REVSTAT – Statistical Journal
Volume 5, Number 1, March 2007, 115–136

PENALIZED TRIMMED SQUARES AND A MODIFICATION OF SUPPORT VECTORS FOR UNMASKING OUTLIERS IN LINEAR REGRESSION [*]

Authors: G. Zioutas
– General Department, Faculty of Technology,
Aristotle University of Thessaloniki, Greece
zioutas@eng.auth.gr

A. Avramidis
– General Department, Faculty of Technology,
Aristotle University of Thessaloniki, Greece

L. Pitsoulis
– General Department, Faculty of Technology,
Aristotle University of Thessaloniki, Greece

Abstract:

- We consider the problem of identifying multiple outliers in linear regression models. We propose a penalized trimmed squares (PTS) estimator, where penalty costs for discarding outliers are inserted into the loss function. We propose suitable penalties for unmasking the multiple high-leverage outliers. The robust procedure is formulated as a Quadratic Mixed Integer Programming (QMIP) problem, computationally suitable for small sample data. The computational load and the effectiveness of the new procedure are improved by using the idea of ϵ*-insensitive* loss function from support vector machines regression. The 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 effects.

Key-Words:

- *robust regression; mixed integer programming; penalty method; least trimmed squares; identifying outliers; support vector machines.*

AMS Subject Classification:

- 62F35, 62J99, 90C99.

[*]This research has been partially funded by the Greek Ministry of Education under the program Pythagoras II.

1. INTRODUCTION

In linear regression models data often contain outliers and bad influential observations. It is important to identify these observations and eliminate them 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 masking outliers and the problem of identifying such cases becomes more difficult, due to masking effects.

The approaches to outlier identification can be separated into two categories: direct approaches and indirect approaches using residuals from the robust fit. Among famous direct approaches, Hadi and Simonoff [9] presented a procedure where it is attempted to separate the data into a set of "clean" data points (of size $k = (n+p-1)/2$) 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. Atkinson [1] 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. Peña and Yohai [14] 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. 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. A famous estimator that preserves high breakdown point (HBP) is the least trimmed squares LTS estimator of Rousseeuw and Leroy [16], that minimize the sum of the k, (coverage $k \geq [(n+p-1)/2]$) smallest squared residuals. But is well known that the LTS loses efficiency. Some better proposals obtain high breakdown points and simultaneously improve the efficiency of the LTS estimator. Among them are the S estimators of Rousseeuw and Yohai [18], the MM estimators of Yohai [24] Simpson, Ruppert, and Carroll [20] and Coakley and Hettmansperger [7], 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 [12] and Stefanski [21] argue that no estimator with a breakdown point greater than $1/n$, can have high finite-sample efficiency in the presence of extreme leverage points. All these improvements to LTS achieve high breakdown 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. Moreover, the LTS method requires the coverage k or equivalently the number $n-k$ of the most likely outliers that produces the largest reduction in the residual sum of square when deleted. Unfortunately, this knowledge of k is typically unknown, Gentleman and Wilk [8].

In this article we propose a different approach penalized trimmed squares PTS, which does not require presetting the number $n-k$ 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 squared 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 priori. 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 [17]. 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 breakdown 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 all of good high-leverage points, if possible, in the data sample, otherwise most of them. Second, to improve the computation time by bringing together the PTS loss function and the idea of ϵ-insensitive loss function from support vector machines, Vapnik [23]. The support vectors have the advantage to reduce the complexity, as usually not all observations but only the support vectors contribute to the predictions, see Christmann [4]. Residuals within the interval $(-\epsilon, \epsilon)$ are ignored in the loss function, and those points outside the so-called ϵ-tube define the regression line. The mathematical programming formula gains the sparseness property and as a result the computation time is significantly reduced. Besides, the effectiveness of the robust regression method is improved, since noisy training data are ignored. For the support vector machines, Suykens et al. [22] and Christmann and Steinwart [5], have emphasized among other properties and the advantage of being robust. Both of the new estimators PTS and ϵ-insensitive PTS have shown robustness against all type of outliers reasonable high breakdown point and well efficiency. The PTS formula has the advantage to remove the outliers and it suffers little from masking 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 datasets and when the computation time is not a problem, we recommend as robust regression procedure the PTS. For moderate data sets the ϵ-insensitive PTS procedure is faster and successful.

In Section 2, we start from the LTS objective function and afterwards the PTS procedure is described. Moreover, the masking problem is described and a suitable penalty function is searched. A mathematical programming formula QMIP is developed in Section 3, for obtaining a PTS estimate. In Section 4, a support vector machines technique is developed with the new ϵ-insensitive loss function. Some benchmark examples are studied in Section 5. The performance of the new estimators PTS and IPTS are tested using Monte-Carlo simulation study in Section 6. Finally, conclusions and future research are addressed in Section 7.

2. TRIMMED SQUARES REGRESSION

We consider the linear regression model with p independent variables

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \ ,$$

where $\boldsymbol{y}$ is the $n\times1$ vector of the response variable $\boldsymbol{y} = (y_1, y_2, ..., y_n)^T$, $\boldsymbol{X}$ is a full rank $n\times p$ matrix of the $p\times1$ vectors of explanatory variables, $\boldsymbol{x}_i = (\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, ..., \boldsymbol{x}_{i,p})$, for $i = 1, 2, ..., n$, $\boldsymbol{\beta}$ is a $p\times1$ vector of unknown parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^T$, and $\boldsymbol{u}$ is a $n\times1$ vector $\boldsymbol{u} = (u_1, u_2, ..., u_n)^T$ of iid random errors with expectation zero and variance σ^2. We observe a sample $(y_i, x_{i,1}, x_{i,2}, ..., x_{i,p})$, for $i = 1, 2, ..., n$, and construct an estimator for the unknown parameters $\boldsymbol{\beta}$. The Least Squares Estimator is defined by minimizing the squared error loss function

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \boldsymbol{u}_i^2 \ .$$

Unfortunately, points that are far from the predicted line (outliers) are overemphasized. Least Squares Estimators are very sensitive to outliers. We wish to construct a robust estimator for the parameter $\boldsymbol{\beta}$, in the sense that the influence of any observation $(\boldsymbol{x}_i, y_i)$ on the sample estimator is bounded.

Rousseeuw and Leroy [16], introduced the Least Trimmed Squares LTS estimator, which fits the best subset of k observations, removing the rest $n-k$ observations. The LTS estimator is defined by minimizing:

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \sum_{i=1}^{k} u_i^2 \ , \\ &\text{s.t.} \quad u_{(1)}^2 < u_{(2)}^2 < u_{(3)}^2 < ... < u_{(k)}^2 \ , \end{aligned} \tag{2.1}$$

where k is the coverage, $k > n/2$ chosen a priori, to maximize the so called breakdown point, $k = (n+p-1)/2$. The estimator has high breakdown point but loses efficiency, since $n-k$ observations have to be removed from the sample even they are not outliers. In real applications the coverage k is unknown. The exact computation of LTS is difficult. Given coverage k, we have to find the best set from all combinations (n,k). The exact algorithm for LTS is a combinatory one, and is suitable for small data sets, i.e. $n < 50$. Fast probabilistic algorithms have been developed for larger samples. In the following proposed robust procedures we consider only exact solutions.

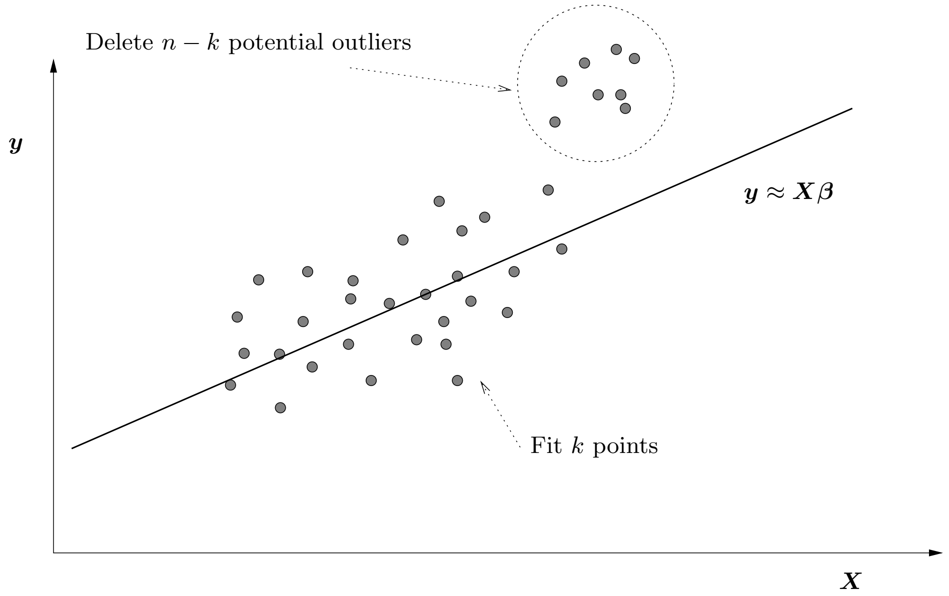
Delete $n-k$ potential outliers
$\boldsymbol{y}$
$\boldsymbol{y} \approx \boldsymbol{X\beta}$
Fit k points
$\boldsymbol{X}$

Figure 1: LTS fitting with coverage k. (In practice the coverage k is unknown).

A problem with the LTS method is that the size $n-k$ of the outlier subset is rarely known. We propose a new approach that does not require presetting the number $n-k$ of outliers to delete from the data set. The basic idea is to insert fixed penalty costs into the loss function for possible deletion. Thus, only observations that produce reduction larger than their penalty costs are deleted from the data set. The penalty costs are defined a priori, in the following section the definition of the penalized trimmed squares estimator PTS is formalized and suitable penalties for multiple high-leverage outliers are proposed. In this work, the PTS estimator is defined over those k observations out of n with the largest maximum likelihood estimation (MLE) fit. We consider as most likely outliers the subset of the observations that produces significant reduction in the residual sum of square when deleted. The proposed PTS estimator minimizes the total sum of squared residuals which is split into two parts; the sum of the k squared residuals in the clean data and the sum of the penalties for deleting the rest $n-k$

observations,

$$\min_{\boldsymbol{\beta},k}\Big(S_k(\boldsymbol{\beta}) + S_{n-k}(\boldsymbol{\beta})\Big) ,$$

$$\text{(2.2)} \qquad \text{or equivalently} \quad \min_{\boldsymbol{\beta},k}\left(\sum_{i=1}^{k} u_i^2 + (n-k)\times(c\sigma)^2\right) ,$$

where, $(c\sigma)^2$ can be interpreted as a *penalty* cost for deleting an observation, σ is a robust residual scale, taken from LTS, and c is a cut-off parameter. The estimator performance is very sensitive to the penalties defined a priori, which regulate the robustness and the efficiency of the estimator. The choice of the robust scale σ plays an important role in the coverage of the PTS estimator. If we wish to obtain an initial clean subset from the PTS estimator (coverage 51%), we choose as scale σ the square root of the minimum mean squared residuals resulted from LTS with the same coverage. Alternatively, in order to delete only the bad outliers, we could get the normalized robust scale σ from the LTS estimator. The minimization problem (2.2) is convex, as it will be proved in Section 3, therefore a global minimum exists. Given that the LTS estimate for coverage k converges to the unique optimum solution of (2.1), the following proposition is useful.

Proposition 2.1. *If the PTS estimator for given penalty $(c\sigma)^2$ converges to the solution $(\boldsymbol{\beta}_{PTS}, k)$, then for the same coverage k the LTS estimator yields the equal estimate $\boldsymbol{\beta}_{LTS} = \boldsymbol{\beta}_{PTS}$.*

Proof: For given penalty $(c\sigma)^2$, the PTS is defined by solving the minimization problem (2.2), and the resulted global minimum is

$$S_{k,PTS} \;=\; S_k(\boldsymbol{\beta}_{PTS}) + (n-k)\times(c\sigma)^2 \;.$$

From the resulted coverage k of the PTS solution, the LTS leads to a unique minimum $S_k(\boldsymbol{\beta}_{LTS})$. Increasing this sum by a constant $(n-k)\times(c\sigma)^2$ yields the unique global minimum sum $S_k(\boldsymbol{\beta}_{LTS}) + (n-k)\times(c\sigma)^2$, which is the same with $S_{k,PTS}$, since both are global minimum. Therefore, both estimates $\boldsymbol{\beta}_{PTS}$ and $\boldsymbol{\beta}_{LTS}$ coincide. □

As a consequence of Proposition 2.1, the PTS estimator can be considered as high breakdown estimator, for small penalty cost $(c\sigma)^2$. For instance, asymptotically under Gaussian conditions, minimizing (2.2) with penalty cost of $c \approx 0.7$, the solution of (2.2) converges to the LTS estimator with high breakdown point $\approx 49\%$. Increasing the parameter c, we obtain better efficiency with reasonable robustness. We have found that for $c=3$, the PTS estimator works well for the catastrophic outliers and this value has been used in the simulation and the examples. Moreover, the PTS estimate is the OLS estimate of the "clean" data subset k. PTS can be approached equivalently by solving the problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_{c\sigma}(u_i) \ ,$$

$$\rho_{c\sigma}(u_i) \ = \ \begin{cases} u_i^2 & \text{for } |u_i| < c\sigma\sqrt{1-h_i}\,, \\ (c\sigma)^2 & \text{for } |u_i| \geq c\sigma\sqrt{1-h_i}\ , \end{cases} \tag{2.3}$$

where the leverages h_i are introduced in the following paragraph. The PTS loss function is simple, for large residual u_i the sum of squared residuals is less rapidly increasing. An interpretation of constant penalizing for big residuals is that the observation $(\boldsymbol{x}_i, y_i)$ does not influence further the regression fitting and can be considered as a deleted one.

As it is known from robust literature, Atkinson and Riani [2], a transformation of residuals that has been useful for outlier diagnostics, is the square of adjusted residual, $\frac{u_i^2}{1-h_i}$, where h_i $(0 < h_i < 1)$ measures the leverage of the i^{th} observation, $h_i = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$. The **general principle** of PTS estimator (2.3) is to delete an observation if its reduction in the sum of squared errors, $S_k(\boldsymbol{\beta})$, is larger than the penalty cost $\frac{u_i^2}{1-h_i} > (c\sigma)^2$. In the solution of the minimization problem (2.3), every residual in the clean data subset has an upper bound $|u_i| < c\sigma(\sqrt{1-h_i})$. 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. Besides, as it 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. Peña and Yohai [14] point out that the individual leverage h_i of each point might be small, whereas the final residual u_i may appear very close to 0, and this is a masking problem.

2.1. Masking problem and choice of penalties

For y-outliers and even for few $\boldsymbol{x}$-outliers the PTS estimator has successful performance. Unfortunately, masking problem arises when there is a group of high leverage points in the same direction. In a set of identical high leverage outliers, the leverage of each outlier is masked; the h_i might be small (Peña and Yohai [13]), $h_i \ll 1$. Deleting a masked leverage point, the reduction in the sum of squared residuals may be small $\frac{u_i^2}{1-h_i} \ll (c\sigma)^2$. In order to eliminate the distortion of the masking problem appropriate penalties for high-leverage observations are searched in this work to unmask the multiple outliers and delete bad high-leverage outliers. Most methods for multiple outlier detection as Hadi and Simonoff [9], Peña and Yohai [14], 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 PTS procedure we follow a similar principle, we propose to down-weight the penalties using information from:

1) The initial leverage of each data point $(\boldsymbol{x}_i, y_i)$, $h_i = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$.

2) The leverage of each point $(\boldsymbol{x}_i, y_i)$ as it joins the clean data subset taken from MCD with coverage k (Rousseeuw and Van Driessen [17]), is $h_i^* = \boldsymbol{x}_i^T(\boldsymbol{X}_{k+1}^T\boldsymbol{X}_{k+1})^{-1}\boldsymbol{x}_i$, which can be considered as the leverage at the clean data set of coverage k. From robust literature, it is expected that $h_i^* \ge h_i$ for the potential x_i-outliers, i.e. for points not included in $\boldsymbol{X}_k$. For the remaining points, which are included in $\boldsymbol{X}_k$, we take $h_i^* = h_i$.

In a bounded influence estimate we wish for every data point $(\boldsymbol{x}_i, y_i)$, $|u_i| \le c\sigma\sqrt{1-h_i^*}$. This can be obtained by weighting the penalty as $\frac{1-h_i^*}{1-h_i}(c\sigma)^2$. Applying the proposed robust function (2.3) to the initial data set

$$\rho_{(1-h_i^*)(1-h_i)c\sigma}(u_i) \;=\; \begin{cases} u_i^2 & \text{for } |u_i| < c\sigma\,\frac{\sqrt{1-h_i^*}}{\sqrt{1-h_i}}\,\sqrt{1-h_i} = c\sigma\sqrt{1-h_i^*}\,, \\[2ex] \dfrac{\sqrt{1-h_i^*}}{\sqrt{1-h_i}}(c\sigma)^2 & \text{for } |u_i| \ge c\sigma\frac{\sqrt{1-h_i^*}}{\sqrt{1-h_i}}\sqrt{1-h_i} = c\sigma\sqrt{1-h_i^*}\,. \end{cases}$$

The above argument leads to the choice of penalty down-weighting with

$$w_i \;=\; \min\left\{1,\, \frac{\sqrt{1-h_i^*}}{\sqrt{1-h_i}}\right\}\,. \tag{2.4}$$

Therefore, the deleting penalties become $(c_i\sigma)^2$, where $c_i = c\,w_i$. For minimizing the penalty loss function in (2.2), a quadratic mixed integer programming formula is used as it is developed in the next paragraph.

3. QMIP FORMULA FOR THE PTS

The new estimator PTS is defined from the solution of the problem (2.2) or (2.3). In order to minimize the penalty loss function in a robust regression, Zioutas and Avramidis [25] proposed a quadratic mixed integer programming formula, called QMIP:

$$
\begin{aligned}
\text{(3.1)} \qquad & \min_{\boldsymbol{\beta},u_i,s_i,\delta_i} \; \sum_{i=1}^{n} \left(u_i^2 + \delta_i (c\,w_i\,\sigma)^2\right) \;, \\
& \text{s.t.} \quad \boldsymbol{x}_i^T \boldsymbol{\beta} + u_i \;\geq\; y_i - s_i \\
& \qquad\;\; \boldsymbol{x}_i^T \boldsymbol{\beta} - u_i \;\leq\; y_i + s_i \\
& \qquad\;\; s_i \leq \delta_i M \\
& \qquad\;\; \delta_i \colon \;\text{zero-one variable} \\
& \qquad\;\; u_i, s_i \geq 0 \quad \text{for}\;\; i=1,...,n \;,
\end{aligned}
$$

where, s is the pulling distance for moving an outlier towards the regression line, $\boldsymbol{\delta}$ is a zero-one decision vector, to indicate which observations must be removed and M is an upper limit of the residuals u_i, $i=1,...,n$. 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:

$$
\begin{aligned}
& \min_{\boldsymbol{\beta}} \;\; \boldsymbol{u}^T \boldsymbol{u} + \boldsymbol{\delta}^T \boldsymbol{p} \;, \\
& \text{s.t.} \quad \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \;\geq\; \boldsymbol{y} - \boldsymbol{s} \\
& \qquad\;\; \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{u} \;\leq\; \boldsymbol{y} + \boldsymbol{s} \\
& \qquad\;\; \boldsymbol{s} \leq \boldsymbol{\delta} M \\
& \qquad\;\; \boldsymbol{u}, \boldsymbol{s} \geq \boldsymbol{0} \;,
\end{aligned}
$$

where, $\boldsymbol{p} = \left((c\,w_1\sigma)^2, (c\,w_2\sigma)^2, ..., (c\,w_n\sigma)^2\right)^T$, $\boldsymbol{u} = (u_1, ..., u_n)^T$, $\boldsymbol{s} = (s_1, ..., s_n)^T$, $\boldsymbol{y} = (y_1, ..., y_n)^T$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_n)^T$ and the matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n]^T$. This problem has linear constraints and a convex quadratic objective function, since the Hessian of $\boldsymbol{u}^T\boldsymbol{u}$ has nonnegative eigenvalues (and it is therefore positive semidefinite). Therefore we have a convex program, which will have a unique global optimum solution according to the Karush–Kuhn–Tucker optimality conditions [3]. Considering that there is a finite number of possible $\boldsymbol{\delta}$, we can conclude that a global optimum solution to the problem exist. Hence, the quadratic mixed integer programming formula (3.1) is convex; therefore, a unique global optimum solution can be obtained for the given data, which is an estimate of the PTS.

In the present work, the solution of the QMIP formula obtained by the Fort/QMIP algorithm, Mitra et al. [11]. Computationally, the PTS estimation is suitable for small number of observations, $n < 50$, otherwise it could be extremely intensive. In the next paragraph we propose an ϵ-insensitive PTS procedure where the QMIP formula gains sparseness and it becomes computationally reasonable even for larger data sets.

4. SUPPORT VECTORS TOLERANT REGRESSION

4.1. ϵ-Insensitive loss function

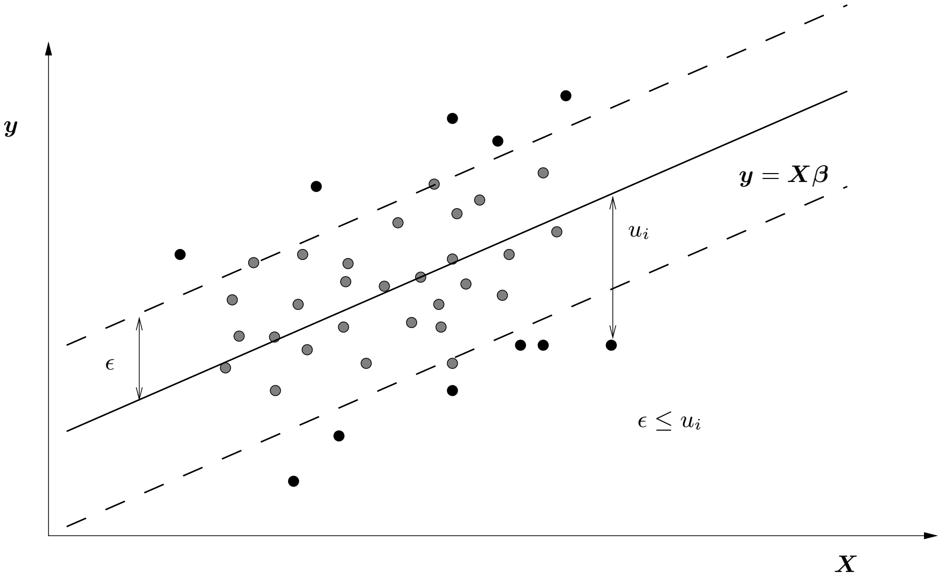

Figure 2: ϵ-insensitive tolerant regression. Only the points outside the tube enter the stochastic term. Points close to actual regression have ϵ loss.

In order, to improve the computation time we use the idea of ϵ-insensitive loss function from support vector machines, proposed by Vapnik [23]. In the ϵ-insensitive loss function small errors are not penalized and it is attempted to fit a tube with radius ϵ to the data, by ignoring (tolerating) small errors, $u < \epsilon$,

$$|y - f(x)|_\epsilon = |y - \boldsymbol{x}^T\boldsymbol{\beta}|_\epsilon = \max\bigl(0, |y - \boldsymbol{x}^T\boldsymbol{\beta}| - \epsilon\bigr) . \tag{4.1}$$

Small errors (below some $\epsilon > 0$) are not penalized in the loss function. The accuracy parameter ϵ controls the number of points outside the tube with radius ϵ. The Support Vectors Regression (SVR) based on the ϵ-insensitive loss function has the advantage to offer sparseness of the solution, Vapnik [23] and Schölkopf and Smola [19]. Christmann and Steinwart [5], [6] proved that kernel methods including SVR have good robustness properties for classification and regression problems if these kernel methods use a bounded and universal kernel and a loss function with bounded first derivative.

We adapt the support vectors technique to our approach modifying the ϵ-insensitive loss function in a squared form, and all the errors smaller than ϵ are penalized with a constant value ϵ^2. Thus, the proposed ϵ-insensitive loss function becomes

$$(4.2)\qquad (y-f(x))^2_\epsilon \;=\; (y-\boldsymbol{x}^T\boldsymbol{\beta})^2_\epsilon \;=\; \max\bigl[\epsilon^2,\, (y-\boldsymbol{x}^T\boldsymbol{\beta})^2\bigr] \;,$$

where, the accuracy parameter ϵ controls the number of points outside the tube, and trades off a potential loss in prediction accuracy with gain of sparseness property and faster solutions.

We bring together, the loss functions of the new ϵ-insensitive and the Penalized Trimmed Squares. Thus, a new estimator called IPTS can yield by solving the problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_{\epsilon,c_i\sigma}(u_i) \;,$$

$$(4.3)\qquad \rho_{\epsilon,c_i\sigma}(u_i) \;=\; \begin{cases} \epsilon^2 & \text{for } |u_i| \leq \epsilon\,, \\ u_i^2 & \text{for } \epsilon < |u_i| < c_i\sigma\sqrt{1-h_i}\,, \\ (c_i\sigma)^2 & \text{for } |u_i| \geq c_i\sigma\sqrt{1-h_i}\;, \end{cases}$$

where $c_i\sigma = \max\{\epsilon,\, c_i\sigma\}$. Under Gaussian conditions good efficiency could be obtained for $\epsilon = 0.612\,\sigma$, Schölkopf and Smola [19]. From our empirical results $\epsilon = \sigma$ was a good choice for faster computation and efficiency. The minimization of the loss function (4.3) is equivalent to the following constraint optimization problem QMIP

$$(4.4)\qquad \begin{aligned} &\min_{\boldsymbol{\beta},u_i,s_i,\delta_i} \sum_{i=1}^{n} \bigl(u_i^2 + \delta_i(c\,w_i\,\sigma)^2\bigr) \;, \\ &\text{s.t.}\quad \boldsymbol{x}_i^T\boldsymbol{\beta} + u_i \;\geq\; y_i - s_i \\ &\qquad\;\; \boldsymbol{x}_i^T\boldsymbol{\beta} - u_i \;\leq\; y_i + s_i \\ &\qquad\;\; u_i \geq \epsilon \\ &\qquad\;\; s_i \leq \delta_i M \\ &\qquad\;\; \delta_i\colon\;\; \text{zero-one variable} \\ &\qquad\;\; u_i, s_i \geq 0 \quad \text{for } i = 1,...,n \;, \end{aligned}$$

where $c\,w_i\,\sigma = \max\{\epsilon,\, c\,w_i\,\sigma\}$, δ_i is a zero-one decision variable, to indicate which observations must be deleted. The IPTS formula is convex, see Section 3, therefore a unique optimum solution can be found and the IPTS is estimated. The tolerance constraint of the above formula leads to sparsity. It should be noted that due to the third constraint any residual smaller than ϵ penalizes the objective function with ϵ^2. A final note must be made regarding the sparseness of the above formula (4.4). All points inside the ϵ-tube do not contribute to the solution: we could remove any one of them, and still obtain the same solution.

The new mathematical programming formula is still convex, see Section 3, and therefore the unique global optimum solution of the convex problem (4.4) yields an estimation of IPTS. In the same solution those $\delta_i = 1$ flag the deleted outliers. This way of identifying outliers with the IPTS, guarantees faster numerical solvability.

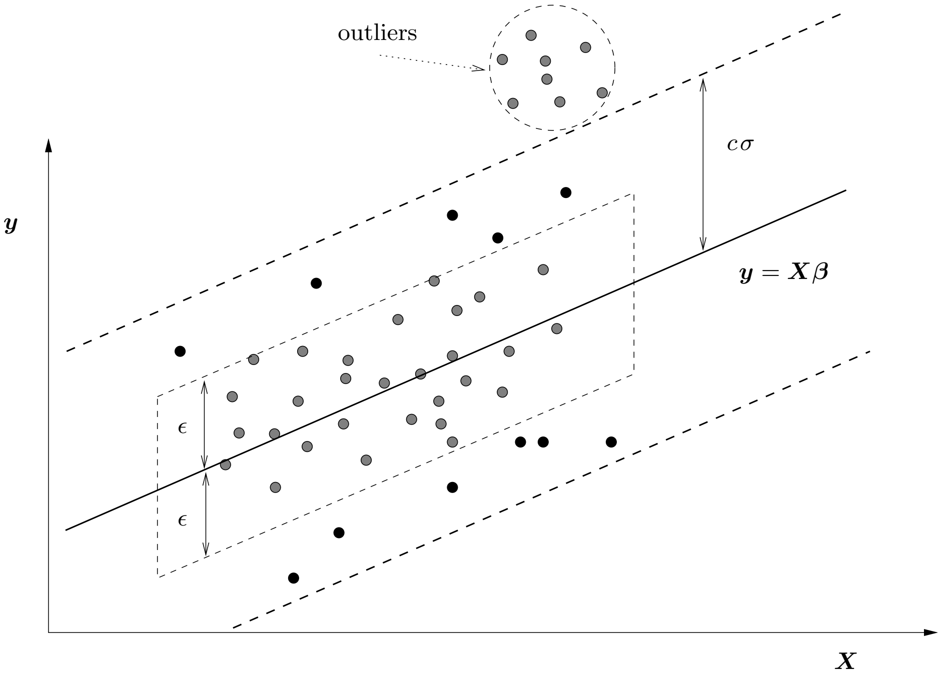

Figure 3: IPTS regression. Appropriate emphasis is given on medium residuals (risk part). De-emphasize small or big errors.

4.2. The Algorithm of IPTS Procedure for large data sets

The parameter ϵ 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 ϵ could be used. Besides, as the size of the data set increases, it would be reasonable to increase the sparseness of the mathematical formula (4.4) in order to reduce the computational time. It should be noted that small changes in the parameter ϵ might increase the sparseness without affecting the correct identification of the outliers. However, as the radius ϵ increases, efficiency of the IPTS estimator may be lost. Therefore, for large data sets, we propose an algorithm of the IPTS procedure which is described briefly as follows:

- Step 1. Estimate the robust scale σ and leverage h_i^*, and determine the penalty costs $(c_i\sigma)^2$.
- Step 2. Solve the QMIP formula for the IPTS estimator.
- Step 3. Remove the detected outliers from the data i.e. points for which $\delta_i = 1$ in Step 2.
- Step 4. Estimate OLS on the clean data set. This is the final IPTS estimator.

Following these steps we obtain the IPTS estimator, which shows good performance as it is illustrated via Examples from literature and Monte Carlo Study in the next sections. More steps could improve further the IPTS estimator by reincluding deleted observations similar to Hadi and Simonoff [9]. However, this is not the goal of the present work.

5. EXAMPLES

The PTS and IPTS procedures have the advantage to remove the outliers and suffers less from masking effects. This is shown by means of real examples or artificial data sets encountered in the literature. The first four data sets, discussed by Rousseeuw and Leroy [16], have become standard "benchmark" data sets for detecting outliers in regression. The high breakdown estimators like LMS, LTS, the MM or its improved versions and the identification procedures of Hadi and Simonoff [9] 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 [16]; in all cases we got good results.

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 breakdown estimators and correctly flags the outliers. Also, our estimators the PTS and IPTS correctly identify the true outliers.

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 IPTS procedure both flags correctly the outliers.

Modified Wood Gravity Data. We next analyze the five predictors data set, based on real data but modified by Rousseeuw [15] 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.

Hawkins, Bradu and Kass Data. The data generated by Hawkins et al. [10] 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 $Y = -0.9525 + 0.1492\,X_1 + 0.1968\,X_2 + 0.1793\,X_3$, 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 $Y = -0.524 + 0.2723\,X_1 + 0.0552\,X_2 - 0.1876\,X_3$, 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 $\sigma = 0.61$ from the LTS and down-weighting the penalty cost with weights w_i from (2.4), rejects only the first 10 points as outliers, which are known as the bad leverage points. More specifically, the IPTS estimate gives $Y = -0.6599 + 0.2393X_1 + 0.0598X_2 - 0.1026X_3$, and its computation time is much faster than the PTS procedure.

New Artificial Data. These data have been created by Hadi and Simonoff [9], 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 0.5. The depended variable was then created to be consistent with the model $y = x_1 + x_2 + u$ with $u \sim N(0, 1)$. The first 3 cases (1–3) were 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 breakdown methods like LTS, MM 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 [9], PTS estimator and IPTS identify correct the clean set $4-25$, with each of the cases $1-3$ having residuals greater than 3.78.

6. MONTE CARLO RESULTS

In this section we perform Monte Carlo experiments to evaluate the performance of our robust procedure and compare it 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 $p = 2$, $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$, where the coefficient values are $\beta_1 = 1.20$, $\beta_2 = -0.80$ and a zero constant term $\beta_0 = 0.0$. We prefer the Gauss distribution for the iid error term $u \sim N(0, \sigma^2 = 16^2)$, while x_{1i} and x_{2i} are iid values drawn also from normal distributions $N(\mu = 20, \sigma^2 = 6^2)$ and $N(\mu = 30, \sigma^2 = 8^2)$ respectively. We 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). An extra value 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. The robust scale estimate σ from LTS with coverage $k = 28$ is used throughout the simulation study. We report the results only of the available well-known robust high breakdown methods. The methods examined are, therefore, five different types of robust estimators: the LTS estimator with coverage $k = [(n+p-1)/2]$, the MM and S1S estimators using in both initially the LTS regression estimate, the proposed PTS estimator solving the QMIP in (3.1), the proposed IPTS estimator solving the QMIP in (4.4). We run all of the computer programs on a 1200 Mhz Athlon AMD Processor. The computations for the robust estimators LTS and MM were carried out using the S-Plus package, while S1S estimator has been computed by the S1S algorithm given in Coakley and Hettmansperger [7]. 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.

All of the following conclusions were supported by careful examination of the individual estimates. Tables 1, 2, 3 and 4 display results concerning the performance of the four robust estimators corresponding to the following cases: Table 1, based on data contaminated by "bad" and "good" high leverage points. Table 2, based on data contaminated only by "good" leverage points. Table 3, based on data contaminated by "bad" high leverage outliers. Table 4, based on data contaminated by "bad" high leverage outliers (heavier contamination).

Table 1: x-outliers 6, "good" leverage points 4, y-outliers 6, $n = 50$.
True: $\beta_0 = 0.0$, $\beta_1 = 1.20$, $\beta_2 = -0.80$.

Estimator	LTS	MM	S1S	PTS	IPTS$_{\epsilon=0.8\sigma}$
Mean estimate of β_0	-0.67	1.82	8.54	0.03	$-$**1.12**
Mean estimate of β_1	1.01	0.98	0.96	1.13	**1.21**
Mean estimate of β_2	-0.68	-0.75	-0.75	-0.81	$-$**0.80**
Mean absolute error of $\hat{\beta}_0$	7.76	5.96	9.53	3.89	**2.82**
Mean absolute error of $\hat{\beta}_1$	0.34	0.27	0.34	0.14	**0.05**
Mean absolute error of $\hat{\beta}_2$	0.15	0.09	0.08	0.07	**0.06**
Mean square error of $\hat{\boldsymbol{\beta}}$	98.91	71.53	146.05	25.43	**14.78**
Norm of bias of $\hat{\boldsymbol{\beta}}$	7.78	5.97	9.54	3.90	**2.82**
Trace of covariance	98.41	68.18	73.05	25.42	**13.54**
Mean square fitting error (true value σ^2=256)	353	314	344	275	**263**
Computation Time (secs)				11	**3**

Table 1 presents the measures of the performance criteria for the four estimators in the presence of bad and good high leverage outliers. Taking account all the performance criteria, the PTS and IPTS outperform the other estimators. In this Table, we see that IPTS outperform the PTS estimator and the IPTS procedure is faster, as it was expected. As far as the computation time of MM, LTS and S1S concern, these are not shown in Tables 1, 2, 3 and 4. This is these estimates results from probabilistic solutions. As it has been mentioned in the previous sections, the PTS and IPTS estimates are the exact solution of QMIP formulas. Therefore, the computation time between probabilistic and exact solutions is not comparable. 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" high leverage points, Table 2, the IPTS estimator outperforms the other estimators. The performance of PTS, MM, S1S and LTS was reasonable well with PTS much better. Of course, one can improve the efficiency of the robust estimates, but at the cost of losing robustness and outlier detection.

Table 2: “good” leverage points 6, $n = 50$.
True: $\beta_0 = 0.0$, $\beta_1 = 1.20$, $\beta_2 = -0.80$.

Estimator	LTS	MM	S1S	PTS	IPTS$_{\epsilon=0.8\sigma}$
Mean estimate of β_0	0.53	−1.16	−2.26	−1.67	**−1.26**
Mean estimate of β_1	1.17	1.20	1.20	1.21	**1.21**
Mean estimate of β_2	−0.77	−0.75	−0.91	−0.75	**−0.77**
Mean absolute error of $\hat{\beta}_0$	7.55	3.02	3.66	2.88	**2.79**
Mean absolute error of $\hat{\beta}_1$	0.08	0.04	0.09	0.04	**0.03**
Mean absolute error of $\hat{\beta}_2$	0.10	0.07	0.10	0.07	**0.06**
Mean square error of $\hat{\boldsymbol{\beta}}$	76.93	18.02	22.88	15.80	**14.91**
Norm of bias of $\hat{\boldsymbol{\beta}}$	7.55	3.03	3.66	2.88	**2.79**
Trace of covariance	76.65	16.67	17.81	13.02	**13.33**
Mean square fitting error (true value σ^2=256)	308	266	268	263	**262**
Computation Time (secs)				9	**2**

In case of only bad high leverage contamination, shown in Table 3, the penalized trimmed squares approach has shown remarkable improvement in both robustness and efficiency, with IPTS the best. As a final conclusion of Tables 1, 2, 3 and taking account all the performance criteria, the IPTS procedure improves reasonable the performance of the PTS. Also, the IPTS procedure is faster.

Table 3: “bad” leverage points 6, $n = 50$.
True: $\beta_0 = 0.0$, $\beta_1 = 1.20$, $\beta_2 = -0.80$.

Estimator	LTS	MM	S1S	PTS	IPTS$_{\epsilon=0.8\sigma}$
Mean estimate of β_0	4.95	1.04	3.06	0.91	**0.02**
Mean estimate of β_1	0.87	1.04	0.81	1.10	**1.15**
Mean estimate of β_2	−0.77	−0.74	−0.82	−0.76	**−0.76**
Mean absolute error of $\hat{\beta}_0$	11.42	5.46	6.94	4.11	**3.92**
Mean absolute error of $\hat{\beta}_1$	0.44	0.22	0.42	0.17	**0.13**
Mean absolute error of $\hat{\beta}_2$	0.22	0.12	0.17	0.10	**0.10**
Mean square error of $\hat{\boldsymbol{\beta}}$	229.69	48.59	103.16	27.24	**22.61**
Norm of bias of $\hat{\boldsymbol{\beta}}$	11.45	5.47	6.99	4.13	**3.93**
Trace of covariance	205.12	47.48	93.67	26.41	**22.60**
Mean square fitting error (true value σ^2=256)	378	298	327	282	**274**
Computation Time (secs)				9	**2.9**

The most fruitful result concerning the IPTS procedure is presented in Table 4. Data are heavy contaminated by bad high leverage outliers. A masking problem arises affecting the performance of the other robust estimators. The IPTS procedure with $\epsilon = 1.5\,\sigma$ has improved significantly the performance criteria and the computation load as well.

Table 4: "bad" leverage points 10, y-outliers 6, $n = 50$.
True: $\beta_0 = 0.0$, $\beta_1 = 1.20$, $\beta_2 = -0.80$.

Estimator	LTS	MM	S1S	PTS	IPTS$_{\epsilon=1.5\sigma}$
Mean estimate of β_0	0.03	-0.97	6.39	-1.14	**$-$1.69**
Mean estimate of β_1	0.77	0.76	0.79	1.15	**1.16**
Mean estimate of β_2	-0.57	-0.51	-0.52	-0.74	**$-$0.74**
Mean absolute error of $\hat{\beta}_0$	9.46	7.42	12.48	5.12	**4.37**
Mean absolute error of $\hat{\beta}_1$	0.56	0.54	0.65	0.21	**0.18**
Mean absolute error of $\hat{\beta}_2$	0.28	0.31	0.30	0.10	**0.09**
Mean square error of $\hat{\boldsymbol{\beta}}$	128.07	87.84	202.22	57.56	**30.50**
Norm of bias of $\hat{\boldsymbol{\beta}}$	9.50	7.51	12.51	5.15	**4.25**
Trace of covariance	127.84	86.63	161.21	56.25	**27.63**
Mean square fitting error (true value σ^2=256)	456	432	490	293	**277**
Computation Time (secs)				13	**0.5**

For large data sets, we could increase the radius ϵ in order to earn computation time, and following the algorithm of subsection 4.2, we obtain reasonable efficiency. In Tables 5 and 6, the success in outlier detection is obvious in large data sets as also the reduction of the computation time of the IPTS estimator as we increase the tube radius.

Table 5: Large artificial data set, 500 points in $\mathbb{R}^2$ including 120 outliers.

Estimator	LTS	PTS	IPTS$_{\epsilon=1.5\sigma}$	IPTS$_{\epsilon=2.0\sigma}$	IPTS$_{\epsilon=2.5\sigma}$
Deleting outlier success	95%	95%	95%	95%	95%
Computation time (sec.)	3800	3800	2500	681	**21**

Table 6: Hawkins et al. [10] artificial data, 75 points in $\mathbb{R}^3$ including 10 outliers.

Estimator	LTS	PTS	IPTS$_{\epsilon=1.5\sigma}$
Deleting outlier success	100%	100%	100%
Computation time (sec.)	255	255	**1.4**

7. CONCLUSIONS AND FUTURE WORK

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. 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 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 a modification of ϵ-insensitive technique from Support Vector Machines we have improved significantly the computational time and the effectiveness of the proposed estimator. However, the computational load of the IPTS estimator is still heavy for large data sets ($n > 100$), since the IPTS procedure is based on Quadratic Mixed Integer Programming which is partly a combinatorial problem. 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 sample data the added computational complexity is worth the potential benefits. Further improvements in the penalized procedure are a subject of ongoing research; for example, determine possible better choices of the penalties and continue the method in a second stage to reconsider the outliers, following one step MM-type procedure. Concerning the computation effort, further research is needed to improve the computational time for large size sample data by determining possible better choice of the ϵ-insensitive size for the IPTS procedure or implementing probabilistic techniques, similar to LTS or others known from robust literature. 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 procedure.

REFERENCES

[1] Atkinson, A. (1994). Fast very robust methods for the detection of multiple outliers, *Journal of the American Statistical Association*, **89**, 1329–1339.

[2] Atkinson, A. and Riani, M. (2000). *Robust Diagnostic Regression Analysis*, John Wiley, Berlin.

[3] Bazaraa, M.; Shevali, H. and Shelty, C. (1993). *Nonlinear Programming: Theory and Algorithms*, John Wiley, New York.

[4] Christmann, A. (2004). *On properties of support vector machines for pattern recognition in finite samples.* In "Statistics for Industry and Technology, Theory and Applications of Recent Robust Methods", Birkhäuser Verlag, Basel, 49–58.

[5] Christmann, A. and Steinwart, I. (2004). On robustness properties of convex risk minimization methods for pattern recognition, *Journal of Machine Learning Research*, **5**, 1007–1034.

[6] Christmann, A. and Steinwart, I. (2005). *Consistency and robustness of kernel based regression*, technical report, University of Dortmund, SFB-475, TR-01/05, submitted.

[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. and 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] 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.

[11] Mitra, G.; Guertler, M. and Ellison, F. (2003). *Algorithms for the solution of large-scale quadratic mixed integer programming (QMIP) models.* In "International Symposium in Mathematical Programming".

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

[13] Peña, D. and Yohai, V.J. (1995). The detection of influential subsets in linear regression using an influence matrix, *Journal of the Royal Statistical Society, Series B (Methodological)*, **57**(1), 145–156.

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

[15] Rousseeuw, P.J. (1984). Least median of squares regression, *Journal of the American Statistical Association*, **79**, 871–880.

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

[17] Rousseeuw, P.J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, **41**, 212–223.

[18] Rousseeuw, P.J. and Yohai, V.J. (1984). *Robust regression by means of S-estimators.* In "Robust and Nonlinear Time Series Analyses" (J. Franke, W. Hardle and R.D. Martin, Eds.), Springer Verlag, 256–272.

[19] Schölkopf, B. and Smola, A. (2000). *Learning with Kernels*, MIT Press.

[20] 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.

[21] Stefanski, L.A. (1991). A note on high-breakdown estimators, *Statistics and Probability Letters*, **11**, 353–358.

[22] 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.

[23] Vapnik, V.N. (1998). *Statistical Learning Theory*, John Wiley, New York.

[24] Yohai, V.J. (1987). High breakdown-point and high efficiency robust estimates for regression, *Annals of Statistics*, **15**, 642–656.

[25] Zioutas, G. and Avramidis, A. (2005). Deleting outliers in robust regression with mixed integer programming, *Acta Mathematicae Applicatae Sinica*, **21**, 323–334.