35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Continuous Wavelet Transform based Continuum Regression for Quantitative Analysis of Surface-enhanced Raman Spectra

Shuo Li[1], Mingon Kang[1], James O. Nyagilo[2], Baoju Zhang[3], Xiaoyong Wu[3], Digant P. Dave[2] and Jean Gao[1]

Abstract— **Surface-enhanced Raman spectroscopy (SERS) has been a routine method used as an analytical tool to do the quantitative analysis of materials. The difficulties mainly come from the inherent instable backgrounds of Raman signals, which unexpectedly increase the intensities of Raman spectra and from the high dimension small sample number problem of Raman data sets, which demands the ability of feature extraction from the regression methods. Targeting at removing the instable background meanwhile extracting the Raman peaks and taking full use of the information of Raman peaks to extract features, we design a new framework that combines new continuum regression (NCR) with continuous wavelet transform (CWT) to do the quantitative analysis of Raman spectra. The experiment results show its performance beats the state of the art methods.**

I. INTRODUCTION

Relying on Raman scattering, Raman spectroscopy has been regarded as one of the most sensitive techniques for chemical analysis giving the unique spectral fingerprint of every chemical compound. When the monochromatic laser light interacts with molecular vibrations or other excitations, the energy of the laser photons will be 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 [1]. With the development of the SERS-nanoparticles, normally a silver or gold colloid or a substrate containing silver or gold, which are designed to enhance the inherently weak magnitude of Raman scattering, surface-enhanced Raman spectroscopy (SERS) 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 has been applied by Cheung *et al*. [3] to quantify the banned food dye; by Lai*et al*.[4] to analyze sulfa drugs; by Strickland and Batt [5] to detect carbendazim and by Stokes *et al*. [6], Graham and Faulds [7] and Zhang *et al*. [8] to detect DNA sequence. It also has been used in the field of biomedical diagnostics, especially in the application of cancer detection research [9], [10], [11]. 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.

In order to estimate the amount of the nanoparticles and so the amount of receptor proteins in cancer cells, the so called Quantitative Analysis of Raman Spectrum (QARS), which is from intensities of the Raman spectrum of one compound to determine the mixing concentration of each component, is the key job. The mixture spectrum of a compound approximately equals to the summation of all the pure spectrum [9], besides, within certain range of concentrations, the intensities of Raman spectrum are approximately linearly related to the concentration of each pure component [10]. Based on these two properties, two QARS models are commonly used: Direct Classical Least Squares (DCLS) and Multivariate Calibration model (MC). Li *et al*. [14], [15] showed MC and the state of the art MC methods, Partial least squares regression (PLSR), are usually superior to DCLS, because of the inherent instable problem of Raman spectra.

In reality, the Raman signals collected from Raman spectroscopy unavoidably contain background intensities, which disturb the QARS. Li *et al*. [16] presented a continuous wavelet transform (CWT) based PLSR method, which can effectively extract Raman peaks (spectrum) and remove backgrounds. From the feature extraction point of view, PLSR assigns higher weights to the Raman shifts that have both big variances of intensities and high correlations with concentrations. But the proportions of two criteria in the objective are fixed to be equal, which limits the flexibility of the model. Continuum regression (CR) methods [17], [18], [19] can adjust the proportions of two criteria and assign weights to Raman shifts in a flexible way. In this paper, a new framework combining CWT and NCR method [19] for the first time is presented, which can effectively solve the instable background problem of Raman spectra and reasonably assign weights to all Raman shifts for the QARS. The performance is better than CWT-PLSR, baseline correction based PLSR and CR methods.

II. METHOD

In this section we will first introduce the MC model and describe the objective functions of three latent variable regression (LVR) methods; then list three CR methods and discuss their flexibility of combining two criteria; in the end explain the principles of peak extraction by using CWT and describe the details of CWT based CR method.

A. Multivariate Calibration Model and LVR Methods

Based on the two properties of Raman spectrum mentioned in section one, the MC model [14] can be used to learn

Shuo Li, Mingon Kang and Jean Gao are with the Computer Science and Engineering Department, University of Texas at Arlington, Arlington, USA Emails: {shuo.li, gao}@uta.edu

[2]James O. Nyagilo and Digant P. Dave are with the Bioengineering Department, University of Texas at Arlington, Arlington, USA Emails: {james.nyagilo, ddave}@uta.edu

[3]Baoju Zhang and Xiaoyong Wu are with School of Physics and Electronics Information, Tianjin Normal University, Tianjin, China

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

the relation between the intensities of Raman spectra and the concentrations of components: $\mathbf{Y} = \mathbf{X\Theta} + \mathbf{E}$, with the $N \times D_x$ matrix $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]^T$ representing N Raman spectra of compounds, the $N \times D_y$ matrix $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N]^T$ representing the ground truth mixing concentrations of D_y pure components in each compound, $\mathbf{\Theta}$ being the $D_x \times D_y$ matrix of coefficients need to be found and $\mathbf{E}$ being the error matrix. Then given a new mixture spectrum $\mathbf{x}$, the concentrations of each component can be predicted as $\hat{\mathbf{y}}^T = \mathbf{x}^T\mathbf{\Theta}$. Since normally N is much smaller than D_x, LVR methods [20] are usually used to solve this high dimensional, collinearity multivariate regression problem. Variables of $\mathbf{X}$ (Raman shifts) are linearly combined into the low dimensional latent variables (LVs) $\mathbf{T} = \mathbf{XW} = [\mathbf{t}_1, ..., \mathbf{t}_K]$, with $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_K]$ are weights vectors; then the regression is done between $\mathbf{Y}$ and $\mathbf{T}$. To find $\mathbf{W}$, LVR methods have different objective functions. Principle component regression (PCR) [21] maximizes variance of LVs: $\max_{\mathbf{w}_i} var(\mathbf{t}_i)$. Reduced-rank regression (RRR) [22] maximizes the correlation between LVs and concentrations: $\max_{\mathbf{w}_i} ||corr(\mathbf{t}_i, \mathbf{Y})||^2$. PLSR is to maximize covariance between LVs and concentrations:

$$\max_{\mathbf{w}_i} ||cov(\mathbf{t}_i, \mathbf{Y})||^2 = \max_{\mathbf{w}_i} ||corr(\mathbf{t}_i, \mathbf{Y})||^2 var(\mathbf{t}_i), \quad (1)$$

which is explained in [14], [15] as for both best representing spectra and best approximating concentration and is a compromise between PCR and RRR.

B. Continuum Regression: Reasonably Using Raman Peaks

Similar to LVR methods, CR methods also need to find the weights $\mathbf{W}$ and then latent variables $\mathbf{T}$. But different with PLSR, who fixes the proportions of $var(\mathbf{t}_i)$ and $corr(\mathbf{t}_i, \mathbf{Y})$ in (1), CR methods combine both criteria in the objective function with an adjustable weight parameter.

de Jong presented a PCovR method [17], whose objective function is described as:

$$\text{obj. } \min_{\mathbf{T}} \alpha||\mathbf{X} - \mathbf{TP}_x||^2 + (1-\alpha)||\mathbf{Y} - \mathbf{TP}_y||^2. \quad (2)$$

When $\alpha = 0$, it is RRR; when $\alpha = 1$, it is PCR; when $0 < \alpha < 1$, it is a compromise between two. The limitations are that it does not include PLSR and the objective function is the summation of two criteria (explained in [17]).

Another CR method is called SCR [18] (or called canonical ridge analysis in [23]), whose objective function is:

$$\text{obj. } \max_{\mathbf{w}_i} \frac{||cov(\mathbf{Xw}_i, \mathbf{Y})||^2}{(1-\alpha)||\mathbf{Xw}_i||^2 + \alpha||\mathbf{w}_i||^2}. \quad (3)$$

When $\alpha = 0$, it equals to RRR, when $\alpha = 1$, it equals to PLSR. The limitation of SCR is that it only compromises between RRR and PLSR, it can not achieve PCR.

To overcomes both limitations, Li *et al*. [19] presented a new continuum regression (NCR) method whose objective function is:

$$\begin{aligned} &\text{obj. } \max_{\mathbf{w}_i} [\mathbf{w}_i^T(\mathbf{X}^T\mathbf{X})^{1-\alpha}\mathbf{w}_i]^{-1}(\mathbf{w}_i^T\mathbf{X}^T\mathbf{YY}^T\mathbf{Xw}_i) \\ &\text{s.t. } \mathbf{t}_i^T\mathbf{t}_j = 0, i = 1, ..., K, j = 1, ..., i-1. \end{aligned} \quad (4)$$

When $\alpha = 0$, it is RRR; when $\alpha = 1$, it is PLS; when $\alpha = \infty$, the portion of $\mathbf{w}_i^T\mathbf{X}^T\mathbf{YY}^T\mathbf{Xw}_i$ can be ignored, and it becomes PCR.

These different objective functions of LVR and CR methods decide the different weights they assign to Raman shifts (RS). PCR gives more weights to the RS that have bigger variances of the intensities. But these RS are not guarantied to be the Raman peaks which should also be correlated with concentrations. Random peaks or noisy peaks, instead of weak Raman peaks, may get more weights. RRR gives more weights to the RS that are more correlated with the concentrations. But it may ignore the main Raman peaks (peaks that have high intensities), and give more weights to some weak peaks or even background. PLSR gives higher weights to the RS that have both big variances of intensities and high correlations with concentrations, which are more likely to be the positions of main Raman peaks. By controlling the parameter α, CR methods can adjust the proportion of each criterion in the objective. When the optimized α is found, the weights are given to the RS in a reasonable way that more important Raman peaks get more weights.

C. CWT: Raman Peaks Extraction

LVR and CR methods can not solve the inherent instable background problem of Raman signals [16], which is mainly because of the emission of fluorescence [24] and instrumental factors [25]. Li *et al*. [16] presented a CWT-PLSR method to solve the problem. CWT [26] is described as

$$\mathbf{C}(a, b) = \int_R x(\tau)\psi_{a,b}(\tau)d\tau, \quad (5)$$

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 function, $a = 1, 2, ..., s$ is the scale, $b = 1, 2, ..., D_x$ is the translation, $\psi(\tau)$ is the mother wavelet function and $\mathbf{C}(a, b)$ is the 2D matrix of wavelet coefficients.

Li *et al*. [16] showed that if the baseline is assumed to be slowly changing and monotonic in the peak support region, the noises are random noises and the mother wavelet function is an even function, after the CWT, the baseline and the noises can be automatically removed. If the mother wavelet is treated as a mask function, the integration in (5) 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 RS. For peaks extraction purpose, Mexican hat function is chosen 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. 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 the heights of peaks.

D. CWT-NCR

In order to take full use of the Raman peak information and give reasonable weights to RS, we combine CWT

and NCR into the CWT-NCR algorithm (summarized in Algorithm 1), which includes 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 CR components number K, the training part is:
1. For every Raman signal (each row of $\mathbf{X}$), get its CWT coefficients $\mathbf{C}$ in (5) with Mexican hat mother wavelet [27];
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 NCR, whose algorithm is described in [19] and return the matrix of 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 $\mathbf{y}$.

Algorithm 1 CWT-NCR Algorithm

Input: $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{x}$, K, s, α
Output: $\mathbf{y}$

1: **for** $i = 1$ to N **do**
2: $\quad\mathbf{C} = CWT(\mathbf{X}(i,:), s)$;
3: $\quad\mathbf{D}(i,:) = Mean(\mathbf{C})$;
4: **end for**
5: $\mathbf{\Theta} = NCR(\mathbf{D}, \mathbf{Y}, K, \alpha)$;
6: $\mathbf{c} = CWT(\mathbf{x}, s)$;
7: $\mathbf{d} = Mean(\mathbf{c})$;
8: $\mathbf{y} = (\mathbf{d} - Mean(\mathbf{D}))\mathbf{\Theta} + Mean(\mathbf{Y})$;

III. EXPERIMENT

To evaluate the effectiveness of the new framework CWT-NCR for quantitative analysis of Raman spectrum, in this section, we compare seven methods PCR [21], RRR [22], PLS2 [12], SIMPLS [13], PCovR [17], SCR [18] and NCR [19], combined with CWT and two baseline correction methods: linear programming baseline correction [28] and iteratively curve-fitting baseline correction [29], testing on three real Raman signal data sets.

A. Data Sets

The Raman signals are collected from the Raman spectroscopy with $20\times$, 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, from 1044 Raman shifts, we extract 896 (71th-966th). All nano-tags are made from 54.67nm Au nano-particles, coated with dyes: DTTC and Cresyl violet (CV) (in data set one); HITC and IR140 (in data set two); DOTC, DTTC, HITC and IR140 (in data set three). All pure nano-tag solutions are made with a concentration of $1.1e^{10}$ nanotags/ml. Then with 11 mixing volume ratios (shown in Fig. 1) we mix two pure nano-tags solutions in the first two data sets, with 21 mixing volume ratios $\{(25\% : 25\% : 25\% : 25\%), (20\% : 25\% : 25\% : 25\%), \ldots, (0 : 25\% : 25\% : 25\%), (25\% : 20\% : 25\% : 25\%), \ldots, (25\% : 25\% : 25\% : 0)\}$, we mix four pure nano-tags solutions in the third data set, and get three groups of mixture nano-tag solution samples. These mixing volume ratios can be treated as relative concentrations of each pure nano-tags. From each sample, 5 duplicate Raman signals are collected, with 20s time interval. So for data set one and two, we have 55 mixture signals, and for data set three, we have 105. In order to reduce the influence of instability of Raman signals, we also get the average signals by taking average of each 5 duplicates. These average signals are shown in Fig. 1.

B. Experiment Design

In order to evaluate the performance of each method, we design a cross-validation method as follow: each average signal of 5 duplicates is treated as the testing sample once and all the other duplicate signals with different mixing ratios are treated as training samples.

Root Mean Square 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.

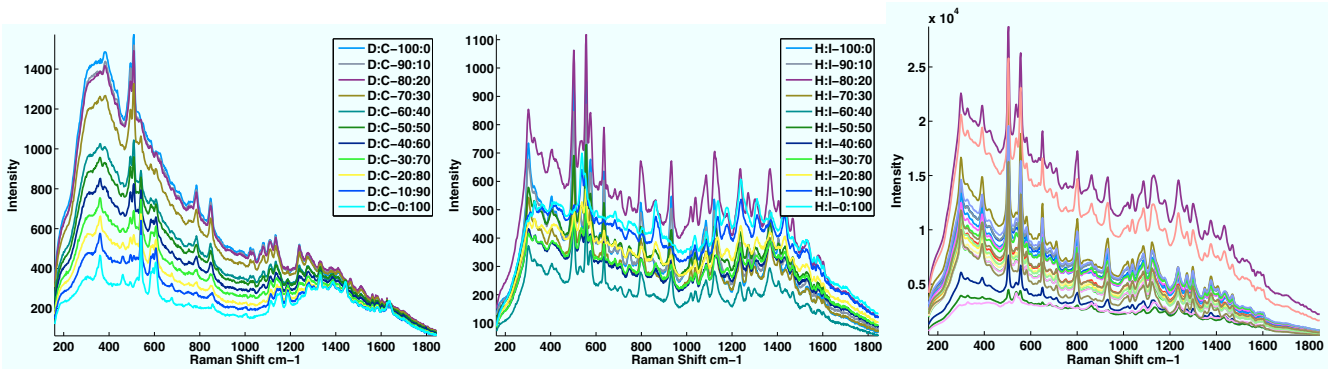
To maximize the performance of all methods, several parameters needs to be optimized, including the polynomial curve-fitting order of the baseline correction methods $pOrder$, maximum wavelet scale numbers s, α of CR methods, the component number of LVR and CR methods K. Different values are tested ($pOrder$ is from 3 to 10; s is from 1 to 20; $\alpha = \{0, 0.05, 0.1, 0.15, ..., 0.95, 1\}$ for PCovR and SCR; $\alpha = \{0, 0.05, 0.1, 0.15, ..., 0.95, 1, 2, 4, 6, 8, 10\}$ for NCR; K is from 1 to 30), the one giving the lowest RMSE is returned as the optimized parameter.

C. Results and Discussion

In this part, we show the RMSE corresponding to the optimized parameters and the optimized component number K in Table I. From columns of each data set, we can see methods based on baseline correction methods are better than without any preprocessing, methods based on CWT are better than those based on baseline correction methods. From rows of Table I we can see PLSR methods (PLS2 and SIMPLS) are normally better PCR and RRR; SCR and PCovR are usually better PLSR; NCR are the best. The results of CWT-NCR are always the best.

IV. CONCLUSIONS

Raman spectroscopy has been regarded as one of the most sensitive techniques that can provide the unique spectral information of analytes. QARS has been the key job of many biomedical applications. This paper presents a new framework that combines CWT and NCR. The advantage is it can effectively extracts the heights information of Raman peaks, reasonably assigns weights to Raman peaks and increases the predicting accuracy of the MC model. The limitation of the framework is there are three parameters need to be decided for each data set, which complicates the



(a) Data set 1: DTTC-CV(D:C) (b) Data set 2: HITC-IR140(H:I) (c) Data set 3: DOTC-DTTC-HITC-IR140

Fig. 1: Average Raman signals and mixing volume ratios. D:C-90:10, for example, means the ratio of mixing volumes of DTTC and CV is 90% : 10%.

TABLE I: The results are shown as: RMSE (K). Columns title **ori** means results without any preprocessing, **bc1** is iteratively curve-fitting baseline correction, **bc2** is linear programming baseline correction, **cwt** is continuous wavelet transform.

Methods	Data Set One				Data Set Two				Data Set Three			
	ori	**bc1**	**bc2**	**cwt**	**ori**	**bc1**	**bc2**	**cwt**	**ori**	**bc1**	**bc2**	**cwt**
PCR	1.66(10)	1.60(19)	1.40(9)	1.41(3)	3.92(4)	3.06(3)	3.51(4)	3.48(16)	4.20(11)	2.81(14)	2.93(20)	2.72(9)
RRR	1.80(30)	1.66(28)	1.37(27)	1.39(3)	3.72(17)	3.06(8)	3.70(9)	3.21(13)	4.22(28)	3.15(22)	3.01(30)	2.82(24)
PLS2	1.67(5)	1.50(3)	1.28(3)	1.43(3)	4.06(9)	3.04(3)	3.46(4)	3.13(3)	4.22(11)	2.72(24)	2.75(21)	2.83(20)
SIM	1.68(5)	1.51(3)	1.26(3)	1.46(3)	4.05(4)	3.14(3)	3.49(4)	3.40(3)	4.18(11)	2.73(23)	2.77(21)	2.83(18)
PCovR	1.66(10)	1.53(30)	1.32(28)	1.41(3)	3.60(25)	3.06(3)	3.23(7)	2.55(16)	4.19(11)	2.81(14)	2.91(30)	2.68(9)
SCR	1.66(7)	1.50(3)	1.28(6)	1.34(3)	3.77(10)	3.04(3)	3.28(6)	3.00(9)	4.21(11)	2.72(30)	2.75(21)	2.67(16)
NCR	1.37(18)	1.30(18)	1.27(3)	**1.21**(3)	3.72(17)	3.03(3)	3.12(4)	**2.49**(24)	4.08(11)	2.70(19)	2.72(20)	**2.66**(19)

quantitative analysis. Our future work is to find good way to decide those parameters.

REFERENCES

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

[2] S. E. J. Bell and N. M. S. Sirimuthu, "Quantitative surface-enhanced Raman spectroscopy," *Chem. Soc. Rev.*, vol. 37, pp. 1012–1024, 2008.

[3] W. Cheung, I. T. Shadi, Y. Xu, and R. Goodacre, "Quantitative analysis of the banned food dye sudan-1 using surface enhanced Raman scattering with multivariate chemometrics," *J. Phys. Chem. C*, vol. 114, pp. 7285–7290, 2010.

[4] K. Lai, F. Zhai, Y. Zhang, X. Wang, B. A. Rasco, and Y. Huang, "Application of surface enhanced Raman spectroscopy for analyses of restricted sulfa drugs," *Sens. and Instrumen. Food Qual.*, vol. 5, pp. 91–96, 2011.

[5] A. Strickland and C. Batt, "Detection of carbendazim by surface-enhanced Raman scattering using cyclodextrin inclusion complexes on gold nanorods," *Anal. Chem.*, vol. 81, pp. 2895–2903, 2009.

[6] R. Stokes, A. Macaskill, P. Lundahl, W. Smith, K. Faulds, and D. Graham, "Quantitative enhanced Raman scattering of labeled DNA from gold and silver nanoparticles," *Small.*, vol. 3, pp. 1593–1601, 2007.

[7] D. Graham and K. Faulds, "Quantitative SERRS for DNA sequence analysis," *Chem. Soc. Rev.*, vol. 37, pp. 1042–1051, 2008.

[8] H. Zhang, M. H. Harpster, H. J. Park, P. A. Johnson, and W. C. Wilson, "Surface-enhanced Raman scattering detection of DNA derived from the west nile virus genome using magnetic capture of Raman-active gold nanoparticles," *Anal Chem.*, vol. 83, pp. 254–260, 2011.

[9] 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*, vol. 105, pp. 5844–5849, 2008.

[10] 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*, vol. 106, pp. 13 511–13 516, 2009.

[11] D. C. Kennedy, K. A. Hoop, L.-L. Tay, and J. P. Pezacki, "Development of nanoparticle probes for multiplex SERS imaging of cell surface proteins," *Nanoscale*, vol. 2, pp. 1413–1416, 2010.

[12] A. Hoskuldsson, "PLS regression methods," *J. Chemometr.*, vol. 2, pp. 211–228, 1988.

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

[14] S. Li, J. Gao, J. O. Nyagilo, and D. P. Dave, "Eigenspectra, a robust regression method for multiplexed Raman spectra analysis," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2010.

[15] S. Li, J. O. Nyagilo, D. P. Dave, and J. Gao, "Probabilistic partial least square regression: a robust model for quantitative analysis of Raman spectroscopy data," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2011.

[16] S. Li, J. O. Nyagilo, D. P. Dave, and J. Gao, "CWT-PLSR for quantitative analysis of Raman spectrum," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2012.

[17] S. de Jong and H. A. Kiers, "Principal covariates regression part i. theory," *Chemometrics Intell. Lab. Syst.*, vol. 14, pp. 155–164, 1992.

[18] E. M. Qannari and M. Hanafi, "A simple continuum regression approach," *J. Chemometr.*, vol. 19, pp. 387–392, 2005.

[19] S. Li, J. Gao, J. O. Nyagilo, and D. P. Dave, "A new continuum regression method for quantitative analysis of Raman spectrum," in *International Conference on Machine Learning and Applications (ICMLA)*, 2012.

[20] A. J. Burnham, R. Viveros, and J. F. MacGregor, "Frameworks for latent variable multivariate regression," *J. Chemometr.*, vol. 10, pp. 31–45, 1996.

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

[22] H. A. L. Kiers and A. K. Smilde, "A comparison of various methods for multivariate regression with highly collinear variables," *Stat. Method. Appl.*, vol. 16, pp. 193–228, 2007.

[23] R. Rosipal and N. Kramer, "Overview and recent advances in partial least squares," *LNCS*, vol. 3940, pp. 34–51, 2006.

[24] 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.*, vol. 56, pp. 1371–1382, 2009.

[25] 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.*, vol. 398, pp. 1827–1832, 2010.

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

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

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

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