.0515	0.0007	0.0069	0.0005	
'andrews'	0.0512	0.0006	0.0006	0.0011	0
'bisquare'	0.0512	0.0006	0.0006	0.0002	0
'cauchy'	0.0513	0.0006	0.0007	0.0001	0
'huber'	0.0575	0.0007	0.0008	0.0001	1
'logistic'	0.0588	0.0008	0.001	0.0003	1
'talwar'	0.051	0.0006	0.0003	0	0
'welsch'	0.0512	0.0006	0.0007	0.0001	0

4.4 Experiment-4

Setup: In this experiment, the data is simulated similar to Experiment-3. However, the outliers form a linear structure. Thus, the methods has to decide the right linear structure based on the majority of the points. **Validation:** The measure, null and alternate hypotheses are similar to Experiment-1. **Results:** The results of this experiment are displayed in Table 5. The columns have similar meaning as described in Experiment-1. From the results, it can be concluded that as the percentage of outliers increase, the number of existing methods favoring the alternate hypothesis increase (at 5% significance level).

Table 4. Experiment-3 results

Case-1: 10% Outliers					
Method	Avg μ	Std μ	Avg Time	Std Time	H_a
'OLS'	10.055	0.9696	0	0.0002	1
'proposed'	0.0655	0.0016	0.0139	0.0036	
'andrews'	0.0634	0.0016	0.0014	0.0014	0
'bisquare'	0.0634	0.0016	0.0012	0.0004	0
'cauchy'	0.064	0.0016	0.0015	0.0003	0
'huber'	0.074	0.0015	0.0018	0.0004	1
'logistic'	0.077	0.0015	0.0022	0.0006	1
'talwar'	0.0623	0.0015	0.0006	0.0001	0
'welsch'	0.0636	0.0016	0.0013	0.0002	0
Case-2: 20% Outliers					
Method	Avg μ	Std μ	Avg Time	Std Time	H_a
'OLS'	15.253	1.5065	0.0001	0.0006	1
'proposed'	0.0671	0.0022	0.0139	0.0033	
'andrews'	0.0645	0.002	0.0015	0.0013	0
'bisquare'	0.0645	0.002	0.0013	0.0004	0
'cauchy'	0.0651	0.0021	0.0017	0.0003	0
'huber'	0.1072	0.0192	0.0032	0.0011	1
'logistic'	0.1235	0.0294	0.004	0.0014	1
'talwar'	3.1556	0.8304	0.0008	0.0002	1
'welsch'	0.0647	0.002	0.0014	0.0002	0
Case-3: 30% Outliers					
Method	Avg μ	Std μ	Avg Time	Std Time	H_a
'OLS'	29.669	20.613	0	0	1
'proposed'	0.0688	0.0017	0.0147	0.0033	
'andrews'	0.9489	0.7712	0.002	0.0014	1
'bisquare'	0.9488	0.7711	0.0017	0.0008	1
'cauchy'	1.5867	1.0927	0.0025	0.0011	1
'huber'	7.4753	5.3805	0.0055	0.0017	1
'logistic'	8.3627	6.1255	0.0086	0.005	1
'talwar'	16.225	12.591	0.0009	0.0004	1
'welsch'	0.9102	0.7206	0.0019	0.001	1

Table 5. Experiment-4 results

Method	Avg μ	Std μ	Avg Time	Std Time	H_a
Case-1: 10% Outliers					
Method	Avg μ	Std μ	Avg Time	Std Time	H_a
'OLS'	14.699	8.5397	0	0	1
'proposed'	0.0652	0.0015	0.0154	0.004	
'andrews'	0.0631	0.0015	0.0015	0.0016	0
'bisquare'	0.0631	0.0015	0.0013	0.0003	0
'cauchy'	0.0636	0.0015	0.0016	0.0003	0
'huber'	0.0658	0.002	0.0018	0.0004	0
'logistic'	0.0681	0.0021	0.0022	0.0005	1
'talwar'	0.062	0.0015	0.0007	0.0001	0
'welsch'	0.0632	0.0015	0.0014	0.0002	0
Case-2: 20% Outliers					
Method	Avg μ	Std μ	Avg Time	Std Time	H_a
'OLS'	22.265	16.762	0	0	1
'proposed'	0.067	0.0022	0.0154	0.004	
'andrews'	0.0645	0.0021	0.0016	0.0013	0
'bisquare'	0.0645	0.0021	0.0014	0.0004	0
'cauchy'	0.0649	0.0021	0.0017	0.0003	0
'huber'	0.0887	0.017	0.003	0.001	1
'logistic'	0.1028	0.0306	0.0037	0.0013	1
'talwar'	0.4209	0.4141	0.0008	0.0002	1
'welsch'	0.0647	0.0021	0.0015	0.0003	0
Case-3: 30% Outliers					
Method	Avg μ	Std μ	Avg Time	Std Time	H_a
'OLS'	30.865	18.579	0	0.0001	1
'proposed'	0.0682	0.0018	0.0153	0.0037	
'andrews'	1.0296	0.9179	0.002	0.0014	1
'bisquare'	1.0326	0.9182	0.0018	0.001	1
'cauchy'	1.8393	1.4973	0.0025	0.0011	1
'huber'	7.7281	5.1359	0.0055	0.0018	1
'logistic'	8.5068	5.6306	0.0096	0.0042	1
'talwar'	16.947	11.303	0.0009	0.0004	1
'welsch'	0.9615	0.7986	0.0019	0.0009	1

5. Conclusion

In this work, a formulation for robust linear regression related to the correntropic loss minimization is presented. The proposed formulation can be approximated as weighted OLS minimization problem. An iterative solution method for the weighted OLS problem, where the weights are adaptive, has been proposed and implemented. Numerical experiments on the simulated data are presented, that compares the proposed method head-to-head with some of the existing methods from the literature. Based on the numerical study, it can be highlighted that the adaptive nature of weights (or the scale parameter) is the key element in handling outliers.

References

1. Bazaraa, M.S., Sherali, H.D., Shetty, C.M.: Nonlinear Programming: Theory and Algorithms. Wiley, Hoboken (2013)
2. Coakley, C.W., Hettmansperger, T.P.: A bounded influence, high breakdown, efficient regression estimator. J. Am. Stat. Assoc. **88**(423), 872–880 (1993)
3. Croux, C., Rousseeuw, P.J., Hössjer, O.: Generalized s-estimators. J. Am. Stat. Assoc. **89**(428), 1271–1281 (1994)
4. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
5. Frank, A., Fabregat-Traver, D., Bientinesi, P.: Large-scale linear regression: development of high-performance routines. Appl. Math. Comput. **275**, 411–421 (2016)
6. Gervini, D., Yohai, V.J., et al.: A class of robust and fully efficient regression estimators. Ann. Stat. **30**(2), 583–616 (2002)
7. Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A.: Robust Statistics: The Approach Based on Influence Functions, vol. 196. Wiley, Hoboken (2011)
8. Handschin, E., Schweppe, F.C., Kohlas, J., Fiechter, A.: Bad data analysis for power system state estimation. IEEE Trans. Power Apparatus Syst. **94**(2), 329–337 (1975)
9. Huber, P.J.: Robust estimation of a location parameter. In: Kotz, S., Johnson, N.L. (eds.) Breakthroughs in Statistics, pp. 492–518. Springer, Heidelberg (1992). https://doi.org/10.1007/978-1-4612-4380-9_35
10. Huber, P.J.: Robust Statistics. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-04898-2_594
11. Jaeckel, L.A.: Estimating regression coefficients by minimizing the dispersion of the residuals. Ann. Math. Stat. 1449–1458 (1972)
12. Lee, Y., MacEachern, S.N., Jung, Y., et al.: Regularization of case-specific parameters for robustness and efficiency. Stat. Sci. **27**(3), 350–372 (2012)
13. Maronna, R.A., Martin, R.D., Yohai, V.J., Salibián-Barrera, M.: Robust Statistics: Theory and Methods (with R). Wiley, Hoboken (2019)
14. MatLab: Largest Matrix Size. https://www.mathworks.com/matlabcentral/. Accessed Sept 2019
15. Naranjo, J.D., Hettmansperger, T.: Bounded influence rank regression. J. Roy. Stat. Soc.: Ser. B (Methodol.) **56**(1), 209–220 (1994)
16. Papageorgiou, G., Bouboulis, P., Theodoridis, S.: Robust linear regression analysis—a greedy approach. IEEE Trans. Sig. Process. **63**(15), 3872–3887 (2015)
17. Raguram, R., Chum, O., Pollefeys, M., Matas, J., Frahm, J.M.: USAC: a universal framework for random sample consensus. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 2022–2038 (2012)
18. Roelant, E., Van Aelst, S., Croux, C.: Multivariate generalized s-estimators. J. Multivariate Anal. **100**(5), 876–887 (2009)

19. Rousseeuw, P., Yohai, V.: Robust regression by means of S-estimators. In: Franke, J., Härdle, W., Martin, D. (eds.) Robust and Nonlinear Time Series Analysis, vol. 26, pp. 256–272. Springer, Heidelberg (1984). https://doi.org/10.1007/978-1-4615-7821-5_15
20. Rousseeuw, P.J.: Least median of squares regression. J. Am. Stat. Assoc. **79**(388), 871–880 (1984)
21. Rousseeuw, P.J.: Multivariate estimation with high breakdown point. Math. Stat. Appl. **8**(283–297), 37 (1985)
22. Rousseeuw, P.J., Hubert, M.: Robust statistics for outlier detection. Wiley Interdisc. Rev.: Data Mining Knowl. Discov. **1**(1), 73–79 (2011)
23. She, Y., Owen, A.B.: Outlier detection using nonconvex penalized regression. J. Am. Stat. Assoc. **106**(494), 626–639 (2011)
24. Siegel, A.F.: Robust regression using repeated medians. Biometrika **69**(1), 242–244 (1982)
25. Singh, A., Pokharel, R., Principe, J.: The C-loss function for pattern classification. Pattern Recogn. **47**(1), 441–453 (2014)
26. Syed, M., Pardalos, P., Principe, J.: Invexity of the minimum error entropy criterion. IEEE Sig. Process. Lett. **20**(12), 1159–1162 (2013)
27. Syed, M.N., Pardalos, P.M., Principe, J.C.: On the optimization properties of the correntropic loss function in data analysis. Optim. Lett. **8**(3), 823–839 (2013). https://doi.org/10.1007/s11590-013-0626-5
28. Syed, M.N., Principe, J.C., Pardalos, P.M.: Correntropy in data classification. In: Sorokin, A., Murphey, R., Thai, M., Pardalos, P. (eds.) Dynamics of Information Systems: Mathematical Foundations, pp. 81–117. Springer, Heidelberg (2012). https://doi.org/10.1007/978-1-4614-3906-6_5
29. Tukey, J.W.: Data analysis, computation and mathematics. Q. Appl. Math. **30**(1), 51–65 (1972)
30. Yohai, V.J., et al.: High breakdown-point and high efficiency robust estimates for regression. Ann. Stat. **15**(2), 642–656 (1987)
31. Yu, C., Yao, W.: Robust linear regression: a review and comparison. Commun. Stat.-Simul. Comput. **46**(8), 6261–6282 (2017)