Assistant Professor Sohel RANA, PhD
East West University, Dhaka-1212, Bangladesh
E-mail: srana_stat@yahoo.com
Lecturer Waleed DHHAN, PhD
Scientific Research Centre, Nawroz University (NZU), Duhok, Iraq
Babylon Municipalities, Ministry of Municipalities and Public Works
Babylon, Iraq
E-mail: w.dhhan@gmail.com
Professor Habshah MIDI, PhD
Faculty of Science and Institute for Mathematical Research
University Putra, Malaysia
E-mail: habshahmidi@gmail.com

FIXED PARAMETERS SUPPORT VECTOR REGRESSION FOR OUTLIER DETECTION

Abstract: *The support vector machine (SVM) is currently a very popular technique of outlier detection as it is a robust model and does not require the data to be of full rank. With a view to evaluate the approximate relationship among the variables, there is necessity to detect outliers that are commonly present in most of natural phenomena before beginning to construct the model. Both of the standard support vector machine(SVM) for regression and modified SV Regression ($\mu-\varepsilon-SVR$) techniques are effective for outlier detection in case of non-linear functions with multi-dimensional inputs; nevertheless, these methods still suffer from a few issues, such as the setting of free parameters and the cost of time. In this paper, we suggest a practical technique for outlier detection by utilising fixed parameters to build SVR model, which reduces computational costs. We apply this technique to real data, as well as simulation data in order to evaluate its efficiency.*

Keywords: *outliers; robustness; sparseness; learning theory; support vector machine.*

JEL Classification: C15

DOI: 10.24818/18423264/52.2.18.16

1. Introduction

The support vector machine (SVM) is currently attracting the attention of many researchers and it has been successfully applied to regression problems (SVR) in addition to classification problems (SVC) (Yang et al., 2004). It is a universal technique to solve problems which are nonlinear, and high-dimensional (full rank or less than full rank) (Williams, 2011). The idea behind the use of the kernel trick in the support vector machine (SVM) is to approximate the non-linear relationship among variables (input space)to a linear form in the feature space (high-dimensional) (Lahiri and Ghanta, 2009). Despite the fact that SVM is a non-parametric method and depends on part of training data, it is still influenced by outliers, which is common issue in real life applications because outliers can be chosen as a part of the support vectors (Chuang et al., 2002).

The majority of data in real applications are vulnerable to noise and outliers that lead to misleading conclusions. According to Hawkins (1980), an outlier can be defined as 'an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism'. Outliers can arise for a number of reasons, such as inherent variability, execution error and measurement error. If the samples contain noise or outliers, the learning technique may try to fit in the undesirable points and this can leads to a skewed approximation function. This is so-called over-fitting problem (Suykens et al., 2002) and can affect the testing error badly; in other words, over-fitting often results in a loss of generalisation ability.

Since the presence of outliers is an abnormal phenomenon, they should be treated (giving down weight) or removed from the complete data set before constructing the approximate model. Because, the target machine (SVR) mostly deals with nonlinear and high-dimensional inputs, conventional methods, such as linear regression, may fail to attain the required efficiency. Moreover, the detection of multiple outliers based on the standard diagnostic methods can be limited due to 'masking' and 'swamping' problems. 'Masking' describes a situation where outliers are incorrectly interpreted as normal points, while 'swamping' is when normal points are incorrectly interpreted as outliers.

Recently, SVM has been successfully applied to detect outliers (Cherkassky and Mulier, 2007). For example, support vector classification was used for outlier detection, as mentioned in Jordaan and Smits (2004). The robustness of SVM with respect to outliers, and the fact that one or more outliers are a portion of the support vectors, makes the technique potentially appropriate for outlier detection (Jordaan and Smits, 2004). Furthermore, SVM has a considerable advantage over the traditional approaches because it is easy to control its free parameters. Anyway, the robustness of support vector machine is not enough to detect outliers easily, as it includes just one angle of the solution triangle (the type of transformation, sparseness and robustness).

DOI: 10.24818/18423264/52.2.18.16

The rest of this paper is organised as follows: in Section 2, a brief description of the Support vector regression and its use for outlier detection is given. Section 3 describes the proposed fixed parameters SVR for outlier detection. In Section 4, the proposed method is evaluated on three real data sets. In section 5, the proposed method is tested on rank-deficient data set (simulation study). Finally, concluding remarks are given in Section 6.

2. SVM regression for outlier detection

In order to understand the SVR methodology, we consider the following regression function of the training data set, $\{(x_1, y_1),...,(x_l, y_l) \mid x \in R^p, y \in R\}$:

$$f(x, w) = w, \Phi(x) + b \tag{1}$$

where x is the space of the independent variables, $\Phi(x)$ is a function, transforms the nonlinear relationship in the original space to be a linear form in a high-dimensional feature space, while w and b are parameters of the weight and the bias of the regression function respectively. Jointly, these parameters are estimated by minimising the next ε -tube loss function (Vapnik, 1995):

$$L_\varepsilon(y_i) = \begin{cases} 0 & ;if \ \mid y_i - f(x, w) \mid \leq \varepsilon \\ \mid y_i - f(x, w) \mid -\varepsilon & ;otherwise \end{cases} \tag{2}$$

The given training data set (x_i, y_i) is to minimise the next convex optimisation problem that is reported by Vapnik (1995):

$$\begin{aligned} &minimize \quad \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} \left(\xi_i + \xi_i^* \right) \\ &subject\,to \quad \begin{cases} y_i - f(x, w) \leq \varepsilon + \xi_i \\ f(x, w) - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0, \quad i = 1, 2, ..., l \end{cases} \end{aligned} \tag{3}$$

where ξ_i and ξ_i^* are the slack variables that give the lower and upper errors, ε is the parameter of the ε - tube loss function and the parameter C represents a trade-off between two directions, the model complexity (flatness) and the quantity of deviations greater than threshold (ε) that are tolerated.

The dual optimisation problem of (3) is to maximise the following convex quadratic problem (Vapnik, 1995):

DOI: 10.24818/18423264/52.2.18.16

$$maximize \quad -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)k(x_i,x_j)-\varepsilon\sum_{i,j=1}^{l}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{l}y_i(\alpha_i-\alpha_i^*) \quad (4)$$

$$subject\ to \quad \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0; \quad \alpha_i,\alpha_i^*\in[0,C]$$

where $k(x_i,x_j)=\left\langle\Phi(x_i),\Phi(x_j)\right\rangle$ is the transformation function (kernel function) (Vapnik, 2000). The final function of SVR can be symbolised as follows:

$$f(x,w)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)k(x_i.x)+b \quad (5)$$

Jordaan and Smits (2004) researched the use of SVR in outlier detection based on its robustness. Later, Nishiguchi et al., (2010) introduced the use of μ−ε−SVR for detection of outliers in case of nonlinear problems with high-dimensional inputs. Although these techniques use the advantages of SVR, it is not easy for non-expert users to set them into practical use due to the difficulty of the tuning of the parameters, as well as the high calculation costs that might result in masking and swamping problems.

The following problems arise when these methods are applied to real life applications: first, the method of Jordaan and Smits (2004) requires high computational costs because detection of an outlier needs several iterations of the optimisation computation, while the other method (Nishiguchi et al., 2010) is only suitable when there are only a few outlier points in the data because the possibility of this technique for detecting and removing outliers is just one outlier point per iteration. Therefore, the computational cost becomes too high when the number of outliers is higher. Second, it mainly depends on trial and error for accurate detection, as it is not clear how users can define the outlier threshold value, which can make this approach difficult for them. Third, according to the machine learning theory, the SVM algorithm has a unique trait that determines its structure (Chuang et al., 2002), which means there is the possibility of the emergence of 'masking' and 'swamping' problems when it uses different values for the ε parameter.

In this paper, our objective is to overcome these drawbacks; we therefore proposed using the SV regression for outlier detection taking into consideration the three angles: the type of transformation, sparseness and robustness.

3. Proposed approach for outlier detection

In order to promote the performance of the standard SVR to detect outliers, we suggest a practical procedure (fixed parameters ε-tube SV Regression) that takes into consideration all three angles(the type of transformation, sparseness and robustness).These three angles produce the so-called triangle of solution; therefore, any

DOI: 10.24818/18423264/52.2.18.16

approach depends only on the question of which is the best. The efficiency of this technique lies in the fact that it requires less time than conventional methods and can detect abnormal points (outliers) without the need to remove them to allow for handling (for instance, minimizes their weights by following robust methods).

In the fixed parameters ε-tube SV Regression, we use the advantage of non-sparseness of the ε-insensitive loss function (the need of all samples).As shown in Ceperic, Gielen and Baric(2014) and Guo, Zhang and Zhang (2010), if the value of threshold ε is very small, then the SV regression model depends on most of the training data, thereby making the resulting solution non-sparse. When the ε parameter greater than zero, it is likely that some of the outliers are not considered as support vectors (fall inside the ε-zone), implying the need for further iterations for detecting outliers correctly. Practically, detection of outliers can be done by using the non-sparse ε -tube loss function(the value of ε parameter equal to zero). The non-sparse ε -tube loss function is defined as follows:

$$L_\varepsilon(y_i) = \begin{cases} 0 & ;if \ | y_i - f(x,w) | \leq 0 \\ | y_i - f(x,w) | & ;otherwise \end{cases} \tag{11}$$

According to that, we can rewrite the convex optimisation problem in (3) as follows:

$$\begin{aligned} &minimize \quad \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} \left(\xi_i + \xi_i^* \right) \\ &subject\,to \quad \begin{cases} y_i - f(x,w) \leq \xi_i \\ f(x,w) - y_i \leq \xi_i^* \\ \xi_i, \xi_i^* \geq 0, \quad i = 1,2,...,l \end{cases} \end{aligned} \tag{12}$$

Consequently, the dual optimisation problem (4) could be rewritten by the following convex quadratic problem:

$$maximize \quad -\frac{1}{2} \sum_{i,j=1}^{l} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) k(x_i, x_j) - \sum_{i,j=1}^{l} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i (\alpha_i - \alpha_i^*) \tag{13}$$

$$subject\ to \quad \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0; \quad \alpha_i, \alpha_i^* \in [0, C]$$

Thus, the prediction SV regression function and the weight vector are represented as follows:

$$f(x,w) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) k(x_i, x) + b \tag{14}$$

$$w = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) \Phi(x_i)$$

DOI: 10.24818/18423264/52.2.18.16

As shown in Rojo-Álvarez et al. (2003), controlling the parameters of SVMs (ε, C and the parameter of kernel) provides insensitivity to outliers or it permits the reduction of the influence of outliers in the optimal solution. Üstün et al. (2005), pointed to that the robustness of the SVR model (14) depends fundamentally on the selection of the C parameter, since the highest α_i and α_i^* values are, by definition of the Lagrange, equal to the value of C parameter. More accurately, a very high C values results in support vectors with a high dispersion among α_i and α_i^* values (14) that produce significant weights. In another words, the highest Lagrange multipliers (α_i and α_i^*) belong to the abnormal sample in the training data set that is deemed an outlier (Jordaan and Smits, 2004).Thereby, the weight vector (14) is increased whenever the value of C parameter is increased, and the presence of outliers. This provides the control on the impact of outliers by controlling the value of the parameter C and the advantages of Kernel function.

Another aspect that should be taken into account is the characteristics of Kernel functions. Williams (2011) pointed to that the algorithm of the SVM is sensitive to the setting option (the type of transformation) and, therefore, users should understand the nature of transformation function (how it works). As the exponential radial basis function (RBF) is the widely used, this paper will focus on a data set for which the RBF kernel is suitable.

3.1. Exponential Radial Basis Function (RBF)

The most common type of kernel is the Gaussian Radial Basis, which could be represented by the following equation:

$$K(x, x_j) = exp\left[\frac{\| x - x_j \|^2}{2h^2}\right] \tag{15}$$

A quick look at the equation (15), we note that output between brackets always a negative value, meaning that the RBF kernel is decreasing exponentially with a start point equal to the upper bound (equal to one). By way of explanation, let x_0 be an outlier, then, the RBF kernel value for pairs (x_0, x_0) will be equal to the upper bound as demonstrated below:

$$K(x_0, x_0) = exp\,[0] = 1 \tag{16}$$

$$K(x_{data\ eest}, x_0) < 1$$

Hence, it can be concluded that the outlier point affects its line more than the influence on the other points and we expect the estimated value (14) for the outlier point to be greatest value among the other estimated values. However, the differences would be

DOI: 10.24818/18423264/52.2.18.16

not clear when the value of the parameter *C* is moderate, implying that we will get a low error. To avoid this case, there are two options: first, we can use very small weights (*C* = 1/100000) to get extremely low estimates, meaning that we will get very high error corresponding to the abnormal point in the data and it will be considered as an outlier. Second, using very high weights (*C* = 100000) produces a very high estimated value corresponding to the unusual sample. Thereby, the use of the highest errors will be failed to detect the outliers because the estimated values are close to its real values. As a result, we would use the estimated values to detect the outliers.

According to the graphics, we can easily observe the points that are far from the majority of the data, however, there are still some difficulties facing the non-expert users. Thus, the criterion of the cut-off point should be used. In order to detect the outlier points correctly, we can utilise robust parameter location (the median) to separate the outliers and the majority of the data. In any clean data such as $|Z_i|$variable, the maximum value of samples is as follows:

$$Max|Z_i| = 2Med|Z_i| + 2\lambda \tag{17}$$

In order to separate the outlier points and the clean points, the equation (17) could be used. However, to use this equation, one needs to estimate the value of λ. Two things should be taken in account to estimate the value of λ parameter, the dispersion of the observations of the variable $|Z_i|$, and the estimated value of the penalising parameter (*C* parameter).In this case, the standard deviation of the robust location parameter (the median) of the variable $|Z_i|$can be used as a predicted value for the parameter λ when the parameter *C* is extremely small and $|Z_i|$are the training errors. As a result of penalising the training error by the parameter we expect high errors (Zong, Liu and Dou, 2006).On the other hand, when the value of the parameter *C* is very high and $|Z_i|$are the predicted values, we can also utilize the standard deviation of the robust location parameter of the variable $|Z_i|$ as a predicted value for the penalizing parameter λ according to the next equation:

$$SST=SSE+SSR \tag{18}$$

The cut-off point in the cases mentioned above could be explained as follows:

$$\begin{aligned} C.P &= 2Med|Z_i| + 2.sd(Med) \\ sd(med) &= \sqrt{\pi\sigma^2 / 2n} \end{aligned} \tag{19}$$

As this approach involves detecting all the outlier points by applying it ones, the computational cost would be less than those of the conventional techniques. Additionally, it is suitable for non-expert users because it introduces fixed set of parameters. In the next two sections, we will utilise the RBF kernel function (15) with

DOI: 10.24818/18423264/52.2.18.16

(h =1, ε = 0, C = 0.0001) as a set of parameters, in the case of using estimation errors to detect outliers, and (h=1, ε =0, C =10000), in the case of using estimated values.

4. Results for real data sets

In order to prove the performance of the proposed approach for outlier detection, we will first discuss the results of real data applications which contains single and multiple outliers. These data sets are the 'wholemeal flour data', the 'international Belgian phone calls data' and the 'Hawkins, Bradu and Kass' data. These real data are chosen because, in numerous previous studies, there is a general consensus on which data points are the true outliers (Maronna et al., 2006; Rousseeuw and Leroy, 1987).

4.1 The copper content data

The first example with two variables is the 24 observations of copper content in wholemeal flour that is sorted in ascending order. The last observation was considered an outlier, as mentioned in most previous studies (Maronna et al., 2006). Figures 1 and2 show the results of applying the fixed parameters SVR graphically to the data set based on estimation errors and estimated values respectively. Table 1, explains the result of applied the proposed method for outlier detection digitally. It is clear that the outlier points are detected correctly.

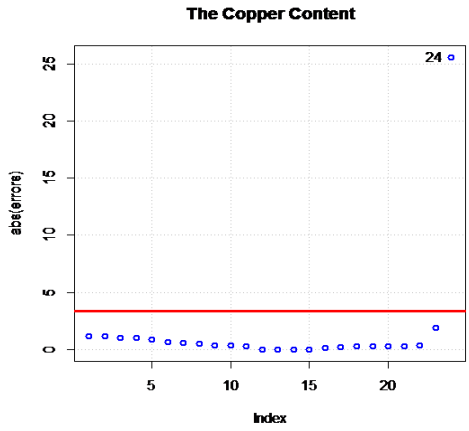

Figure 1. Identification of outliers based on estimated errors.

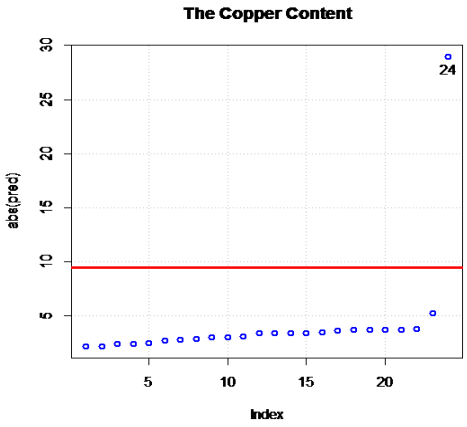

Figure 2. Identification of outliers based on estimated values

DOI: 10.24818/18423264/52.2.18.16

Table 1: The results of applying the proposed method on copper content data set.

Index	Err. (3.34)	Est. (9.48)	Index	Err. (3.34)	Est. (9.48)	SL.	Err. (3.34)	Est. (9.48)
1	1.1848	2.2000	9	0.3547	3.0299	17	0.2148	3.5999
2	1.1847	2.2001	10	0.3547	3.0297	18	0.3148	3.6996
3	0.9847	2.3999	11	0.2847	3.1002	19	0.3148	3.7000
4	0.9847	2.3998	12	0.0148	3.3699	20	0.3148	3.7000
5	0.8847	2.4994	13	0.0149	3.4001	21	0.3148	3.7003
6	0.6847	2.7001	14	0.0148	3.3998	22	0.3848	3.7700
7	0.5847	2.7997	15	0.0148	3.4000	23	1.8948	5.2804
8	0.4847	2.8997	16	0.1148	3.5004	**24**	**25.564**	**28.949**

4.2 Belgian phone data

In the Belgian Statistical Survey, a data set was found containing the total number of international phone calls made between the years 1950 and 1973, which contains heavily contaminated data (Leroy and Rousseeuw, 1987).As shown in Figures 3, 4 and Table 2, the proposed approach is effective to determine outliers correctly by using estimation errors and estimated values respectively.

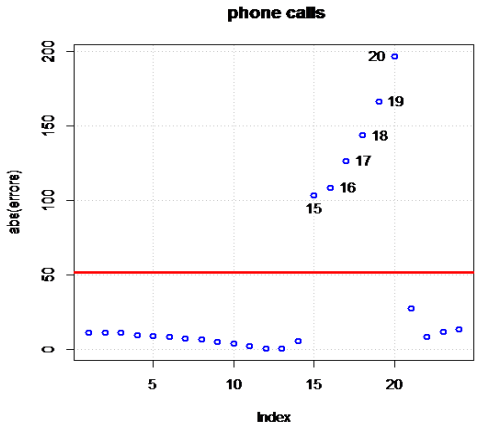

Figure 3. Identification of outliers based on estimation errors.

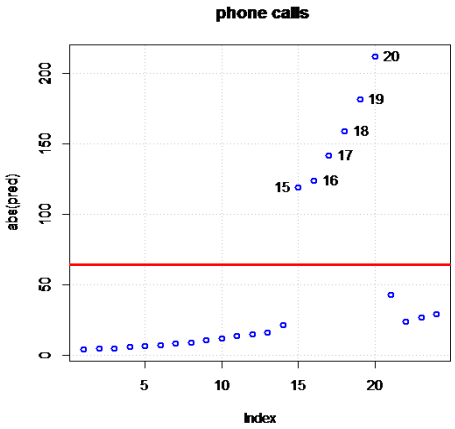

Figure 4. Identification of outliers based on estimated values.

DOI: 10.24818/18423264/52.2.18.16

Table 2: The results of applying the proposed method on phone calls data set.

Index	Err. (51.7)	Est. (64.5)	Index	Err. (51.7)	Est. (64.5)	Index	Err. (51.7)	Est. (64.5)
1	11.099	4.3998	9	4.8998	10.600	**17**	**126.49**	**141.99**
2	10.799	4.7003	10	3.4998	12.000	**18**	**143.49**	**159.00**
3	10.799	4.6996	11	1.9998	13.499	**19**	**166.49**	**181.99**
4	9.5998	5.8999	12	0.5999	14.900	**20**	**196.49**	**211.99**
5	8.8998	6.5998	13	0.5999	16.099	21	27.499	42.999
6	8.1998	7.3002	14	5.6998	21.199	22	8.4998	24.000
7	7.3998	8.0997	**15**	**103.49**	**118.99**	23	11.499	26.999
8	6.6998	8.7999	**16**	**108.49**	**124.00**	24	13.499	29.000

4.3 Hawkins, Bradu and Kass data (HBK)

The last example that has multiple variables is the HBK data set, an artificially constructed data with 10 bad leverage points (the first 10 observations) which affect the regression line and lie far away from it. While the observations11–14 are considered good leverage points that lie near the regression line (Hawkins, 1980).As seen in Figures5 and 6,the proposed method succeeded in detecting outlier points of the data set correctly, whether using estimation errors or estimated values. Table 3 demonstrates the digital results the proposed approach for outlier detection based both of the estimated errors and values.

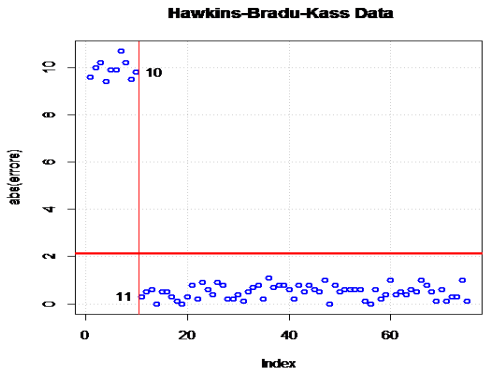

Figure 5. Identification of outliers based on estimation errors.

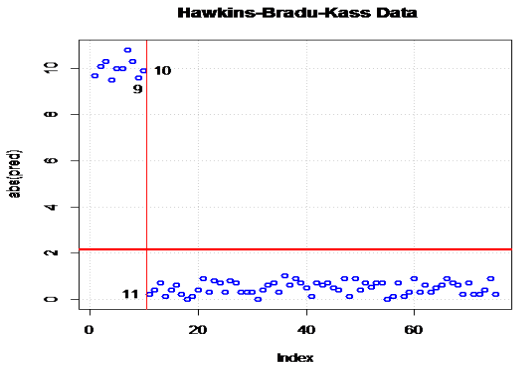

Figure 6. Identification of outliers based on estimated values.

DOI: 10.24818/18423264/52.2.18.16

Table 3: The results of applying the proposed method on HBK data set.

Index	Err. (2.14)	Est. (2.15)	Index	Err. (2.14)	Est. (2.15)	Index	Err. (2.14)	Est. (2.15)
1	**9.5997**	**9.6998**	26	0.8998	0.8002	51	0.5998	0.7003
2	**9.9997**	**10.099**	27	0.7999	0.6999	52	0.5997	0.5001
3	**10.199**	**10.299**	28	0.1999	0.2999	53	0.5999	0.7001
4	**9.3998**	**9.4996**	29	0.1998	0.2998	54	0.6000	0.6998
5	**9.8997**	**10.000**	30	0.3998	0.2995	55	0.0996	0.0003
6	**9.8997**	**9.9999**	31	0.0999	0.0002	56	0.0009	0.1002
7	**10.699**	**10.800**	32	0.4997	0.4003	57	0.5998	0.6999
8	**10.199**	**10.299**	33	0.6999	0.6000	58	0.1999	0.1000
9	**9.4997**	**9.5997**	34	0.7999	0.7001	59	0.3998	0.2996
10	**9.7998**	**9.8999**	35	0.2000	0.2996	60	0.9998	0.9000
11	0.2998	0.1998	36	1.0998	0.9997	61	0.3997	0.3000
12	0.4998	0.3999	37	0.6998	0.6000	62	0.4999	0.5995
13	0.5999	0.6997	38	0.7997	0.9003	63	0.3999	0.2999
14	0.0003	0.1000	39	0.7998	0.7000	64	0.5999	0. 4995
15	0.5000	0.3996	40	0.5996	0.4997	65	0.5000	0.5997
16	0.4999	0.6001	41	0.1999	0.1002	66	0.9996	0.9004
17	0.2997	0.2002	42	0.7999	0.7003	67	0.7998	0. 6998
18	0.0999	0.0002	43	0.4999	0.6001	68	0.4999	0.5996
19	0.0009	0.1004	44	0.7998	0.7000	69	0.0998	0.2004
20	0.3000	0.4000	45	0.5999	0.4999	70	0.6001	0.6998
21	0.7999	0.9004	46	0.4998	0.3996	71	0.1000	0.1999
22	0.2000	0.2998	47	0.9998	0.9003	72	0.2996	0.1999
23	0.8999	0.8000	48	0.0003	0.0998	73	0.3001	0.4004
24	0.6001	0.6998	49	0.7999	0.8998	74	0.9997	0.8995
25	0.4000	0.3003	50	0.4998	0.3997	75	0.0999	0.2001

5. Results for the simulation studies

We consider two types of simulation studies, where the first simulation deals with rank deficient data and the second simulation checks the reliability of the

DOI: 10.24818/18423264/52.2.18.16

proposed method by detecting the correct number of outliers. The simulation studies were done by R software.

5.1 Simulation I

In order to elucidate the efficiency of the proposed approach for data that are rank deficient, we can examine the following example when the sample sizes less than the explanatory variables (n = 25 and p = 30):

$$y = X\beta + r \tag{20}$$

where x_{ij}, b_j and r_i are sampled randomly according to the standard normal distribution (Friedman et al., 2010). Five points (1–5) were replaced by arbitrary large numbers equal to 35 to be five artificially bad leverage points (outlying in X and Y directions). In Figures 7, 8 and Table 4, we can see that the proposed method clearly succeeded to detect outliers for rank-deficient data.

5.2 Simulation II

In this section, we report a simulation study to assess the reliability of the proposed fixed parameters SVR method and compare it with the robust Mahalanobis Distance (RMD) technique in terms of the correct identification, masking and swamping problems. The evaluation of these techniques is based on the rate of correct detection of bad observations and the rate of masking and swamping effects. A good approach is the one that has higher percentage of correct detection of bad leverage points with smaller rates of masking and swamping. Here, experiments are designed

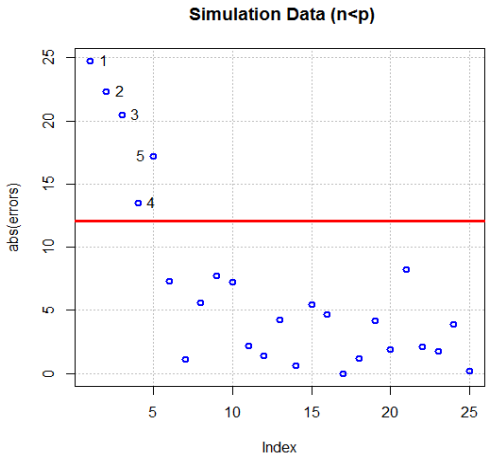

Figure 7. Identification of outliers based on estimation errors

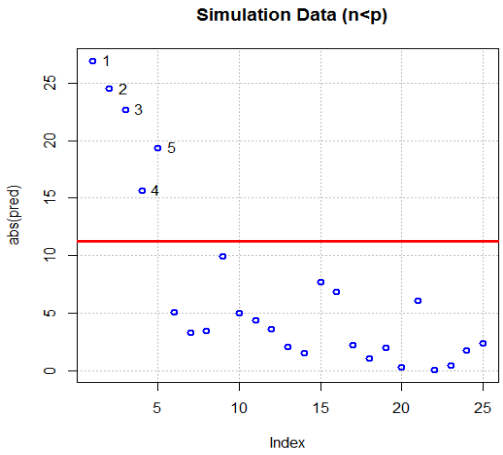

Figure 8. Identification of outliers based on estimated values.

DOI: 10.24818/18423264/52.2.18.16

Table 4: The results of applying the proposed method of rank deficient data set.

Index	Err (12.1)	Est (11.28)	Index	Err (12.1)	Est (11.28)	Index	Err (12.1)	Est (11.28)
1	**24.727**	**26.919**	10	7.2161	5.0243	19	4.1741	1.9824
2	**22.310**	**24.503**	11	2.1656	4.3576	20	1.8869	0.3049
3	**20.482**	**22.674**	12	1.4271	3.6191	21	8.2437	6.0517
4	**13.496**	**15.688**	13	4.2231	2.0312	22	2.1174	0.0747
5	**17.198**	**19.390**	14	0.6486	1.5434	23	1.7471	0.4447
6	7.2770	5.0851	15	5.4824	7.6745	24	3.9175	1.7253
7	1.1063	3.2982	16	4.6570	6.8492	25	0.1855	2.3777
8	5.6248	3.4329	17	0.0000	2.1925	-	-	-
9	7.7410	9.9331	18	1.1761	1.0157	-	-	-

for two sets of explanatory variables based on linear and nonlinear models. The first set is based on the nonlinear model in (21) with two predictors (p = 2), while the second set is based on the general linear regression model in (20) with three predictors (p = 2)

$$y = \sqrt{x_1^2 + x_2^2 + r_i} \tag{21}$$

The explanatory variables are generated randomly from a uniform distribution with mean zero and variance one, while the additive residuals r_i are generated from standard normal distribution. In each experiment, different size of samples (n = 20, 40, 100 and 150) and different percentages of contamination (θ = 0.05, 0.10, 0.15 and 0.20) are used. The bad leverage observations are generated based on the position j of the first $n \times \theta$ observations for both x and y variables. In order to generate these points, the first observations in each explanatory variable is kept fixed at10 + j, which appear later in the dependent variable based on the used model automatically. The comparison results based on1000 replications of this simulation study are summarized in Tables (5) and (6). These tables demonstrate the percentage of correct detection of bad leverage points, and the rates of masking and swamping for all possible combinations of p, n and θ.

It is interesting to note the results from Tables (5) and (6) that the proposed FP-SVR method consistently displays higher rate of detection of BLP with almost negligible swamping and masking rates for all combinations of values of p, n and θ .

DOI: 10.24818/18423264/52.2.18.16

On the other hand, the RMD presents higher rate of detection of BLP but its swamping effect is very high compared with the proposed method which indicates the superiority of the proposed method. The results of the study show that the proposed FP-SVR technique performs best overall of the RMD method.

Table 5: Percentage of correct identification of BLP, masking and swamping for simulation data with two predictors

Cont. level	*n*	% Correct detection		% Masking		% Swamping	
		RMD	FP-SVR	RMD	FP-SVR	RMD	FP-SVR
5%	20	100	100	0	0	30	0.23
	40	100	100	0	0	10.6	0.50
	100	100	100	0	0	11.7	0.43
	150	93.3	93.3	6.7	6.7	1.25	0.29
10%	20	100	100	0	0	25	0.01
	40	100	100	0	0	5.9	0.08
	100	100	100	0	0	10.4	0.05
	150	100	100	0	0	0.71	0.03
15%	20	100	100	0	0	20	0.01
	40	100	100	0	0	4.6	0.01
	100	100	100	0	0	4.4	0
	150	97.8	97.8	2.2	2.2	0.89	0
20%	20	100	100	0	0	5	0
	40	100	100	0	0	6.2	0
	100	100	100	0	0	1.1	0
	150	100	100	0	0	0.7	0

Table 6: Percentage of correct identification of BLP, masking and swamping for simulation data with three predictors

Cont. level	*n*	% Correct detection		% Masking		% Swamping	
		RMD	FP-SVR	RMD	FP-SVR	RMD	FP-SVR
5%	20	100	100	0	0	30	0
	40	100	100	0	0	9.5	0
	100	100	100	0	0	5.6	0.01
	150	94.3	94.3	6.7	6.7	0.8	0
10%	20	100	100	0	0	5	0
	40	100	100	0	0	4.4	0
	100	100	100	0	0	2.3	0
	150	100	100	0	0	0.7	0

DOI: 10.24818/18423264/52.2.18.16

Table 6 continue

15%	20	100	100	0	0	10	0
	40	100	100	0	0	2.8	0
	100	100	100	0	0	0.7	0
	150	97.7	97.7	2.22	2.22	0.8	0
20%	20	100	100	0	0	0	0
	40	100	100	0	0	3	0
	100	100	100	0	0	0.1	0
	150	100	100	0	0	0.7	0

6. Conclusion

In this study, we proposed a practical approach to detecting outliers in the case of full rank and rank-deficient (less than full rank) by using fixed parameters of the SV regression. This technique has advantages over the previous methods as it diminishes computational cost and it succeeded in introducing a fixed set of the free parameters, making it appropriate for non-expert users. Moreover, outliers could be detected without the need to remove and replace them. The efficiency of this approach was tested on real and simulated data sets.

REFERENCES

[1] **Ceperic, V., Gielen, G., Baric, A. (2014),** ***Sparse E-Tube Support Vector Regression by Active Learning;*** *Soft Computing,* 18(6), 1113-1126;

[2] **Cherkassky, V., Mulier, F.M. (2007),** ***Learning from Data: Concepts, Theory, and Methods;*** New Jersey: *John Wiley & Sons*;

[3] **Chuang, C., Su, S., Jeng, J., Hsiao, C. (2002),** ***Robust Support Vector Regression Networks for Function Approximation with Outliers;*** *Neural Networks, IEEE Transactions,* 13(6), 1322–1330;

[4] **Friedman, J., Hastie, T., Tibshirani, R. (2010),** ***Regularization Paths for Generalized Linear Models via Coordinate Descent;*** *Journal of Statistical Software,* 33(1), 1–22;

[5] **Guo, G., Zhang, J., Zhang, G. (2010),** ***A method to Sparsify the Solution of Support Vector Regression;*** *Neural Computing and Applications,* 19(1), 115–122;

[6] **Hawkins, D.M. (1980),** ***Identification of Outliers;*** New York: *Springer;*

[7] **Jordaan, E.M.,Smits G. F. (2004),** ***Robust Outlier Detection Using SVM Regression. Neural Networks;*** *2004 Proceedings. 2004 IEEE International Joint Conference on IEEE,* 2017-2022;

DOI: 10.24818/18423264/52.2.18.16

[8] **Lahiri, S.K.,Ghanta, K.C. (2009),** ***Support Vector Regression with Parameter Tuning Assisted by Differential Evolution Technique: Study on Pressure Drop of Slurry Flow in Pipeline;*** *Korean Journal of Chemical Engineering* 26(5), 1175–1185;
[9] **Maronna, R.A., Martin, R.D., Yohai, V.J. (2006),** ***Robust Statistics: Theory and Methods;*** New York: *John Wiley & Sons*;
[10] **Nishiguchi, J., Kaseda, C., Nakayama, H., Arakawa, M., Yun Y.(2010),** ***Modified Support Vector Regression in Outlier Detection*****;** *Neural Networks (IJCNN); The 2010 International Joint Conference on IEEE*; Barcelona, Spain, 2750–2754;
[11] **Rojo-Álvarez, J.L., Martínez-Ramón, M., Figueiras-Vidal, A.R., García-Armada, A., Artés-Rodríguez A. (2003),** ***A Robust Support Vector Algorithm for Nonparametric Spectral Analysis;*** *IEEE Signal Processing Letters*, 10(11), 320–323;
[12] **Rousseeuw, P.J., Leroy, A.M. (1987),** ***Robust Regression and Outlier Detection;*** New York: *John Wiley & Sons;*
[13] **Smola, A.J.,Schölkopf, B. (2004),** ***A Tutorial on Support Vector Regression;*** *Statistics and Computing*, 14(3), 199–222;
[14] **Suykens, J.A., De Brabanter, J., Lukas, L., Vandewalle, J. (2002),** ***Weighted Least Squares Support Vector Machines: Robustness and Sparse Approximation;*** *Neurocomputing*, 48(1), 85–105;
[15] **Üstün, B., Melssen, W., Oudenhuijzen, M., Buydens, L. (2005),** ***Determination of Optimal Support Vector Regression Parameters by Genetic Algorithms and Simplex Optimization;*** *Analytica Chimica Acta*, 544(1), 292–305;
[16] **Vapnik, V. (1995),** ***The Nature of Statistical Learning Theory; Data Mining and Knowledge Discovery;*** New York: *Springer;*
[17] ***Vapnik, V.*** **(2000),** ***The Nature of Statistical Learning Theory;*** New York: *Springer;*
[18] **Williams, G. (2011),** ***Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery;*** New York: *Springer;*
[19] **Yang, H., Huang, K., Chan, L., King, I.,Lyu, M. R. (2004),** ***Outliers Treatment in Support Vector Regression for Financial Time Series Prediction;*** *Neural Information Processing*, 1260–1265;
[20] **Zong, Q., Liu, W., Dou L,(2006),** ***Parameters Selection for SVR Based on PSO. Intelligent Control and Automation;*** *2006.WCICA 2006.The Sixth World Congress on IEE*, 2811–2814.

DOI: 10.24818/18423264/52.2.18.16