Quantitative Analysis of Near-Infrared Spectra by Wavelet-based Interferences Removal and Least Squares Support Vector Regression

Yingqiang Ding
College of Information Engineering, Zhengzhou University, Zhengzhou, PR China
Email: cnpowerfoot@gmail.com

Dan Peng
College of Grain Oil and Food Science, Henan University of Technology, Zhengzhou, PR China
Email: pengdantju@gmail.com

Abstract—**A new hybrid algorithm named EODT-LS-SVR based on least squares support vector regression (LS-SVR) with wavelet-based EODT algorithm as preprocessed tools is proposed for removing the interferences and developing the quantitative models with high precision in near-infrared (NIR) spectra. EODT-LS-SVR algorithm is composed of two steps. In the first step, the preprocessing algorithm named EODT, which combines the ideas of wavelet packet transform (WPT), orthogonal signal correction (OSC) and information theory, is employed for the characteristic extraction of analyte information through multi-scale analysis. Entropy-based baseline signal removing (EBSR) algorithm is applied to remove the baseline of the spectra based on information theory with WPT-based analysis, and then the information orthogonal to the concentrations of analyte is removed by OSC algorithm in each frequency band of spectra. In the second step, LS-SVR method coupled with grid search and particle swarm optimization (PSO) technique for parameters optimization is used to enhancing the quality of regression models. EODT-LS-SVR algorithm was validated by two NIR spectral datasets, one used for measuring the fat concentration of milk and the other used for measuring the oil content of corn. The comparison of prediction results demonstrated that the performance of calibration models developed by EODT-LS-SVR algorithm is better than that developed by other conventional algorithms, showing the high efficiency and the high quality for quantitative model development in NIR spectra of complex samples.**

Index Terms—**Quantitative analysis, Entropy, Least squares support vector regression, Interference information, Near-infrared spectroscopy, multi-scale analysis.**

I. INTRODUCTION

As one of the most powerful analytical techniques, near-infrared (NIR) spectroscopy is being extensively applied in agriculture, petrochemical, textile, food, pharmaceutical industry, biotechnology, and medicine [1-3]. Due to the characteristics of rapidity, simplicity and nondestructive or non-invasive measurement, NIR analysis has been taking a very important role both in laboratory analysis and on-line analysis [4]. Especially, with the development of chemometrics and the use of chemometrical methods in modeling, NIR techniques have been being considered as the standard procedures for many quantitative analyses [5].

In NIR analysis, a large number of chemical and physical properties are related with the spectral data through chemometrical method, therefore the efficiency of quantitative NIR method is dependent on chemometrical calibration model. Usually, there are two steps in building the chemometrical calibration model [6, 7]. The first step is to preprocess the NIR spectra for the purpose of removing interference information, and the second step is to develop the calibration model based on the preprocessed data and the multivariate methods. Thus, in order to build a NIR calibration model with high precision and strong robustness, it is necessary to select or develop suitable algorithms in the two steps above.

In fact, NIR spectra include not only the analyte information (the information of substance to be analyzed or tested), but also the interference information (irrelevant information) in most applications [8, 9]. Typically, the interferences, which are mainly composed of the noise, background, scattering, absorbance of other components and so on, are the dominant sources of spectra variation [10]. In other words, the presence of interferences deteriorates the problem of extraction of relatively weak analyte information in NIR spectra. Consequently, the NIR analysis would generate poor results. Classical preprocessing methods for interferences removal are to fit the interference information as a line or a polynomial curve. But in many cases, the interference information is not fitted as "ideal" as a line or a polynomial curve, indicating the interferences are varying and non-constant [11, 12]. Then, other kinds of algorithms have been proposed for interferences removal, such as derivative, standard normal variation (SNV) [13], multiplicative scatter correction (MSC) [14], orthogonal

doi:10.4304/jcp.7.4.880-889

signal correction (OSC) [15, 16], discrete wavelet transform (DWT) [17, 18], net analyte signal (NAS) [19], and so on.

OSC, which is developed by Wold and coworkers [15], works by removing the parts linearly unrelated (orthogonal) to the response of the calibration model. It is an efficient algorithm for removing synthetical noise coming from background, scattering, instrument, measurement conditions variation, etc. DWT and its generalization [17], the discrete wavelet packet transform (DWPT), are relatively new and promising mathematical tools for the pretreatment of NIR spectra. Through separating a spectrum into a series of frequency contributions, they can provide information in time and frequency domain. Compared with DWT, DWPT is able to decompose a signal into low frequency band and high frequency band in finer and more flexible way. These characteristics of DWPT make it possible to perform feature extraction, signal denoising, and background analysis by selecting the corresponding wavelet coefficients in the wavelet domain [20, 21]. However, among these algorithms, none of them can give the appropriate result independently in most applications, especially in the application of multi-components analysis. Moreover, these preprocessing algorithms sometimes may discard useful information other than interference information and lead to the reduction of predictive ability of the calibration model. According to the complexity of interferences in NIR spectra, some kinds of combination of these preprocessing methods can be tried to improve the performance.

Additionally, in the analysis of NIR data, a reliable quantitative and qualitative calibration model is also a matter of primary importance for characterization of unknown samples, so it is crucial to study the methods for model construction. Common multivariate methods for building the calibration model include multiple linear regression (MLR) [22], principal component regression (PCR) [23], partial least squares (PLS) [24], artificial neural network (ANN) [25], etc. PLS, which is based on linear models, is the most widely used calibration method and served as the satisfactory solution in most case. However, PLS cannot work well all the time, especially in situations where a nonlinear relationship is clearly present or the linear relationship between the spectra and the measured property is deteriorated badly by the interferences. In such cases, ANN, which is based on multilayer perceptrons calculation, is proposed and applied with relative more success. Nevertheless, ANN suffers several critical drawbacks such as the non-reproducibility problem due to the existence of many local minima caused by the multilayer perceptrons gradient based training method [26]. Besides, the ANN method excessively relies on train sample data, but in most cases, the limited number of sample data and the dimension reduction of spectra for computation simplification would also greatly weaken the prediction ability of ANN model. To achieve better performance, a promising methodology called least squares support vector regression (LS-SVR) [27, 28], which is originated from support vector machines regression (SVR) [29], has been introduced to perform multivariate regression. Instead of the quadratic programming, LS-SVR only requires solving a set of linear equations to lead to global model with simplified computation. The main advantages of LS-SVR include the superior generalization and accurate prediction, indicating that the LS-SVR has the outstanding power to deal with nonlinearity problems as well as linearity problems. Due to these merits of LS-SVR method, it is very suitable for NIR spectroscopy analysis and rapidly gains a lot of successful applications [28].

In this paper, a new hybrid algorithm, which is the combination of a wavelet-based preprocessing algorithm and LS-SVR method, is proposed to improve the prediction ability of NIR quantitative analysis in the presence of interferences. The preprocessing algorithm, which takes advantages of entropy-based baseline signal removing (EBSR) algorithm, OSC and DWPT, is also a combinational algorithm with the name of EODT. Consequently, the total algorithm in this paper is denoted as EODT-LS-SVR. In order to validate the effectiveness of EODT-LS-SVR algorithm, two real NIR spectral datasets were analyzed by different processing methods for measuring the concentrations of analyte, and the results indicated that the EODT-LS-SVR algorithm can develop the NIR quantitative model with better prediction ability.

II. THEORY AND ALGORITHMS

EODT-LS-SVR algorithm focuses on two steps. One is for the removal of interference information from mass spectra by EODT algorithm. The other is for the calibration models establishment by LS-SVR algorithm based on the preprocessed data. In this section, basic remarks about algorithms used in our study are presented. Some references on detail are also included.

A. Discrete Wavelet Packet Transform

DWPT decomposes a signal into localized contribution labeled by a scale and a position parameter, and each of the contributions at different scale represents the information of different frequency contained in the original signal [17]. The decomposition of DWPT is executed through a convolution of the signal with various scales and translations of a mother wavelet. As shown in Fig. 1, a signal denoted as $\mathbf{S}$ or $\boldsymbol{s}_{\mathbf{0,0}}$ is fully decomposed up to L levels, where $\boldsymbol{s}_{j,i}$ represents the ith wavelet packet coefficients in jth level decomposition. The WPT decomposing arithmetic can be described as [18]

$$\begin{cases} \mathbf{s}_{j+1,2i} = \mathbf{H}\mathbf{s}_{j,i} \\ \mathbf{s}_{j+1,2i+1} = \mathbf{G}\mathbf{s}_{j,i} \end{cases} \tag{1}$$

where $\mathbf{H}$ and $\mathbf{G}$ are the low-pass filter and the high-pass filter, $j = 1, 2 \ldots L$, and $i = 1, 2 \ldots 2^j\text{-}1$. Denote L as the largest number of decomposition level. By the decomposition, $\mathbf{S}$ can be split into 2^j individual frequency contributions.

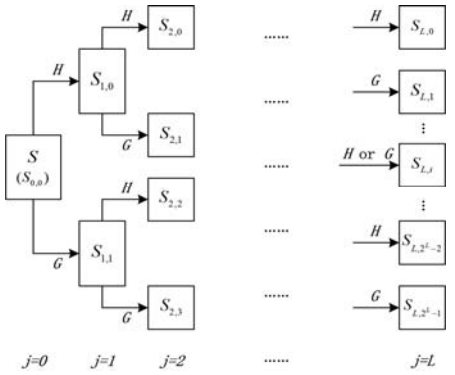

Figure 1. Principle of wavelet packet decomposition.

After decomposition, the spectra denoising can be implemented by thresholding operation in wavelet domain. In this paper, the wavelet packet coefficients are processed using soft thresholding, and the SURE method proposed by Donoho based on stein's unbiased risk estimation is selected [30]. By utilizing inverse WPT, the denoised coefficient also can be converted back into the original domain. Denote the denoised $s_{j,i}$ as $s'_{j,I}$, and the inverse DWPT can be described as

$$\mathbf{s}'_{\mathbf{j,i}} = \mathbf{H}^{*}\mathbf{s}'_{\mathbf{j+1,2i-1}} + \mathbf{G}^{*}\mathbf{s}'_{\mathbf{j+1,2i}} \tag{2}$$

where $\mathbf{H}^{*}$ and $\mathbf{G}^{*}$ are the pairing operators of $\mathbf{H}$ and $\mathbf{G}$. Thus, the $\mathbf{p}_{L,i}$, which represents contribution of the individual frequency bands, is calculated by executing the Equation 2 L times with only $\boldsymbol{s}'_{L,i}$. Then the signal $\mathbf{S}'$, which is obtained by applied denoising to $\mathbf{S}$, can be computed as

$$\mathbf{S}' = \sum_{i=0}^{2^{L}-1} \mathbf{p}_{\mathbf{L,i}} \,. \tag{3}$$

B. Entropy-based Baseline Signal Removing (EBSR) Algorithm

In general, the baseline signal, which is located in low frequency band, is non-constant, varying, component-dependent and different from sample to sample in NIR spectra. In many types of practical application, the varying baseline can evenly dominate the variation of spectra and makes the analysis process more complicate. Here, EBSR algorithm aims at removing the baseline irrelevant to the measured properties and maintaining the informative spectra variations at the same time based on DWPT and information entropy theory [31].

According to the frequency component $\boldsymbol{p}_{L,i}$, its information entropy can be calculated as:

Step1. Build a linear regression model as

$$\mathbf{p}_{\mathbf{L,i}} = \mathbf{Y} \cdot \mathbf{g}_{\mathbf{i}} + \boldsymbol{\varepsilon}_{\mathbf{i}} \tag{4}$$

where $\mathbf{Y}$ is the $m\times1$ matrix with a measured property in m samples, $\boldsymbol{g}_i$ is the regression matrix and $\boldsymbol{\varepsilon}_i$ is the residual error matrix. Using least square method, the $\boldsymbol{g}_i$ can be computed as

$$\mathbf{g}_{\mathbf{i}} = (\mathbf{Y}^{T}\mathbf{Y})^{-1}\mathbf{Y}^{T}\mathbf{p}_{\mathbf{L,i}} \,. \tag{5}$$

Then the information matrix IM_j is calculated as

$$IM_{i} = \mathbf{g}_{\mathbf{i}}^{\mathbf{T}}(\mathbf{g}_{\mathbf{i}}\mathbf{g}_{\mathbf{i}}^{\mathbf{T}})^{-1}\mathbf{g}_{\mathbf{i}} \,. \tag{6}$$

Step2. Calculate the information entropy corresponding to Y as [31]

$$I_{L,i} = -\sum q_{k}^{(i)} \log(q_{k}^{(i)}) \tag{7}$$

where $q_k^{(i)}$ is the kth diagonal element of IM_i.

Assume that $\mathbf{X}$ is the $m\times n$ spectral data matrix with n wavelength points in m samples. The principle of EBSR can be summarized as

Step1. According to (7), compute the information entropy of whole spectra $\mathbf{X}$ as I_{whole}.

Step2. Perform WPT decomposition on $\mathbf{X}$ by the maximum levels (L_{max}), and obtain the $2^{L_{\max}}$ frequency components $\left\{\mathbf{p}_{\mathbf{L}_{\max},\mathbf{i}}\right\}$.

Step3. Set $k = 0$ and $\boldsymbol{g}_{lf,0} = \mathbf{X}$. Compute the entropy of $\boldsymbol{g}_{lf,0}$ as $I_{lf,0}$. Then denote $I_{b,k}$ as

$$I_{b,k} = I_{lf,k} - I_{whole} \,. \tag{8}$$

Step4. Add variable k by one, and update the $\boldsymbol{g}_{lf,k}$ and $\boldsymbol{B}_k$ as

$$\mathbf{g}_{\mathbf{lf,k}} = \mathbf{g}_{\mathbf{lf,k-1}} - \mathbf{p}_{\mathbf{L}_{\max},\mathbf{k-1}} \,,\ \mathbf{B}_{\mathbf{k}} = \sum_{i=1}^{k} \mathbf{p}_{\mathbf{L}_{\max},\mathbf{i-1}} \,. \tag{9}$$

Then recalculate the information entropy $I_{b,k}$.

Step5. If the performance of $I_{b,k}$ is similar to that of $I_{b,k-1}$, the signal $\boldsymbol{B}_k$ can be eliminated as baseline. Then, increase variable k by one and repeat Step4 until the performances of $I_{b,Q}$ is different from that of $I_{b,Q-1}$.

Step6. Remove the components representing baseline, and the left spectra ($\mathbf{X}_{\mathrm{EBSR}}$) can be used to develop multivariate calibration models.

$$\mathbf{X}_{\mathbf{EBSR}} = \mathbf{X} - \mathbf{B}_{\mathbf{Q\text{-}1}} \tag{10}$$

If the low frequency components only represent baseline, the entropy would greatly increase because of the increase of uncertainty. Therefore, if transition of entropy appears between $I_{b,Q}$ and $I_{b,Q-1}$, the component $\mathbf{p}_{\mathbf{L}_{\max},\mathbf{Q\text{-}1}}$ must contain useful information and can't be viewed as baseline.

C. Orthogonal Signal Correction

The goal of OSC is to correct the $\mathbf{X}$ by removing the information orthogonal to $\mathbf{Y}$. The details of OSC algorithm can be described as following [16]:

Step1. Compute the projection of $\mathbf{Y}$ onto $\mathbf{X}$ as

$$\mathbf{Y}_{\mathbf{p}} = \mathbf{X}\mathbf{X}^{+}\mathbf{Y} \tag{11}$$

where $\mathbf{X}^{+}$ is the Moore-Penrose inverse of $\mathbf{X}$ and is equal to $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$.

Step2. Compute the projection of $\mathbf{X}$ onto the orthogonal complement space of $\mathbf{Y}$ as

$$\mathbf{Z} = (\mathbf{I} - \mathbf{Y}_{\mathbf{p}}\mathbf{Y}_{\mathbf{p}}^{+})\mathbf{X} \tag{12}$$

where $\mathbf{I}$ is a identity matrix.

Step3. Compute the score matrix of $\mathbf{ZZ}^{\mathrm{T}}$ through SVD decomposition as *[***T, D***] = SVD* ($\mathbf{ZZ}^{\mathrm{T}}$).

Step4. Compute the weight matrix as

$$\mathbf{W} = \mathbf{X}^{+}\mathbf{T} . \tag{13}$$

Step5. Renew the **T** as

$$\mathbf{T} = \mathbf{XW} . \tag{14}$$

Step6. Compute the loading matrix as

$$\mathbf{G} = \mathbf{X}^{\mathrm{T}}\mathbf{T}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1} . \tag{15}$$

Step7. The corrected spectra $\mathbf{X}_{\mathrm{OSC}}$ can be formulated as

$$\mathbf{X}_{\mathrm{osc}} = \mathbf{X} - \mathbf{TG}^{\mathrm{T}} . \tag{16}$$

To correct again, **X** can be replaced by $\mathbf{X}_{\mathrm{OSC}}$, and then the equations from (11) to (16) will be repeated.

D. Least Squares Support Vector Rregression

The present form of SVR was developed by AT&T Bell Laboratory and has been applied to chemometric issues for discrimination and quantitative prediction. To simplify SVR algorithm as well as achieve robustness and sparseness, Suykens proposed an alternate formulation of the SVR strategy called LS-SVR [27]. In LS-SVR, the e-insensitive loss function is replaced by a squared loss function. While foregoing the benefits of automatic sparseness and perhaps some insensitivity to outliers, LS-SVR can be trained much more efficiently. Like SVR, LS-SVR is also based on a kernel substitution. Therefore, the **X** in (8) should be replaced by an $m \times m$ kernel matrix **K** with the form

$$\mathbf{K} = \begin{pmatrix} k_{1,1} & \cdots & k_{1,m} \\ \vdots & \ddots & \vdots \\ k_{m,1} & \cdots & k_{m,m} \end{pmatrix} \tag{17}$$

where $k_{i,j}$ is calculated by the kernel function. Here, the radial basis function (RBF) kernel was used as kernel function, and the element of *K* is calculated as

$$k_{i,j} = \exp\left(-\left\|\mathbf{x}_{\mathbf{i}} - \mathbf{x}_{\mathbf{j}}\right\|_F^2 \Big/ 2\sigma^2\right) \tag{18}$$

where $\|\cdot\|_F$ denotes the 2-norm, $\mathbf{x_i}$ and $\mathbf{x_j}$ are the NIR spectra of samples, and σ is the kernel width parameter.

After constructing the Lagrangian equation, setting equality constraints, and simplifying, a linear Karush-Kuhn-Tucker (KKT) system can be obtained as

$$\begin{pmatrix} 0 & \mathbf{I}_{\mathbf{m}}^{\mathrm{T}} \\ \mathbf{I}_{\mathbf{m}} & \mathbf{K} + \gamma^{-1}\mathbf{I}_{\mathbf{m}\times\mathbf{m}} \end{pmatrix} \begin{pmatrix} a \\ \mathbf{H} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{Y} \end{pmatrix} \tag{19}$$

where $\mathbf{I_{m\times m}}$ is an $m \times m$ identity matrix, $\mathbf{I_m}$ is a $m \times 1$ vector of ones, γ is the regularization constant, **H** is the vector of model coefficients and *a* is the model bias term. Compared with SVR, LS-SVR has only two parameters (γ, σ) for tuning after simplification. It should be stressed that it is very important to do a careful parameter selection to achieve a good calibration model. Once the two optimized parameters were obtained, the model can be built as

$$y_{new} = \sum_{i=1}^{m} h^{(i)} \exp\left(-\left\|\mathbf{x}_{\mathbf{new}} - \mathbf{x}_{\mathbf{i}}\right\|_F^2 \Big/ 2\sigma^2\right) + a \tag{20}$$

where $h^{(i)}$ is the i*th* element of **H**, $\mathbf{x_{new}}$ is the spectrum of unknown sample, and y_{new} is prediction value.

E. The EODT-LS-SVR Algorithm

EODT-LS-SVR algorithm can be summarized as follows:

Step1. Perform EBSR algorithm on the raw spectral matrix $\mathbf{X_{raw}}$ to search low frequency contribution which can be viewed as baseline signal. Then the processed spectra matrix can be computed as

$$\mathbf{X}^{\mathbf{EBSR}} = \mathbf{X}_{\mathbf{raw}} - \mathbf{B}_{\mathbf{Q\text{-}1}} . \tag{21}$$

Step2. Perform an L_{max} level DWPT decomposition on the matrix $\mathbf{X}^{\mathbf{EBSR}}$. Then the thresholding operation is applied to the wavelet packet coefficients for denoising by SURE method in soft thresholding type [30]. Through inverse DWPT, the new frequency component contributions ($\mathbf{p}_{\mathbf{L_{max}},\mathbf{Q\text{-}1}} \sim \mathbf{p}_{\mathbf{L_{max}},\mathbf{2^{L_{max}}\text{-}1}}$) can be reconstructed from the denoised wavelet packet coefficients.

Step3. Perform OSC algorithm on frequency components $\mathbf{p}_{\mathbf{L_{max}},\mathbf{Q\text{-}1}} \sim \mathbf{p}_{\mathbf{L_{max}},\mathbf{2^{L_{max}}\text{-}1}}$ based on **Y**, and save the OSC-filtered component $\mathbf{p}^{\mathbf{OSC}}_{\mathbf{L_{max}},\mathbf{i}}$, the loading matrix $\mathbf{W}_{\mathbf{L_{max}},\mathbf{i}}$ and $\mathbf{T}_{\mathbf{L_{max}},\mathbf{i}}$.

Step4. Perform summation according to (3), and calculate the filtered spectra as

$$\mathbf{X}^{\mathbf{EODT}} = \sum_{i=Q-1}^{2^L-1} \mathbf{p}^{\mathbf{OSC}}_{\mathbf{L_{max}},\mathbf{i}} . \tag{22}$$

Step5. Based on the matrix $\mathbf{X}^{\mathbf{EODT}}$, perform grid-search technique and particle swarm optimization (PSO) technique to search the optimal parameters (γ_{op}, σ_{op}) [32], with which the LS-SVR model can achieve the smallest error. Then calculate the parameters (a_{op}, $\mathbf{B_{op}}$) corresponding to (γ_{op}, σ_{op}) and build the LS-SVR model using these parameters.

The application of EODT-LS-SVR algorithm on the prediction set $\mathbf{X_{pre}}$ can be summarized as follows:

Step1. After performing DWPT decomposition on the $\mathbf{X_{pre}}$, removing the baseline signal based on the parameter Q determined by EBSR algorithm. Then, perform the SURE-based DWPT denoising algorithm and reconstruct the component contributions ($\mathbf{p}_{\mathbf{L_{max}},\mathbf{Q\text{-}1,pre}} \sim \mathbf{p}_{\mathbf{L_{max}},\mathbf{2^{L_{max}}\text{-}1,pre}}$).

Step2. Using matrix **W** and matrix **T**, the OSC-filtered component can be obtained as

$$\mathbf{p}^{\mathbf{OSC}}_{\mathbf{L_{max}},\mathbf{i},\mathbf{pre}} = \mathbf{p}_{\mathbf{L_{max}},\mathbf{i},\mathbf{pre}} - \mathbf{p}_{\mathbf{L_{max}},\mathbf{i},\mathbf{pre}}\mathbf{W}_{\mathbf{L_{max}},\mathbf{i}}(\mathbf{T}^{\mathbf{T}}_{\mathbf{L_{max}},\mathbf{i}}\mathbf{T}_{\mathbf{L_{max}},\mathbf{i}})^{-1}\mathbf{T}^{\mathbf{T}}_{\mathbf{L_{max}},\mathbf{i}}\mathbf{p}_{\mathbf{L_{max}},\mathbf{i},\mathbf{pre}} \tag{23}$$

Then the EODT-filtered prediction set $\mathbf{X}^{\mathbf{EODT}}_{\mathbf{pre}}$ can be calculated according to (22).

Step3. Using LS-SVR model, the measured property of unknown sample can be predicted as

$$y_{pre,j} = \sum_{i=1}^{m} h_{op}^{(i)} \exp\left(-\left\|\mathbf{x}_{\mathbf{pre,j}} - \mathbf{x}_{\mathbf{i}}\right\|_F^2 \Big/ 2\sigma_{op}^2\right) + a_{op} \quad (24)$$

where $h_{op}^{(i)}$ is the *i*th element of $\mathbf{B_{op}}$, $\mathbf{x_{pre,j}}$ is a spectrum in $\mathbf{X}_{\mathbf{pre}}^{\mathbf{EODT}}$, $\mathbf{x_i}$ is a spectrum in $\mathbf{X}^{\mathbf{EODT}}$ and $y_{pre,j}$ is the predicted value.

III. EXPERIMENTS

A. Sample Preparation

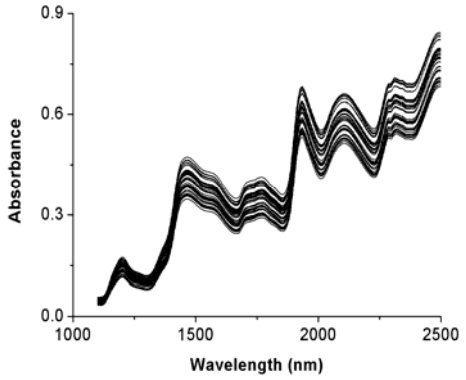

Figure 2. NIR spectra of corn samples.

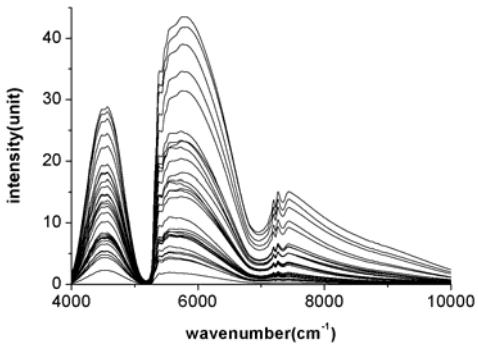

Figure 3. NIR spectra of milk samples.

Two NIR datasets were used to test the performances of EODT-LS-SVR algorithm. One was obtained from Cargill Inc., and also can be downloaded from http://www.eigenvector.com/Data/Corn/index.html. At Cargill, 80 corn samples were scanned from 1100nm to 2498nm, operating at 2nm resolution (Fig. 2). The objective is to predict the oil content from the set of spectra by the mp5 instrument. These corn samples were split into a calibration set and a prediction set with each including 40 samples. The other is the spectral data of 130 Homogenized milk samples supplied by Tianjin University of Science & Technology with the fat concentration ranging from 1.3g dL^{-1} to 5.8g dL^{-1}. NIR transmission spectra (10000cm^{-1} to 4000cm^{-1}) were collected by a Spectrum-GX FT-IR Spectrometer (Perkin-Elmer, USA) equipped with PbS detector and 0.5mm quartz sample cell, operating at 4 cm^{-1} resolution (Fig. 3). The objective is to predict the fat concentrations (measured by Rose-Gottlied method) through these collected spectra. Total of these samples were split into two sets, one used as calibration set including 100 samples and the other with 30 samples used as validation set.

B. Software and Computation

All the aforementioned calculations were performed using Matlab 2006a (The Math Works, Natick, USA). The PSO procedure was developed based on PSO toolbox v0.3. The free LS-SVM lab v1.5 (Suykens, Leuven, Belgium) was implemented with Matlab to develop all the LS-SVM models. The optimal parameters of LS-SVM models for the calibration set were selected based on the squared correlation coefficient (R^2) and the root mean square error of calibration (RMSEC). The accuracy of the calibration model was evaluated by R^2 and the root mean square error of prediction (RMSEP) for validation set.

IV. RESULTS AND DISCUSSIONS

A. Profiles of Information Entropy of Each Frequency Component

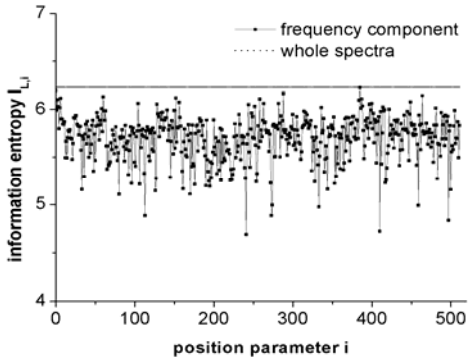

Figure 4. Distribution of information entropy of corn samples.

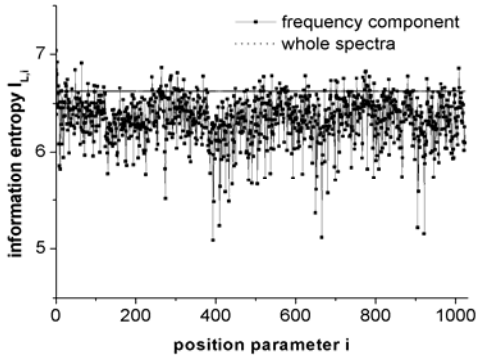

Figure 5. Distribution of information entropy of milk samples.

Spectra collected from most spectroscopic instruments are inherently multiscale in nature owing to contributions from the measured properties related with different localization in time and frequency domains. This phenomenon can be demonstrated by performing WPT decomposition and information entropy calculation on the

spectra. To measure the spectral variations, the raw spectra of calibration set were decomposed to 9 levels (512 frequency components) for corn samples and 10 levels (1024 frequency components) for milk samples using 'db4' mother wavelet. After computing the information entropy of each component ($I_{L,i}$), the distribution profiles of spectral contribution as a function of scale and position parameter were obtained as shown in Fig. 4 and Fig. 5. It can be seen that the information entropy of each component is different, and the information entropy of most of components is lower than that of whole spectra. The reasons are twofold. Firstly, the analyte information contained in raw spectra is not concentrated, but distributes in many frequency components. Secondly, the collected spectra may suffer from large background and noise variations causing the increase of information entropy. In fact, an obvious noise can be found in Fig. 1, especially when the wavelength is around 1400nm (7140cm^{-1}). This is partly due to the strong absorbancy dominated by water component in this region, which also can generate large background variation at the same time. As a result, the uncertainty of whole spectra must increase, followed by the increase of information entropy. Therefore, before building the calibration model, it is necessary to perform preprocessing to improve the signal-to-noise ratio as well as select the useful information through signal processing and multivariate statistics.

B. Effect of EODT Algorithm on Interferences Removal

1) Baseline Removal Based on Information entropy

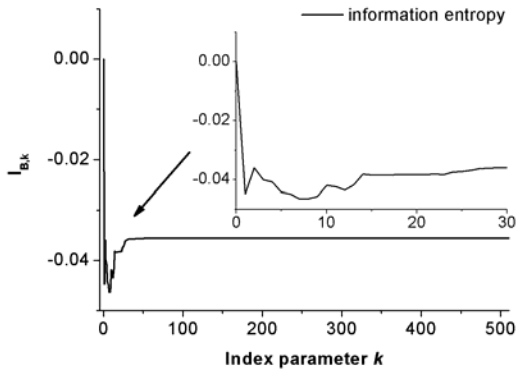

Figure 6. The information entropy for baseline removal of corn samples.

From Fig. 4 and Fig. 5, although the changes in spectral variation of different components can be found, it is difficult to determine how much spectral variation. According to (9), the increase of k corresponds to the gradual reduction of analyte information, while the information entropy loss indicates the importance of corresponding component in whole spectra. To remove baseline signal with minimum information loss, the entropy ($I_{B,k}$) was calculated as shown in Fig. 6 and Fig. 7. With k equaling to 0, the whole spectra were used and no information loss. As presented in Fig. 6, $I_{B,k}$ quickly decreases at first ($k \leq 1$) and then increases when k is equal to 2. Because the subtraction of component $\mathbf{p}_{9,1}$ results into increase of entropy, it must contain analyte information and can't be viewed as baseline signal. Thus, the summation of components B_1 can be identified as the contribution from background for removal. The estimated baseline signal of corn samples was illustrated in Fig. 8. Similarly, for milk samples, the summation of components B_5 can be viewed as baseline signal (Fig. 9).

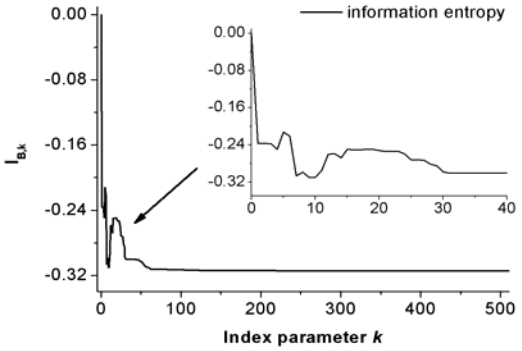

Figure 7. The information entropy for baseline removal of milk samples

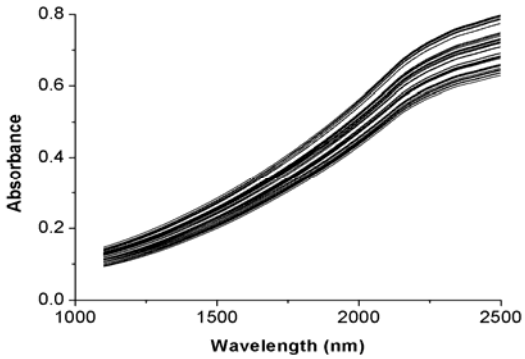

Figure 8. The removed baseline based on EBSR algorithm for corn samples.

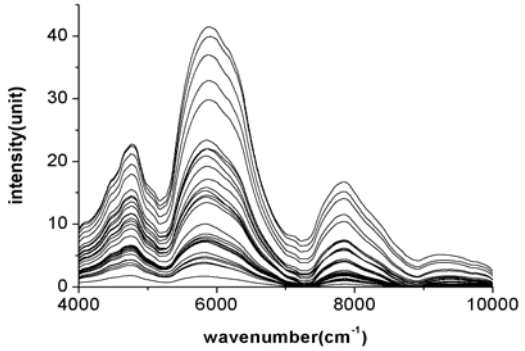

Figure 9. The removed baseline based on EBSR algorithm for milk samples.

2) Interference Information Removed by EODT Algorithm

According to the application of concentration measurement, the analyte information is defined as the information relevant to the concentrations of analyte, while the interference information is defined as the information orthogonal to the concentrations of analyte.

Generally, interference information in NIR spectra mainly includes background (in low frequency), variations caused by other components in milk (in middle frequency), noise from external environment and instrument (in middle and high frequency), and *etc*. After baseline removal by EBSR algorithm, other frequency components (except $\mathbf{B_1}$ or $\mathbf{B_5}$) generated from DWPT decomposition contain both analyte and interference information. In other words, the distributions of interference information are varying in different frequency bands, meaning that the spectra are multi-scale [33, 34]. Classical methods are usually applied globally to the spectra to remove the interference. But the selection of an entire frequency can retain a significant portion of the interference spectra variation. Therefore, the effects are unsatisfactory.

In this paper, the EODT algorithm can takes full advantage of multi-scale analysis through the combination of DWPT decomposition and OSC algorithm. Due to the multi-scale property of interference information, the portions of spectral energy removed by EODT algorithm are also different in each frequency band. Here, the spectral energy is defined as

$$E_s = \sum_{i=1}^{n} s_i^2 \tag{25}$$

where s_i is the intensity of a wavenumber point, and n is the number of points in a spectrum. To show the multi-scale characteristics of EODT algorithm, percentage of the spectral energy removed for corn samples is presented in Fig. 10. It is shown that, on average, 42% of the spectral energy is removed, indicating that only 58% of spectral energy is related to the prediction of analyte concentration. Also, it can be seen that almost all of the spectral energy in the low frequency components ($\mathbf{p_{9,1}}$~$\mathbf{p_{9,80}}$) and middle frequency components ($\mathbf{p_{9,180}}$~$\mathbf{p_{9,230}}$, $\mathbf{p_{9,310}}$~$\mathbf{p_{9,360}}$) are retained. This can be explained by the fact that the majority of useful information is contained in the low and middle frequency band. On the contrary, large portion of spectral energy in high frequency band is removed. This is probably a consequence of that high frequency components mainly include random noise and the OSC algorithm can fully remove this kind of interference information.

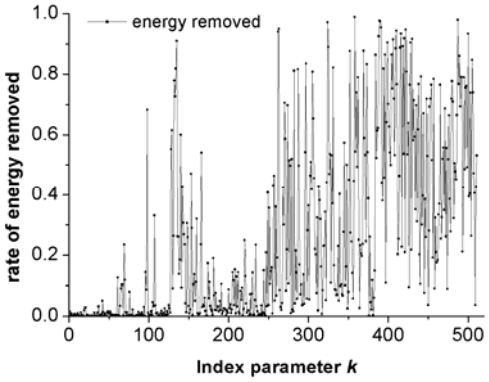

Figure 10. Average percentage of spectral energy removed by EODT for corn samples.

The two spectral datasets processed by EODT algorithm based on analyte concentration matrix are shown in Fig. 11 and Fig. 12. Compared with the spectra in Fig. 2 and Fig. 3, it can be seen that most of the spectral variations are removed and some spectral features related to analyte are extracted.

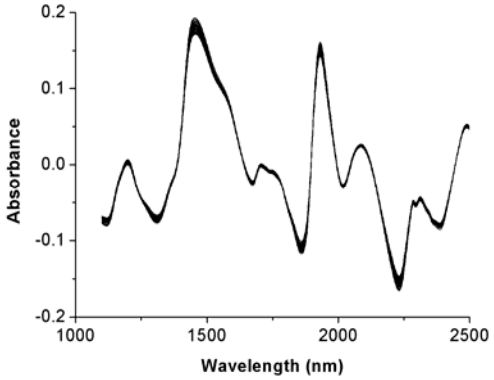

Figure 11. The EODT-filtered spectra of corn samples.

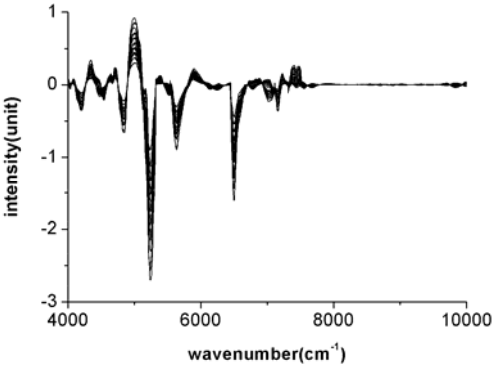

Figure 12. The EODT-filtered spectra of milk samples.

In order to further verify the effectiveness of EODT algorithm for interference information removal, the principal component analysis (PCA) [35] was calculated on the EODT-filtered spectral data of 40 milk samples based on fat concentration matrix. 40 special samples from calibration set are split into 4 subsets with protein concentrations of 1.7g dL^{-1}, 2.9g dL^{-1}, 4.1g dL^{-1} and 5.3g dL^{-1}, respectively. The fat concentrations of 10 samples in each subset range from 1.3g dL^{-1} to 5.8g dL^{-1} with interval 0.5g dL^{-1}. Fig. 13 is the plot of the scores of the first four principal components (PCs) which can explain more than 92% of the total variance. It can be seen that the samples of each subset are clustered together, but the distribution regions of the four subsets are different in PCs. The region of subset 1 is clearly separated from the regions of other three subsets, and only small portions of overlap exist between subset 2 and subset 3 as well as between subset 2 and subset 4. Based on this analysis, there is a clear requirement to develop calibration model for concentration determination by means of multivariate methods.

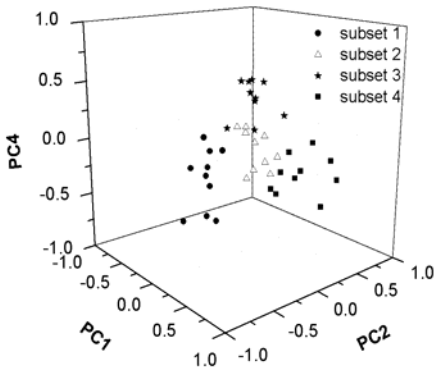

Figure 13. Score plot of principle component analysis PC1, PC2 and PC4 for subset 1 (●), subset 2 (△), subset 3 (★) and subset 4 (■).

C. The Calibration Models

In the model developed by LS-SVR with RBF kernel, the training process for searching the optimal parameters (γ, σ) is a manageable task, similar to the process employed to select the number of factors for PLS model, but in this case for a two-dimensional problem. The searching process of EODT-LS-SVR algorithm comprises two steps:

Step1. With the values of γ in the range of 1-2000 with increment of 20 and σ in the range of 0.01-80 with increment of 0.5, the process of grid searching starts, and the calibration models are developed on the grid points according to (19) and (20). These ranges of parameters are selected from the previous studies where it established the magnitude of the parameters to be optimized. For each combination of parameters (γ, σ), RMSECs for oil and fat are calculated and the optimized parameters can be selected which produced the smaller RMSEC. Through grid searching, the optimized regions of parameters were determined. When γ ranges from 1 to 90 and σ ranges from 0.6 to 19, the RMSEC for oil tends to a minimum. Also, when γ ranges from 40 to 50 and σ ranges from 20 to 26, the RMSEC for fat tends to a minimum.

Step2. In the two optimized parameter regions for fat and protein, PSO technique is used to find the optimal parameters (γ_{op}, σ_{op}) due to its characteristics of fast multi-peak searching and dynamic optimization. Because the dimension of search space is 2, the number of particles is set as 25 [32]. Additionally, the initial position of each particle can be set as the center of the optimized region and the maximum number of iterations is set as 10000. After PSO searching, the optimal parameters for corn were found with the value of γ = 13.90 and σ = 1.18, and the optimal parameters for milk were found with the value of γ = 49.30 and σ = 21.42.

D. Results comparison of different multivariate models

Fig. 14 and Fig. 15 depict the RMSEP curves of PLS models with different preprocessing methods for oil content of corn samples and fat concentration of milk samples in validation set. For the sake of comparison, first ten latent variables were calculated. In 'none-PLS', 'MSC-PLS' and 'OSC-PLS' models, the PLS method is directly applied to the entire spectra without multiscale process. It can be seen from Fig. 14 and Fig. 15, the RMSEP curves for oil and fat using EODT-PLS are much lower than those with other algorithms, indicating that EODT algorithm can effectively remove the interference information and extract the analyte information. In addition, the trends of the two RMSEP curves of EODT-PLS are also much similar to each other. With the increase of number of latent variables, the RMSEP curves descend gradually and then become flat with small jitter.

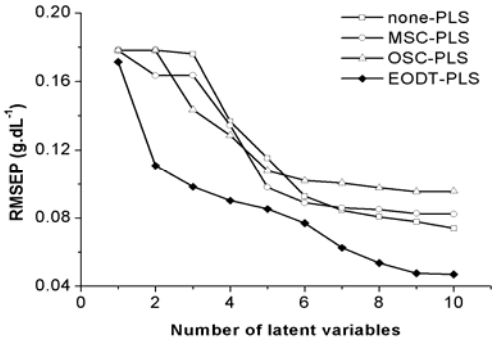

Figure 14. RMSEPs of oil content vs. latent variables.

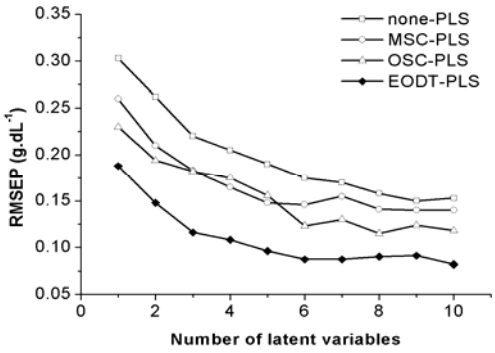

Figure 15. RMSEPs of fat concentration vs. latent variables.

To investigate the prediction accuracy of different models, the best RMSEPs and the corresponding R2 for validation set are presented in Table 1. With the same model parameters used in Table 1, the results of EODT-LS-SVR prediction from the validation set for oil concentration and fat concentration are shown in Fig. 16 and Fig. 17, respectively. As it can be seen, both show a highly linear correlation into the predicted values and the real values. In Table 1, the 'LS-SVR' and 'OSC-LS-SVR' also mean that the LS-SVR method with optimum parameters is directly applied to the entire raw spectra or entire OSC-filtered spectra. As observed from Table 1, when the models are developed by the same multivariate method coupled with different preprocessing algorithms, the RMSEPs obtained by EODT algorithm are much smaller, further illustrating the excellent properties for interferences removal. When the models are developed by the same preprocessing algorithm, the RMSEPs obtained using LS-SVR method outperforms those using PLS method by a considerable margin. This can be explained by the fact that other preprocessing methods can't effectively remove the interference information in raw spectra owing to the multiscale property, and the

linear relationship with respect to analyte is still deteriorated by the remaining interferences. Thus, the nonlinear models are more appropriate.

and milk samples, some appropriate results were acquired for predicting the concentrations of analyte. By comparison against models with different preprocessing

TABLE I. THE PREDICTION RESULTS BASED ON DIFFERENT METHODS.

method	oil concentration		fat concentration	
	R^2	RMSEP(g.dL^{-1})	R^2	RMSEP(g.dL^{-1})
PLS	0.860	0.074	0.869	0.150
LS-SVR	0.867	0.052	0.893	0.134
OSC-PLS	0.913	0.095	0.926	0.118
OSC-LS-SVR	0.882	0.061	0.946	0.102
EODT-PLS	0.981	0.047	0.975	0.087
EODT-LS-SVR	0.987	0.018	0.981	0.073

Only the model yielding the best performance for each algorithm is shown.

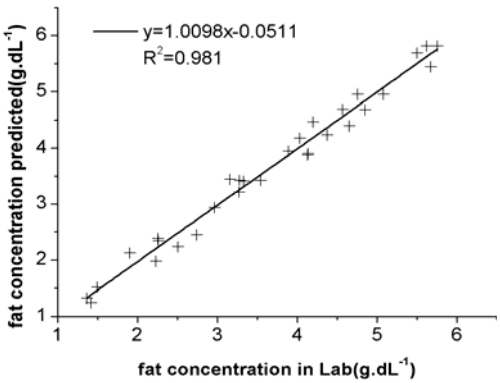

Figure 16. Scatter plot of lineal correlation of the oil concentration prediction vs. laboratory analysis in validation set

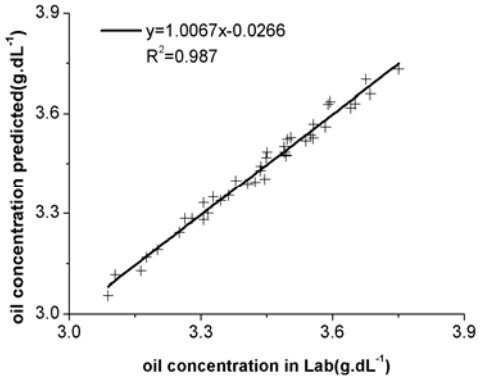

Figure 17. Scatter plot of lineal correlation of the fat concentration prediction vs. laboratory analysis in validation set.

V. CONCLUSIONS

According to the characteristics of NIR spectra, a method named EODT-LS-SVR is proposed to remove interference information and improve the prediction ability of calibration model. In the process of EODT-LS-SVR, DWPT and EBSR algorithm are used to remove the baseline and noise, and then OSC algorithm is applied to eliminate the interference information orthogonal to analyte in each scale. At last, LS-SVR method is used to develop the calibration model. Through the applications of building the NIR quantitative models of corn samples

algorithms and multivariate methods, EODT-LS-SVR-based models can give more accurate prediction result, indicating that the strategy of EODT-LS-SVR is a promising way for quantitative modeling of complex samples.

ACKNOWLEDGMENT

This study is supported by the National Key Technology R&D Program in the 11th Five Years Plan of China (No. 2006BAI03A03), the Key Program for Science and Technology Development of Henan Province (No.102101210600) and the Doctoral Foundation Program of Zhengzhou University.

REFERENCES

[1] C. M. McGoverin, L. C. H. Ho, J. A. Zeitler, C. J. Strachan, K. C. Gordon, and T. Rades, " Quantification of Binary Polymorphic Mixtures of Ranitidine Hydrochloride Using NIR Spectroscopy," *Vib. Spectroscopy*. vol. 41, pp. 225-231, August 2006.

[2] Y. K. Li, X. G. Shao, W. S. Cai, "A Consensus Least Squares Support Vector Regression for Analysis of Near-infrared Spectra of Plant Samples," *Talanta*. vol.72, pp. 217-222, April 2007.

[3] M. Blanco, M. Alcalá, J. M. González, and E. Torras, "Near Infrared Spectroscopy in the Study of Polymorphic Transformations," *Anal. Chim. Acta.* vol. 567, pp. 262-268, May 2006.

[4] N. Li, Y. Wang, K. X. Xu, "Fast Discrimination of Traditional Chinese Medicine According to Geographical Origins with FTIR Spectroscopy and Advanced Pattern Recognition Techniques," *Opt. Express.* vol. 14, pp, 7630-7635, August 2006.

[5] A. Borin, M. F. Ferrao, C. Mello, D. A. Maretto, R. J. Poppi, "Least-squares Support Vector Machines and Near Infrared Spectroscopy for Quantification of Common Adulterants in Powdered Milk," *Anal. Chim. Acta.* vol. 579, pp. 25-32, October 2006.

[6] W. S. Cai, Y. K. Li, X. G. Shao, "A Variable Selection Method Based on Uninformative Variable Elimination for Multivariate Calibration of Near-infrared Spectra," *Chemometr. Intell. Lab. Syst.* vol. 90, pp. 188-194, February 2008.

[7] J. Koljonen, T. E. M. Nordling, J. T. Alander, "A Review of Genetic Algorithms in Near Infrared Spectroscopy and

Chemometrics: Past and Future," *J. Near Infrared Spectrosc*. vol. 16, pp. 189-197, March 2008.
[8] H. W. Tan, S. D. Brown, "Multivariate Calibration of Spectral Data Using Dual-domain Regression Analysis," *Anal. Chim. Acta.* vol. 490, pp. 291-301, August 2003.
[9] Y. L. Liu, S. R. Delwiche, R. A. Graybosch, "Two-dimensional Correlation Analysis of Near Infrared Spectral Intensity Variations of Ground Wheat," *J. Near Infrared Spectrosc*. vol. 17, pp. 41-50, January 2009.
[10] M. B. Seasholtz, B. Kowalski, "The Parsimony Principle Applied to Multivariate Calibration," *Anal. Chim. Acta.* vol. 277, pp. 165-177, May 1993.
[11] D. Chen, X. G. Shao, B. Hu, and Q. D. Su, "A Background and Noise Elimination Method for Quantitative Calibration of Near Infrared Spectra," *Anal. Chim. Acta.* vol. 511, pp. 37-45 , May 2004.
[12] J. Kuligowski, G. Quintás, S. Garrigues, and M. D. L. Guardia, "On-line Gradient Liquid Chromatography–Fourier Transform Infrared Spectrometry Determination of Sugars in Beverages Using Chemometric Background Correction," *Talanta*. vol. 77, pp. 779-785, December 2008.
[13] R. J. Barnes, M. S. Dhanoa, S. J. Lister, "Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra," *Appl. Spectrosc*. vol. 43, pp. 772-777, May 1989.
[14] P. Geladi, D. M. Dougall, H. Martens, "Linearization and Scatter-Correction for Near-Infrared Reflectance of Meat," *Appl. Spectrosc.* vol. 39, pp. 491-500, March 1985.
[15] S. Wold, H. Antti, F. Lindgren, and J. Öhman, "Orthogonal Signal Correction of Near-infrared Spectra," *Chemometr. Intell. Lab. Syst.* vol. 44, pp. 175-185, December 1998.
[16] J. A. Westerhuis, S. D. Jong, A. K. Smilde, "Direct Orthogonal Signal Correction," *Chemometr. Intell. Lab. Syst.* vol. 56, pp. 13-25, April 2001.
[17] B. Walczak, B. V. D. Bogaert, D. L. Massart, "Application of Wavelet Packet Transform in Pattern Recognition of Near-IR Data," *Anal. Chem.* vol. 68, pp. 1742-1747, October 1996.
[18] B. Jawerth, W. Sweldens, "An Overview of Wavelet based Multiresolution Analyses," *SIAM Review.* vol. 39, pp. 377-412, September 1994.
[19] J. Ferré, N. K. M. Faber, "Net Analyte Signal Calculation for Multivariate Calibration," *Chemometr. Intell. Lab. Syst.* vol. 69, pp. 123-136, November 2003.
[20] R. N. F. D. Santos, R. K. H. Galvão, M. C. U. Araujo and E. C. D. Silva, "Improvement of Prediction Ability of PLS Models Employing the Wavelet Packet Transform: A Case Study Concerning FT-IR Determination of Gasoline Parameters," *Talanta.* vol. 71, pp. 1136-1143, February 2007.
[21] F. Ehrentreich, "Wavelet Transform Applications in Analytical Chemistry," *Anal. Bioanal. Chem.* vol. 372, pp. 115-121, January 2002.
[22] K. G. Kowalski, "On the Predictive Performance of Biased Regression Methods and Multiple Linear Regression," *Chemometr. Intell. Lab. Syst.* vol. 9, pp. 177-184, September 1990.
[23] Y. L. Xie, J. H. Kalivas, "Local Prediction Models by Principal Component Regression," *Anal. Chim. Acta.* vol. 348, pp. 29-38, August 1997.
[24] P. Geladi, B. R. Kowalski, "Partial Least-squares Regression: A Tutorial," *Anal. Chim. Acta.* vol. 185, pp. 1-17, January 1986.
[25] I. A. Basheer, M. Hajmeer, "Artificial Neural Networks: Fundamentals, Computing, Design, and Application," *J. Microbiol. Meth.* vol. 43, pp. 3-31, December 2000.
[26] Y. Zhang, Q. Cong, Y. F. Xie, J. X. Yang, and B. Zhao, "Quantitative Analysis of Routine Chemical Constituents in Tobacco by Near-infrared Spectroscopy and Support Vector Machine," *Spectrochim. Acta, Part A.* vol. 71, pp. 1408-1413, December 2008.
[27] J. A. K. Suykens, T. V. Gestel, J. D. Brabanter, B. D. Moor, and J. Vandewalle, *Least-Squares Suport Vector Machines*, World Scientific, Singapore, 2002.
[28] R. P. Cogdill, P. Dardenne, "Least-squares Support Vector Machines for Chemometrics: An Introduction and Evaluation," *J. Near Infrared Spectrosc.* vol. 12, pp. 93-100, December 2004.
[29] A. I. Belousov, S. A. Verzakov, J. V. Frese, "Applicational Aspects of Support Vector Machines," *J. Chemometr.* vol. 16, pp. 482-489, August 2002.
[30] D. L. Donoho, "De-Noising by Soft-Thresholding," *IEEE Trans. Inform. Theory.* vol. 41, pp. 613-627, May 1995.
[31] K. Eckschlager, V. Stepanek, "Information Theory in Analytical Chemistry," *Anal. Chem.* vol. 54, pp. 1115A-1127A, September 1982.
[32] J. Schutte, A. Groenwold, "A Study of Global Optimization Using Particle Swarm," *J. Global. Optim.* vol. 31, pp. 93-108, January 2005.
[33] B. K. Alsberg, A. M. Woodward, D. B. Kell, "An Introduction to Wavelet Transforms for Chemometricians: A Time-Frequency Approach," *Chemometr. Intell. Lab. Syst.* vol. 37, pp. 215-239, June 1997.
[34] B. R. Bakshi, "Multiscale Snalysis and Modeling Using Wavelets," *J. Chemometrics.* vol. 13, pp. 415-434, March 1999.
[35] T. N. Yang, S. D. Wang, "Robust Algorithms for Principal Component Analysis," *Pattern. Recogn. Lett.* vol. 20, pp. 927-933, September 1999.

Yingqiang Ding is lecture at college of information engineering, Zhengzhou University. He has received PhD degree from Tianjin University, and his major is signal processing. His research interests include digital signal processing, information theory and near-infrared spectroscopy technique.

Dan Peng is associate professor in college of grain oil and food science, Henan University of Technology. She has received PhD degree from Tianjin University. Her research interests include food safety, biomedical signal processing and technique of data analysis.