"A Penalized Trimmed Squares Method for Deleting Outliers in Robust Regression"

G. Zioutas, A. Avramidis and L. Pitsoulis

Division of Computational Methods and Computer Programming,
General Department, Faculty of Technology,
Aristotle University of Thessaloniki,
541 24 Thessaloniki, GREECE
{zioutas@eng.auth.gr}

Abstract: We consider the problem of identifying multiple outliers in linear regression models. In robust regression the unusual observations should be removed from the sample in order to obtain better fitting for the rest of the observations. Based on the LTS estimate, we propose a penalized trimmed square estimator PTS, where penalty costs for discarding outliers are inserted into the loss function. We search for suitable penalty costs for multiple high-leverage outliers, which are based on robust leverage and scale. Thus, the best fit for the majority of the data is obtained after eliminating only outliers from the data set. The robust estimation is obtained by minimizing the loss function with a mathematical programming technique, computationally suitable for small sample data. The computational load and the effectiveness of the new procedure are improved by using the idea of *e-insensitive* tube from support vectors machine regression. The PTS loss function is transformed to an *e-Insensitive*, where small errors are ignored, and the mathematical formula gains the sparseness property. The good performance of the PTS estimator allows identification of multiple outliers avoiding masking or swamping effects. We conduct benchmark examples and a simulation study to investigate the procedure's efficiency for robust estimation and power as outlier detection. As a result, the performance of both types of PTS is superior to other methods and is worth the extra computational load.

Keywords: Robust regression, mathematical programming, least trimmed squares, identifying outliers, support vector machines, penalty methods.

1. Introduction

Linear regression models are commonly used to analyze data from many fields of study. These data often contain outliers and influential observations. It is important to identify these observations and bound their influence or eliminate from the data set. If the data are contaminated with a single or few outliers the problem of identifying such observations is not difficult. However, in most cases data sets contain more outliers or a group of influential observations and the problem of identifying such cases becomes more difficult, due to masking and swamping effects.

The approaches to outlier identification separated into two categories: direct approaches and indirect approaches using residuals from the robust fit.

Among famous direct approaches, Hadi and Simonoff (1993) presented a procedure where it is attempted to separate the data into a set of "clean" data points (of size $h=(n+p-1)/2$) and a set of points that contain the potential outliers. The potential outliers are then tested to see how extreme they are relative to the clean subset, using an appropriate diagnostic measure like the adjusted residual, or Cook distance. According to Pena and Yohai (1995) the success of the procedure is based on the initial clean subset of data. The procedure works well for low-leverage outliers but may fail when the sample contains a set of several high-leverage outliers.

Atkison (1994) proposed an identification method of multiple outliers by using a simple forward search starting from initial random subsets. The procedure requires again that at least one of the subsets does not contain high-leverage outliers. Pena and Yohai (1999) proposed a successful fast procedure for detecting group of outliers in many situations, where due to masking effects the usual diagnostics procedures fail. However, they do not claim that their proposal keeps breakdown point of the original estimates. Their procedure has two stages; in the first stage high-leverage points eliminated from the data set irrespective of bad or good leverage points, therefore is susceptible to swamping effects and some effectiveness may be lost. Although in the second stage the efficiency is improved by testing again the potential outliers, some precision may be lost from the first stage. Generally, the key to the success of the above procedures is to obtain a clean initial subset of data.

An indirect approach to outlier identification is through a robust regression estimate. If a robust estimate is relatively unaffected from outliers, then the residuals from the robust fit should be used to flag the outliers.

The generalized M estimators (GM estimators), such as the proposals of Mallows (1975), Hampel (1978), Krasker and Welsch (1982), and other robust estimators were intended to protect the estimator against to response outliers and outlying values of $\mathbf{x}$. In particular, they have influence functions bounded in both $\mathbf{x}$ and y. Unfortunately these bounded influence estimators have small break-down point.

The low break down-point of the GM estimators is a deficiency. Several high break-down point estimators, known as HBP, are based on minimizing a more robust scale estimate than sum of squared residuals, but is well known that lose efficiency Among them is the least median of squares estimator (LMS) off Rousseeuw (1984), which minimizes the median of the squared residuals. Rousseeuw and Leroy (1987) proposed an alternative way to overcome the inefficiency of LMS by flagging points deemed somewhat outlying by LMS, and then performing OLS regression on the clean data set. An alternative estimator that preserves high breakdown is the least trimmed squares LTS of Rousseeuw and Leroy (1987), which minimizes the sum of the h ($h=[(n+p-1)/2]$) smallest squared residuals.

Some better proposals obtain high break down points and simultaneously improve the efficiency of the HBP estimators. Among them are the S estimators of Rousseeuw and Yohai (1984), the MM estimators of Yohai (1987) and Yohai and Zamar (1988), which combine good asymptotic efficiency under the normal linear model with HBP. These estimators, uses a less efficient high-breakdown method as an initial estimate, and then uses an M estimation strategy based on the redescending ψ function. Although they have achieved good asymptotic properties, may have low finite-sample efficiencies if the design contains high leverage points. Morgenthaler (1989) and Stefanski (1991) argue that no estimator with a break down point greater than 1/n can have high finite-sample efficiency in the presence of extreme leverage points.

In order to obtain simultaneously bounded influence, high break down and efficient regression estimators, Coakley and Hettmansperger (1993), Simpson, Ruppert, and Carroll (1992) proposed the one step GM estimates starting with high break down initial estimator and using the Schweppe schema with Mallows weights.

All these improvements to LTS achieve high break down-point, improve the efficiency and have the bounded influence property. However, these estimators are based mainly on the initial LTS regression coefficient value. In practice, their performance depends heavily on the precision of the initial

coefficient estimates. Sometimes, in data contaminated by high-leverage outliers, a bad initial coefficient value does not lead to a good final robust estimation.

Gentleman and Wilk (1975) defined the K most likely outliers as the subset of K observations that produces the largest reduction in the residual sum squares when deleted. Unfortunately, this method requires knowledge of K, which is typically unknown.

In this article we propose a different approach (penalized trimmed squares), which does not require presetting the number (K or h) of outliers to delete from the data set. The new estimator PTS is defined by minimizing a convex objective function (*loss function*), which is the sum of square residuals and penalty costs for discarding bad observations. The robust estimate is obtained by the unique optimum solution of the convex mathematical formula called QMIP.

The PTS estimator is very sensitive to the penalties defined a priory. In fact, these penalty costs are a function of the robust scale σ and leverage of the design points provided by the LTS and minimum covariance determinant MCD of Rousseeuw and Van Driessen (1999). In particular, these penalties in the loss function regulate the robustness and the efficiency of the estimator.

The main purpose of the presented paper is first to construct a regression estimator that has high break down point combined with good efficiency. For this purpose appropriate penalties for high-leverage observations are developed so as to unmask the multiple outliers and delete bad high-leverage outliers whereas keeping most of good high-leverage points in the data sample.

Second, to improve the computation time by bringing together the PTS loss function and the idea of *e-insensitive* loss function from support vector machine, Vapnick (1998). The support vectors have the advantage to reduce the complexity, as usually not all observations but only the support vectors contribute to the predictions, Christmann (2004). Residuals within interval ($-e$, e) ignored in the loss function, and those points outside the so-called e-tube define the regression line. The mathematical programming formula gains the sparseness property and as a result the computation time is reduced. Besides, the effectiveness of the robust regression method is improved, since noisy training data are ignored. For the support vector machine Christmann and Steinwart (2004), Suykens et al. (2002) have emphasized among other properties and the advantage of being robust.

Both of the new estimators PTS and *e-insensitive* PTS have shown robustness against all type of outliers reasonable high break-down point and well efficiency. The PTS formula has the advantage to remove the outliers and it suffers less from masking or swamping effects. Generally, the proposed estimator has the ability to handle a group of outliers. This is shown by means of Examples and Monte Carlo Study. For small or moderate datasets and when the computation time is not a problem, we recommend either PTS or *e-insensitive* PTS procedure.

In section 2, the PTS method is described, we present the penalty loss function and the PTS estimator is defined. In section 3, a mathematical programming formula QMIP is developed for minimizing the penalty loss function and obtaining the PTS estimate. In section 4, we describe the *e-insensitive* PTS procedure. The performance of the new estimators is tested with benchmark examples in section 5. The features of PTS and IPTS estimators are illustrated with a Monte-Carlo simulation study in section 6. Finally, conclusions, future research and computational issues are addressed in section 7.

1.1. Notation

We consider the linear regression model with p independent variables

$$\mathbf{y} = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{u}\,, \tag{1.1}$$

where $\mathbf{y}$ is the $n\times 1$ vector of the response variable $\mathbf{y}=(y_1, y_2, \ldots, y_n)^{\mathrm{T}}$, $\mathbf{X}$ is a full rank $n\times p$ matrix of the $p\times 1$ vectors of explanatory variables, $\mathbf{x_i}=(x_{i,1}, x_{i,2},\ldots,x_{i,p})$, for i=1, 2, …, n, $\boldsymbol{\beta}$ is a $p\times 1$ vector of unknown parameters $\boldsymbol{\beta}=(\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$, and $\mathbf{u}$ is a $n\times 1$ vector $\mathbf{u}=(u_1, u_2, \ldots, u_n)^{\mathrm{T}}$ of random errors with expectation zero and variance σ^2. We observe a sample $(y_i, x_{i,1}, x_{i,2}, \ldots,x_{i,p})$, for i=1, 2, …, n, and we wish to construct a robust estimator in the sense that the influence of any observation (x_i, y_i) on the sample estimator is bounded.

The least squares estimate of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

and let $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)^{\mathrm{T}}$ be the vector of fitted values given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ is the Hat matrix and $\hat{\mathbf{u}} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n)^{\mathrm{T}}$ is the vector of least squares

residuals given by

$$\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

The residual mean square is given by $\hat{\sigma}^2 = \mathrm{SSE}/(\mathrm{n}-\mathrm{p})$, where $\mathrm{SSE} = \hat{\mathbf{u}}^{\mathrm{T}}\hat{\mathbf{u}}$ is the residual sum of squares**.**

The reduction of the residual sum squares due to the deletion of an observation *i* is

$$\mathrm{SSE} - \mathrm{SSE}_{(\mathrm{i})} = \frac{\mathrm{u}_{\mathrm{i}}^2}{1-\mathrm{h}_{\mathrm{i}}} \tag{1.2}$$

where h_i ($0< h_i<1$) measure the leverage of the corresponding observation and is the diagonal elements of the hat matrix H,

$$h_{\mathrm{i}} = \mathbf{x}_{\mathrm{i}}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_{\mathrm{i}} \tag{1.3}$$

It is obvious from (1.2) that deletion of high leverage point (large h_i) leads to reduction of SSE greater than u_i^2. As it has been suggested by Hadi and Simonoff (1993), Pena and Yohai (1999), and Atkison and Riani (2000), a transformation of residuals that has been useful for outlier diagnostics, is the adjusted residual, which is defined as

$$\frac{\mathrm{u}_{\mathrm{i}}}{\sqrt{1-\mathrm{h}_{\mathrm{i}}}} \tag{1.4}$$

2. The Penalty Trimmed Squares Method

As mentioned earlier, a general approach to outlier detection is the form of clean subset of the data that is presumably free of outliers, and then tests the outlying ness of the remaining points relative to the clean data. One way of finding the clean subset is to determine the subset of size *K*, the deletion of which produces the largest reduction in the residual sum of squares. This is essentially equivalent to finding the subset of size *n-K* with the minimum sum of squares.

A problem with this approach is that the size *K* of the outlier subset is rarely known. In order to remove only the outliers in a robust regression, we proposed a new approach, where it is not required knowledge of the cardinality *K* of the subset of outliers. The basic idea is to insert fixed penalty costs into the loss function for possible deletion. Thus, observations that produce reduction larger than their penalty costs are deleted from the data set.

The penalty costs are defined a priory, in the following section the definition of the penalized trimmed square estimator PTS is formalized and suitable penalties for multiple high-leverage outliers are searched.

2.1. The objective function of PTS

We consider that the proposed procedure is based on the LTS estimator. Rousseeuw and Leroy (1987), introduces the LTS which search to find the subset of the data of size *h*, yielding the smallest sum of the *h* squared residuals,

$$\min_{\boldsymbol{\beta}} imize[S_h(\boldsymbol{\beta}) = \sum_{\mathrm{i}=1}^{h}(u_{\mathrm{i}}^2)] \tag{2.1}$$

where *h* is chosen a priori, to maximize the so called break-down point, $h=(n+p-1)/2$. The estimator has high break down point but loses efficiency, since *n-h* observations have to be removed from the sample even they are not outliers.

In order to remove only the outliers in a robust regression, the basic idea is to define as most likely outliers the subset of the observations that produces significant reduction in the residual sum of squares when deleted. Penalty costs are inserted into the objective function (2.1) in order to delete only those observations, which cause significant reduction.

The proposed PTS estimator minimizes the total sum, the sum of the *h* (*h* is not given a priory) square residuals in the clean data and the sum of the penalties for deleting the rest *n-h* observations,

$$\min_{\boldsymbol{\beta},\, h} imize[S_n(\boldsymbol{\beta}) = S_h(\boldsymbol{\beta}) + \mathrm{S}_{n\text{-}h}(\boldsymbol{\beta}) = \sum_{\mathrm{i}=1}^{h}(u_{\mathrm{i}}^2) + \sum_{\mathrm{i}=h+1}^{n}(c_i\,\sigma)^2] \tag{2.2}$$

where, the size h of the clean subset is unknown, σ is a robust residual scale taken from the LTS estimator and the value $(c_i\sigma)^2$ can be interpreted as a *penalty* cost for deleting the i observation, it is defined a priory. The PTS converges to LTS, by decreasing the penalty cost. For instance, asymptotically under Gaussian conditions, minimizing (2.2) with a small constant penalty cost, about $c_i \approx 0.7$, the solution of (2.2) leads to LTS estimator.

Proposition 1. If the PTS estimator for given constant penalty $(c\sigma)^2$ leads to the solution (**β,** h), then given h the LTS estimator yields the same estimate **β.**

Proof: The PTS is defined by minimizing $S_h(\boldsymbol{\beta}) + \mathrm{S}_{n\text{-}h}(\boldsymbol{\beta})$, which has a unique global optimum solution (**β** h). Thus, the sum S_h is the smallest and unique among all subsets with h observations, since the second sum $\mathrm{S}_{n\text{-}h}$ depends only on the number n-h. Given h from PTS, the LTS leads to the minimization of the sum of h square residuals S_h, therefore in both estimators S_h is the same sum. □

As a consequence of Proposition 1, it can be considered that the PTS estimate is equivalent with the OLS applying on the subset of the clean data of size h, yielding the same sum squares S_{h}.

2.2. The Loss Function and Definition of the PTS estimator

We referred on the deletion diagnostic methods, Riani and Atkison (2000), to examine the importance of individual observations to inference about the fit. From robust literature is known that an observation at a point of high leverage would often have a small least squares residual. However, if the observation is deleted and the model refitted to the n-1 observations, the resulting reduction in the sum squares errors will be large if the observation is a high-leverage outlier. The general principle of PTS estimator is to delete an observation if its reduction (1.2) is larger than the penalty cost,

$$\frac{u_i^2}{1-h_i} > (c_i\sigma)^2 .$$

This leads to the conclusion that in the solution of the minimization problem (2.2) every residual in the clean data subset has an upper bound

$$|u_i| < c_i\sigma(\sqrt{1-h_i}) . \tag{2.4}$$

However, as the number of the observations to be deleted increases, there is a combinatorial explosion of the number of deleted subsets to be considered, which can lead to difficulties in interpretation. In the proposed approach many potential outliers considered at once without using the exact expression of multiple deletion diagnostics.

Finally, the proposed estimator PTS can be defined equivalently by solving the problem

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \sum_{\mathrm{i}=1}^{\mathrm{n}} \rho_{c_i\sigma}(u_{\mathrm{i}}) \tag{2.5}$$

where $\rho_{c_i\sigma}$ is the loss function defined as follows

$$\rho_{c_i\sigma}(u_{\mathrm{i}}) = \begin{cases} u_i^2 & for \quad u_i < c_i\,\sigma\sqrt{1-h_i} \\ (c_i\sigma)^2 & for \quad u_i \geq c_i\,\sigma\sqrt{1-h_i} \end{cases} \tag{2.6}$$

The proposed loss function as it is shown in (2.6) is simple, for large residual u_i ($u_i \geq c_i\sigma$) the sum square residuals is less rapidly increasing, since the square residual u_i^2 is reduced to the fixed value $(c_i\sigma)^2$. An interpretation of this is that the observation ($\boldsymbol{x}_{\mathrm{i}}$,$y_{\mathrm{i}}$) does not influence further the regression fitting and can be considered as a deleted one.

It is emphasized for the loss function (2.5), that the upper bound of the residuals corresponding to high-leverage points is reduced significantly

$$|u_i| < cw_i\ \sigma\sqrt{1-h_i}\ , \tag{2.7}$$

where $c_i=cw_i$, and w_i is a function of robust leverage for down-weighting the penalty cost In the next paragraph We propose suitable weights w_i for the high-leverage points in order to eliminate the distortion of the masking problem.

Obviously, the penalty cost $(c_i\sigma)^2$ controls the robustness and efficiency of the estimator (2.5), and therefore a suitable penalty has to be searched.

2.3. The masking problem and choice of penalty cost

Without loss of generality, consider that the penalty cost in the PTS loss function is constant for each observation point, $(c_i\sigma)^2=(c\sigma)^2$. Thus, solving the problem (2.5) low-leverage outliers will be deleted, due to their large residuals ($u>c\sigma$) or equivalently their significant reduction in the sum squares is greater than the penalty cost $(c\sigma)^2$.

Similarly, for a single (or perhaps few) bad high-leverage outlier the leverage h_i is large, therefore its deletion will reduce the sum square residuals significantly due to $u^2/(1-h_i)$.

But, as is known the leverage value h_i can be distorted by the presence of collection of points which individually have small leverage values but collectively forms a high leverage group. In a set of identical high-leverage outliers, Pena and Yohai (1995), showed that the individual leverage h_i of each point might be small, whereas the final residual u_i may appear very close to 0. This implies that the reduction in the sum squares corresponding to these points may be very small, consequently, as the reduction is smaller than the cost,

$$\frac{u_i^2}{1-h_i} << (c\sigma)^2 ,$$

each outlier of that group remains in the clean subset.

Therefore, a masking problem arises, when a group of κ (identical) high-leverage outliers appear in the same direction (same predictor variable), and the estimator of solving (2.2) may fail to delete the high-leverage outliers. The difficulty arises from the fact that the leverage of each outlier in the group is masked. Even, deletion of the whole group will cause significant reduction in the sum squares, the most likely is to be smaller than the total penalty cost $\kappa(c\sigma)^2$. In case of a group of κ high-leverage outliers, the benefit of its deletion is like to delete one outlier, whereas the penalty cost is like to delete κ outliers. Therefore, there is need to reduce the penalty cost for the multiple high leverage points.

Most methods for multiple outlier detection Riani and Atkison (2000), Chatterjee and Hadi (1988), Hadi and Simonoff (1993), Pena and Yohai (1999), seek to divide the data into two parts, a larger "clean data" part and the outliers. The clean data are then used for the estimation of useful parameters. In the new procedure we follow similar principle, if the proper value of penalty cost for removing a high leverage outlier from the clean data set was known there would no be difficulty in detecting the bad group of high leverage outliers in the initial data set.

More specifically, suppose that h_i represents the leverage of each masked high-leverage outlier

$$h_i = x_i^{\mathrm{T}}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}x_i ,$$

and h_i^* represents its leverage as it joins the subset of the m clean observations

$$h_i^* = x_i^{\mathrm{T}}(\mathrm{X}_{(m+1)}^{\mathrm{T}}\mathrm{X}_{(m+1)})^{-1}x_i \tag{2.8}$$

where $\mathrm{X}_{(m+1)}$ is the $(m+1)\times p$ matrix that contains the m clean data and the joint i outlier. The masked outlier in the original data set is a single high-leverage outlier as it joins the clean data set. Applying the proposed robust procedure (2.5) with fixed penalty $(c\sigma)^2$, probably the single high-leverage outlier will be deleted as it was developed before, since its final residual exceeds the upper bound

$$|u_i| < c(\sqrt{1-h_i^*})\sigma .$$

Our intention is to obtain the same upper bound for the masked outlier in the initial data set. This can be done by applying the proposed loss function to the initial data set and modifying the basic penalty cost to

$$\frac{1-h_i^*}{1-h_i}(c\sigma)^2$$

for the corresponding *i* observation. Thus, the multiple high-leverage points can be unmasked by down weighting the penalty cost in the loss function with the weights w_i,

$$w_i = \frac{1 - h_i^*}{1 - h_i}. \tag{2.9}$$

A priory estimation of h_i^* for each observation can be obtained from the clean data set resulted from the MCD procedure, Rousseeuw and Van Driessen (1999). Alternatively a useful estimation of h_i^* could be obtained by down-weighting the high-leverage points in the X matrix with Hampel-Mallows design-weights (1978), or Krasker-Welsch design-weights (1982), or Coakley and Hettmansperger design-weights (1993).

In this study, the proposed design weights (2.9) have been obtained with any of the alternative h_i^*, and all of them work well. However, often in robust design procedures good high-leverage points removed from the clean data and as a consequence the value h_i^* is overestimated. As a result, in the PTS procedure good high-leverage points might be deleted from the data set, and efficiency will be reduced. In the following, we suggest lower bound for penalties derived from LTS.

As masking is especially produced by high-leverage observations, in the proposed procedure the lower penalties concern only high-leverage points. In any case, the penalties should not be lower than $(c_L\,\sigma)^2$, where for this penalty cost the PTS converges to LTS, and no masking problem continue to exist.

Generally, the penalty cost is reduced reasonable for high-leverage points. Based on the above argument, the penalty cost of the loss function (2.5) is modified to

$$(c_i\sigma)^2 = \max[(c_L\sigma)^2, (cw_i\sigma)^2]. \tag{2.10}$$

We have found that c_L=1.35 and c=3 works well and these values have been used in the simulation and examples.

The lower bound in (2.10) yields a desirable property; good high leverage points remain in the clean data set if their *y* values fit well the pattern of data and their residual is smaller than $1.35\sigma\sqrt{1-h_i^*}$. Therefore, the lower bound of the penalty cost is very important for identifying the good leverage points and removing the swamping problem.

2.4. Properties of PTS

The good robust properties for the PTS estimator are based on the penalty cost $(c_i\sigma)^2$ for deleting an outlier. Since, the proposed penalty (2.10) depends on the robust design matrix and robust residual scale, the yielded estimator inherits:

- ***Robustness***, because, the penalty cost is discounted (down-weighted) for high-leverage points, their influence is bounded due to (2.7), and the masking problem is eliminated.
- ***Efficiency,*** because, the proposed robust procedure discards only catastrophic observations, since there is penalty $(c_i\sigma)^2$ for each deletion. High leverage point (**x**-outlier) is not deleted if its final residual is smaller than $c_i\sigma$. Besides, it is obtained the best model fit to the clean data points.
- ***High Break-Down Point,*** because, PTS estimator takes high break-down estimate for residual scale σ and design weights $w(x_i)$, from the LTS and MCD procedure.

We do not claim that our proposed PTS keeps the 50% break-down point of the original LTS. However, the proposed approach succeeds in detecting groups of outliers in many situations where, due to masking effect, the usual diagnostic procedures and robust estimates fail. This is shown by means of a Monte Carlo study and benchmark examples, our procedure obtained high-break down point with reasonable well efficiency.

3. Mathematical Programming Formulas

Before presenting the mathematical formula for computing the PTS estimator, it would be convenient to describe briefly previous formulas for general robust procedure developed by Camarinopoulos and Zioutas (2002), Mangasarian and Musicant (2000), Zioutas (2004). The robust regression of Huber-type is formulated as a mathematical programming problem. The mathematical formula has a

continuous and convex objective function and the constraints form a closed convex set, Arthanari and Dodge (1993).

3.1. Quadratic Programming Formula for General Robust Procedure

For given scale parameter σ, an M-estimator of Huber-type, Huber (1981), could be defined by minimizing the sum of less rapidly increasing functions of the residuals. If ε_i indicates the size of shorten big residuals u_i and $2c_i\sigma\varepsilon_i$ can be considered as the penalty cost of pulling y_i towards its fitted value for a distance ε_i, the robust estimator could be defined by solving an equivalent problem,

$$\underset{\boldsymbol{\beta},\,\varepsilon_i,u_i}{\text{minimize}} \sum_{i=1}^{n}(u_i^2 + 2c_i\sigma\varepsilon_i) \tag{3.1}$$

subject to constraints:

$$\begin{aligned}
&\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2 + u_i + \varepsilon_i \geq y_i \\
&\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2 - u_i - \varepsilon_i \leq y_i \\
&\boldsymbol{\beta}_{1i}, \boldsymbol{\beta}_{2i}, u_i, \varepsilon_i \geq 0 \qquad\qquad for \quad i = 1,..., n
\end{aligned}$$

In this formula all *decision variables* are positive, these are:

- the unknown regression coefficients, $\boldsymbol{\beta}_1^{\mathrm{T}}=(\beta_{11},...\beta_{1p})$, $\boldsymbol{\beta}_2^{\mathrm{T}}=(\beta_{21},...,\beta_{2p})$, therefore, an estimation of parameter $\boldsymbol{\beta}$ is obtained by solving (3.1) and setting $\boldsymbol{\beta}=\boldsymbol{\beta}_1-\boldsymbol{\beta}_2$,
- the residuals u_i ,
- and the pulling distances ε_i .

The first two constraints describe geometrically the regression line for the cases of positive or negative residuals, since the decision variables are constrained to positive values in the mathematical programming formula.

The quadratic programming problem (3.1) has a continuous and convex objective function and the constraints form a closed convex set Arthanari and Dodge (1993), Mangasarian and Musicant (2000). Therefore, during the minimization process the simplex method searches the feasible solutions on the extreme points or corners of the convex region. The iterative method stops at the extreme point $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mathbf{u}, \boldsymbol{\varepsilon})$, which is the unique global minimum, since the objective function is continuous and convex.

3.2. QMIP Formula for the PTS

The new estimator PTS is defined from the solution of the problem (2.2) or (2.5). In order to minimize the penalty loss function in a robust regression, Zioutas and Avramidis (2005) proposed a quadratic mixed integer programming formula, called QMIP. Fixed penalty costs replace the Huber's pulling cost $2c_i\sigma\varepsilon_i$ in the objective function. For the PTS estimator, the mathematical formula 3.1 is transformed to a quadratic mixed integer programming formula,

$$\underset{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,u_i,\varepsilon_i,\delta_i}{\text{minimize}} \quad \sum_{i=1}^{n} (u_i^{\,2} + \delta_i (c_i\sigma)^2) \tag{3.2}$$

subject to constraints:

$$\begin{aligned}
&\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2 + u_i \geq y_i - \varepsilon_i \\
&\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2 - u_i \leq y_i + \varepsilon_i \\
&\varepsilon_i \leq \delta_i\, M \\
&\delta_i : zero - one \text{ variable} \\
&\boldsymbol{\beta}_{1i}, \boldsymbol{\beta}_{2i}, u_i, \varepsilon_i \geq 0 \qquad\qquad for \quad i = 1,..., n
\end{aligned}$$

where δ_i is a *zero-one* decision variable, to indicate which observations must be deleted.

For δ_i=1, the point (x_i, y_i) is deleted, because the third constraint allows the pulling distance ε_i to take any value (upper limit *M,* the maximum regression residual u), therefore the y_i value is pulled towards the fitted value by a distance ε_i, so as the residual is reduced to zero, $u_i = 0$. In this case the objective function increases by a fixed penalty cost $(c_i\sigma)^2$.

For δ_i=0, the third constraint restricts the pulling distance ε_i to zero, thus, the y_i value is not removed.

It is noted here, that this mathematical formula can be adapted for the LTS estimator, by removing the penalties from the objective function and constraining the sum of δ_i with *n-h*, which is the maximum number of deleting points in the LTS estimator.

The quadratic mixed integer programming formula (3.2) still it is convex, therefore, a unique global optimum solution can be obtained for the given data, which is an estimate of the PTS. The convexity and the unique optimum solution for the problem (3.2) are presented specifically in Appendix A.

The solution of the QMIP formula, in the present work, is obtained by the Fort/QMIP algorithm, Mitra *et al* (2003). Computationally, the PTS estimation is suitable for small number of observations, n<100, otherwise it could be extremely intensive. In the next paragraph we propose an *e-insensitive* PTS procedure where the QMIP formula gains sparseness and it becomes computationally reasonable even for larger data sets.

4. The *e-Insensitive* PTS Procedure

The resulted estimator of this procedure is basically the PTS estimator ignoring the informations of observations which corresponds to small residuals, u_i<*e,* see Fig. 1. This idea comes from Support Vector Machine Regression (SVMR), Vapnik (1998), in order to reduce complexity and gain possible even better rubstness properties, Cristmann (2004), Suykens et al. (2002).

The new IPTS estimate is obtained in a two phases method. The first phase is devoted for identifying outliers in the initial data set. A tube radius *e* is fitted to the original data set by minimizing the *e-insensitive* PTS loss function. In the second phase we apply the OLS on the clean data set and the obtained estimate we call IPTS (from e-*Insensitive* PTS).

Based on the IPTS estimation a further step could be done following the MM or S1S procedure to reconsider the deletion of possible good leverage outliers.

4.1. The e-Insensitive PTS Loss Function

In order to improve the computational time we bring from the support vector machine the idea of the *e-insensitive* tube or (sphere)

FIGURE 1. (*Right here****)***

Vapnik (1998) devised the so-called *e-insensitive* loss function,

$$|y - f(x)|_e = \left|y - \mathbf{x}^{\mathbf{T}}\boldsymbol{\beta}\right|_e = \max\left|0, y - \mathbf{x}^{\mathbf{T}}\boldsymbol{\beta} - \mathrm{e}\right|, \tag{4.1}$$

which does not penalize errors below some e>0. In *SV* regression, a desired accuracy *e* is specified a priori. It is then attempted to fit a tube with radius *e* to the data. At each point $\boldsymbol{x_i}$ it is allowed an error *e*, everything above *e* is captured (represents) in residual u_i, which is penalized in the loss function. The case of choosing *e*=0 corresponds to OLS –regression.

Using the *e-insensitive* loss function, only the points outside the *e*-tube enter the stochastic term of the regression model, whereas the points close to actual regression have zero loss. Using the *e-insensitive* loss function the resulted robust estimator gains resistance. Local movements of target values of points inside the *e*-tube do not influence the regression, Scholkopf and Smola (2002), Vapnik (1998).

When noise corrupts all the data, “close” points may be wrong due to noise only- noisy data can be fitted “too well” which leads to poor generalization on future data, Musicant and Mangasarian (2000). Regression line should be influenced by real data not noise, thus errors from these points should be ignored.

Smola and Scholkopf, (1998) suggested square form of the loss function (4.1). Similarly, we adapt the sv’s idea to our approach modifying the *e-insensitive* loss function in a square form, and all the errors

below e are penalized with a constant value e^2. Thus, the proposed *e-insensitive* loss function is becoming

$$\left(y - f(x)\right)_e^2 = \left(y - \mathbf{x}^{\mathbf{T}}\boldsymbol{\beta}\right)_e^2 = \max[e^2, (y - \mathbf{x}^{\mathbf{T}}\boldsymbol{\beta})^{\mathbf{2}}]\,. \tag{4.2}$$

Finally, the *e-Insensitive* PTS estimation is obtained by solving a modified problem of (2.5) and (2.6),

$$\underset{\beta}{\text{minimize}} \quad \sum_{\mathrm{i}=1}^{\mathrm{n}} \rho_{e,c_i\sigma}(u_{\mathrm{i}}^2)\ , \tag{4.3}$$

where $\rho_{e,c_i\sigma}$ is the loss function defined as follows:

$$\rho_{e,c_i\sigma}(u_{\mathrm{i}}^2) = \begin{cases} e^2 & for \quad |u_i| \le e \\ u_i^2 & for \quad e < |u_i| < c_i\sigma\sqrt{1-h_i} \\ (c_i\sigma)^2 & for \quad |u_i| \ge c_i\sigma\sqrt{1-h_i} \end{cases} \tag{4.5}$$

The main difference with the loss function (2.5) of the PTS estimator lies on the fact that small residuals are ignored with a fixed penalty cost e^2.

The accuracy parameter e controls the number of points outside the tube (called support vectors). We have to trade off a potential loss in prediction accuracy with gain of sparseness property and faster solutions. Under normal conditions a good efficiency could be obtained for e=0,6120σ, Scholkopf and Smola (2002).

As the objective of this stage is to obtain preliminary robust estimate, in our Monte Carlo study we have used e=σ. In general, for larger sample sizes we recommend increasing this constant for faster solution. But, there must remain enough data points outside the tube (support vectors) for a reasonable fitting of the tube in the data and consequently correct identification of the outliers. Of course, the number of support vectors is related and with the regression coefficient degree p, Steinwart (2002) gives lower asymptotical bounds on the number of support vectors, i.e., on the data points with non-vanishing coefficients, and investigates the asymptotic behavior. However, the value e for the *e-insensitive PTS* is beyond the scope of this work and it remains an issue for further research.

4.2. The QMIP Formula for the e-insensitive PTS Estimation

The minimization of the loss function (4.3) is equivalent to the following constraint optimization problem

$$\underset{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, u_i, \varepsilon_i, \delta_i}{\text{minimize}} \quad \sum_{i=1}^{n} (u_i^{\,2} + \delta_i (c_i\sigma)^2) \tag{4.6}$$

subject to:

$$\begin{aligned} & \mathbf{x}_{\mathrm{i}}^{\mathrm{T}}\boldsymbol{\beta}_1 - \mathbf{x}_{\mathrm{i}}^{\mathrm{T}}\boldsymbol{\beta}_2 + u_i \ge y_i - \varepsilon_i \\ & \mathbf{x}_{\mathrm{i}}^{\mathrm{T}}\boldsymbol{\beta}_1 - \mathbf{x}_{\mathrm{i}}^{\mathrm{T}}\boldsymbol{\beta}_2 - u_i \le y_i + \varepsilon_i \\ & u_i \ge e \\ & \varepsilon_i \le \delta_i M \\ & \delta_i : zero - one \text{ variable} \\ & \boldsymbol{\beta}_{1i}, \boldsymbol{\beta}_{2i}, u_i, \varepsilon_i \ge 0 \qquad for \quad i = 1,..., n \end{aligned}$$

Note that due to the third constraint any residual smaller than e penalizes the objective function with e^2.

A final note must be made regarding the sparseness of the above formula (4.6). All points inside the e-tube do not contribute to the solution: we could remove any one of them, and still obtain the same solution. The parameter e can be useful for the desired accuracy and sparseness.

In present case, however, our main goal is the identification of the outliers and faster computation, therefore larger values for the parameter e could be used. Besides, as the size of the data set increases, it would be reasonable to increase the sparseness of the mathematical formula (4.6) in order to reduce the computational time. It should be noted that small changes in the parameter e might increase the

sparseness without affecting the correct identification of the outliers.

The new mathematical programming formula is still convex see Appendix A, and therefore the unique global optimum solution of the convex problem (4.6) yields an estimation of *e-insensitive* PTS. In the same solution those δ_i=1 flag the deleted outliers. This way of identifying outliers with the *e-insensitive* PTS, guarantees faster numerical solvability.

4.3. The Algorithm of e-Insensitive PTS Procedure

The algorithm of the *e-insensitive* PTS procedure can be described briefly as follows:

- *STEP 1.* Estimate the robust scale σ and leverage h_i^* (2.7), and determine the penalty costs $(c_i\sigma)^2$.
- *STEP 2.* Solve the QMIP formula (4.6) for the *e-insensitive* PTS estimate, remove from the data the trimmed outliers.
- *STEP 3*. Estimate OLS on the clean data.
- *STEP 4.* Based on the OLS, reconsider the outliers, following one step MM or S1S schema.
- *STEP 5.* Estimate OLS on the final clean data set.

Following these steps we obtain the IPTS estimator, which inherits the high break-down point and efficiency as it is illustrated via Examples from literature and Monte Carlo Study in the next sections.

5. Examples

In this section we describe the application of the proposed procedures to the analysis of several real or artificial data sets encountered in the literature. The first four data sets, discussed by Rousseeuw and Leroy (1987), have become standard "benchmark" data sets for detecting outliers in regression. The high break down estimators like LMS, LTS, the MM or its improved versions and the identification procedures of Hadi *et al*. correctly identify the outliers for these four data sets. Both of our proposals PTS and IPTS identify the true outliers correctly as significantly outlying. Further, the proposed procedures in this article have been tested with many other examples of Rousseeuw and Leroy (1987); in all cases we got good results.

5.1. Telephone Data

We start with the data, which relate the number of telephone calls in Belgium to the variable year, for 24 years. Cases 15-20 are unusually high; cases 14 and 21 are marginal. The outliers draw the OLS regression line upwards, masking the true outliers, while swamping in the clean cases 2-24 as too low. The MM estimator is similar to the other high break down estimators and correctly flags the outliers. Also, our estimators the PTS and *e-insensitive* PTS correctly identify the true outliers.

5.2. The Stars Data

This set consists of 47 measurements of the logarithm of effective temperature at the surface of a star and the logarithm of the light intensity of the star. Although there is a direct relationship between the two variables for most of the stars, the four red giants (cases 11, 20, 30, and 34) have low temperature with high light intensity, and a scatter plot shows them as clear outliers and leverage points. The OLS- and M-estimate lines are very similar, being drawn toward the outliers are masked. The bounded influence estimator is less sensitive to the outliers than are the OLS and M estimators, having (small) positive slope, but the outliers are still masked.

The high-breakdown estimators LTS and MM find the true relationship if the efficiency level is set lower than the typical 95%(for efficiencies up to 80-90%). Considering stronger efficiency the MM estimator fails for this data. Application of the PTS and *e-insensitive* PTS procedure both flags correctly the outliers.

5.3. Modified Wood Gravity Data

We next analyze the five predictors data set, based on real data but modified by Rousseeuw (1984) to contain outliers at cases 4, 6, 8, and 19. All of the identification methods discussed above, as well, the OLS, M, and bounded influence estimates, fail to identify the outliers. The MM estimator is successful for this data, with the true outliers having large residuals. The proposed PTS and IPTS estimators are also successful.

5.4. Hawkins, Bradu, and Kass Data

The data generated by Hawkins *et al*.[6] for illustrating the merits of a robust technique. This artificial data set offers the advantage that at least the position of the good or bad leverage points is known. The Hawkins, Bradu and Kass data consists of 75 observations in four dimensions. The first ten observations is a group of identical bad leverage points, the next four points are good leverage while the remaining are good data. The problem in this case is to fit a hyperplane to the observed data. Plotting the regression residuals from the model obtained from the standard OLS estimator, the bad high-leverage point data are masked and do not show up from the residual plot.

Some robust methods not only fail to identify the outliers, but they also swamp in the good cases 11-14. The MM estimate is

$$Y=-0.952503+.149238X_1+0.196807X_2+0.179251X_3,$$

which means that the true outliers are masked, whereas cases 11-14 are swamped in. Less efficient versions of the MM (up to 80%) give results similar to LTS and correctly flag the outliers. The LTS estimate is

$$Y=-0.524+0.272316X_1+0.055229X_2-0.1876X_3$$

and correctly flags the outliers. An initial estimate of robust design weights reveals the first 14 points of this data set as high leverage points. Application of the PTS and IPTS to these data, starting with robust scale estimate about σ=0.61 from the LTS and down-weighting the penalty cost with weights w_i from (2.8), rejects only the first 10 points as outliers, which are known as the bad leverage points. More specifically, the IPTS estimate gives

$$Y=-0.65989+0.239316X_1+0.059829X_2-0.10256X3,$$

and its computation time is much more faster than the PTS procedure.

5.6. New Artificial Data

These data have been created by Hadi and Simonoff, (1993), in order to illustrate the performance of various robust methods in outlier identification. The two predictors were originally created as uniform (0, 15) and were then transformed to have a correlation of .5. The depended variable was then created to be consistent with the model $y=x_1+x_2+u$. with u~N(0, 1). The first 3 cases1-3 was contaminated to have predictor values around (15, 15), and to satisfy y= x_1+x_2+4. Scatterplots or diagnostics have failed to detect the outliers.

Many identification methods fail to identify the three outliers. Some bounded influence estimates have largest absolute residual at the clean case 17, indicating potential swamping. The LMS regression line in cases 6, 11, 13 17 and 24 yields larger absolute LMS residual values than the true outliers The more efficient high break down methods like LTS, MM, S1S do identify the three outliers as the most outlying cases in the sample, but the residuals are to small to be considered significantly outliers. In contrast, robust methods proposed by Hadi and Simonoff (1993), PTS estimator and IPTS identify correct the clean set 4-25, with each of the cases 1-3 having residuals >3.78.

6. Monte Carlo Results

6.1. Simulation Design

In this section we perform a Monte Carlo experiments to evaluate the performance of our robust procedure and compare with the well-known methods discussed in this article. To carry out one simulation run, we proceeded as follows. The distributions of independent variables and errors and the values of parameters are given. The observations, y_i, were obtained following the regression model second degree p=2,

$$y_i = \beta_0 + \beta_1 \mathrm{x}_{1\mathrm{i}} + \beta_2 \mathrm{x}_{2\mathrm{i}} + u_i$$

where the coefficients values are β_1=1.20, β_2=-0.80 and a zero constant term β_0=0.0. We prefer the Gauss distribution for the error term $u \sim N(0, \sigma^2=16)$, and $\mathrm{x}_{1\mathrm{i}}$ and $\mathrm{x}_{2\mathrm{i}}$ are values drawn also from

normal distributions $N(\mu=20, \sigma=6)$ and $N(\mu=30, \sigma=8)$ respectively.

We select a sample size of $n=50$ and consider that the sample may contain three types of outliers, regression outliers ("bad" high-leverage points), "good" high-leverage points, and response outliers (y-outliers), which arose as follows:

An extra values is drawn from the uniform distribution $U(a=80, b=220)$ and

- for the regression outlier is added to x_{1i} or x_{2i},
- for the "good" leverage point is added to x_{1i} or x_{2i} but the value of the dependent variable y_i follows their contamination , according to the above regression model,
- for the response outlier is added to y_i.

All simulation results are based on 100 replications enough to obtain a relative error <10% with a reasonable confidence level of at least 90% for all the simulation estimates.

We report the results only of the available well-known robust high break down methods. Thus the methods that break down in the presence of only few high-leverage outliers (M, OLS,) are not discussed. The methods examined are, therefore, four different types of robust estimators:

- the LTS estimator,
- the MM estimator,
- the S1S estimator,
- the proposed estimators PTS and IPTS.

A small problem arises when comparing different approaches. How should the cutoffs be chosen? All the constants of the robust estimates are tuned so that to protect the estimator against high contamination (up to 40% or 50%). Thus, for all methods we use the default values for high break down point.

We run all of the computer programs on a1200 Mhz Athlon AMD Processor. The computations for the robust estimators LTS and MM were carried out using the S-Plus package and the MINITAB code for the estimator S1S, which is available in the paper off Coakley and Hettmansperger [6]. The simplex iterations for the QMIP solution were carried out on the same machine using the solver FortMP/QMIP-Fortran Code provided by CARISMA, Brunel University, U.K., 2003.

6.2. Results

We start our comments examining carefully the individual replications. As an example, the sample of coefficient estimates of the first 20 replications, displayed in Table 1, is of interest. The data are rather heavy contaminated with high-leverage, good leverage points and y-outliers (~20%). In the most games all robust estimators had a good performance yielding good regression coefficient estimates. We emphasize on the 2^{th}, 4^{th}, 12^{th}, 14^{th}, 19^{th}, 21^{th} and 23^{th} replication, which are the worst cases in this sample. In the 2^{th} case the S1S improves little over the LTS estimator, our PTS and IPTS estimators have given better estimates. In the 12^{th} replication the S1S estimator improved the LTS estimation, the MM, PTS and IPTS have improved even better. The 12^{th} replication is of grate interest, since the MM and PTS have improved the LTS estimation yielding almost the same estimates.

***Table 1**. (Right here)*

Generally, it should be noted from the Table 1 that the MM and PTS estimators often yield very close coefficient values. The PTS outperforms only in the cases where the MM starts with wrong initial coefficient estimate from LTS estimator. This is verified, also, and from the rest replications through this Mode-Carlo study.

The performance of the robust estimators is measured by Monte Carlo estimates (actual calculations) of the following criteria: mean estimation of $\hat{\beta}_i$, mean absolute error of $\hat{\beta}_i$, norm of the bias of $\hat{\boldsymbol{\beta}}$, and mean square fitting error. The simulation results are summarized in Tables 2, 3, 4.

All of the following conclusions were supported by careful examination of the individual estimates:

The Tables 2, 3 and 4 display results concerning the performance of the five robust estimators corresponding to the following cases:

- Table 2, based on clean data contaminated only by "good" leverage points
- Table 3, based on data contaminated only by regression outliers (contamination 12%).
- Table 4, based on data contaminated by regression outliers, "good" leverage points and

response outliers (contamination 32%).

Not surprisingly, most of the methods are more effective in the case of clean data. For the simulation conducted over clean data contaminated only by "good" leverage points, Table 2, the IPTS estimator outperforms the other estimators. The performance of PTS, MM and S1S was reasonable well with IPTS slight better. Of course, one can improve the efficiency of the robust estimates, but at the cost of losing robustness and outlier detection.

Tables 2-4 present the measures of the performance criteria for the five estimators. Taking account all the performance criteria mentioned above the PTS and IPTS estimator outperforms, in the presence of "good" leverage points (Table 2), in the presence of regression outliers (Table 3), and in the presence of all types of outliers like bad high-leverage, good high-leverage and low-leverage outliers (Table 4). In these Tables, we see that MM and S1S procedure outperform the LTS estimators as it was expected from the asymptotic results.

Table 2. *(Right Here)*

Comparing the results between Table 3 and Table 4, note that all robust estimators have improve their performance in Table 4, even the data are more contaminated. This is because good high-leverage points exist in the Table 4. Among the estimators, MM, LTS and IPTS have much better improvements. Especially for the IPTS estimator, the improvement is significant due to its resistance to swamping effects. Whereas

Table 3. *(Right Here)*

Table 4. *(Right Here)*

Also in all Tables 2-4, it is observed that the new estimators PTS and IPTS compared to others have decreased the norm of bias keeping efficiency in good levels. Consequently, fitting the regression coefficient estimates to clean data, lower mean square fitting error has been obtained by the PTS and IPTS estimators.

Given that for the true regression model the mean square fitting error is 254, we claim that a significant improvement about 20% has been obtained with the new robust approach.

Table 5. *(Right Here)*

In Table 5 we show the percentage p_1 of detecting the true outliers in all Monte Carlo replications. In the same Table we measure the swamping, the percentage p_2 of the false outliers where good leverage points identified as outliers. The PTS and IPTS procedures outperform to the other methods, with the IPTS having the advantage to obtain promising p_1 and keeping swamping p_2 at the smallest value. Among the other estimators, MM show reasonable well effectiveness in outlier detection but generally is subject to more swamping.

Since in most cases the real data may be contaminated with unknown number of both types of **x**-outliers and response outliers, our approach could improve the performance criteria for a finite sample, compared to the best competitors in robust regression.

The proposed methods PTS and IPTS provide the best overall performance. They are of the few approaches not to break down in more difficult problems of multiple high-leverage outliers; also, it is clear that they are attractive outlier identification methods. If a choice were to be made between PTS and IPTS, the IPTS is of more interest. It is somewhat more powerful when there are high leverage outliers, is not swamping and is computationally faster.

Regarding the computation time, to determine the performance of the proposed methods, generally there was heavy computational load comparing with the LTS, MM S1S and other robust procedures. In the above simulation study for n=50, p=3 and contamination 20-30% each replication took about 15 seconds for the PTS procedure and about 1.5-2 seconds for the IPTS procedure. The IPTS computationally was superior to PTS, ten times faster in average. But, as we increase the contamination for the same size of data set and model, the computation load starts to become heavier. Another software package CPLEX-ILOG specialized for QMIP problems has found about ten times faster. There for, an IPTS estimate could be obtained in reasonable time for moderate data size. However, the proposed procedures for PTS and IPTS do not recommend for large size regression problems.

7. Final Remarks

The PTS estimate procedure based on robust residual scale and leverage from the LTS and MCD respectively, can be used successfully in regression problems. Through benchmark Examples and Monte Carlo simulation the proposed estimators have shown robustness against all type of outliers, high break down-point, and reasonable well efficiency.

The robust estimates presented in this article give directly a useful diagnostic tool to identify multiple outliers. The penalized procedure has the advantage to remove the catastrophic outliers and it does not suffer from masking or swamping problems. Generally, the proposed estimator PTS has the ability to handle effectively a group of outliers.

The new estimator PTS is obtained through a convex quadratic mixed integer programming formula (QMIP). The computational effort to solve this formula is heavy. Following the *e-insensitive* technique from Support Vector Machines we have improved the computational time and the effectiveness of the proposed estimator. However, the proposed procedure is slower than the original statistical methods.

Based on the above optimum criteria and results, we conclude that the PTS estimator outperforms in many circumstances and is reasonable for both regression and response outliers. Therefore, it is accessed that for small or medium sample data (n<100) the added computational complexity is worth the potential benefits.

Further improvements in the penalized procedure are a subject of ongoing research; for example to:

- determine possible better choices of the penalties,
- continue the method in a second stage to reconsider the outliers, following one step MM-type or S1S-type procedure.

Concerning the computation effort further research is needed to improve the computational time for medium or bigger size sample data by:

- using specialized solvers for the proposed QMIP formulas,
- determining possible better choice of the *e-insensitive* size for the IPTS procedure,
- implementing re-sampling techniques, similar to LTS or others known from robust literature.

Moreover, from this study, one can gain much by using mathematical programming for robust regression over the ordinary ones. The quadratic mixed integer programming method offers a systematic approach to robust estimation and offers the flexibility in being able to involve some other aspects. For instance, to compound other weight functions for down-weighting simultaneously and other influential points, or to put constraints for the parameter vector $\boldsymbol{\beta}$ if it is necessary**.**

As a final remark, since the number of outliers in a medium sample data is not known, we recommend the use of the PTS or IPTS estimator, which provides a good protection against bias without sacrificing much efficiency.

APPENDIX A: The Convexity of QMIP Formula

We have the following quadratic mixed integer program (QMIP)

$$PTS(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\delta}): \qquad \underset{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\delta}}{\text{minimize}} \sum_{i=1}^{n}(u_i^{*2}+\delta_i(c_i\sigma)^2)$$

subject to:

$$\begin{aligned}
&\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2+u_i^{*}\geq y_i-\varepsilon_i\\
&\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2-u_i^{*}\leq y_i+\varepsilon_i\\
&\varepsilon_i\leq\delta_i M\\
&\delta_i\in\{0,1\}\\
&\boldsymbol{\beta}_{1\mathrm{i}},\boldsymbol{\beta}_{2\mathrm{i}},u_i^{*},\varepsilon_i\geq 0 \qquad \text{for } i=1,\ldots,n
\end{aligned}$$

where $c_i\sigma$ is constant. Given any fixed $\boldsymbol{\delta}\in\{0,1\}^n$ from the 2^n possible ones, and using matrix notation we have the following mixed integer quadratic problem

$$PTS(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2): \qquad \underset{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}{\text{minimize}}\ \mathbf{u}^{\mathrm{T}}\mathbf{u}$$

subject to:

$$\mathbf{X}\boldsymbol{\beta}_1 - \mathbf{X}\boldsymbol{\beta}_2 + \mathbf{u} \geq \mathbf{y} - \boldsymbol{\varepsilon}$$
$$\mathbf{X}\boldsymbol{\beta}_1 - \mathbf{X}\boldsymbol{\beta}_2 - \mathbf{u} \geq \mathbf{y} + \boldsymbol{\varepsilon}$$
$$\boldsymbol{\varepsilon} \leq \boldsymbol{\delta} M$$
$$\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mathbf{u}, \boldsymbol{\varepsilon} \geq \mathbf{0}$$

where $\mathbf{p} = ((c_1\sigma)^2, (c_2\sigma)^2, ..., (c_n\sigma)^2)^{\mathrm{T}}$, $\mathbf{u} = (u_1^*, ..., u_n^*)^{\mathrm{T}}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^{\mathrm{T}}$, $\mathbf{y} = (y_1, ..., y_n)^{\mathrm{T}}$, $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{bmatrix}$ and $\boldsymbol{\beta}_i = (\beta_{i1}, ..., \beta_{in})^{\mathrm{T}}$ for $i = 1,2$. The problem $PTS(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ has linear constraints and a convex quadratic objective function, since the Hessian of $\mathbf{u}^{\mathrm{T}}\mathbf{u}$ has nonnegative eigenvalues (and it is therefore positive semi-definite). Therefore we have a convex program, which will have a unique global optimum solution according to the KKT optimality conditions, Bazaraa and Sherali (1993). Considering that there is a finite number of possible $\boldsymbol{\delta}$, we can conclude that a global optimum solution to $PTS(\boldsymbol{\beta}, \boldsymbol{\delta})$ exist.

References

[1] Arthanari, T.S., and Dodge, Y., (1993), Mathematical Programming in Statistics, John Wiley & Sons, Inc.

[2] Bazaraa, M.S., Sherali, H.P., Shetty, C.M., (1993), *Nonlinear Programming: Theory and Algorithms*, 2nd edition, John Wiley & Sons.

[3] Camarinopoulos, L., and Zioutas, G., (2002), "Formulating Robust Regression Estimation as an Optimum Allocation Problem", *Journal of Statistical Computation and Simulation*, vol. 72, No 9, 687-705.

[4] Chatterjee, S., and Hadi, A.S., (1988), *Sensitivity Analysis in Linear Regression*, New York, Chapman and Hall.

[5] Cristmann A., (2004), "On properties of Support Vector Machines for Pattern Recognition in Finite Samples", Statistics for Industry and Technology, 49-58, Birkhauser Verlag Basel/Switzerland.

[6] Christmann A., Steiwart I., (2004) „On Robustness Properties of Convex Risk Minimization Methods for Pattern Recognition", *Journal of Machine Learning Research*, Vol. 5, pp. 1007-1034.

[7] Coakley, C.W., and Hettmansperger, T.P., (1993), "A Bounded Influence, High Breakdown, Efficient Regression Estimator", *Journal of the American Statistical Association*, 88, 872-880.

[8] Gentleman, J.F., & Wilk, M.B., (1975), "Detecting Outliers, II, Supplementing the direct analysis of residuals", Biometrics, 31, 387-410.

[9] Hadi, A.S., and Simonoff, J.S., (1993), "Procedures for the identification of multiple outliers in linear models", *Journal of the American Statistical Association*, 88, 1264-1272.

[10] Hampel, F.R., (1978), "Optimally bounding the gross error sensitivity and influence of position in factor space", *Proceedings of the ASA Statistical Computing Section*, ASA, Washington, D.C., 59-64.

[11] Hawkins, D.M., Bradu, D., and Kass, G.V., (1984), "Location of several outliers in multiple regression data using elemental sets", *Technometrics*, 26, 197-208.

[12] Huber, P.J., (1981), *Robust Statistics*, John Wiley, New York.

[13] Krasker, W.S., and Welsch, R.E., (1982), "Efficient Bounded-Influence Regression Estimation", *Journal of the American Statistical Association*, 77, 595-604.

[14] Mallows, C.L., (1975), "On Some Topics in Robustness", *unpublished memorandum*, Bell Telephone Laboratories, Murray Hill, New Jersey.

[15] Mangasarian, O.L, Musicant, D.R., (2000), "Robust Linear and Support Vector Regression", *IEEE Transactions on Pattern Analysis and Machine* Intelligence, 22, 950-955.

[16] Mitra, G., Guertler, M., and Ellison, F., (2003), "Algorithms for the Solution of Large-Scale Quadratic Mixed Integer Programming (QMIP) models", *International Symposium on Mathematical Programming*, August 2003, Brunel, U.K.

[17] Morgenthaler, S., (1989), "Comment on Yohai and Zamar", *Journal of the American Statistical Association*, 84, 636.

[18] Musicant, D.R., Mangasarian, O.L, "Support Vector Regression", International Symposium on Mathematical Programming, August 10, 2000.

[19] Pena, D., and Yohai, V.J., (1995), "The detection of influential subsets in linear regression using an influence matrix". *Journal of the Royal Statistical Society*, 47, 145-156.

[20] Pena, D., and Yohai, V.J., (1999), "A procedure for robust estimation and diagnostic in regression", *Journal of the American Statistical Association*, 94, 174-188.

[21] Riani, M., and Atkison, A., (2000), *Robust Diagnostic Regression Analysis*, Springer, Berlin.

[22] Rousseeuw, P.J., (1984), "Least Median of Squares Regression", *Journal of the American Statistical Association*, 79, 871-880.

[23] Rousseeuw, P.J., and Van Driessen, K., (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator", *Technometrics*, 41, 212-223.

[24] Rousseeuw, P.J., and Leroy, A.M., (1987), *Robust Regression and Outlier Detection*, Wiley: New York.

[25] Rousseeuw, P.J., and Yohai, V.J., (1984), "Robust Regression by Means of S-estimators", *in Robust and Nonlinear Time Series Analyses*, (Lecture Notes in Statistics No. 26), eds. J. Franke, W. Hardle, and Martin R.D., Springer Verlag, 256-272.

[26] Scholkopf, B., and Smola, A., (2002), *Learning with* Kernels, MIT Press.

[27] Simpson, D.J., Ruppert, D., and Carroll, R.J., (1992), "On One Step GM Estimates and Stability of Inferences in Linear Regression", *Journal of the American Statistical Association*, 87, 439-450.

[28] Smola, A.J., Scholkopf, B., (1998), "On Kernel-Based Method for Pattern Recognition, Regression, Approximation and Operator Inversion" *Algorithmica*, 22, 211-231.

[29] Stefanski, L.A., (1991), "A Note on High-Breakdown Estimators", *Statistics and Probability Letters*, 11, 353-358.

[30] Steiward I., (2002), " Consistency of Support Vector Machines and Other Regulized Kernel Machines" *IEEE Transactions on Information Theory*, Vol. 51, pp. 128-142, 2005.

[31] Suykens J.A.K, Brabanter J., Lukas L., and Vandewalle J. (2002) " Weighted Least Squares Support Vector Machines: Robustness and Sparse Approximation" *Neurocomputing* 48, 85-105.

[32] Vapnick, V.N., (1998), *Statistical Learning Theory*, Willey, New York.

[33] Yohai, V.J., (1987), "High Breakdown-point and High Efficiency Robust Estimates for Regression, *Annals of Statistics*, 15, 642-656.

[34] Yohai, V.J., and Zamar, R.Z., (1988), "High Breakdown-point Estimates of Regression by Means of Minimization of an Efficient Scale", *Journal of the American Statistical Association*, 83, 406-413.

[35] Zioutas, G., (2004), "Quadratic Mixed Integer Programming Models in Minimax Robust Regression Estimators", Statistics for Industry and Technology, Theory and Applications of Recent Robust Methods, 387-400, Birkhauser Verlag Basel/Switzerland.

[36] Zioutas, G., and Avramidis, A., (2005), "Deleting Outliers in Robust Regression with Mixed Integer Programming", *Acta Mathematicae Applicatae Sinica*, 21, 323-334.

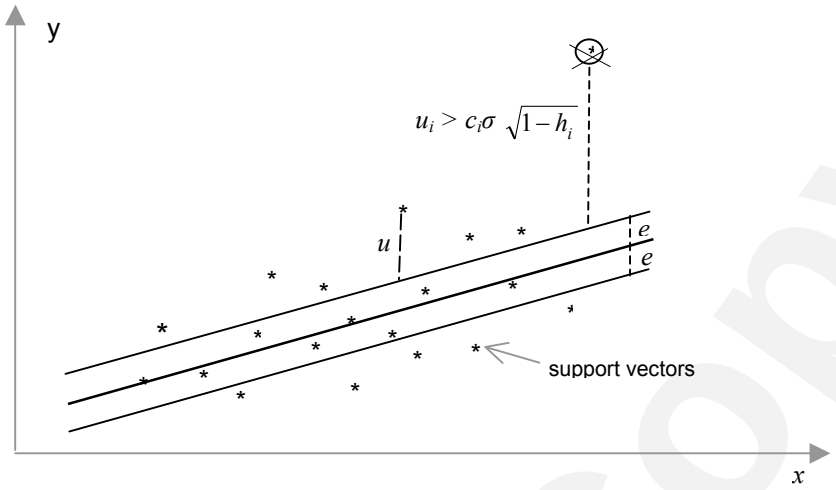

Figure 1. In *SV* regression, a desired accuracy *e* is specified a priori. It is attempted to fit a tube with radius *e* to the data.

Table 1

A sample of 25 replications.
No. of regression outliers 3, y-outliers 3, "good" leverage points 2
regression parameters: $\beta_1 = 1.20$, $\beta_2 = -0.80$

replication	LTS		S1S		MM		PTS		IPTS	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
1	1.16	-0.80	0.95	-0.73	1.22	-0.80	1.21	-0.80	1.22	-0.80
2	**0.06**	**-0.44**	**0.16**	**-0.80**	**0.06**	**-0.75**	**1.13**	**-0.85**	**1.13**	**-0.85**
3	1.17	-0.80	1.07	-0.77	1.19	-0.85	1.19	-0.84	1.19	-0.84
4	**0.19**	**-0.74**	**0.24**	**-0.91**	**1.14**	**-0.70**	**1.14**	**-0.76**	**1.14**	**-0.76**
5	1.16	-0.89	1.13	-0.94	1.18	-0.88	1.18	-0.88	1.18	-0.88
6	1.06	-0.70	0.99	-0.68	1.06	-0.72	1.04	-0.86	1.17	-0.86
7	1.14	-0.72	1.17	-0.76	1.13	-0.70	1.12	-0.74	1.25	-0.75
8	1.22	-0.77	0.95	-0.73	1.23	-0.84	1.23	-0.84	1.23	-0.84
9	1.25	-0.22	1.02	-0.59	1.28	-0.64	1.27	-0.81	1.27	-0.81
10	1.68	-0.65	1.36	-0.64	1.09	-0.89	1.09	-0.61	1.10	-0.72
11	1.73	-0.87	1.18	-0.84	1.25	-0.83	1.25	-0.83	1.25	-0.82
12	**1.16**	**-0.07**	**1.04**	**-0.85**	**1.20**	**-0.77**	**1.21**	**-0.76**	**1.20**	**-0.77**
13	0.78	-0.81	0.96	-0.73	1.16	-0.73	1.15	-0.73	1.15	-0.73
14	**-0.09**	**-0.62**	**0.00**	**-0.66**	**0.01**	**-0.60**	**1.25**	**-0.80**	**1.13**	**-0.67**
15	1.64	-0.75	1.32	-0.73	1.14	-0.71	1.14	-0.71	1.14	-0.71
16	1.16	-0.79	1.14	-0.78	1.20	-0.85	1.28	-0.87	1.19	-0.85
17	1.28	-0.84	1.65	-0.84	1.29	-0.86	1.28	-0.87	1.29	-0.86
18	1.37	-0.41	1.29	-0.69	1.39	-0.49	1.38	-0.88	1.27	-0.90
19	**1.20**	**-0.81**	**1.13**	**-0.80**	**0.10**	**-0.70**	**1.19**	**-0.86**	**1.17**	**-0.79**
20	1.17	-0.81	1.14	-0.74	1.23	-0.75	1.23	-0.75	1.23	-0.75
21	**0.14**	**-0.52**	**0.37**	**-0.56**	**0.21**	**-0.50**	**1.26**	**-0.64**	**1.19**	**-0.59**
22	1.24	-0.48	1.68	-0.69	1.26	-0.80	1.26	-0.88	1.26	-0.88
23	**-0.09**	**-0.65**	**0.39**	**-0.74**	**0.05**	**-0.74**	**0.04**	**-0.74**	**1.26**	**-0.78**
24	1.18	-0.82	1.18	-0.79	1.20	-0.77	1.21	-0.80	1.21	-0.80
25	1.25	-0.94	0.55	-0.81	1.23	-0.88	1.23	-0.88	1.23	-0.88

Table 2

No. of outliers 0, "good" leverage points 3,
regression parameters: $\beta_0 = 0.00$, $\beta_1 = 1.20$, $\beta_2 = -0.80$

Estimators	LTS	S1S	MM	PTS	IPTS
$\hat{\beta}_0$	0.531	-2.259	-1.160	-1.666	-1.256
$\hat{\beta}_1$	1.170	1.202	1.204	1.214	1.214
$\hat{\beta}_2$	-0.771	-0.909	-0.754	-0.751	-0.768
Mean square error of $\hat{\boldsymbol{\beta}}$	76.929	22.883	18.020	15.802	14.905
Mean absoute error of $\hat{\beta}_0$	7.547	3.656	3.023	2.877	2.792
Mean absoute error of $\hat{\beta}_1$	0.081	0.085	0.036	0.035	0.026
Mean absoute error of $\hat{\beta}_2$	0.100	0.102	0.071	0.067	0.060
Norm of bias of $\hat{\boldsymbol{\beta}}$	7.551	3.661	3.027	2.879	2.794
Trace of covaiance	76.645	17.809	16.673	13.023	13.327
Mean square fitting error	307.97	268.43	266.31	263.45	261.70

Table 3

No. of regression outliers 3, "good" leverage points 0,
regression parameters: $\beta_0 = 0.00$, $\beta_1 = 1.20$, $\beta_2 = -0.80$

Estimators	LTS	S1S	MM	PTS	IPTS
$\hat{\beta}_0$	4.946	8.541	1.817	0.906	0.022
$\hat{\beta}_1$	0.874	0.962	0.980	1.102	1.146
$\hat{\beta}_2$	-0.767	-0.940	-0.750	-0.761	-0.761
Mean square error of $\hat{\boldsymbol{\beta}}$	229.692	146.051	71.532	27.242	22.609
Mean absoute error of $\hat{\beta}_0$	11.415	9.532	5.960	4.114	3.924
Mean absoute error of $\hat{\beta}_1$	0.438	0.342	0.266	0.167	0.126
Mean absoute error of $\hat{\beta}_2$	0.217	0.079	0.088	0.097	0.101
Norm of bias of $\hat{\boldsymbol{\beta}}$	11.445	9.544	5.974	4.127	3.933
Trace of covaiance	205.121	73.082	68.179	26.410	22.604
Mean square fitting error	377.73	344.06	314.24	281.96	273.66

Table 4

No. of regression outliers 3, y-outliers 3, "good" leverage points 2,
regression parameters: $\beta_0 = 0.00$, $\beta_1 = 1.20$, $\beta_2 = -0.80$

Estimators	LTS	S1S	MM	PTS	IPTS
$\hat{\beta}_0$	-0.671	3.057	1.041	-0.272	-1.115
$\hat{\beta}_1$	1.008	0.805	1.035	1.158	1.210
$\hat{\beta}_2$	-0.677	-0.822	-0.740	-0.800	-0.796
Mean square error of $\hat{\boldsymbol{\beta}}$	98.914	103.158	48.589	22.767	14.780
Mean absoute error of $\hat{\beta}_0$	7.757	6.938	5.457	3.102	2.815
Mean absoute error of $\hat{\beta}_1$	0.340	0.423	0.222	0.102	0.046
Mean absoute error of $\hat{\beta}_2$	0.154	0.169	0.117	0.060	0.058
Norm of bias of $\hat{\boldsymbol{\beta}}$	7.775	6.987	5.474	3.107	2.822
Trace of covaiance	98.412	93.665	47.475	22.691	13.536
Mean square fitting error	353.34	326.87	298.25	271.85	262.74

TABLE 5
No. of regression outliers 3, "good" leverage points 2, y-outliers 3,
p_1: Fraction of detecting bad leverage outliers.
p_2: Fraction of false outliers, detecting good leverage points (swamping)

Estimators:		PTS	IPTS	MM	LTS	S1S
Detecting percentage	p_1:	**0,90**	0,86	0,81	0,60	0,67
Swamping percentage	p_2:	0,11	**0,07**	0,14	0,16	0,20