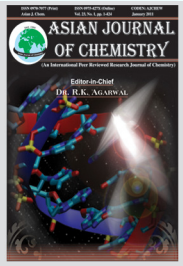
Asian Journal of Chemistry; Vol. 23, No. 3 (2011), 992-1000

ASIAN JOURNAL OF CHEMISTRY

www.asianjournalofchemistry.co.in



Comparison of Least Square Support Vector Machine-Based Calibration Methods in Diesel Property Analysis by Near Infrared Spectroscopy

TU-NAN DAN and LIAN-KUI DAI*

State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310027, Zhejiang, P.R. China

*Corresponding author: E-mail: lkdai@iipc.zju.edu.cn

(*Received*: 25 January 2010; *Accepted*: 28 October 2010) AJC-9220

Partial least square (PLS) and least square support vector machine (LSSVM) are widely applied in NIR calibration modeling. The aim of this paper is to compare three kinds of LSSVM based calibration methods in diesel property by NIR analysis, which are regular LSSVM, LSSVM with feature extraction by principal component analysis (PCA) and PLS. Thirty nine diesel samples with known properties which include calculated cetane index, density, total sulfur and T50 (boiling point at 50 % recovery), are collected from a refinery in China. Their near infrared spectra are measured by a spectrometer with the wavelength range of 900-1700 nm. They are divided into a calibration subset with 29 samples and a validation subset with 10 samples. The above LSSVM based calibration methods as well as PLS are employed to build models with the calibration samples and tested with the validation samples. Experimental results show that the LSSVM with PLS feature extraction presents the best performances in all of the calibration methods. Its root mean squared error of prediction (RMSEP) of diesel calculated cetane index, density, total sulfur and T50 are 0.45, 3.19, 0.0482 and 3.90, respectively and the corresponding multiple correlation coefficients of prediction (R^2) are 0.953, 0.924, 0.974 and 0.954, respectively.

Key Words: NIR spectroscopy, Support vector machine, Diesel, Quantitative analysis.

INTRODUCTION

Diesel is a widely used vehicle fuel which needs accurate quantitative analysis. There are several properties that affect the quality of diesel such as cetane number, density, total sulfur and T50 (boiling point at 50 % recovery) *etc*. However, the standard measurement methods such as ASTM D-613 and ASTM D-6890 are expensive, complicated, time consuming and sometimes pollutive to the environment. For example, cetane number, one of the most important indices to evaluate the combustion performance of diesel oil, is measured by ASTM D-613 method. This method requires a special industry test engine and operates under standard conditions. The analytical equipment is expensive and needs frequent maintenance and the analytical procedure is time-consuming. Therefore, the traditional method is not suitable for fast analysis required by modern petroleum industries.

Near infrared (NIR) spectroscopy technology is an indirect analytical method, which has been well developed since the late 1980s. Near infrared spectroscopy has been successfully used in the property quantitative analysis and classification of gasoline[1-5], diesel[6,7], biodiesel[8,9] and alcohol fuel[10,11].

In near infrared spectroscopy, multivariate calibration methods are widely applied in quantitative analysis. Multivariate calibration methods can be divided into two parts *i.e.*, linear and non-linear methods. Linear calibration methods include multiple linear regression (MLR), principal component regression (PCR), partial least square (PLS) regression, *etc*. and non-linear calibration methods include standard support vector machine (SVM)[12], least square support vector machine (LSSVM)[13,14], artificial neural network (ANN)[15,16], *etc*.

Linear and non-linear calibration methods have different features. In linear calibration methods, model training and parameter optimization are relatively easy and the model structure is simple. However, when dealing with strong non-linear relationship between instrument responses and predicted properties, linear model cannot predict accurately. In order to improve model performance, local modeling approach can sometimes be used, but it is still unsatisfactory if there are few training samples in the neighbor of the test sample. Non-linear calibration methods can overcome the above problem, which have been widely used to build near infrared calibration models. Least square support vector machine is one of the most commonly used non-linear calibration methods in recent years, because it has better prediction performances than artificial neural network and costs little calculation than standard support vector machine.

In general, there are strong non-linear relationship between the near infrared spectral data for a set of samples and their properties, so LSSVM can directly be used to build the calibration model. This paper is to compare three kinds of LSSVM based calibration methods in diesel property NIR analysis, which are regular LSSVM, LSSVM with feature extraction by PCA and PLS. The above LSSVM based calibration methods as well as PLS are employed to build models with the calibration samples and tested with the validation samples. Experimental results show that the LSSVM with PLS feature extraction presents the best performances in all of the calibration methods.

Theory

Least square support vector machine (LS-SVM): Given a calibration set of N data points $\{x_k, y_k\}_{k=1}^{N}$, , where $x_k \in R^m$ is the regression vector and $y_k \in R$ is the output. It can be constructed to estimate the unknown function between the regression vector and output as follow form:

$$y = w^T\varphi(x) + b \tag{1}$$

where the vector w and the constant b are the parameters to be identified, $\varphi(x)$ is a non-linear function which map the input space R^m to a higher dimension feature space. According to the structural risk minimization principle, we can define optimization problem as follows:

$$\min_{w,b,e} J(w,b,e) = \frac{1}{2}w^Tw + \frac{1}{2}\gamma\sum_{k=1}^{N} e_k^2 \tag{2}$$

subject to $$y_k = w^T\varphi(x_k) + b + e_k,\ k = 1, \ldots, N \tag{3}$$

where e_k is the error between the actual output and the predictive output of the k^{th} data. The Lagrangian function could be established

$$L(w,b,e,\alpha) = J(w,b,e) - \sum_{k=1}^{N}\alpha_k\{w^T\varphi(x_k) + b + e_k - y_k\} \tag{4}$$

where $\alpha_k \in R,\ k = 1, \ldots, N$ are Lagrangian multipliers. The solution of α and b could be given by computing following equations:

$$\begin{bmatrix} 0 & I_v^T \\ I_v & \Omega + \gamma^{-1}I \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ i \end{bmatrix} \tag{5}$$

where $Y = [y_1, \ldots, y_N]^T$, $1_v = [1, \ldots, 1]^T$, $\alpha = [\alpha_1, \ldots, \alpha_N]^T$,

and $$\Omega_{ij} = \varphi(x_1)^T\varphi(x_j) = K(x_i, x_j),\ i, j = 1,2, \ldots, N \tag{6}$$

the LSSVM model can be given by

$$y(x) = \sum_{i=1}^{N}\alpha_i K(x, x_i) + b \tag{7}$$

where $K(x, x_i),\ i = 1, \ldots, N$ is any kernel function satisfying the Mercer condition.

RBF kernel function is one of the most commonly used kernel function, it can be expressed as

$$K(x_i, x_j) = \exp\left(-\frac{\| x_i - x_j \|^2}{\sigma^2}\right) \tag{8}$$

where σ^2 is the width parameter which controls the kernel function radial scope.

Since σ^2 changes too much when dealing with different calibration samples, the following RBF kernel function is proposed in this paper,

$$K(x_i, x_j) = \exp\left(-\frac{\| x_i - x_j \|^2}{AD\cdot\sigma^2}\right) \tag{9}$$

$$AD = \frac{1}{N}\sum_{i=1}^{N}\| x_{train}(i) - \bar{x}_{train} \|^2 \tag{10}$$

where $\bar{x}_{train}$ is the center of the calibration subset samples.

Least square support vector machine with feature extraction by principal component analysis and partial least square: Mostly, the spectral data is with high-dimension. The spectral dimension should be reduced by feature extraction, not only because dimension reduction can decrease the computation time, but also because feature extraction can eliminate noise information and improve model prediction accuracy. Two feature extraction methods, PCA and PLS, are usually applied in LSSVM model. For simplicity, the resulting models are named PCA-LSSVM and PLS-LSSVM[17]. Their detail structures are shown in Fig. 1, where X_{train} is the instrument response matrix of calibration subset samples, Y_{train} is their predicted property vector, T_{train} is the principal component matrix of X_{train}, P_{train} is the loading matrix of X_{train}, W_{train} is the correlation coefficient matrix of X_{train} and Y_{train}, X_{test} is the instrument responses of a validation sample, T_{test} is the principal component vector of X_{test} and $Y_{predict}$ is the property prediction of the validation sample.

In the PCA method, the PCA score vector of the validation sample spectral data is firstly calculated and then fed to the LSSVM model. Since the PCA score contains only the spectral information, the correlation between the PCA score and the predicted property is not strong. In the PLS method, PLS score is used as the input of LSSVM model instead of PCA score. Partial least square score contains both of the spectral information and the predicted property information, so it has strong correlation with the predicted property.

Suppose X_{train} is an $N \times M$ matrix, Y_{train} is an $N \times 1$ vector, N is the number of calibration samples, M is the dimension of spectral data and f is the factor number of PCA and PLS feature, then the main processes of PLS-LSSVM and PCA-LSSVM can be expressed as follows.

Partial least square-least square support vector machine modeling and prediction algorithm: Step-1: Obtain the principal component matrix T_{train}, the loading matrix P_{train} and the correlation coefficient matrix W_{train} by applying PLS1 algorithm. **Step-2:** Calculate the LSSVM model coefficient α and b in eqn. 7 by T_{train} and Y_{train}. **Step-3:** For a validation sample with the spectrum X_{test}, compute T_{test} and $Y_{predict}$ by the follow equations

$$T_{test} = X_{test}W_{train}(P_{train}^TW_{train})^{-1},$$

$$Y_{predict} = \sum_{k=1}^{N}\alpha_k K(T_{train}(k), T_{test}) + b \tag{11}$$

where T_{train} is the k^{th} row of T_{train}.

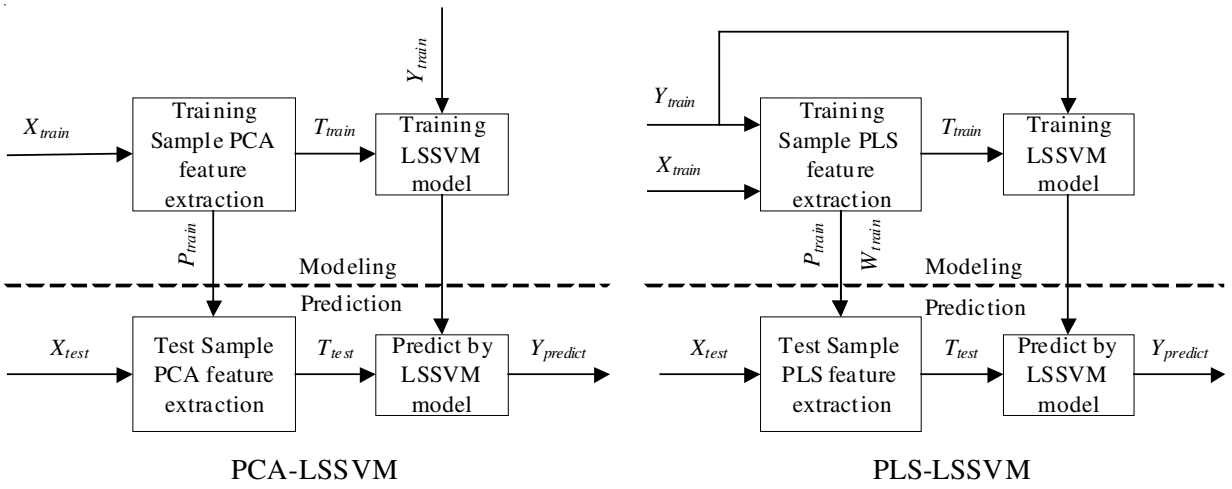


Fig. 1. Structures of principal component analysis-least square support vector machine and partial least square-least square support vector machine

Principal component analysis-least square support vector machine modeling and prediction algorithm: Step-1: Obtain the principal component matrix T_{train}, the loading matrix P_{train} by applying PCA algorithm. **Step-2:** Calculate the LSSVM model coefficient α and b in eqn. 7 by T_{train} and Y_{train}. **Step-3:** For a validation sample with the spectrum X_{test}, compute T_{test} and $Y_{predict}$ by the follow equations

$$T_{test} = X_{test} P_{train},$$

$$Y_{prodict} = \sum_{k=1}^{N} \alpha_k K(T_{train}(k), T_{test}) + b \quad (12)$$

where T_{train} is the k^{th} row of T_{train}.

EXPERIMENTAL

A group of 39 diesel samples were obtained from a refinery in China. The diesel properties include calculated cetane index, density, T50 (boiling point at 50 % recovery, ºC) and total sulfur content and their actual values were measured by ASTM reference methods.

The NIR spectra were obtained by a BTC261E spectrometer (B &WTEK, USA) over a wavelength range of 900-1700 nm. The nominal spectral resolution is 1.5 nm and the cell material is quartz (10 mm optical path). The dark spectra and the reference spectra are obtained first. Both of the dark and reference spectrums were measured 20 times. Their average spectra are used to calculate the absorbance spectra of samples. Each sample is scanned 20 times with 2 m integral time and the average stands for its spectrum.

Spectral preprocessing: The original absorbance spectra should be preprocessed to improve the signal noise ratio (SNR). In this case, the following preprocessing steps are used. Firstly, the absorbance spectra in the wavelength range between 1000 and 1600 nm are selected. Secondly, Savitzky-Golay convolution method[18] is used to obtain the second derivative spectra and the polynomial filter width is 45 nm. At last, the second derivative spectra are calculated by standard normal variation (SNV)[19] approach. The spectra before and after the preprocessing are shown in Fig. 2.

Calibration subset samples selection and model evaluation criterion: All 39 samples are divided into two subsets: the calibration subset and the validation subset. In order to ensure that the calibration subset has the similar range and distribution of reference value as the validation subset, all samples are sorted by the reference value, 29 samples are selected at equal intervals to be calibration subset and the remaining 10 samples to be validation subset.

In order to compare different calibration models, root mean squared error of prediction (RMSEP) and the correlation coefficient (R^2) are chosen to evaluate model performances. The RMSEP and R^2 are defined as follows:

$$RMSEP = \sqrt{\frac{1}{np}\sum_{k=1}^{np}(y(k) - y_p(k))^2},$$

$$R^2 = 1 - \frac{\sum_{k=1}^{np}(y(k) - y_p(k))^2}{\sum_{k=1}^{np}(y(k) - \bar{y})^2} \quad (13)$$

where y(k) and $y_p(k)$ represent the actual value and predictive value of the kth sample, np is the number of the validation subset samples, $\bar{y}$ is the average value of y(k) in the validation subset.

Software: The software platform of the experiment in this paper is Matlab 7.5. It works on a PC computer with 1.80 GHz Intel processor and Windows XP operation system. All the programs used in this paper are written by us.

RESULTS AND DISCUSSION

To compare the prediction results, we apply PLS, LSSVM, PCA-LSSVM and PLS-LSSVM to build calibration models with the calibration subset samples, respectively and then evaluate these models with the validation subset samples. The kernel function of LSSVM, PCA-LSSVM and PLS-LSSVM is the improved RBF function defined in eqn. 9.

The influences of the model parameters on the calibration models are shown in Figs. 3-6. The standard error of leave-one-out cross-validation (SECV) is introduced as the evaluation criterion. It is found that all of the calibration model parameters are interacted with each others, we chose those parameters

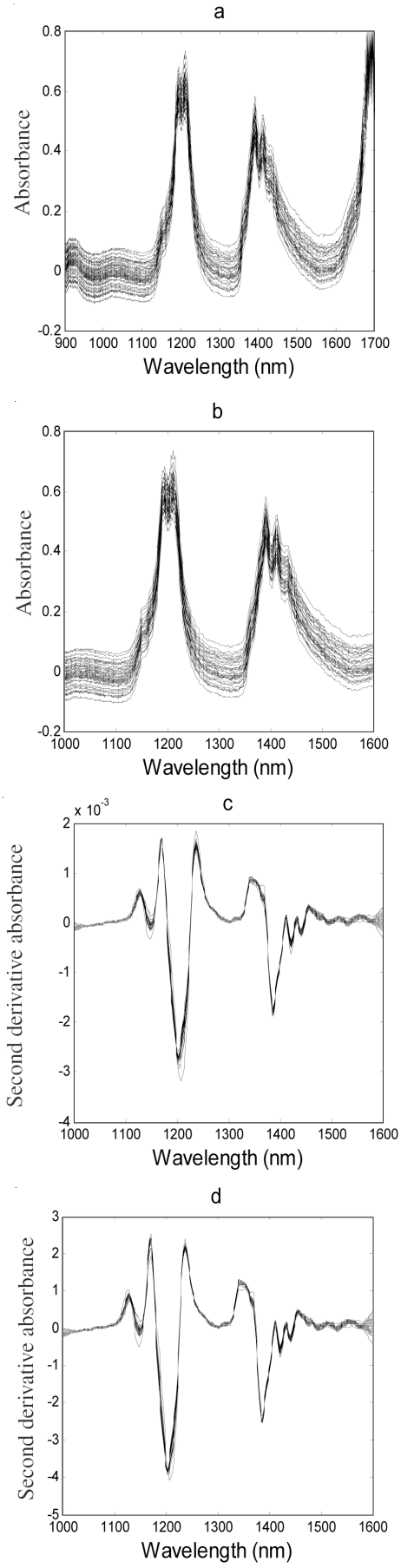


Fig. 2. (a) Original absorbance spectra, (b) spectra after wavelength selection, (c) second derivative spectra and (d) spectra treated by SNV

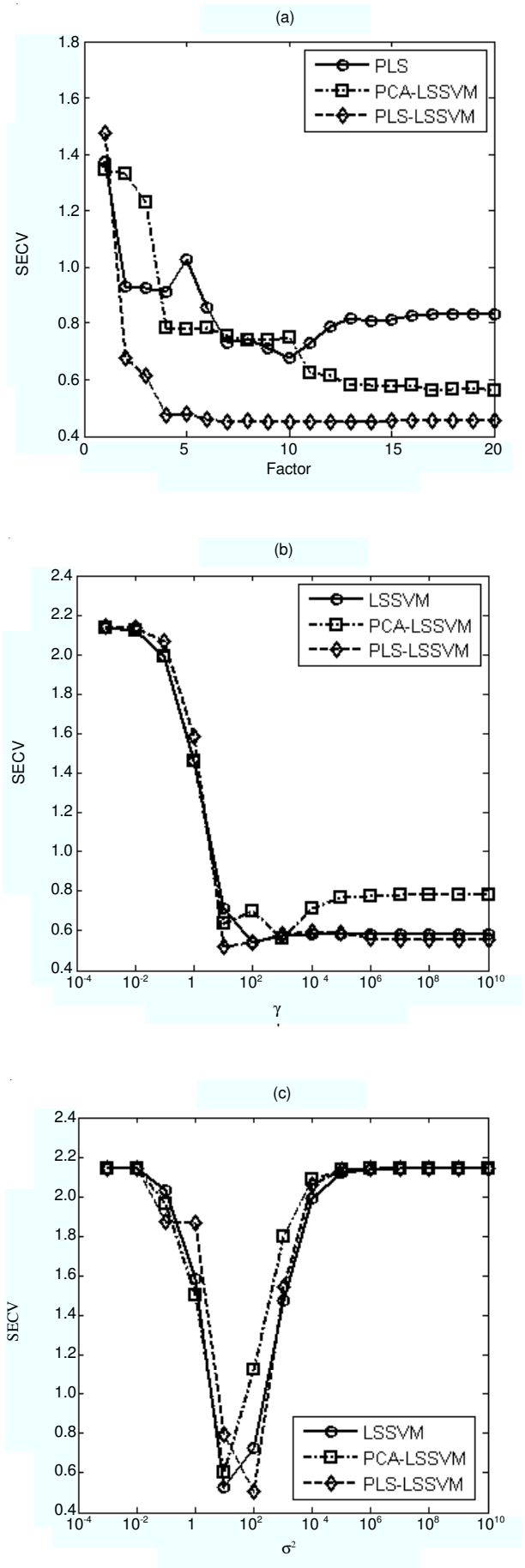


Fig. 3. Influence of parameters on NIR cetane models

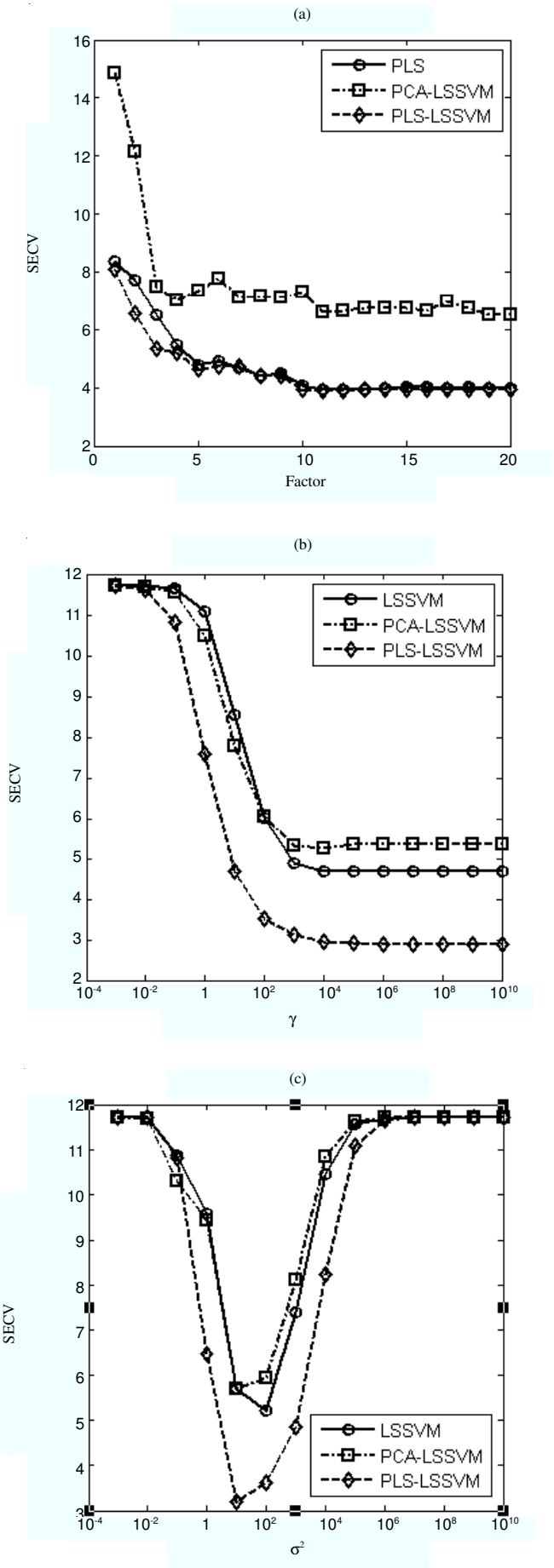


Fig. 4. Influence of parameters on NIR density models

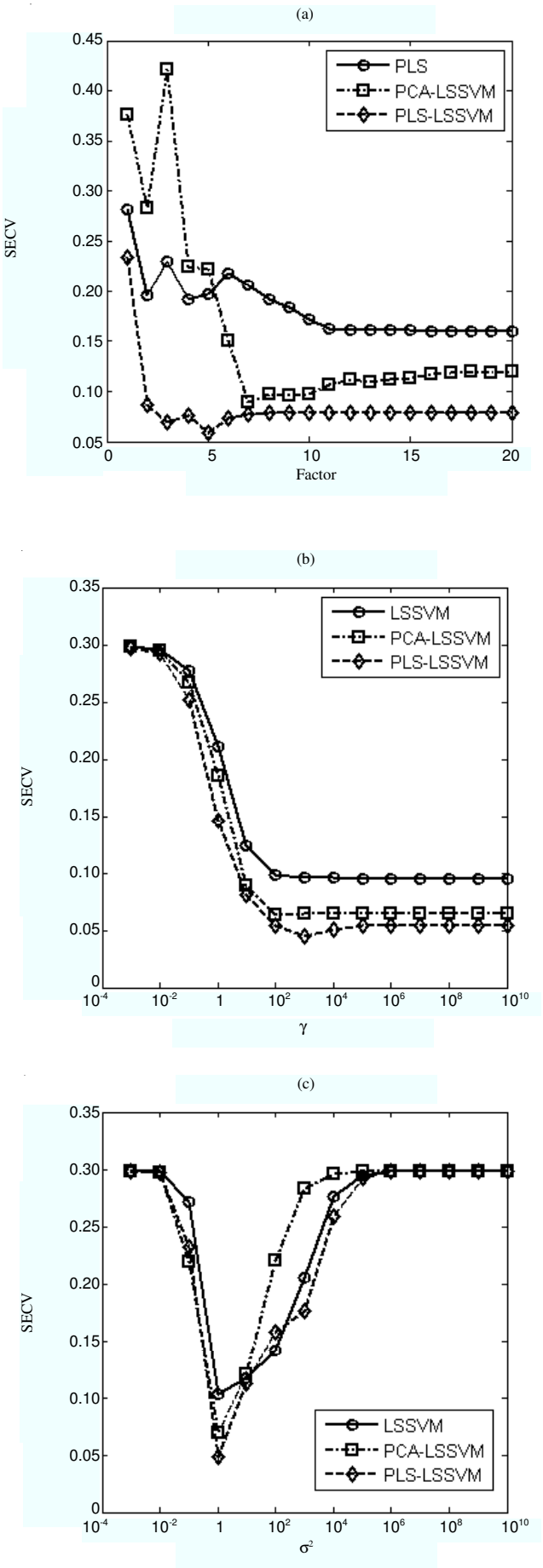


Fig. 5. Influence of parameters on NIR total sulfur models

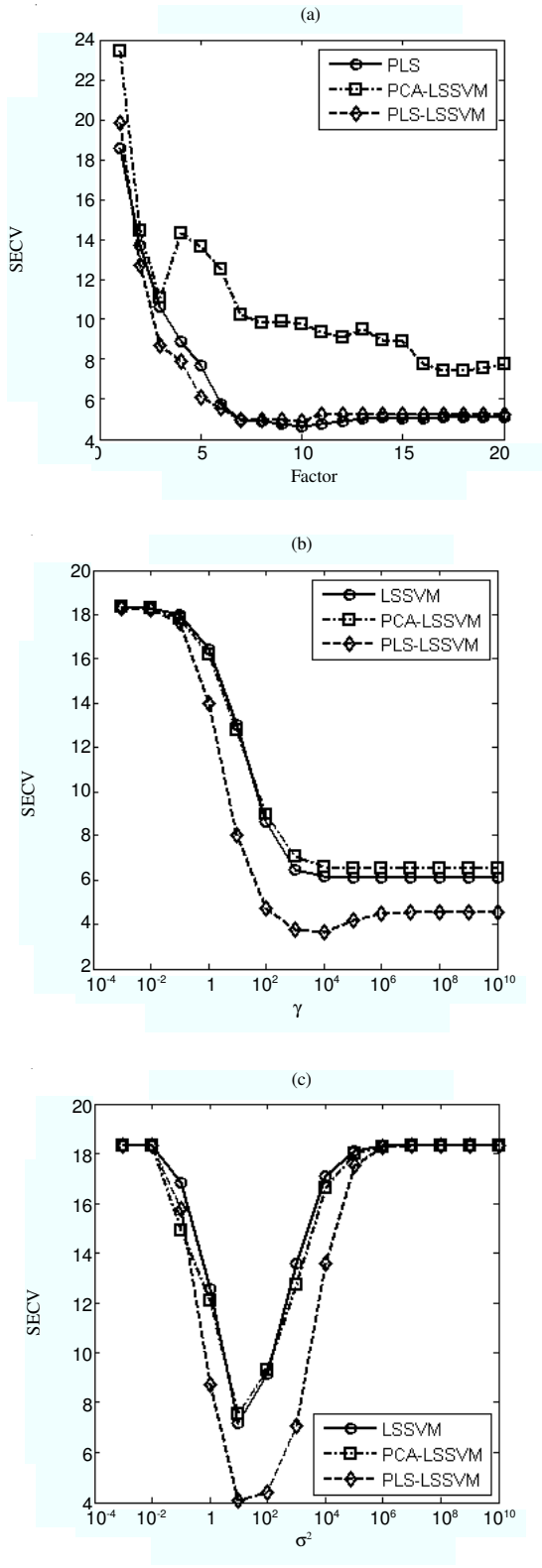


Fig. 6. Influence of parameters on NIR T50 models

which make the SECV of calibration models be relatively low and stable. The optimized model parameters are shown in Table-1.

TABLE-1
FOUR CALIBRATION MODEL PARAMETERS

Diesel property	PLS	LSSVM		PCA-LSSVM			PLS-LSSVM		
	f	y	σ^2	f	y	σ^2	f	y	σ^2
Calculated cetane index	10	50	4	13	15	7	10	18	40
Density (kg/m^3)	11	500	42	11	300	17	11	500	10
Total sulfur (%, w/w)	16	500	2	7	30	1	5	200	1
T50 (ºC)	10	300	12	14	500	7	10	500	20

Table-2 shows the R^2 and the SEP of four calibration methods and the scatter diagrams of real and predictive values of diesel predicted properties are shown in Figs. 7-10, respectively. Results show that PLS-LSSVM performs best in all of the four models. When predicting density and T50, PLS and PLS-LSSVM models present better performances than LSSVM and PCA-LSSVM models, however, when predicting cetane and sulfur, all three LSSVM-based models perform better than PLS model, especially PLS-LSSVM model. This is because that the NIR spectra of diesel have strong linearity with its density and T50 properties and have serious non-linearity with the cetane and sulfur properties. Partial least square-least square support vector machine combines linear and non-linear calibration model. The linear part is to enhance the correlation coefficients of the NIR spectra and the predicted properties and the non-linear part is to find the best fitting function. Therefore, it can be concluded that PLS-LSSVM calibration model performs well in both of the linear and non-linear cases.

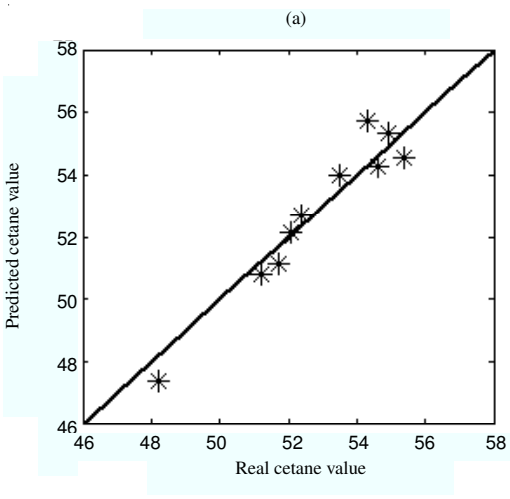


TABLE-2
DIESEL PROPERTIES PREDICTED BY FOUR CALIBRATION MODELS

Diesel property	PLS		LSSVM		PCA-LSSVM		PLS-LSSVM	
	R^2	RMSEP	R^2	RMSEP	R^2	RMSEP	R^2	RMSEP
Calculated cetane index	0.893	0.68	0.944	0.49	0.921	0.58	0.953	0.45
Density (kg/m^3)	0.923	3.22	0.807	5.09	0.766	5.60	0.924	3.19
Total sulfur (%, w/w)	0.782	0.1392	0.897	0.0959	0.947	0.0688	0.974	0.0482
T50 (ºC)	0.961	3.60	0.845	7.16	0.832	7.45	0.954	3.90

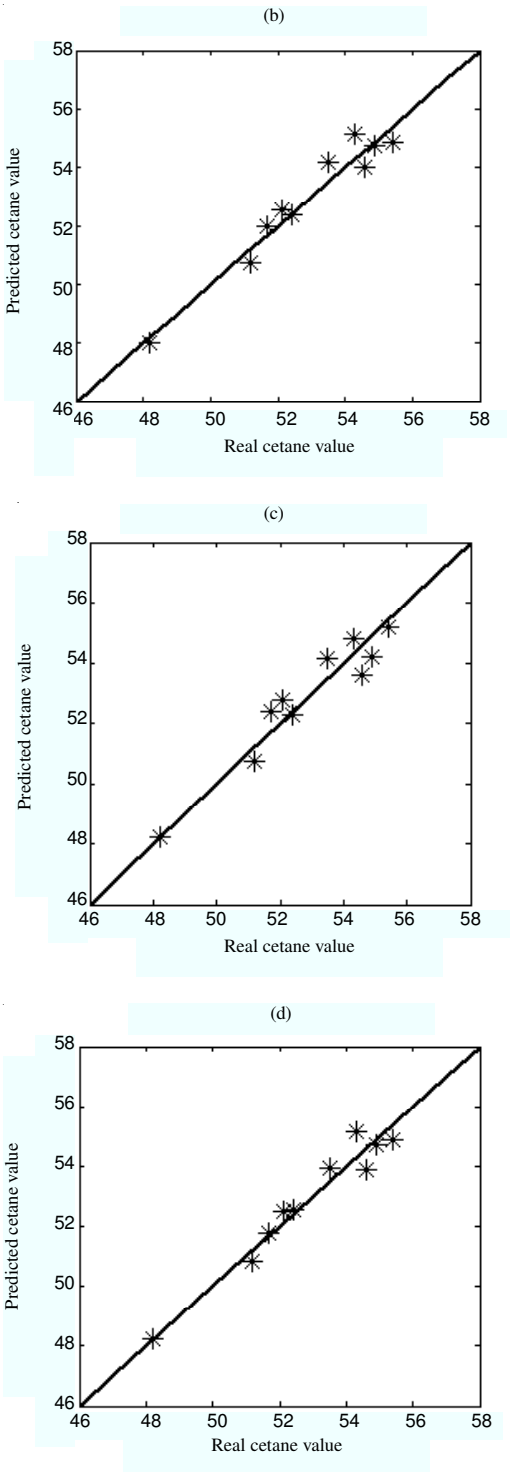


Fig. 7. Scatter plots of real and predicted cetane value (a) PLS (b) LSSVM (c) PCA-LSSVM (d) PLS-LSSVM

Conclusion

In this work, we have compared three LSSVM based calibration methods and PLS by building NIR calibration model to predict diesel properties. Experimental results show that the LSSVM with PLS feature extraction presents the best performances in all of the calibration methods. To improve the correlation coefficient between the predicted properties and NIR spectra, it is very useful to introduce feature extraction even for non-linear calibration methods

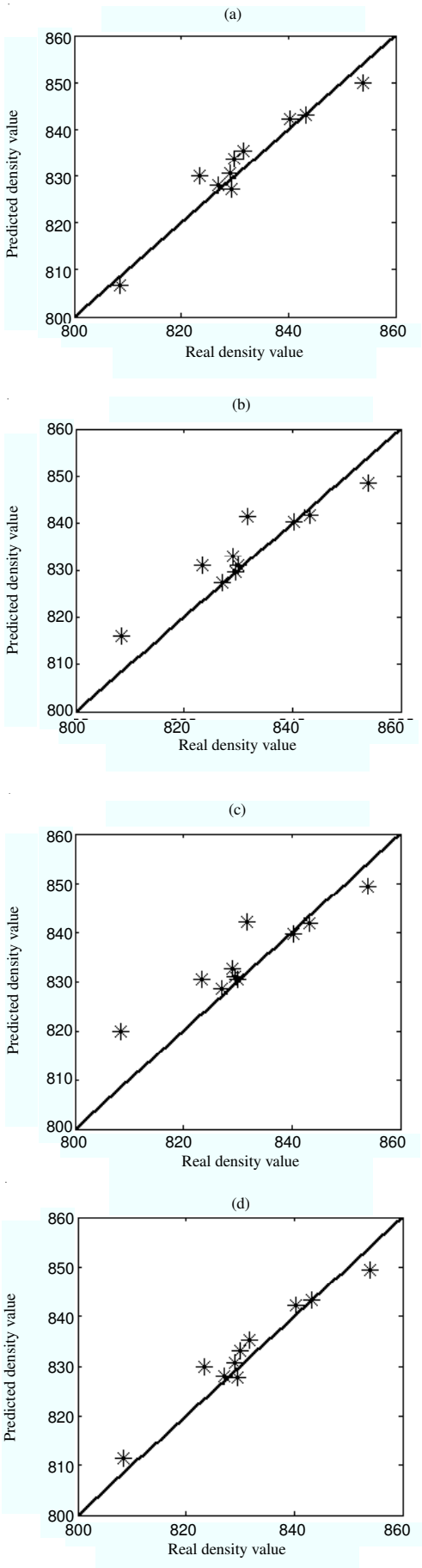


Fig. 8. Scatter plots of real and predicted density value (a) PLS (b) LSSVM (c) PCA-LSSVM (d) PLS-LSSVM

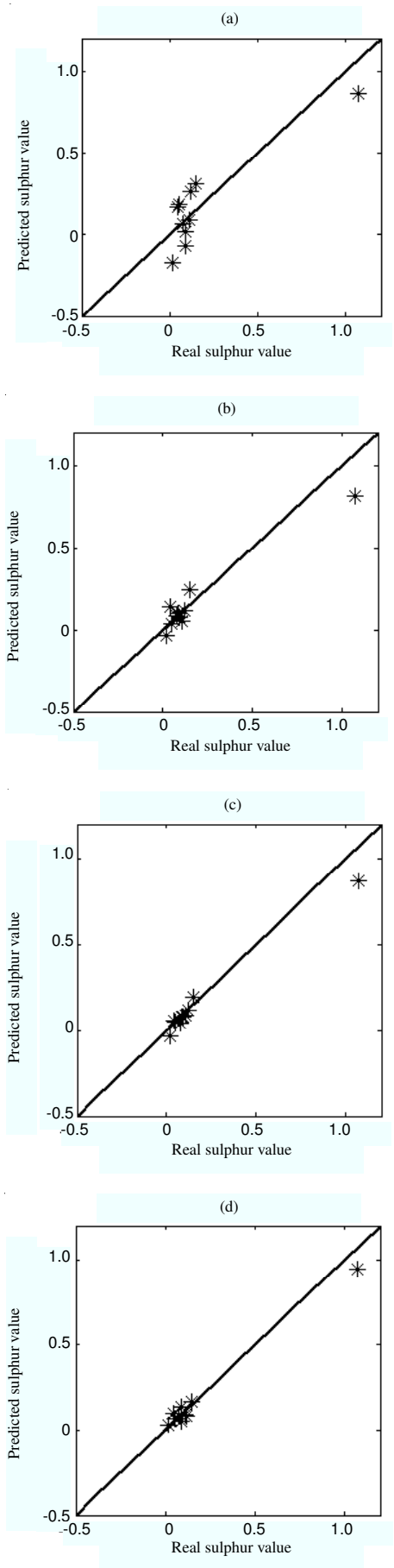


Fig. 9. Scatter plots of real and predicted total sulfur value (a) PLS (b) LSSVM (c) PCA-LSSVM (d) PLS-LSSVM

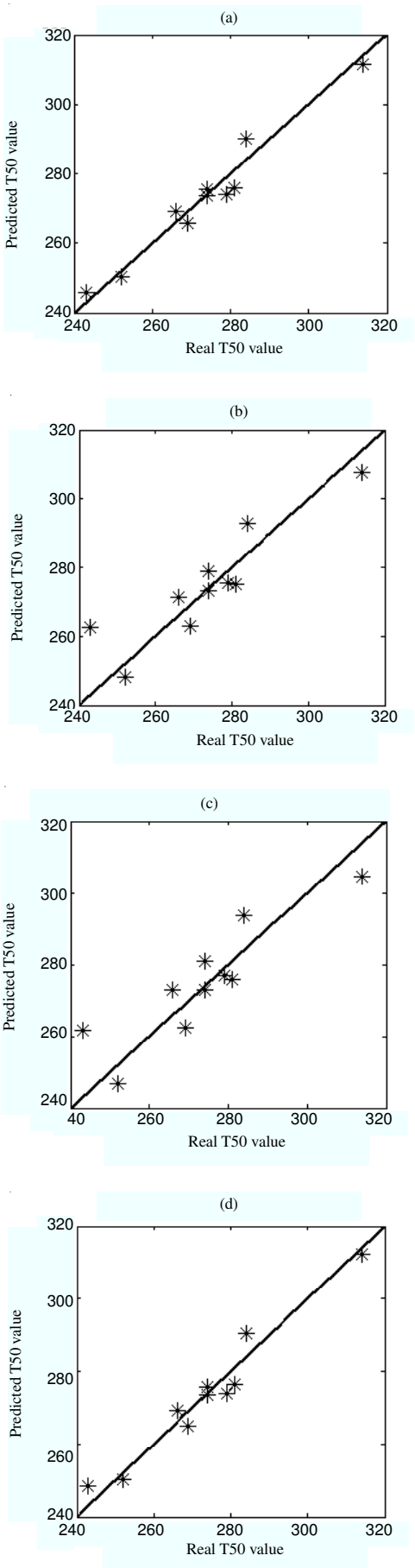


Fig. 10. Scatter plots of real and predicted T50 value (a) PLS (b) LSSVM (c) PCA-LSSVM (d) PLS-LSSVM

ACKNOWLEDGEMENTS

This work was supported by the National High Technology Research and Development Program ("863" Program) of China Grant No. 2006AA040309.

REFERENCES

1. J.J. Kelly, C.H. Barlow, T.M. Jinguji and J.B. Callis, *Anal. Chem*., **61**, 313 (1989).
2. S. Amat-Tosello, N. Dupuy and J. Kister, *Anal. Chim. Acta*, **642**, 6 (2009).
3. R.M. Balabin, R.Z. Safieva and E.I. Lomakina, *Chemom. Intell. Lab. Syst*., **88**, 183 (2007).
4. K. Brudzewski, A. Kesik, K. Kolodziejczyk, U. Zborowska and J. Ulaczyk, *Fuel*, **85**, 538 (2006).
5. R.M. Balabin, R.Z. Safieva and E.I. Lomakina, *Fuel*, **87**, 1096 (2008).
6. F.S.G. Lima and L.E.P. Borges, *J. Near Infrared Spectrosc*., **10**, 269 (2002).
7. V.O. Santos, F.C.C. Oliveira, D.G. Lima, A.C. Petry, E. Garcia, P.A.Z. Suarez and J.C. Rubim, *Anal. Chim. Acta*, **547**, 188 (2005).
8. L. Lira, F. Vasconcelos, C. Pereira, A. Paim, L. Stragevitch and M. Pimentel, *Fuel*, **89**, 405 (2010).
9. M. Pimentel, G. Ribeiro, R. Cruz, L. Stragevitch, J. Filho and L. Teixeira, *Microchem. J*., **82**, 201 (2006).
10. H. Fernandes, I. Raimundo, C. Pasquini and J. Rohwedder, *Talanta*, **75**, 804 (2008).
11. M. Sohn, D. Himmelsbach, F. Barton, C. Griffey, W. Brooks and K. Hicks, *Appl. Spectrosc*., **61**, 1178 (2007).
12. V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York (1995).
13. J.A.K. Suykens and J. Vandewalle, *Neural Processing Lett*., **9**, 293 (1999).
14. F. Chauchard, R. Cogdill, S. Roussel, J.M. Roger and V. Bellon-Maurel, *Chemom. Intell. Lab. Syst*., **74**, 141 (2004).
15. M.-Y. Liu, Y. Meng, J.-F. Li, H.-T. Zhang and H.-Y. Wang, *Chin. J. Chem*., **17**, 1209 (2006).
16. Y. Dou, Y.-Q. Sun and Y.-L. Ren, *Anal. Chim. Acta*, **528**, 55 (2005).
17. J.-F. Lv and L.-K. Dai, Proceedings of the 6th WCICA, 5228 (2006).
18. A. Savitzky and M.J.E. Golay, *Anal. Chem*., **36**, 1627 (1964).
19. R.J. Barnes, M.S. Dhanoa and S.J. Lister, *Appl. Spectrosc*., **43**, 772 (1989).