Al-Mustansiriyah Journal of Science
ISSN: 1814-635X (print), ISSN:2521-3520 (online)
Volume 30, Issue 1, 2019
DOI: http://doi.org/10.23851/mjs.v30i1.388

Research Article Open Access

Robust Variable Selection Technique for Single Index Support Vector Regression Model

Waleed Dhhan[1], Thaera Alameer[2*]
[1] Babylon Municipalities, Ministry of Municipalities and PublicWorks, Babylon, IRAQ
[2] Department of Statistics, College of Administration and Economics, Mustansiriyah University, IRAQ
*Correspondent author email: tha_alameer@uomustansiriyah.edu.iq

ArticleInfo

Received
03/12/2017

Accepted
02/01/2018

Published
15/08/2019

Abstract

The single index support vector regression model (SI-SVR) is a useful regression technique used to alleviate the problem of high-dimensionality. In this study, we propose a robust variable selection technique for the SI-SVR model by using vital method to identify and minimize the effects of outliers in the data set. The effectiveness of the proposed robust variable selection technique of the SI-SVR model is explored by using various simulation examples. Furthermore, the suggested method is tested by analyzing a real data set which highlights the utility of the proposed methodology.

Keywords: Single-index model, Support vector regression, Variable selection, High-dimensional, Outliers.

الخلاصة

انموذج متجهات دعم الانحدار ذات المؤشر الواحد هو تقنية انحدار مفيدة للتخفيف من مشكلة الابعاد العالية. في هذه الدراسة نقترح تقنية اختيار المتغيرات الحصينة لانموذج متجهات دعم الانحدار ذات المؤشر الواحد وذلك باستخدام طريقة فعالة لتحديد وتقليل آثار القيم المتطرفة في مجموعة البيانات. اذ تم استكشاف فعالية تقنية اختيار المتغيرات الحصينة لانموذج متجهات دعم الانحدار ذات المؤشر الواحد المقترحة باستخدام امثلة محاكاة مختلفة. علاوة على ذلك، تم اختبار الطريقة المقترحة من خلال تحليل مجموعة بيانات الحقيقية والتي تسلط الضوء على فائدة المنهجية المقترحة.

Introduction

The single index model (SIM) is a promising regression technique has been proposed with a smooth unknown link function $F(.)$, to alleviate the problem of high-dimensionality (Ichimura, 1993). In the last few years, it has gained much attention of many researchers because of its excellent performance to address the problem of high dimensionality. The main idea of the single-index model is to achieve dimension reduction by finding one variable which aggregates the dimension of predictors. it is semi-parametric model involves of two parts, parametric and nonparametric which result in hybrid assumptions. These hybrid assumptions of the SIM are weaker than the parametric assumptions and the same time stronger than the fully nonparametric model assumptions which provide high accuracy of parametric model and the flexibility of a nonparametric model (Horowitz, 2009). Therefore, the SIM minimizes the risk of mis-specification relative to the parametric model and at the same time avoid drawbacks of fully nonparametric models such as the lack of capability of extrapolation, and the difficulty of interpretation. In general, the SIM combines the high precision of the parametric model with the flexibility of the nonparametric model. However, the high accuracy of the parametric part will be uncertain when the data sets have redundant variables. Furthermore, it cannot be applied for less than full rank data. To avoid these limitations, the variable selection procedure has been proposed by Dhhan et al (2017). They succeeded to propose the Elastic net for single index support vector regression (ENSI-SVR) efficiently. Variable selection is very important technique in regression models estimation and it has been used to estimate the

models of the data sets which have tens or hundreds or more of variables are available (Guyon, and Elisseeff, 2003). In most applications, the data sets consist of redundant predictor which leads to deteriorate the generalization ability of the model. Hence, variable selection is used to reduce the number of original variables by selecting their significant subset which still keeps the generalization ability in comparison with that of the original predictors.

The ENSI-SVR as proposed by Dhhan et al (2017) used the semi-parametric least squares (SLS) and weighted semi-parametric least squares (WSLS) to estimate the vector of parameters β and the SVR tool to evaluate the unknown link function $F(.)$, of the SIM. This technique is represented an extension of the SIM which was proposed by Ichimura et al. (1993). Nevertheless, this approach is not robust against outliers which appear in most real life applications.

Generally, classical methods are not resistant to outlying samples; even one outlier can destroy the estimator. An outlier can be defined as an observation which is significantly different from the majority of the data points. These data points which deviates so much from the other points are affect badly to the model estimation. Robust statistics have recently appeared as a family of techniques to estimate regression models in the presence of outliers. it work on to minimize the effects of outliers in the data sets by giving down weights of abnormal data points that may reach zero at times.

In this paper, we proposed an extension of the ENSI-SVR model of Dhhan et al. (2017) by considering the fixed parameter support vector regression method (FP-SVR) for detecting and minimizing the effects of outliers in the data (Dhhan et al. 2015). The advantage of using the FP-SVR is that it can detect outliers and bad leverage points efficiently.

The rest of this paper is organized as follows: A brief description of the proposed method based on FP-SVR is given in Section 2. In Section 3 and 4, we explore the effectiveness of the proposed method using real and simulation data sets. Finally, the discussion of the results is summarized in Section 5.

Materials and Methods

The ENSI-SVR method

The Elastic Net Single index support vector regression as proposed by Dhhan et al. (2017) is to achieve simultaneous variable selection, parameter estimation and mean regression estimation. Let we have the following SIM to illustrate the ENSI-SVR.

$$y = F(x^T\beta_{ENSI}, r) \tag{1}$$

where y is the dependent variable and, x is the vector of predictors, β_{ENSI} is the vector of parameters, r is the vector of residuals, $F(\cdot)$ is the unknown nonparametric link function

The vector of parameters β_{ENSI}, is estimated by using the following minimization problem

$$\begin{aligned}\hat{\beta}_{ENSI} = \arg\min_{\beta_{ENSI}} \{&\|y - f(x^T\beta_{ENSI})\|^2 \\ &+ \lambda_2\|\beta_{ENSI}\|_2^2 \\ &+ \lambda_1\|\beta_{ENSI}\|_1\}\end{aligned} \tag{2}$$

where $f(x^T\beta_{ENSI})$ is the unknown link function estimator of $F(x^T\beta_{ENSI})$, λ_1 and λ_2 correspond to the penalties of Lasso and Ridge regression, respectively that control the amount of regularization applied to the estimation.

As the vector of parameters β_{ENSI} is inside the link function, the first step is to estimate it efficiently. For this aim, the semi-parametric least squares (SLS) which is used (Ichimura 1993) since it provides $1/\sqrt{n}$ - consistency rate.

After estimating the vector of parametrs $\hat{\beta}_{ENSI}$, we can calculate the single index as

$$x^* = x^T\hat{\beta}_{ENSI} \tag{3}$$

The second step of this method is to use the support vector regression to evaluate the nonparametric link function $F(.)$.

$$f(x^*) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)k(x^*{}_i.x^*) + b \tag{4}$$

Robust Proposed method

In this article, the FP-SVR method has been employed to identify and minimize the effects of outliers and bad leverage points (Dhhan et al. 2015). The advantage of this technique is that it succeeded to introduce fixed optimal

parameters of the SVR model (5). It recommended using the parameters of the SVR model as: C *=100000,* ε *=0, and h=1.*

$$z = \hat{y}_{SVR} = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) k(x_i . x) + b \, , \alpha_i, \alpha_i^* \epsilon \, [0, C] \quad (5)$$

According to this approach, any point with a z value greater than the cut-off point (6) is considered to be an outlier

$$CP = 2Median|z| + 2\sqrt{\pi var(z)/2n} \quad (6)$$

In order to prevent the effects of outliers and bad leverage points, we can use the following weight function which is built based the FP-SVR

$$\tau_i = min[1, CP/z] \quad (7)$$

Based on τ function, we can rewrite Eq. (4) to achieve the final robust ENSI-SVR function.

$$f(x^*) = \sum_{i=1}^{n} \tau_i \, (\alpha_i - \alpha_i^*) k(x^*_i . x^*) + b \quad (8)$$

Real case study

In this subsection, the near infrared spectroscopy (NIR) data has been used to evaluate the proposed method (RESI-SVR) in case of rank deficient data. This data consists of 235 Variables (p=235) contain the first derivatives of near infrared spectroscopy absorbance values with 166 alcoholic fermentation mashes (n=166) of various feedstock (wheat, rye and corn), and two variables (output) containing the concentration of ethanol and glucose. In this example we have used the ethanol as dependent variable (Y) with all 235 input variables (X).

The results of applying ENSI-SVR and RESI-SVR are summarized in Table 1 and Figure 1. with combination of parameters (C, ε and h) by three values for each parameter.

According to these results, the proposed method achieved low levels of MSE for different values of the parameters, whereas the ENSI-SVR method achieved high levels of MSE for all of the parameters values which reflects the superiority of the proposed RESI-SVR method.

Table 1: The MSE of ENSI-SVR and RESI-SVR methods for NIR data

Parameters		EN-SVR			RE-SVR		
		C=1	*C=50*	*C=100*	*C=1*	*C=50*	*C=100*
h =0.5	*ε =0.0*	0.8702	0.5264	0.5529	0.0870	0.0536	0.0532
	ε =0.1	0.8543	0.5504	0.5218	0.0954	0.0530	0.0581
	ε =0.2	0.9234	0.5403	0.5257	0.0928	0.0590	0.0555
h =1	*ε =0.0*	1.2340	0.5982	0.6477	0.4234	0.0508	0.0627
	ε =0.1	1.1220	0.7571	0.7501	0.3220	0.0737	0.0753
	ε =0.2	1.6254	0.8247	0.7410	0.2254	0.0814	0.0841
h =5	*ε =0.0*	2.4288	2.1234	0.6770	0.2288	0.2214	0.2970
	ε =0.1	2.6321	2.4147	0.8486	0.2121	0.2127	0.2466
	ε =0.2	2.9360	1.7871	0.7927	0.2860	0.1879	0.1329

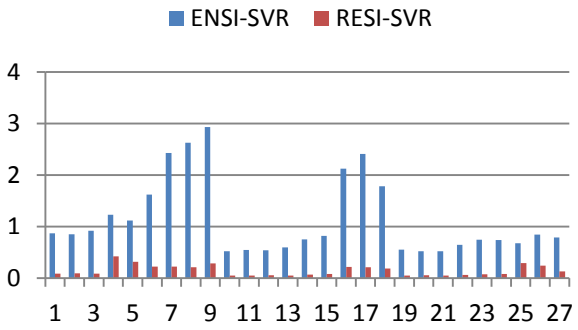

Figure 1: The MSE of ENSI-SVR and RESI-SVR methods for NIR data

Simulations examples

In this section, the effectiveness of the proposed method RESI-SVR is investigated using two simulation data sets to get rid the problem of high-dimensional in the presence of outliers in terms of the accuracy of prediction and dimension reduction. The proposed method evaluated using the prediction risk (MSE) which is averaged over 100 replications. The models are built using the training samples which account for 70% of the data sets and it testied using 30% of the data sets as testing samples. Both of two examples are contaminated with 10% percentage. The

residual term r is contaminated by arbitrary large number equal to 100.

Simulation I

In order to illustrate the proposed robust method, let we have the single index model, $y = f(x^T\beta) + r$ with 20 predictors (Hu et al. 2013; Wu et al. 2010). The nonlinear function is $f(\cdot) = 5\cos(\cdot) + \exp(-\cdot^2)$, and the vector of parameters is $\beta = \beta_0/\|\beta_0\|$, $\beta_0 = (\beta_1, \beta_2, \dots, \beta_{20})^T$, β_1=β_2,…,= $\beta_5 = 1$, β_6=β_7,…,= $\beta_{20} = 0$. The vector of predictors x is sampled from uniform distribution with parameters [0,1], while the vector of residuals r is generated from the exponential distribution with parameter $[\theta = 1/2]$. In this example, three different sample sizes have been
used to compare the proposed method with ENSI-SVR method. The results of applying the compared methods are summarized in Table 2. It consists of the prediction risk (MSE) of proposed method compared with ENSI-SVR by using two different values of the SVR parameters.

Table 2: The MSE of ENSI-SVR and RESI-SVR methods for 20 predictors.

n	*Parameters*	EN-SVR				RESI-SVR			
		ε =0		ε =0.2		ε =0		ε =0.2	
		C=1	*C=10*	*C=1*	*C=10*	*C=1*	*C=10*	*C=1*	*C=10*
25	*h=1*	3.5298	6.6202	3.9282	7.1773	1.6879	2.2529	1.7068	2.8122
	h=5	3.0298	6.1202	3.2282	7.8605	1.4924	1.1865	1.5088	0.7576
50	*h=1*	6.1741	9.0908	7.1643	6.0451	2.2791	1.6667	2.1321	1.2417
	h=5	5.1907	8.8321	6.1543	8.5620	2.2575	1.4572	2.1986	1.0803
100	*h=1*	9.1630	7.8721	6.6541	5.6871	5.0146	2.8546	5.0702	2.6952
	h=5	9.8762	6.9501	8.0932	5.0567	5.3805	4.3200	6.1506	4.5703

According to Figure 2, the proposed RESI-SVR is achieved the minimum values of the MSE compared to the ENSI-SVR method for all combinations of samples and SVR parameters. This refers to the superiority of the proposed RESI-SVR method over ENSI-SVR method.

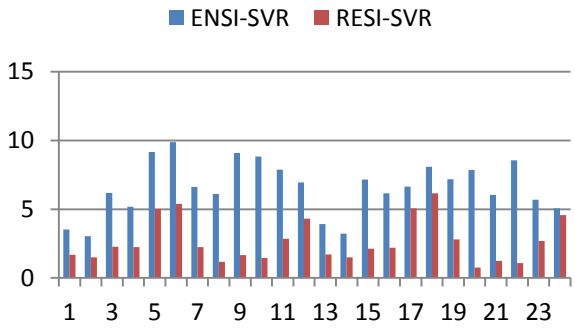

Figure 2: The MSE of ENSI-SVR and RESI-SVR methods for 20 predictors.

Simulation II

In the example, we have used the single index model, $y = sin(x^T\beta) + 0.1\,r$ (Peng and Huang, 2011). It should be noted that 40 predictors and three sample sizes have been used (25, 50, and 100) to evaluate the proposed method. The first sample size is less than full rank since the number of predictors; p is larger than sample size, n. The vector of predictors x and the vector of residuals r are generated based on standard normal distribution. The set of parameters that used in this example is $\beta = \beta_0/\|\beta_0\|$,$\beta_0 = (\beta_1, \beta_2, \dots, \beta_{40})^T = (3, 1.5, 0, 0, 2, 0, \dots\dots 0)^T$. The results of applying methods of RESI-SVR and ENSI-SVR are presented in Table 3. These results are illustrated graphically in Figure 3. According to Table and Figure 3, the proposed RESI-SVR method is achieved values of the MSE lower than the ENSI-SVR method for all combinations of samples and SVR parameters which reflects the superiority of the proposed RESI-SVR over ENSI-SVR method.

Table 3: The MSE of ENSI-SVR and RESI-SVR methods for 40 predictors.

n	***Parameters***	**ENSI-SVR**				**RESI-SVR**			
		ε =0		**ε =0.2**		**ε =0**		**ε = 0.2**	
		C=1	***C=10***	***C=1***	***C=10***	***C=1***	***C=10***	***C=1***	***C=10***
25	***h=1***	**14.738**	**12.354**	**14.619**	**12.441**	**1.1010**	**1.6543**	**5.1012**	**5.1012**
	h=5	**15.032**	**13.436**	**14.985**	**13.363**	**5.0091**	**3.6081**	**1.5176**	**1.5176**
50	***h=1***	**6.9837**	**5.9738**	**6.7114**	**5.8625**	**0.9517**	**0.9338**	**4.8702**	**3.9890**
	h=5	**9.2996**	**6.2061**	**8.6109**	**6.4301**	**0.9765**	**2.9876**	**1.9807**	**1.9112**
100	***h=1***	**3.3431**	**4.2109**	**3.8902**	**4.8931**	**4.0231**	**0.8891**	**2.7311**	**2.0612**
	h=5	**3.1160**	**1.7726**	**3.8724**	**2.3910**	**1.7320**	**1.6210**	**1.0987**	**1.0012**

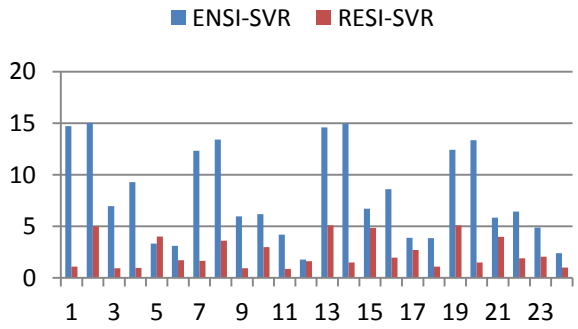

Figure 3: The MSE of ENSI-SVR and RESI-SVR methods for 40 predictors.

Conclusions

The robust variable selection for the single-index model, RESI-SVR has been proposed to achieve simultaneously outlier detection and dimension reduction. The key to the success of the robust proposed method is the use of the FP-SVR to detect and minimize outliers and leverage points. The proposed method, RESI-SVR is tested using simulation and real data sets with same set of parameters and sample sizes. Finally, the comparative results refer to the superiority of our proposed method, RESI-SVR over existing ENSI-SVR method to dispose of abnormal data and reduce the curse of high dimensionality.

References

[1] Dhhan,W., Rana, S. and H. Midi (2015), Non-sparse ε-Insensitive Support Vector Regression for Outlier Detection; Journal of Applied Statistics, 42(8), 1723-1739.

[2] Dhhan, W., Rana, S. Alshaybawee, T. and H. Midi (2017), Elastic Net For Single Index Support Vector Regression Model." Economic Computation & Economic Cybernetics Studies & Research 51(2).

[3] Guyon, I. & Elisseeff, A. (2003), An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 3, 1157-1182.

[4] Horowitz, J. L. (2009), Semiparametric and Nonparametric Methods in Econometrics. Springer.

[5] Hu, Y., Gramacy, R. B. & Lian, H. (2013), Bayesian Quantile Regression for Single-index Models. Statistics and Computing, 23(4), 437-454.

[6] Ichimura, H. (1993), Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models. Journal of Econometrics, 58(1), 71-120.

[7] Peng, H. & Huang, T. (2011), Penalized Least Squares for Single Index Models. Journal of Statistical Planning and Inference, 141(4), 1362-1379.

[8] Wu, T. Z., Yu, K. & Yu, Y. (2010), Single-index Quantile Regression. Journal of Multivariate Analysis, 101(7), 1607-1621.