CWT-PLSR for Quantitative Analysis of Raman Spectrum

Shuo Li*, James O. Nyagilo†, Digant P. Dave† and Jean Gao*
*Computer Science and Engineering Department, The University of Texas at Arlington,
Arlington, Texas, USA. E-mail: {shuo.li, gao}@uta.edu
†Bioengineering Department, The University of Texas at Arlington,
Arlington, Texas, USA. E-mail: {james.nyagilo, ddave}@uta.edu

***Abstract*—Quantitative analysis of Raman spectra using Surface Enhanced Raman scattering (SERS) nanoparticles has shown the potential and promising trend of development in vivo molecular imaging. Partial Least Square Regression (PLSR) methods are the state-of-the-art methods. But they rely on the whole intensities of Raman spectra and can not avoid the instable background. In this paper we design a new CWT-PLSR algorithm that uses mixing concentrations and the average continuous wavelet transform (CWT) coefficients of Raman spectra to do PLSR. The average CWT coefficients with a Mexican hat mother wavelet are robust representations of the Raman peaks, and the method can reduce the influences of the instable baseline and random noises during the prediction process. In the end, the algorithm is tested on three Raman spectrum data sets with three cross-validation methods, and the results show its robustness and effectiveness.**

***Keywords*-CWT; PLSR; Quantitative Analysis; Raman Spectrum;**

I. Introduction

Raman scattering or Raman effect is the physical phenomenon when the monochromatic laser light interacts with molecular vibrations or other excitations, resulting in the energy of the laser photons being shifted upwards or downwards. The shifts in energy are referred as Raman frequencies or Raman shifts. A characteristic range of Raman shifts, which give their unique spectral information of a particular molecule, are collectively referred to as the Raman spectrum of that molecule [13]. But the inherently weak magnitude of Raman scattering limits the sensitivity and as a result, the biomedical applications of Raman spectroscopy. The development of the surface-enhanced Raman spectroscopy or scattering (SERS) offers an exciting opportunity to overcome this serious signal to noise problem inherent in Raman spectroscopy [13]. It is regarded as one of the most sensitive techniques that can provide the spectral fingerprint of every chemical compound, and has been a routine method used as an analytical tool in food industry, pharmaceutical, chemical and biological community [2] to investigate the composition of materials. It also has been used in the field of biomedical diagnostics, especially in the application of cancer detection research [11], [18]. Antibody conjugated nanoparticles, which can be attached to specific proteins in cancer cells, are injected into body. Cancer can be diagnosed by detecting large amount of such nanoparticles gathered inside body by quantitative analysis of Raman spectrum.

Quantitative analysis of spectrum is also called spectroscopic calibration, which is mainly to determine the values of the chemical or physical properties (e.g. concentrations of the pure components in the mixture materials) of the analyte from its measured spectrum. The state-of-the-art methods are Partial Least Square Regression (PLSR) methods [19], which are developed from the partial least squares (PLS) technique [14]. PLSR methods include PLS1, PLS2 and SIMPLS [16]. PLS1 is for one dimensional response matrix and PLS2 is for multidimensional response matrix. Both are based on the NIPALS algorithm, which is described in [9], [16]. [9] gives a deep analysis of PLS2 and [16] gives a good picture of PLS2. SIMPLS [4] improves PLS2 by avoiding the deflation of original data matrixes. Though PLSR methods are better than least square regression model [12], they can not avoid the instable background problem. Continues wavelet transform (CWT) is an effective way to extract the peak information and automatically remove the background [5]. To use both peak shape information of Raman spectrum and the correlation between peak heights and concentrations, we design a CWT based PLSR (CWT-PLSR) that uses CWT coefficients of Raman spectra to do PLS regression. It is robust to random noises and instable baseline and the performance is better than traditional PLSR and baseline correction based PLSR.

In section two, we introduce traditional PLSR methods and analysis the limitation of these methods for quantitative analysis of Raman spectra. To solve these limitations, in section three, we describe the CWT-PLSR method. In section four, we evaluate CWT-PLSR with three Raman spectrum data sets and three different cross-validation methods, and compare it with other regression methods.

II. PLSR Calibration

In this section, we first introduce the full spectrum calibration model and the latent space calibration model for quantitative analysis of Raman spectrum; then we describe the basic ideas of two PLSR methods: PLS2 and SIMPLS, as well as the prediction process; in the end, we analyze the limitations of PLSR.

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

A. Multivariate Calibration Models

The Raman spectrum of a compound is approximately equals to the summation of Raman spectrum of each pure component in that compound [11]. Also under certain range of concentrations, the intensities of Raman spectrum are approximately linearly related to the concentration of each pure component [18]. So the relation between the intensities of Raman spectra and the concentrations of components of compounds can be represented as the full spectrum calibration model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Theta} + \mathbf{E}, \tag{1}$$

with the $N \times D_x$ matrix $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]^T$ being N mixture Raman spectra of compounds, and the $N \times D_y$ matrix $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N]^T$ representing the mixing concentrations of D_y pure components in each compound. Both $\mathbf{X}$ and $\mathbf{Y}$ are preprocessed as zero-mean matrixes. $\boldsymbol{\Theta}$ is the $D_x \times D_y$ matrix of coefficients. $\mathbf{E}$ is the error matrix. Usually for Raman spectrum data sets, N is much smaller than D_x, if the Raman spectra are thought as data points in a high dimensional space, they stay in a low dimensional subspace. It is reasonable to project them into a low dimensional subspace (latent space):

$$\mathbf{T} = \mathbf{X}\mathbf{W}, \tag{2}$$

with columns of the $D_x \times K$ matrix $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_K]$ are the projection directions, and the $N \times K$ score matrix $\mathbf{T} = [\mathbf{t}_1, ..., \mathbf{t}_K]$ is a low dimensional representation of $\mathbf{X}$. And the previous full spectrum calibration model becomes the latent space calibration model:

$$\mathbf{Y} = \mathbf{T}\mathbf{Q} + \mathbf{E}, \tag{3}$$

with $\mathbf{Q}$ is the matrix of regression coefficients.

B. PLSR Methods

The goal of PLSR is to find those K projecting directions that make the unrelated score vectors $\{\mathbf{t}_i\}_{i=1}^{K}$ most representing $\mathbf{X}$ and most correlating to $\mathbf{Y}$ simultaneously:

$$\begin{aligned}
&\text{obj. } \max_{\mathbf{w}_i}\{||\sqrt{var(\mathbf{t}_i)}corr(\mathbf{t}_i, \mathbf{Y})||^2 = ||cov(\mathbf{t}_i, \mathbf{Y})||^2 \\
&\qquad = \mathbf{w}_i^T\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w}_i\} \\
&\text{s.t. } \mathbf{w}_i^T\mathbf{w}_i = 1;\ \mathbf{t}_i^T\mathbf{t}_j = 0 \\
&\text{for } i = 1, \ldots, K \text{ and } j = 1, \ldots, (i-1).
\end{aligned} \tag{4}$$

Here $\mathbf{t}_i = \mathbf{X}\mathbf{w}_i$, $var(\mathbf{t}_i)$ is the variance of $\mathbf{t}_i$, $corr(\mathbf{t}_i, \mathbf{Y})$ is the vector of correlation coefficients between $\mathbf{t}_i$ and each column of $\mathbf{Y}$, $cov(\mathbf{t}_i, \mathbf{Y})$ is the vector of covariances between $\mathbf{t}_i$ and each column of $\mathbf{Y}$, $||\mathbf{a}||$ is Euclidean norm of vector $\mathbf{a}$ and K is the component number. For $i = 1$, $\mathbf{w}_1$ is the first eigenvector of $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$ corresponding to the biggest eigenvalue. For $i = 2, \ldots, K$, because of the unrelated constraint $\mathbf{t}_i^T\mathbf{t}_j = 0$, there is no closed form solution for Equation (4). PLS2 and SIMPLS use different ways to satisfy this constraint.

PLS2 iteratively deflates $\mathbf{X}$ to get residual matrix $\mathbf{X}_i$ and get the corresponding projection direction $\mathbf{r}_i$ of the residual matrix:

$$\text{obj. } \max_{\mathbf{r}_i} \mathbf{r}_i^T\mathbf{X}_i^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}_i\mathbf{r}_i, \text{ s.t. } \mathbf{r}_i^T\mathbf{r}_i = 1, \tag{5}$$

with $\mathbf{X}_i$ is got from a deflation process in Algorithm 1, in which, $eig(\mathbf{A})$ means getting the first eigenvector of matrix $\mathbf{A}$ corresponding to the biggest eigenvalue. The projecting direction $\mathbf{r}_i$ is to project residual matrix $\mathbf{X}_i$, and the projecting directions of original data set $\mathbf{X}$ can be calculated as $\mathbf{W} = \mathbf{R}(\mathbf{P}^T\mathbf{R})^{-1}$ [16] or $\mathbf{W} = \mathbf{P}(\mathbf{P}^T\mathbf{P})^{-1}$ [12]. Hoskuldsson [9] proved that after the deflation, it will satisfy the unrelated constrain $\mathbf{t}_i^T\mathbf{t}_j = 0$ in Equation (4).

Algorithm 1 PLS2 Deflation Process

```
for $i = 1$ to $K$ do
    $\mathbf{r}_i = eig(\mathbf{X}_i^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}_i)$; % Get projection directions $\mathbf{r}_i$
    $\mathbf{t}_i = \mathbf{X}_i\mathbf{r}_i$; % Get score vectors $\mathbf{t}_i$
    $\mathbf{p}_i = \mathbf{X}_i^T\mathbf{t}_i/(\mathbf{t}_i^T\mathbf{t}_i)$; % Get loading vectors $\mathbf{p}_i$
    $\mathbf{X}_{i+1} = \mathbf{X}_i - \mathbf{t}_i\mathbf{p}_i^T$; % Get the residual matrices of $\mathbf{X}_i$
end for
Store $\mathbf{R} = [\mathbf{r}_1, ..., \mathbf{r}_K]$; $\mathbf{P} = [\mathbf{p}_1, ..., \mathbf{p}_K]$
```

Instead of finding projection directions of residual matrixes $\mathbf{X}_i$, SIMPLS directly finds projection directions of the original matrix $\mathbf{X}$ by projecting the cross covariance matrix $\mathbf{X}^T\mathbf{Y}$ on to orthogonal subspace $\mathbf{P}_i^{\perp} = \mathbf{I} - \mathbf{P}_{i-1}\mathbf{P}_{i-1}^{+}$ iteratively to satisfy the unrelated constrain, with $\mathbf{P}_{i-1}^{+} = (\mathbf{P}_{i-1}^T\mathbf{P}_{i-1})^{-1}\mathbf{P}_{i-1}^T$ is the Moore-Penrose inverse of $\mathbf{P}_{i-1}$ and $\mathbf{P}_{i-1} = [\mathbf{p}_1, \ldots, \mathbf{p}_{i-1}]$ is the loading matrix. The objective function of SIMPLS in each iteration is:

$$\text{obj. } \max_{\mathbf{w}_i} \mathbf{w}_i^T\mathbf{P}_i^{\perp}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{P}_i^{\perp}\mathbf{w}_i, \text{ s.t. } \mathbf{w}_i^T\mathbf{w}_i = 1. \tag{6}$$

$\mathbf{w}_i$ can be solved as the first eigenvector of $\mathbf{P}_i^{\perp}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{P}_i^{\perp}$ corresponding to the biggest eigenvalue. de Jong [4] proved that these $\mathbf{w}_i$ satisfy the constrain $\mathbf{w}_i^T\mathbf{X}^T\mathbf{X}\mathbf{w}_j = 0$ in Equation (4). And the detail algorithm can be found in [4].

After finding the projection directions $\mathbf{W}$ and the score matrix $\mathbf{T} = \mathbf{X}\mathbf{W}$, the matrix of regression coefficients $\mathbf{Q}$ in Equation (3) can be calculated as:

$$\mathbf{Q} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{Y} \tag{7}$$

And the relation between $\mathbf{X}$ and $\mathbf{Y}$ can be expressed as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Theta} + \mathbf{E}_y = \mathbf{X}\mathbf{W}\mathbf{Q} + \mathbf{E}_y, \tag{8}$$

When given a testing spectrum $\mathbf{x}$, we can predict the mixing concentrations $\mathbf{y}$ as:

$$\mathbf{y} = (\mathbf{x} - Mean(\mathbf{X}))\boldsymbol{\Theta} + Mean(\mathbf{Y}), \tag{9}$$

with $Mean(\mathbf{X})$ is the mean vector of the rows of $\mathbf{X}$.

C. Limitations of PLSR

The signals we got from Raman spectroscopy (Raman signals) mainly contain two parts: Raman spectrum (peaks) that have the fingerprint property to different materials and instable background which makes Raman signals irreproducible. Figure 1 shows two groups of Raman signals collected from two compound samples with components have different mixing concentrations. We can see the duplicates Raman signals generated from the same sample have different backgrounds. This inherent instable background is mainly because of the emission of fluorescence [7]. Besides, some instrumental factors, like variations in laser power or wavelength, optical train variations or irreproducible sample placement, and the change of position and angel of Ag or Au sol attached on analyte molecules during time [19], will also give instable spectrum. Traditional PLSR only considers the intensities information of Raman signals without separating the Raman spectrum from the instable background, which affects the quantity prediction accuracy.

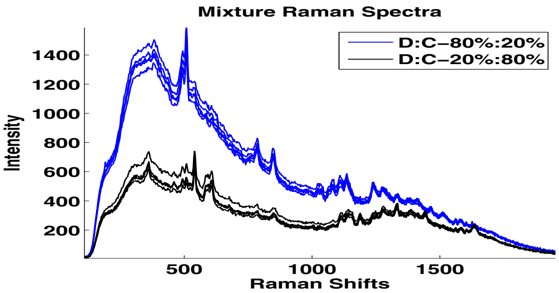

Figure 1: Instable background demonstration: Raman signals of two different samples. Signals with the same color are five duplicate Raman signals of one sample collected at different time.

III. CWT-PLSR

In this section, we introduce a continuous wavelet transform (CWT) based PLSR algorithm which can accurately remove the background and extract the Raman spectrum.

CWT [3] can be described as:

$$\mathbf{C}(a,b) = \int_R x(\tau)\psi_{a,b}(\tau)d\tau, \qquad (10)$$

with $x(\tau)$ is one Raman signal, τ is the time variable, here means different Raman shifts, $\psi_{a,b}(\tau) = \frac{1}{\sqrt{a}}\psi(\frac{\tau-b}{a})$ is any scaled and translated wavelet, $a = 1, 2, ..., s$ is the scale, $b = 1, 2, ..., D_x$ is the translation and $\mathbf{C}(a,b)$ is the 2D matrix of wavelet coefficients. Figure 2 shows the CWT coefficients of one Raman signal. The brightness of the figure represents the intensities of coefficients. At peak positions, the corresponding CWT coefficients are high, and the coefficients are increasing as the increasing of scales.

The CWT-PLSR algorithm includes two parts: training (modeling) part and testing (predicting) part. Given training data: mixture Raman signals $\mathbf{X}$ and mixing concentrations $\mathbf{Y}$, maximum wavelet scale s and PLS components number K, the training part is:
1. For every Raman signal (each row of $\mathbf{X}$), get its CWT coefficients $\mathbf{C}$ in Equation (10) with Mexican hat mother wavelet [5];
2. Calculate the average coefficients of $\mathbf{C}$ along the scale dimension as $Mean(\mathbf{C}) = \frac{1}{s}\sum_{a=1}^{s}\mathbf{C}(a,b)$, and store them in one row of matrix $\mathbf{D}$;
3. Instead of using $\mathbf{X}$, using $\mathbf{D}$ and $\mathbf{Y}$ to do PLSR, and return the PLSR coefficients $\mathbf{\Theta}$;
Then given a testing Raman signal $\mathbf{x}$, the testing part is:
1. Get the CWT coefficients $\mathbf{C}$ of $\mathbf{x}$, and calculate its average coefficients $\mathbf{d}$;
2. Estimate the mixing concentrations $\mathbf{y}$.
The whole algorithm is summarized in Algorithm 2.

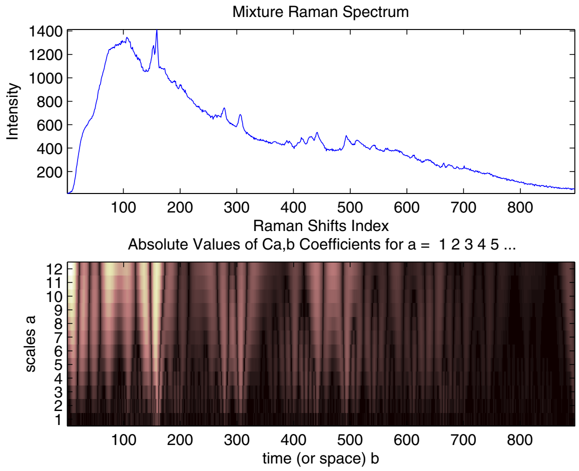

Figure 2: Scores of continuous wavelet transform on one Raman signal at different scales.

Algorithm 2 CWT-PLSR Algorithm

```
Input: $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{x}$, $K$, $s$
Output: $\mathbf{y}$
for $i = 1$ to $N$ do
    $\mathbf{C} = CWT(\mathbf{X}(i,:), s)$;
    $\mathbf{D}(i,:) = Mean(\mathbf{C})$;
end for
$\mathbf{\Theta} = PLSR(\mathbf{D}, \mathbf{Y}, K)$;
$\mathbf{C} = CWT(\mathbf{x}, s)$;
$\mathbf{d} = Mean(\mathbf{C})$;
$\mathbf{y} = (\mathbf{d} - Mean(\mathbf{D}))\mathbf{\Theta} + Mean(\mathbf{Y})$;
```

If we assume the baseline is slowly changing and monotonic in the peak support region, the baseline of the peak can be locally approximated as a constant G plus an odd function $B(\tau)$ defined in the peak support region and with the peak center as the original point [5]. The intensities of the Raman spectrum $x(\tau)$ at any region $[\tau_1, \tau_2]$ can be represented as following:

$$x(\tau) = P(\tau) + B(\tau) + G + E(\tau);\ \tau \in [\tau_1, \tau_2], \tag{11}$$

If the support region $[\tau_1, \tau_2]$ is the region of the Raman peak, $P(\tau)$ is the real Raman peak; otherwise if it is the region of the background, $P(\tau) = 0$. $B(\tau)$ is the background function with zero mean, G is a constant, $E(\tau)$ is the random noise. The coefficients in Equation (10) can be rewritten as:

$$\begin{aligned}\mathbf{C}(a,b) = &\int_R P(\tau)\psi_{a,b}(\tau)d\tau + \int_R B(\tau)\psi_{a,b}(\tau)d\tau \\ &+ \int_R G\psi_{a,b}(\tau)d\tau + \int_R E(\tau)\psi_{a,b}(\tau)d\tau.\end{aligned} \tag{12}$$

Because the wavelet function $\psi_{a,b}(\tau)$ is zero-mean, the third term in Equation (12) is zero. And for symmetric wavelet, like Mexican Hat wavelet, $B(\tau)$ is an even function, so the second term is zero. Also the zero-mean random noise function $E(\tau)$ tends to be canceled out by, so the fourth term tends to be zero. Thus, only the term with real peak $P(\tau)$ is left in Equation (12). That is to say, as long as the background is slowly changing and locally monotonic in the peak support region with random noise, it will be automatically removed in calculating the CWT coefficients.

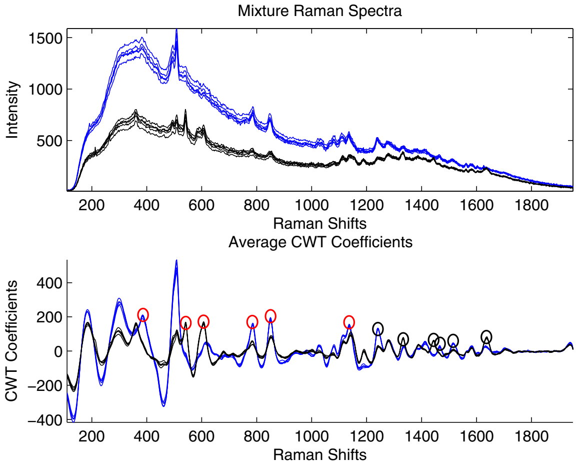

Figure 3: Average CWT coefficients along different scales, red circles show the stable extracted Raman peaks, black circles show weak peaks.

If the mother wavelet is treated as a mask function, the integration in Equation (10) is essentially a pattern matching, and the coefficients $\mathbf{C}$ are scores that measure how much the shapes of the signal matching to the mask function with different scales, at each Raman shift. For peaks extraction purpose, Mexican hat function is chosen here as the mother wavelet, since it has the shape of a peak. Then the positions at Raman peaks tend to have high scores and backgrounds tend to have low scores. And at smaller scales, the scores measure the shape in narrow ranges; at bigger scales, the scores measure the peak shape in wider ranges. So the mean values of these scores along different scales will give a robust estimation of peaks. Figure 3 shows the average CWT coefficients along different scales of different Raman signals. It removes the instable background and extracts the relative stable Raman spectrum. Also because the high intensity backgrounds are removed, the weak peaks can be used more efficiently.

IV. Experiment

To evaluate the effectiveness of CWT-PLSR for quantitative analysis of Raman spectrum, in this section, we compare it with PLSR methods (PLS2 and SIMPLS), baseline correction based PLSR and other latent variables regression methods, testing on three Raman signal data sets and using three cross validation methods.

A. Data Sets and Cross Validation

The Raman signals we use are collected from the Raman spectroscopy with 20×, 0.4_{NA} lens and $785nm$ laser wavelength. Raman Shifts range from -79.65cm^{-1} to 2071.80cm^{-1} with 1044 values. To avoid the influence of the strong intensity from Rayleigh Scattering and for CWT calculation purpose, from 1044 Raman Shifts, we extract 896 (71th-966th). All nano-tags are made from 54.67nm Au nano-particles, coated with the dyes (summarized in Table I). All pure nano-tag solutions are made with a concentration of $1.1e^{10}$ nanotags/ml. Then with 11 and 21 mixing volume ratios (shown in Figure 4), pure nano-tag solutions are mixed into mixture solutions. From each solution, 5 duplicate Raman spectra are collected, with 20s time interval. So for each data set, we have mixture signals $\mathbf{X} \in \Re^{N\times 896}$, N is summarized in Table I. And the mixing volume ratios of all duplicates $\mathbf{Y} \in \Re^{N\times D_y}$ can be treated as relative concentrations of each pure nano-tags. Also to reduce the influence of instability of Raman signals, we can get the average signals $\bar{\mathbf{X}} \in \Re^{\frac{N}{5}\times 896}$ by taking average of each 5 duplicates. These average signals together with the mixing volume ratios $\bar{\mathbf{Y}} \in \Re^{\frac{N}{5}\times D_y}$ are shown in Figure 4.

Table I: Data sets summary

Data set	Dyes on nano-tags	D_y	N
1	DTTC, CV	2	11 × 5
2	HITC, IR140	2	11 × 5
3	DOTC, DTTC, HITC, IR140	4	21 × 5

In order to take fully use of all three data sets, we design three cross-validation methods to evaluate the predicting ability of methods: Cross Validation on Duplicate testing spectra (CVD): all 5 duplicate spectra of the same mixing ratio are treated as the testing samples, and all the other mixture signals are treated as the training samples. Iteratively, until every duplicate is treated as the testing

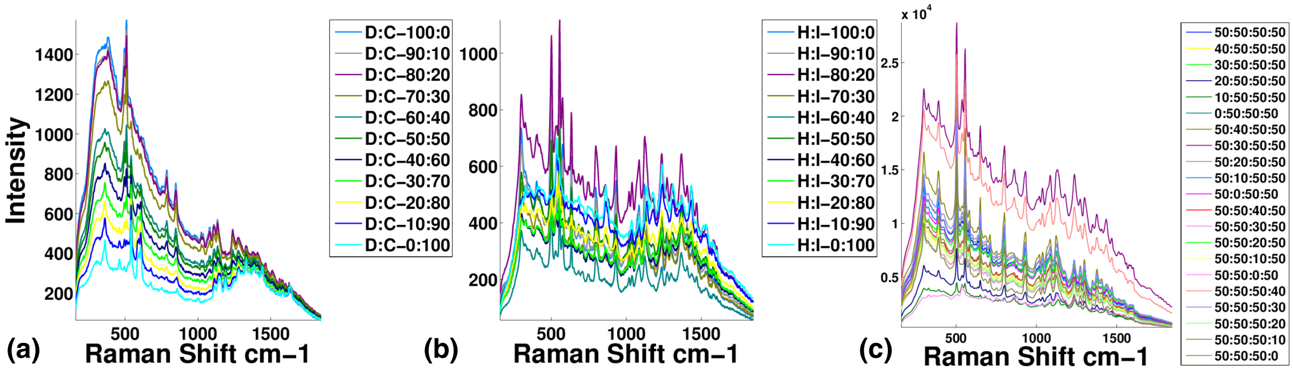

Figure 4: Average Raman spectra of mixture nano-tags with different mixing volume ratios in three data sets. These ratios are: (a) DTTC(D):CV(C); (b) HITC(H):IR140(I); (c) DOTC:DTTC:HITC:IR140.

sample once. Cross Validation on Average testing spectra (CVA): the average spectrum of the 5 duplicates with the same ratio is treated as the testing sample and all the other duplicate spectra are treated as training samples. Iteratively, until every average spectrum is treated as the testing sample once. Cross Validation on Average testing Average training spectra (CVAA): the average spectrum of the 5 duplicates with the same ratio is treated as the testing sample and all the other average spectra are treated as training samples. Iteratively, until every average spectrum is treated as the testing sample once.

Square Root of Mean Squares Error (RMSE) is used as the criterion for evaluating the prediction accuracy. It is defined as: $RMSE = (\sum_{i=1}^{N}\sum_{j=1}^{D_y}(\hat{y}_{i,j} - y_{i,j})^2/ND_y)^{1/2}$, with $\hat{y}_{i,j}$ and $y_{i,j}$ are the estimated ratio and ground truth ratio respectively of the ith sample and the jth dye.

B. Experimental Setting

We compare CWT-PLSR with the following methods: Ridge Regression (RR) [8]; Principle Component Regression (PCR) [10]; Orthonormalized PLS (OPLS) [17]; PLS2 [9]; SIMPLS(SIM) [4]; linear programming baseline correction [1] based PLSR (P-PLS2 and P-SIM) and iteratively curve-fitting baseline correction [6] based PLSR (I-PLS2 and I-SIM). RR needs to add a parameter κ to $(\mathbf{X}^T\mathbf{X})^{-1}$ of the least square solution of (1) to solve the singularity problem: $\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$, κ needs to be a small number and we can simply set it as 0.1. Similar to RR, OPLS also needs a parameter to remove the singularity problem, and it is also set as 0.1.

To maximize the performance of all the latent variable regression methods, the component number K needs to be optimized for each data set and each cross validation method. Every possible component number is tested, the one giving the lowest RMSE returns as the optimized one K^*. The results are in Table II.

Beside optimized component number, other parameters also need to be optimized. For baseline correction based PLSR methods (P-PLS2, P-SIM, I-PLS2 and I-SIM), the polynomial curve-fitting order p needs to be decided. For CWT-PLSR (PLS2 and SIM), the optimized wavelet scale numbers s needs to be found (PLS2-s and SIM-s). All of the optimized parameters are found by the same way as K^*, and summarized in Table III.

C. Results and Analysis

Table III: Optimized parameters

Parameter	Data Set 1			Data Set 2			Data Set 3		
	D	A	AA	D	A	AA	D	A	AA
P-PLS2-p	5	5	5	6	11	7	7	7	7
P-SIM-p	10	10	10	11	11	11	10	10	10
I-PLS2-p	8	9	8	7	7	7	10	9	9
I-SIM-p	8	8	12	9	9	7	10	10	9
PLS2-s	15	15	14	30	30	28	10	10	9
SIM-s	15	15	14	30	30	29	8	8	8

In Table II, we show the results of different regression methods, using three data sets and three cross-validation methods, and the bold face represents the best result. Most of the results of latent variable regression methods (PLS2, SIM, PCR, OPLS) are better than those of RR, which means the latent space calibration model is better than the full spectrum calibration model. Baseline correction based PLSR (P-PLS2, P-SIM, I-PLS2 and I-SIM) are usually better than PLSR (PLS2 and SIM), since they reduce the influence of the instable backgrounds. But they are not always better because locally the baselines may not be perfect backgrounds, and the hard cut of these baselines will lose Raman peak information (as mentioned in section II-C). The results of CWT-PLS methods (CWTPLS2 and CWTSIM) are always better than other methods, and the optimized component numbers of them are lower and more stable than other methods. These because CWT-PLS methods more effectively reduce the instable background and random noises, and extract more useful peak information.

The results of CVA and CVAA tend to be better than CVD, since the average testing signals will reduce the

Table II: RMSE and K^* for each method, on three data sets and three cross-validation methods (CVD, CVA and CVAA). The results are shown as: RMSE (K^*).

Methods	Data Set One			Data Set Two			Data Set Three		
	CVD	**CVA**	**CVAA**	**CVD**	**CVA**	**CVAA**	**CVD**	**CVA**	**CVAA**
RR	2.94 (/)	1.61 (/)	2.34 (/)	4.36 (/)	4.23 (/)	4.41 (/)	8.91 (/)	9.67 (/)	11.69 (/)
PCR	2.81 (6)	1.63 (9)	2.25 (7)	4.14 (5)	4.04 (5)	4.12 (6)	8.22 (13)	8.99 (13)	10.15 (10)
OPLS	2.94 (1)	1.54 (2)	2.08 (2)	4.20 (2)	4.09 (2)	4.41 (1)	8.91 (4)	9.67 (4)	11.69 (4)
PLS2	2.93 (8)	1.59 (9)	2.20 (4)	4.22 (5)	4.11 (5)	3.86 (4)	7.90 (10)	8.65 (10)	9.75 (10)
SIM	2.72 (4)	1.61 (9)	2.27 (4)	4.13 (4)	4.03 (4)	3.90 (4)	8.30 (10)	9.06 (10)	10.68 (8)
P-PLS2	2.78 (5)	1.47 (5)	1.56 (5)	4.27 (6)	3.99 (4)	3.74 (4)	5.78 (21)	5.50 (21)	6.09 (19)
P-SIM	2.84 (5)	1.61 (5)	2.21 (4)	4.52 (10)	4.29 (4)	4.30 (4)	5.97 (21)	5.67 (21)	6.48 (11)
I-PLS2	2.70 (3)	1.50 (3)	1.92 (4)	3.99 (4)	3.75 (4)	3.86 (4)	5.71 (24)	5.44 (24)	6.18 (19)
I-SIM	2.66 (3)	1.51 (3)	1.88 (4)	4.56 (6)	4.42 (6)	4.53 (4)	5.73 (24)	5.46 (24)	6.19 (19)
CWTPLS2	**2.56** (3)	**1.43** (3)	**1.45** (3)	**3.28** (3)	**3.16** (3)	**3.13** (3)	5.44 (8)	5.26 (8)	5.31 (8)
CWTSIM	2.65 (3)	1.47 (3)	1.46 (3)	3.39 (3)	3.27 (3)	3.24 (3)	**5.34** (8)	**5.15** (8)	**5.17** (8)

influence of instable backgrounds. And CVA is better than CVAA, since CVA has more training samples than CVAA. The results of data set 1 is the best among three data sets. Because first, the mixing concentrations of two nano-tags are related (summation equals to 100%), if one concentration can be estimated, the other is easy to get; second, data set 1 has one nano-tag (DTTC) dominating the Florence background and this background tends to linearly related to its concentration, so the instable backgrounds in data set 1 do not affect the estimation of the concentration of DTTC; third, the Raman peaks of two nano-tags in data set 1 have less overlaps than the other two data sets, which also decreases the difficulty of prediction. In data set 2 the mixing concentrations of two nano-tags are also related, but there is no dominating Florence background, so the instable mixing background will affect the prediction more than data set 1. Plus, there are more overlaps between Raman peaks of two nano-tags, so the results of data set 2 are worse than data set 1. In data set 3, one nano-tag (HITC) also has a dominated background, but its results are the worst. This because it contains four nano-tags, are there are more overlaps between the Raman peaks in data set 3, so it increases the difficulty of prediction. Also the mixing concentrations of four nano-tags are not related, then the dominated background will affect the prediction of the other three. So baseline correction based methods and CWT-PLSR improve most in data set 3.

V. Conclusions

In this paper, we give a new CWT-PLSR method for quantitative analysis of Raman spectrum. It treats the average CWT coefficients along different scales as the estimation of Raman spectrum (peaks) and combines mixing concentrations to do PLSR. This method can effectively reduce random noises and remove the instable baseline, and extract the Raman peaks. And the experimental results illustrate that it outperforms the direct PLSR and the baseline correction based PLSR, and is a robust method for quantitative analysis of Raman spectrum. CWT-PLSR is a good research direction, and now we are directly using the average CWT coefficients to do PLSR. But these CWT coefficients unavoidably contain valley points with negative intensities. In the future work, we will analyze whether these valley points affect the prediction.

References

[1] S.-J. Baek, A. Park, A. Shen, and J. Hu. A background elimination method based on linear programming for Raman spectra. *J. Raman Spectrosc.*, 42:1987–1993, 2010.

[2] S. Bell and M. Sirimuthu. Quantitative surface-enhanced Raman spectroscopy. *Chem. Soc. Rev.*, 37:1012–1024, 2008.

[3] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA., 1992.

[4] S. de Jong. SIMPLS: an alternative approach to partial least squares regression. *Chemometrics Intell. Lab. Syst.*, 18:251–263, 1993.

[5] P. Du, W. A. Kibbe, and S. M. Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22:2059–2065, 2006.

[6] F. Gan, G. Ruan, and J. Mo. Baseline correction by improved iterative polynomial fitting with automatic threshold. *Chemometrics Intell. Lab. Syst.*, 82:59–65, 2006.

[7] C. Gobinet, V. Vrabie, M. Manfait, and O. Piot. Preprocessing methods of Raman spectra for source extraction on biomedical samples: application on paraffin-embedded skin biopsies. *IEEE Trans. Biomed. Eng.*, 56:1371–1382, 2009.

[8] A. E. Hoerl and R. W. Kennard. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12:69–82, 1970.

[9] A. Hoskuldsson. PLS regression methods. *J. Chemometr.*, 2:211–228, 1988.

[10] I. T. Jolliffe. *Principal Component Analysis, Second Edition.* Springer, 2002.

[11] S. Keren, C. Zavaleta, Z. Cheng, A. de la Zerda, O. Gheysens, and S. S. Gambhir. Noninvasive molecular imaging of small living subjects using Raman spectroscopy. *PNAS*, 105:5844–5849, 2008.

[12] S. Li, J. Gao, J. O. Nyagilo, and D. P. Dave. Probabilistic partial least square regression: a robust model for quantitative analysis of Raman spectroscopy data. In *Proceedings of IEEE BIBM*, 2011.

[13] J. Lombardi and R. Birke. A unified approach to surface-enhanced Raman spectroscopy. *J. Phys. Chem. C*, 112:5605–5617, 2008.

[14] R. Rosipal and N. Kramer. Overview and recent advances in partial least squares. *LNCS*, 3940:34–51, 2006.

[15] H. Wold. *Path models with latent variables: the NIPALS approach*. Academic, 1975.

[16] S. Wold, M. Sjstrma, and L. Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics Intell. Lab. Syst.*, 58:109–130, 2001.

[17] K. Worsley, J. B. Poline, K. J. Friston, and A. C. Evans. Characterizing the response of PET and fMRI data using multivariate linear models. *Neuroimage*, 6:305–319, 1997.

[18] C. L. Zavaleta, B. R. Smith, I. Walton, W. Doering, G. Davis, B. Shojaei, M. J. Natan, and S. S. Gambhir. Multiplexed imaging of surface enhanced Raman scattering nanotags in living mice using noninvasive Raman spectroscopy. *PNAS*, 106:13511–13516, 2009.

[19] L. Zhang, Q. Li, W. Tao, B. Yu, and Y. Du. Quantitative analysis of thymine with surface-enhanced raman spectroscopy and partial least squares (PLS) regression. *Anal. Bioanal. Chem.*, 398:1827–1832, 2010.