*J.Natn.Sci.Foundation Sri Lanka 2022 50 (2): 395 - 404*
DOI: http://dx.doi.org/10.4038/jnsfsr.v50i2.10292

**RESEARCH ARTICLE**

## Regression analysis

# Implementation of adaptive lasso regression based on multiple Theil-Sen Estimators using differential evolution algorithm with heavy tailed errors†

**E Dünder[1*], T Zaman[2], MA Cengiz[1] and K Alakuş[1]**
[1] *Department of Statistics, Faculty of Science, Ondokuz Mayıs University, Turkey.*
[2] *Department of Statistics, Faculty of Science, Çankırı Karatekin University, Turkey.*

Submitted: 18 November 2020; Revised: 29 November 2021; Accepted: 24 December 2021

**Abstract:** The last decade has witnessed that penalized regression methods have become an alternative to classical methods. Adaptive lasso is one type of method in penalized regression and is commonly used in statistical modelling to perform variable selection. Apart from the classical lasso setting, the adaptive lasso requires the coefficient weights inside the target function. The main issue in adaptive lasso is to select the optimal weights in the model since the selected weights have serious impacts on the estimation results. However, there is no compromise for choosing the weights as a universal approach, and they should be chosen properly with the statistical assumptions. When the error terms are heavy-tailed, classical estimation (such as least squares) gives poor results in adaptive lasso because of the lacking robustness. This article deals with the selection of optimal weights in the presence of heavy-tailed errors for the adaptive lasso. To solve the distributional problem, we integrated the Theil-Sen estimation (TSE) approach into the adaptive lasso for heavy-tailed erroneous cases while choosing the weights. During the selection of the optimal tuning parameters, we employed a differential evolution algorithm (DEA) between a range of lambda values. The simulation studies and real data examples confirm the power of our combination of Theil-Sen estimators and differential evolution algorithm in the presence of heavy-tailed errors in the adaptive lasso.

**Keywords:** Adaptive lasso, heavy-tailed errors, Theil-Sen estimators, weight vector selection.

## INTRODUCTION

The classical lasso solution does not consider the marginal effects of the regression coefficients during the penalization. Ignorance of the nature of each coefficient can cause poor results. To overcome this problem, adaptive lasso is proposed by Zou (2006). The researchers published many articles to improve the adaptive lasso. Lu *et al.* (2012) studied the robustness of adaptive lasso under misspecification. Huang *et al.* (2008) investigated the initial estimators for adaptive lasso and made a comparison with classical lasso. Chen and Chan (2011) performed subset selection in subset autoregressive moving-average using adaptive lasso for time series analysis. Ren and Zhang (2013) employed model selection in vector autoregressive models via adaptive lasso. Yoon *et al.* (2017) presented a new algorithm based on adaptive lasso with ARMA-GARCH errors in linear regression analysis. Foster *et al.* (2013) used adaptive lasso for monotone single-index models to estimate the index parameter. Hui *et al.* (2015) proposed a new criterion for the selection of the optimal tuning parameter in adaptive lasso and presented the efficiency of this alternative criterion. Zhang and Xiang (2015) studied the oracle properties of adaptive group lasso in high dimensional cases.

* Corresponding author (emre.dunder@omu.edu.tr; https://orcid.org/0000-0002-6640-5284)
† This paper is an extended version of the study presented in 'International Conference on Information Complexity and Statistical Modelling in High Dimensions with Applications IC-SMHD-2016', 18–21 May, 2016

The main issue in adaptive lasso is the choosing optimal weights. Generally ordinary least squares (OLS) and ridge estimators are used for the weights. Alternatively, Qian and Yang (2013) proposed a weight selection method based on standard errors of regression coefficients. Fan *et al.* (2014) suggested using a robust two-step method for the estimation of the weight vector.

In previous studies, the adaptive lasso estimators are adopted using different loss functions such as Huber-M (Lacroix & Zwald, 2011), least absolute deviation (Wang *et al.*, 2007) etc. Machkour *et al.* (2020) developed a robust estimator for adaptive lasso for contaminated data. Li and Wang (2020) implemented a least absolute deviation (LAD) based adaptive lasso approach for robust estimation of change point models.

These types of studies only focused on changing the loss function, but also the adaptive beta coefficients can be changed to improve the performance of the models when the dataset contains some problems like heavy-tailed error terms. The classical regression solutions fail to produce useful results when the error terms are heavy-tailed and robust estimators can be used as an alternative solution for adaptive lasso. Instead of changing the loss function, we indicate that the usage of robust estimators in adaptive lasso for weight selection enables to fix the results in the presence of heavy-tailed errors.

It is troublesome to decide which weights should be used in adaptive lasso when heavy-tailed errors emerge. Heavy-tailed errors occur frequently in statistical models and cause problematic cases (Barros *et al.*, 2008; Zhou & Wu, 2011; Honda, 2013). Robust techniques are widely used to eliminate the drawbacks of heavy-tailed errors.

This study proposes to use TSE as the weights in adaptive lasso when heavy-tailed errors exist. The robustness property of the TSE brings promise for successful accomplishment. This study also utilizes the DEA to select tuning parameters in adaptive lasso models, which is one of the most used heuristic optimization techniques.

## MATERIALS AND METHODS

### Adaptive lasso

Adaptive lasso is a type of regularization method which selects and estimates the regression coefficients simultaneously by considering a weight vector (Zou, 2006). Instead of classical lasso solutions, adaptive lasso assigns weights to each regression coefficient, so it considers the marginal effects of predictors. Adaptive lasso estimators β* are obtained such that

$$\beta^* = \arg\min \|y - X\beta\|^2 + \lambda \sum_{j=1}^{p} w_j \Big| \quad ...(1)$$

where '*w*' represent the weight vector for each regression coefficient, and λ shows the tuning parameter. Equation (1) can be solved efficiently as a convex optimization problem. The solution highly depends on the selected weight vector.

As an alternative to existing *w* estimates for the weights, we will use $\beta^{TSE} = (\beta_1^{TSE}, \beta_2^{TSE}, ... \beta_p^{TSE})$ which represents the regression coefficients obtained with TSE. In the first phase, we obtain the $\beta^{TSE}$ estimates and put it into the weight vector. Because TSE handles the robustness issue, the adaptation of TSE improves the results of adaptive lasso models in the presence of outliers or heavy tailed errors. Also, the properties of TSE satisfy the main assumptions of the adaptive lasso weights, so TSE becomes an attractive approach.

### Theil Sen estimators

The method, put forward by Theil in 1950, is one of the most commonly used slope finding methods (Theil, 1950). Theil's method, which is used to find the slope of a vector, is based on the calculation of the median of slope values calculated from an observation pair, $(x_i, y_i)$ and $(x_j, y_j)$ (Hussain & Sprent, 1983). The $(x_1, y_1), \dots, (x_n, y_n)$ we have, consist of n observation pairs. The $x_i$ values are known, different, independent of each other, and sorted as $x_1 < x_2 < ... < x_n$. In the regression model, the variances of $\varepsilon_i{}'$s consist of $\sigma_\varepsilon^2$, and random errors resulted from a symmetric continuous distribution, whose median is zero and originates from the same distribution (Rao & Gore, 1982). In Theil's method in simple linear regression, $\beta_0$ and $\beta_1$ should be estimated in such a way that median of the error term must be zero (Maritz, 1979). Estimation of $\beta_1$, $\hat{\beta}_1$ while $i < j$ and $(x_i \neq x_j)$, is a weight median of all $N = \binom{n}{2}$ curve estimations of $S_{ij} = \frac{y_j - y_i}{x_j - x_i}$ (Wang & Yu, 2005).

In the literature, Theil (1950) suggested the median of common paired slopes as an estimator of the slope parameter in a simple linear regression model. Later, Sen (1968) expanded on this estimation to find the ties. Theil-Sen estimation is, with a 29.3% breakdown point, robust and has a limited effect function and high asymptotic efficiency. For this reason, it is competing well with other slope estimators (Sen, 1968; Dietz, 1989;

Wilcox, 1998). Wang (2005) explored the behaviour of TSE when the covariate is random. Also, Sen (1968) obtained consistence and asymptotic distribution of TSE. Also, when error distribution is discontinuous at a certain point, TSE is found to be super-efficient. Even though most of its good characteristics are interpreted clearly, many statisticians tried to expand on it (Oja & Niinima, 1984; Zhou & Serfing, 2008). Since TSE is formulated only for a simple linear model, it is underdeveloped and rarely used. While for TSE to be expanded to a multiple linear model is obvious and attractive, this is technically hard and slows down the generalization and exploration processes. Dang *et al.* (2008) suggested several different methods of using the multivariate median in expanding the TSE of a parameter in a simple linear model to a multivariate linear model. Multivariate medians generalize univariate medians of the variables. For the essential concept, see Small (1990). According to the study of Dang *et al.* (2008), it is applicable for any multivariate median identified through depth functions.

**Multivariate Theil-Sen estimation**

Let us discuss a multivariate linear regression model where $\xi \geq 1$. Through following the above mentioned procedure, first, the estimation of $\zeta = (\alpha, \beta^T)^T$ can be found as the solution of equation (1)

$$Y_j - \alpha - X_j^T \beta = 0, i \in z_{\xi+1} = \{j_1, \ldots, j_{p+1}\} \quad ...(2)$$

Here, $z_{\xi+1}$ is the $(\xi+1)$ subset of $\{1, \ldots, n\}$. If matrix $(\xi+1) \times (\xi+1)$ is $(X_z : z \in z_{\xi+1})$, it can be invertible. This estimation is represented with $\hat{\zeta}_{z_{\xi+1}}$ to accentuate dependency to $\xi+1$ observations. Then, the natural expansion of TSE from a simple linear regression model to a multivariate regression model becomes a multivariate median as below:

$$\hat{\zeta}_n = Mmed\{\hat{\zeta}_{\mathbf{z}_{\xi+1}} : \forall\, \mathbf{z}_{\xi+1}\} \quad ...(3)$$

Here, $\hat{\zeta}_{z_{\xi+1}}$ is at the same time the estimate of least squares of $\zeta$ depending on observations $\{(X_j, Y_j) : j \in \mathbf{z}_{\xi+1}\}\ \xi+1$. At this angle, $r$ different arbitrary combinations of $\{(X_j, Y_j) : j \in z_r\}$ may be chosen which means, here, $\xi + 1 \leq r \leq n$ and it constitutes the least squares estimation of $\hat{\zeta}_{z_r}$. Then, the multivariate TSE of parameter $\zeta$ will naturally be the multivariate median of all possible least squares estimates, and is explained as follows:

$$\hat{\zeta}_n = Mmed\{\hat{\zeta}_{z_r} : \forall\, \mathbf{z}_r\} \quad ...(4)$$

The least squares estimate is as below:

$$\hat{\zeta}_z = (X_{\mathbf{z}}^T X_{\mathbf{z}})^{-1} X_{\mathbf{z}}^T Y_{\mathbf{z}} \quad ...(5)$$

There are as many cases as $r$ combinations of total $n$ for estimating the regression parameters with the multivariate Theil-Sen estimator (MTSE) method (Dang *et al.*, 2008). The regression coefficients are estimated by the least square method, considering these $r$ combination values. For each calculated value of $r$ combinations, in total, there will be as many regression coefficient values as the number of $r$ combinations of $n$. Next, the spatial median of the estimated regression coefficients will be calculated simultaneously. This spatial median value obtained depending on the regression coefficients will be the multiple linear regression estimation values of the MTSE method. Here, the combination count may differ depending on independent variable count and sample size. So, a certain percentage of obtained combination counts can be taken. In the literature, in the study of Dang *et al.*, (2008), the process is done by taking one percent of this combination count, and it is used as an estimator.

Let us mention some asymptotic characteristics of TSE. Firstly, Sen (1968) showed that if $x_i,\ i = 1, 2, \ldots, n$ are known non-identical constants and $F$ is an accumulated distribution function, TSE estimation is asymptotically normal. Wang (2005) explored asymptotic characteristics of TSE estimation on the assumption of random variable $x_i$ and researched the strong consistency of TSE estimation and asymptotic distribution under the assumption that $\varepsilon$ and $X$ are independent of each other and $F$ is arbitrary (continuous or discontinuous).

TSE satisfies two main assumptions for adaptive lasso weights: consistency and asymptotic normality. The consistency and asymptotic normality properties of TSE can be found in Peng *et al.* (2008), so the proofs are omitted.

**Selection of the tuning parameters with differential evolution algorithm**

In this part, we present the working scheme of the DEA and describe the tuning parameter selection process. DEA and employing steps are described in the following parts.

**Differential evolution algorithm**

Differential evolution algorithm (DEA) is a heuristic optimization method that can give efficient results in continuous-valued problems (Mayer *et al.,* 2005; Storn & Price, 1997). DEA is especially useful where the parameters are defined in a completely arranged space (Karaboga, 2004). Another significant superiority of DEA to other intuitional methods is that it can be encoded easily (Mayer *et al.*, 2005). There have been several studies about this from the day it was developed. Some of these studies are about improvement of the algorithm (Hrstka & Kuverova, 2004; Bergey & Ragsdale, 2005; Sun *et al.*, 2005; Becerra & Coello, 2006). Fundamentally, the working scheme of DEA is based on genetic algorithm. Crossover, mutation, and selection operators are used in DEA similar to genetic algorithm. In the optimization problem, the solution sets improve with the help of operators through iterations.

Instead of applying each operator differently to all populations respectively, chromosomes are handled one by one, and a new individual in three other randomly chosen chromosomes is obtained. During this process, mutation and crossover operators are used. The fitting of the new chromosome and the existing chromosome are compared and the better one is carried over to the next population as a new individual. In DEA, after objective function, variables and constraints are determined with the following steps.

**Encoding and initial population**

The number of variables belonging to the problem determines the size of chromosomes and this is expressed with *D. NP* represents the chromosome number determined by the user. *NP* should be selected as more than three because at least three are required to produce new chromosomes, other than the existing ones. In the starting phase, the initial population is produced, which consists of the number *NP* of *D* sized chromosome (Karaboga, 2004).

**Mutation**

In DEA, except for the chromosome used in mutation, three chromosomes are selected which are different from each other. The difference value of first two selected chromosomes is calculated. Then, this difference chromosome is multiplied by the F parameter. The F parameter is selected usually between 0 and 2. As a result, the weighted difference chromosome is added to the third selected chromosome. At the end of the mutation phase, the current chromosomes are constructed, which will be used in the crossover phase.

**Crossover**

In the crossover phase, the trial chromosome, $(u_{i,G+1})$, is produced as a candidate for the new generation by using the difference chromosome obtained through mutation and the previous chromosome, $x_{i,G}$. Each gene belonging to the trial chromosome is chosen with probability *CR* from the difference chromosome and with probability 1–*CR* from the current chromosome. If the random number, which is generated between 0 and 1, is less than *CR*, the gene is selected from $n_{j,i,G+1}$ or from the current chromosome. The aim is taking the gene from the new difference chromosome with a pre-determined rate.

**Fitness function**

A new trial chromosome is obtained by using three different chromosomes with mutation and crossover operators. The criterion for determining the chromosome, which will pass to the new generation ($G = G + 1$), is the fitness values. In this stage, the one which is going to be calculated is the fitness value of $(u_{i,G+1})$. The value of the chromosome is calculated by entering $u_j$ values belonging to the target chromosome $(u_{i,G+1})$.

**Selection**

The new generation is created by evaluating the existing generation and newly produced chromosomes, with the selection operator. The existence probability of the new chromosomes is related to the fitness values. In DEA, due to the comparisons being made one-to-one, complicated selection processes are not necessary. The chromosome with better fitting relative to compared chromosomes is presented as an individual of new generation.

**Ending of algorithm**

The main aim in DEA is to obtain chromosomes which have always better fitness values, so the solutions approximate to optimum. This cycle needs to be kept up until equation $G = G_{max}$ is met. As an ending criterion of the algorithm, the difference between the best and the worst fitting value in the population can be determined, as it will be a very small number such as $10^{-6}$ (Ali & Törn, 2004). For more detail, see Keskintürk (2006).

In our study, the purpose is to find the optimal λ, the tuning parameter value in adaptive lasso. With the scheme mentioned above, DEA is employed to obtain the optimal tuning parameter. The fitness function is the Bayesian information (BIC)

$$BIC = -2LL(M_{adalaso}) + m\log(n) \quad ...(6)$$

where $LL(M_{adalaso})$ indicates the log-likelihood of the adaptive lasso model, 'n' shows sample size, and 'm' shows the number of included predictor variables (Schwarz, 1978). The usage of DEA for selecting the optimal tuning parameter for adaptive lasso is given in Figure 1.

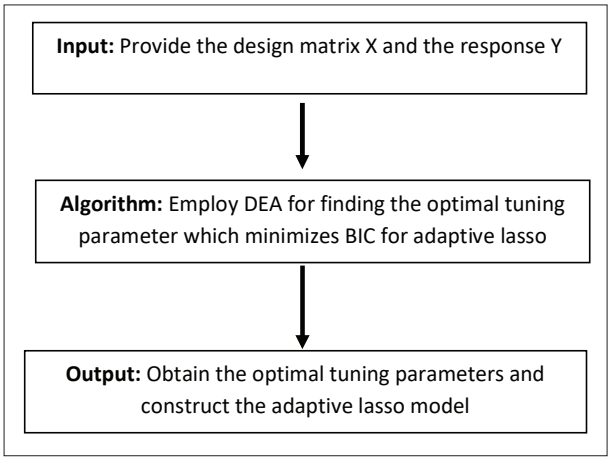


**Figure 1:** The working scheme of DEA for tuning parameter selection in adaptive lasso

## NUMERICAL STUDIES

We conducted some numerical studies with Monte Carlo simulations and two real dataset examples. We measured the correct and incorrect identifications of zero and non-zero coefficients in simulations. In the real dataset examples, we checked the prediction accuracies of the models and sparseness of the regression coefficients. The datasets are compatible for our aims which consist of heavy-tailed errors. All implementations were performed in R software (R Core Team, 2017). Adaptive lasso models were estimated with 'glmnet' and 'sealasso' packages (Friedman *et al.*, 2010; Qian, 2013).

### Monte Carlo simulations

We performed a Monte Carlo simulation study to evaluate the efficiency of the Theil-Sen approach (THEIL). To compare the performance of TSE, we employed three techniques: ordinary least squares (OLS), standard error adjusted lasso (SEA-lasso), and NSEA-lasso (Qian & Yang, 2013). The SEA-lasso and NSEA-lasso methods estimate the adaptive lasso model using the standard errors of the regression coefficients. The detailed description of these two algorithms can be found in Qian and Yang (2013).

The simulation design was constructed similar to McDonald and Galarneau (1975). We generated the datasets by taking account of heavy-tailed errors. Simulation settings include three models, and these models consist of $p$ = 5,10,15 predictors. Predictor variables were generated such that $X_k \sim N(0,1)$, $k$ = 1,2,..,p. The correlation structure was selected similar to McDonald and Galarneau (1975), as the following:

$$X_{ij} = \sqrt{1-\alpha^2}\, z_{ij} + \alpha z_{ip} \quad , i = 1,2,\ldots,n \; , j = 1,2,\ldots,p \quad ...(7)$$

where $z \sim N(0,1)$ and α is a control parameter for the degree of multicollinearity. In all cases, we fix α = 0.50 so as not to account high correlation. Four models are considered with different numbers of variables and beta coefficients. The following model structures are given:

(i) β = (10,0,0,0) (ii) β = (20,20,0,0)
(iii) β = (30,0,0,0,0,0) (iv) β = (40,40,40,0,0,0)

Simulation designs include heavy-tailed errors. The error term is settled in the models with t-distributions as follows:

$$Y = X\beta + \varepsilon \quad ...(8)$$

where $\varepsilon \sim t(u)$ with two different degrees of freedoms (df) u = 1,3. These values are appropriate for illustrating the heavy-tailed errors in regression models. The prediction accuracies are evaluated with mean average errors (MAE) as the following:

$$MAE.ERR = \sum_{i=1}^{n} \left| \frac{Y - \hat{Y}}{Y} \right| \quad ...(9)$$

where $\hat{Y}$ indicates the predicted values. The difference between the true and estimated beta coefficients are measured using

$$\text{MAE. BETA} = \sum_{i=1}^{n} \left| \frac{\beta^{true} - \beta^{est}}{\beta^{true}} \right| \quad ...(10)$$

where $\beta^{true}$ shows the true beta coefficients and $\beta^{est}$ shows the estimated values. In simulation results, 'C' represents the percentage of correctly excluded redundant variables and 'I' represents the percentage of incorrectly excluded necessary variables. Based on this representation, a simulation result is considered as successful for higher 'C' and lower 'I' values for the relevant estimator. The sample size is $n$ = 50, 100, 500 and the number of runs is fixed as 100.

We reported the simulation results in Tables 1, 2, 3, and 4 for each coefficient vector. It is seen that TSE produces lower error values when comparing with OLS, SEA and NSEA. TSE is clearly superior to each estimator in adaptive lasso. Especially, TSE provides distinctive results while keeping the relevant variables in the adaptive lasso model so 'I' values are lower for TSE. Almost in all cases, TSE can identify the true zeros more successfully. TSE does not fall into trap of identifying non-zero coefficients wrongly as much as OLS, SEA and NSEA. Clearly, the distinction becomes visible when heavy tailed errors are dominant with df = 1. All the methods can perform well for df = 3 but the error of

**Table 1:** Simulation results for the first simulation design

| n | Method | df = 1 | | | | df = 3 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | C | I | MAE.PRED | MAE.BETA | C | I | MAE.PRED | MAE.BETA |
| 50 | OLS | 96.66667 | 24 | 2.88193 | 1.38816 | 100 | 0 | 0.80793 | 0.06559 |
| | TSE | 98.33333 | 1 | 2.77776 | 0.90512 | 100 | 0 | 0.80580 | 0.06535 |
| | SEA | 97 | 24 | 2.87251 | 1.37146 | 100 | 0 | 0.80823 | 0.06578 |
| | NSEA | 97 | 24 | 2.87374 | 1.37714 | 100 | 0 | 0.80790 | 0.06579 |
| 100 | OLS | 97 | 26 | 7.94813 | 2.86781 | 100 | 0 | 0.82572 | 0.04784 |
| | TSE | 98.66667 | 2 | 7.58659 | 2.33777 | 100 | 0 | 0.82496 | 0.04753 |
| | SEA | 96.33333 | 26 | 7.96082 | 2.81579 | 100 | 0 | 0.82571 | 0.04787 |
| | NSEA | 96.66667 | 26 | 7.93016 | 2.84215 | 100 | 0 | 0.82573 | 0.04784 |
| 500 | OLS | 96.66667 | 21 | 4.18052 | 1.02275 | 100 | 0 | 0.89192 | 0.02366 |
| | TSE | 97.33333 | 0 | 3.70729 | 0.84695 | 100 | 0 | 0.89171 | 0.02393 |
| | SEA | 96.66667 | 21 | 4.18416 | 1.02505 | 100 | 0 | 0.89192 | 0.02366 |
| | NSEA | 96.33333 | 21 | 4.18663 | 1.02250 | 100 | 0 | 0.89192 | 0.02366 |

**Table 2:** Simulation results for the second simulation design

| n | Method | df = 1 | | | | df = 3 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | C | I | MAE.PRED | MAE.BETA | C | I | MAE.PRED | MAE.BETA |
| 50 | OLS | 98.50000 | 6.50000 | 4.76758 | 9.59747 | 100 | 0 | 0.33313 | 0.13234 |
| | TSE | 98.50000 | 0.50000 | 4.87711 | 10.01887 | 100 | 0 | 0.33290 | 0.12936 |
| | SEA | 98.50000 | 7 | 4.80389 | 9.61851 | 100 | 0 | 0.33302 | 0.13279 |
| | NSEA | 98.50000 | 7 | 4.81359 | 9.61466 | 100 | 0 | 0.33298 | 0.13218 |
| 100 | OLS | 97.50000 | 13.50000 | 2.10379 | 2.19424 | 100 | 0 | 0.37776 | 0.08206 |
| | TSE | 98.50000 | 1 | 2.38788 | 1.45866 | 100 | 0 | 0.37740 | 0.08410 |
| | SEA | 97.00000 | 13.50000 | 2.13894 | 2.17986 | 100 | 0 | 0.37609 | 0.08234 |
| | NSEA | 97.00000 | 13.50000 | 2.13756 | 2.18251 | 100 | 0 | 0.37611 | 0.08230 |
| 500 | OLS | 96.00000 | 12 | 5.29656 | 10.49660 | 100 | 0 | 0.43585 | 0.04034 |
| | TSE | 99.00000 | 0.50000 | 5.22106 | 9.44882 | 100 | 0 | 0.43586 | 0.04047 |
| | SEA | 96.00000 | 12.00000 | 5.29905 | 10.49529 | 100 | 0 | 0.43580 | 0.04034 |
| | NSEA | 96.00000 | 11.50000 | 5.30284 | 10.46473 | 100 | 0 | 0.43581 | 0.04034 |

**Table 3:** Simulation results for the third simulation design

| N | Method | df = 1 | | | | df = 3 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | C | I | MAE.PRED | MAE.BETA | C | I | MAE.PRED | MAE.BETA |
| 50 | OLS | 97.60000 | 2 | 8.78802 | 5.25607 | 100 | 0 | 0.19078 | 0.03952 |
| | TSE | 98.20000 | 0 | 8.36007 | 4.80112 | 100 | 0 | 0.19077 | 0.04004 |
| | SEA | 97.60000 | 2 | 8.81944 | 5.24011 | 100 | 0 | 0.19076 | 0.03952 |
| | NSEA | 97.60000 | 2 | 8.81940 | 5.24408 | 100 | 0 | 0.19075 | 0.03942 |
| 100 | OLS | 97.60000 | 5 | 1.33447 | 1.19090 | 99.80000 | 0 | 0.38087 | 0.02943 |
| | TSE | 98.20000 | 0 | 1.22619 | 0.71966 | 100 | 0 | 0.38231 | 0.02869 |
| | SEA | 97.40000 | 6 | 1.33815 | 1.22553 | 99.80000 | 0 | 0.38087 | 0.02943 |
| | NSEA | 97.60000 | 5 | 1.33151 | 1.18611 | 99.80000 | 0 | 0.38087 | 0.02937 |
| 500 | OLS | 97.60000 | 3 | 1.42711 | 1.89066 | 100 | 0 | 0.30288 | 0.01275 |
| | TSE | 98.20000 | 0 | 1.42585 | 1.28176 | 100 | 0 | 0.30287 | 0.01318 |
| | SEA | 97.60000 | 3 | 1.42659 | 1.88826 | 100 | 0 | 0.30288 | 0.01273 |
| | NSEA | 97.60000 | 3 | 1.42674 | 1.88867 | 100 | 0 | 0.30288 | 0.01273 |

**Table 4:** Simulation results for the fourth simulation design

| n | Method | df = 1 | | | | df = 3 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | C | I | MAE.PRED | MAE.BETA | C | I | MAE.PRED | MAE.BETA |
| 50 | OLS | 98.66667 | 0.01667 | 1.23140 | 1.18738 | 100 | 0 | 0.10927 | 0.12188 |
| | TSE | 99.66667 | 0 | 1.16676 | 1.03551 | 100 | 0 | 0.10945 | 0.11721 |
| | SEA | 98.66667 | 0.01667 | 1.24083 | 1.21144 | 100 | 0 | 0.10862 | 0.12225 |
| | NSEA | 98.66667 | 0.01667 | 1.23637 | 1.20414 | 100 | 0 | 0.10859 | 0.12201 |
| 100 | OLS | 99.33333 | 0.02333 | 1.84601 | 1.24895 | 100 | 0 | 0.12763 | 0.07688 |
| | TSE | 100 | 0.00333 | 1.79133 | 0.75646 | 100 | 0 | 0.12778 | 0.07566 |
| | SEA | 99.33333 | 0.02333 | 2.09537 | 1.24958 | 100 | 0 | 0.12756 | 0.07686 |
| | NSEA | 99.33333 | 0.02333 | 2.04209 | 1.24971 | 100 | 0 | 0.12756 | 0.07674 |
| 500 | OLS | 99.66667 | 0.03000 | 0.60514 | 1.52497 | 100 | 0 | 0.16789 | 0.03883 |
| | TSE | 100 | 0 | 0.58041 | 0.61665 | 100 | 0 | 0.16787 | 0.03859 |
| | SEA | 99.66667 | 0.03000 | 0.60765 | 1.52754 | 100 | 0 | 0.16797 | 0.03890 |
| | NSEA | 99.66667 | 0.03000 | 0.60741 | 1.52692 | 100 | 0 | 0.16796 | 0.03889 |

the coefficients was estimated better with TSE. From the above results, the chosen weights with TSE obviously improve the correctness of the estimated adaptive lasso model.

## Real dataset examples

In this part, we have analysed some real datasets to evaluate the predictive performances of estimators. Meanwhile we presented the sparsity of the regression coefficients. For the application we used Alcohol and Boston datasets which are available in R software, 'MASS' and 'robustbase' packages (Venables & Ripley, 2002; Maechler *et al.*, 2016). These datasets contain heavy-tailed errors, so the real datasets are conformable for the computations. The Alcohol dataset contains 44 observations and 6 independent variables. This dataset represents the physicochemical characteristics of 44 aliphatic alcohols. The purpose is to predict the solubility of the alcohols. The Boston dataset consists of 506 observations and 13 explanatory variables. This dataset is related to the housing values in the suburbs of Boston.

The aim of Boston dataset is to predict the crime rates by town.

Predictive performance is measured by cross validation technique. The datasets are divided into two parts as 'test-train.' Train sets contain 80% and test sets contain 20% of the data, respectively.

**Table 5:** Regression coefficients for Alcohol dataset

| Regressor | OLS | TSE | SEA | NSEA |
|---|---|---|---|---|
| SAG | 0 | 0 | 0 | 0 |
| V | 0 | -0.01365 | 0 | 0 |
| logPC | -1.07144 | -1.42559 | -1.23636 | 0 |
| P | 0 | 0 | 0 | -0.67092 |
| RM | -0.17295 | 0 | -0.15721 | 0 |
| Mass | 0 | 0 | 0 | 0 |

**Table 6:** Regression coefficients for Boston dataset

| Regressor | OLS | TSE | SEA | NSEA |
|---|---|---|---|---|
| Ht | 0 | 0 | 0 | 0.00948 |
| Wt | 1.09996 | 1.08407 | 0.79137 | 1.10635 |
| LBM | -1.24933 | -1.22963 | -0.92578 | -1.26896 |
| RCC | 0 | 0 | 0 | 0 |
| WCC | 0 | 0 | 0 | 0 |
| Hc | 0 | 0 | 0 | 0 |
| Hg | -0.13658 | -0.24795 | 0 | 0 |
| Ferr | 0 | 0 | 0 | 0 |
| BMI | 0 | 0.01639 | 0 | 0 |
| SSF | 0 | 0 | 0.05241 | 0 |

**Table 7:** Predictive performance results

| Method | AIS-Error | Alcohol-Error |
|---|---|---|
| OLS | 0.06430 | 0.25919 |
| TSE | 0.06337 | 0.15656 |
| SEA | 0.06922 | 0.26448 |
| NSEA | 0.06861 | 0.23996 |

The regression coefficients are shown for each weighting method in Tables 5 and 6. It is seen that there is not much difference among the methods in terms of sparseness. The predictive performance of each approach is given in Table 7. In experimental results, TSE performs better than OLS, SEA, and NSEA in terms of prediction for new observations. It should be pointed out that TSE accomplishes well for each dataset when used in adaptive lasso.

## CONCLUSION

Adaptive lasso has become a useful alternative solution in terms of penalized regression modelling. The weights that are used in adaptive lasso have a significant effect on the regression coefficient estimates. Classical OLS and ridge estimators fail to obtain accurate parameter estimates when the regression model includes heavy-tail errors. To improve the achievement of adaptive lasso, we propose to use TSE as the weight vector. Also, we benefited from the power of heuristic optimization while finding the optimal tuning parameter. DEA is used for this task.

According to simulations studies, TSE provides more accurate parameter estimates when used as the weight in adaptive lasso. Moreover, the prediction accuracy is better than the other three weight selection approaches. One of the advantages of TSE emerged in the identification of zero and non-zero coefficients. OLS, SEA, and NSEA weight selection methods choose non-zero values as zero much more than TSE. Also, TSE is more capable of determining the zero coefficients correctly. These findings exhibit the superiority of TSE within adaptive lasso in terms of variable selection. In real dataset examples, TSE weights demonstrate preferable performance for out of sample prediction. The sparsity of regression coefficients do not differ much among weight selection approaches.

Emergence of heavy tailed errors leads to trouble, and it is hard to overcome this serious problem in regression analysis. Although robust methods provide a good way to solve this problem, there is not much research for lasso type regression methods. Within the use of lasso methods in cases with heavy tailed errors, it is possible to eliminate the redundant variables. Our proposed robust weight selection scheme gives completely successful results when using TSE and DEA within adaptive lasso. Also, our approach shows that the use of robust estimators in adaptive lasso weights significantly enhances the model selection accuracy. This result can encourage researchers to adopt different estimators for adaptive weights instead of completely changing the loss functions. Our study uncovers that the changing of the loss function is not the only way of improving the robustness of the adaptive lasso model. Even though the target function includes squared loss term, the results ameliorate heavy-tailed

errors with TSE. In our view, researchers should pay attention to this result in further studies.

Moreover, theoretical aspects of TSE satisfy the conditions of weight usage in adaptive lasso regression analysis because TSE supplies the oracle properties of consistency and asymptotic normality. Consequently, we only demonstrated the applicability of TSE with adaptive lasso and concluded the success with several numerical examples. We suggest using TSE when there are many predictors for further practical studies to accomplish model selection in the presence of heavy-tailed errors.

### Conflict of Interest

All authors declared that there is no conflict of interest involved in this study.

## REFERENCES


Ali M.M. & Törn A. (2004). Population set-based global optimization algorithms: some modifications and numerical studies. *Computer and Operations Research* **31**: 1703–1725.
DOI: https://doi.org/10.1016/S0305-0548(03)00116-3

Barros M., Paula G.A. & Leiva V. (2008). A new class of survival regression models with heavy-tailed errors: robustness and diagnostics. *Lifetime Data Analysis* **14**(3): 316–332.
DOI: https://doi.org/10.1007/s10985-008-9085-1

Becerra R.L. & Coello C.A.C. (2006). Cultured differential evolution for constrained optimization. *Computer Methods in Applied Mechanics and Engineering* **195**(33–36): 4303–4322.
DOI: https://doi.org/10.1016/j.cma.2005.09.006

Bergey P.K. & Ragsdale C. (2005). Modified differential evolution: a greedy random strategy for genetic recombination. *Omega* **33**: 255–265.
DOI: https://doi.org/10.1016/j.omega.2004.04.009

Chen K. & Chan K.S. (2011). Subset ARMA selection via the adaptive Lasso. *Statistics and its Interface* **4**: 197–205.
DOI: https://doi.org/10.4310/SII.2011.v4.n2.a14

Dang X., Peng H., Wang X. & Zhang H. (2008). Theil-Sen estimators in a multiple linear regression model. Available at *https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.508.3461*

Dietz E.J. (1989). Teaching regression in a nonparametric statistics course. *The American Statistician* **43**: 35–40
DOI: https://doi.org/10.1080/00031305.1989.10475606

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

Foster J.C., Taylor J.M. & Nan B. (2013). Variable selection in monotone single-index models via the adaptive LASSO. *Statistics in Medicine* **32**(22): 3944–3954.
DOI: https://doi.org/10.1002/sim.5834

Friedman J., Hastie T. & Tibshirani R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1): 1–22.
DOI: https://doi.org/10.18637/jss.v033.i01

Honda T. (2013). Nonparametric quantile regression with heavy-tailed and strongly dependent errors. *Annals of the Institute of Statistical Mathematics* **65**(1): 23–47.
DOI: https://doi.org/10.1007/s10463-012-0359-8

Hrstka O. & Kucerova A. (2004). Improvements of real coded genetic algorithms based on differential operators preventing premature convergence. *Advances in Engineering Software* **35**: 237–246.
DOI: https://doi.org/10.1016/S0965-9978(03)00113-3

Huang J., Ma S. & Zhang C.H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica* **18**(4): 1603–1618.

Hui F.K., Warton D.I. & Foster S.D. (2015). Tuning parameter selection for the adaptive lasso using ERIC. *Journal of the American Statistical Association* **110**(509): 262–269.
DOI: https://doi.org/10.1080/01621459.2014.951444

Hussain S.S. & Sprent P. (1983). Non-parametric regression. *Journal of the Royal Statistical Society: Series A (General)* **146**(2): 182–191.
DOI: https://doi.org/10.2307/2982016

Karaboga D. (2004). *Artificial Intelligence Optimization Algorithms*. Atlas Release Publication, İstanbul, Turkey.

Keskintürk T. (2006). Differential development algorithm. *Istanbul Commerce University Journal of Science* **5**(9): 85–99.

Lacroix L.S. & Zwald L. (2011). Robust regression through the Huber's criterion and adaptive lasso penalty. *Electronic Journal of Statistics* **5**: 1015–1053.
DOI: https://doi.org/10.1214/11-EJS635

Li Q. & Wang L. (2020). Robust change point detection method via adaptive LAD-LASSO. *Statistical Papers* **61**(1): 109–121.
DOI: https://doi.org/10.1007/s00362-017-0927-3

Lu W., Goldberg Y. & Fine J.P. (2012). On the robustness of the adaptive lasso to model misspecification. *Biometrika* **99**(3): 717–731.
DOI: https://doi.org/10.1093/biomet/ass027

Machkour J., Muma M., Alt B. & Zoubir A.M. (2020). A robust adaptive Lasso estimator for the independent contamination model. *Signal Processing* **174**: 107608.
DOI: https://doi.org/10.1016/j.sigpro.2020.107608

Maechler M., Rousseeuw P., Croux C., Todorov V., Ruckstuhl A., Salibian-Barrera M., Verbeke T., Koller M., Conceicao E.L.T. & Palma M.A. (2016). robustbase: Basic robust statistics r package version 0.92-7. Available at *http://CRAN.R-project.org/package=robustbase*

Maritz J.S. (1979). On Theil's method in distribution-free regression. *Australian Journal of Statistics* **21**(1): 30–35.
DOI: https://doi.org/10.1111/j.1467-842X.1979.tb01116.x

Mayer D.G., Kinghorn B.P. & Archer A.A. (2005). Differential

evolution – an easy and efficient evolutionary algorithm for model optimization. *Agricultural Systems* **83**: 315–328. DOI: https://doi.org/10.1016/j.agsy.2004.05.002

McDonald G.C. & Galarneau D.I. (1975). A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association* **70**(350): 407–416. DOI: https://doi.org/10.1080/01621459.1975.10479882

Oja H. & Niinimaa A. (1984). *On Robust Estimation of Regression Coefficients, Research Report*. Department of Applied Mathematics and Statistics, University of Oulu, Finland.

Peng H., Wang S. & Wang X. (2008). Consistency and asymptotic distribution of the Theil-Sen estimator. *Journal of Statistical Planning and Inference* **138**: 1836–1850. DOI: https://doi.org/10.1016/j.jspi.2007.06.036

Qian W. (2013). sealasso: Standard error adjusted adaptive lasso. R package version 0.1-2. Available at *https://CRAN.R-project.org/package=sealasso*

Qian W. & Yang Y. (2013). Model selection via standard error adjusted adaptive lasso. *Annals of the Institute of Statistical Mathematics* **65**(2): 295–318. DOI: https://doi.org/10.1007/s10463-012-0370-0

R Core Team (2017). R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. Available at *https://www.R-project.org/*

Rao K.S.M. & Gore A.P. (1982). Nonparametric tests for intercept in linear regression problems. *Australian Journal of Statistics* **24**(1): 42–50. DOI: https://doi.org/10.1111/j.1467-842X.1982.tb00806.x

Ren Y. & Zhang X. (2013). Model selection for vector autoregressive processes via adaptive lasso. *Communications in Statistics-Theory and Methods* **42**(13): 2423–2436. DOI: https://doi.org/10.1080/03610926.2011.611317

Schwarz G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**(2): 461–464. DOI: https://doi.org/10.1214/aos/1176344136

Sen P.K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association* **63**(324): 1379–1389. DOI: https://doi.org/10.1080/01621459.1968.10480934

Small C. (1990). A survey of multidimensional medians. *International Statistical Review/Revue Internationale de Statistique* **58**: 263–277. DOI: https://doi.org/10.2307/1403809

Storn R. & Price K. (1997). Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization* **11**(4): 341–359. DOI: https://doi.org/10.1023/A:1008202821328

Sun J., Zhang Q. & Tsang E.P.K. (2005). DE/EDA: A new evolutionary algorithm for global optimization. *Information Sciences* **169**: 249–262. DOI: https://doi.org/10.1016/j.ins.2004.06.009

Theil H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Indagationes Mathematicae* **12**(85): 173.

Venables W.N. & Ripley B.D. (2002). *Modern Applied Statistics with S*. 4th Edition. Springer, New York, USA. DOI: https://doi.org/10.1007/978-0-387-21706-2

Wang X. (2005). Asymptotic of the Theil-Sen estimator in a simple linear regression model with a random covariate. *Journal of Nonparametric Statistics* **17**: 107–120. DOI: https://doi.org/10.1080/1048525042000267743

Wang X. & Yu Q. (2005). Unbiasedness of the Theil–Sen estimator. *Nonparametric Statistics* **17**(6): 685–695. DOI: https://doi.org/10.1080/10485250500039452

Wang H., Li G. & Jiang G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business and Economic Statistics* **25**(3): 347–355. DOI: https://doi.org/10.1198/073500106000000251

Wilcox R. (1998). Simulations on the Theil-Sen regression estimator with right censored data. *Statistics and Probability Letters* **39**(1): 43–47. DOI: https://doi.org/10.1016/S0167-7152(98)00022-4

Yoon Y.J., Lee S. & Lee T. (2017). Adaptive lasso for linear regression models with ARMA-GARCH errors. *Communications in Statistics-Simulation and Computation* **46**(5): 3479–3490.

Zhang C. & Xiang Y. (2015). On the oracle property of adaptive group Lasso in high-dimensional linear models. *Statistical Papers* **57**(1): 249–265. DOI: https://doi.org/10.1007/s00362-015-0684-0

Zhou W. & Serfling R. (2008). Multivariate spatial U-quantiles: A Bahadur–Kiefer representation, a Theil–Sen estimator for multiple regression, and a robust dispersion estimator. *Journal of Statistical Planning and Inference* **138**(6): 1660–1678. DOI: https://doi.org/10.1016/j.jspi.2007.05.043

Zhou Z. & Wu W.B. (2011). On linear models with long memory and heavy-tailed errors. *Journal of Multivariate Analysis* **102**(2): 349–362. DOI: https://doi.org/10.1016/j.jmva.2010.09.009

Zou H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**(476): 1418–1429. DOI: https://doi.org/10.1198/016214506000000735