Talanta 83 (2011) 1401–1409

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

An emphatic orthogonal signal correction-support vector machine method for the classification of tissue sections of endometrial carcinoma by near infrared spectroscopy

Jiajin Zhang[a], Zhuoyong Zhang[a,*], Yuhong Xiang[a], Yinmei Dai[b], Peter de B. Harrington[c]

[a] Department of Chemistry, Capital Normal University, Beijing 100048, China
[b] Beijing Obstetrics and Gynecology Hospital, Affiliated Capital University of Medical Science, Beijing 100006, China
[c] Center for Intelligent Chemical Instrumentation, Clippinger Laboratories, Department of Chemistry and Biochemistry, Ohio University, Athens, OH 45701-2979, USA

ARTICLE INFO

Article history:
Received 12 July 2010
Received in revised form 9 November 2010
Accepted 11 November 2010
Available online 18 November 2010

Keywords:
NIRS
EOSC
SVM
Cancer detection
Chemometrics

ABSTRACT

A new application of emphatic orthogonal signal correction (EOSC) for baseline correction of near infrared spectra from reflectance measurements of tissue sections is introduced. EOSC was evaluated and compared with principal component orthogonal signal correction (PC-OSC) by using support vector machine (SVM) classifiers. In addition, some exemplary synthetic data sets were created to characterize EOSC coupled to SVM for classification. Orthogonal experimental design coupled with analysis of variance (ANOVA) was used to determine the significant parameters for optimization, which were the OSC method and number of components for the model. EOSC combined with the SVM gave better predictions with respect to a larger number of components and was not as susceptible to overfitting the data as the classifier built with PC-OSC data. These results were supported by simulations using synthetic data sets. EOSC is a softer signal correction approach that retains more signal variance which was exploited by the SVM. Classification rates of $93 \pm 1\%$ were obtained without orthogonal signal correction with the SVM. PC-OSC and EOSC data gave similar peak prediction accuracies of $94 \pm 1\%$. The key advantages demonstrated by EOSC were its resistance to overfitting, fine-tuning capability or softness, and the retention of spectral features after signal correction.

1. Introduction

Endometrial cancer is the most common malignancy of the female genital system [1]. Endometrial cancer causes menstrual disorder, ovarian tumors, and polycystic ovary syndrome. Although the etiological factors causing endometrial carcinoma are unknown, the reported risk factors include menstrual [2], fecund [3], hormonal [4], metabolic [5], dietary [6], etc.

The main diagnostic methods for endometrial cancer include cytological examination, B-ultra examination, diagnostic curettage, hysteroscopy diagnosis, retroperitoneal lymph node imaging, and nuclear magnetic resonance (NMR) imaging. Among them the hysteroscopy diagnosis can only be used for the cases for which diagnosis cannot be confirmed. For those cases additional endometrial biopsy is needed and NMR is used for non-damage diagnosis. However, even with these additional diagnostic methods, the accuracy is still insufficient while the cost is increased. Recently the development of non-destructive examination methods such as magnetic resonance imaging (MRI) combined with chemometrics for disease diagnosis has garnered attention [7–9].

There have been several examples of the application of non-destructive examination by diagnostic models. For breast cancer diagnosis, images of cells obtained from fine needle aspiration were classified using support vector machines (SVMs), radial basis function (RBF) networks, and self-organizing maps (SOMs) [10]. An approach based on the implementation of multiclass SVMs with error correcting output codes (ECOCs) was reported for diagnosis of erythemato-squamous disease from patient symptoms [11]. Huang et al. [12] constructed a hybrid SVM-based strategy with feature selection to render a diagnosis between the breast cancer and fibroadenoma and to find the important risk factor for breast cancer. Probabilistic neural network (PNN) and SVM neural network models have also been investigated for classifying normal and abnormal hysteroscopy images of the endometrium based on texture analysis for the early detection of gynecological cancer [13].

Near infrared spectroscopy (NIRS) is a spectroscopic method which uses the near infrared region of the electromagnetic spectrum (13,000–4000 cm^{-1}). Typical applications of NIRS include pharmaceutics [14], medical diagnostics [15], food [16], and agrochemical quality control [17]. NIR spectra are useful for char-

* Corresponding author.
E-mail address: gusto2008@vip.sina.com (Z. Zhang).

0039-9140/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2010.11.020

acterization and identification of complex matrices [18]. Cancer cells and tissues differ in chemical composition from normal tissues, and this difference is the basis of cancer diagnosis by NIRS. NIRS may also detect differences in cell morphology between normal and cancerous cells. The morphology may affect the scattering of the NIR radiation that manifests in baseline variations that correlate with the tissue class. An effective classification procedure for diagnosis of prostate cancer from NIRS measurements has been reported [19].

NIRS reflectance measurements of complex samples such as tissue sections are beset with varying backgrounds and baselines that can deter chemometric methods from working effectively [20]. Orthogonal signal correction (OSC) methods provide a means of removing complex background variations while retaining the signal. The various methods all work in principle by creating basis sets that are orthogonal to the signal which is defined in this case by the different classes of tissue. The methods differ in the approach that they use to define the basis sets. NIR spectra are frequently plagued with unwanted variances such as baseline variances especially for large-scale studies and for complex samples. Various OSC methods have been developed in recent years after it was first introduced by Wold [21,22]. The objective of this work is to evaluate an arcane background correction method for cancer diagnosis by SVM classification of NIR measurements of tissue samples. An emphatic orthogonal signal correction (EOSC) method developed by Wu is described [23,24]. This method has not yet been used in the field of chemistry and a new application of EOSC is presented in this paper. The principal component orthogonal signal correction (PC-OSC) method [20,22] which is the least constrained and simplest of OSC correction methods was used as a reference method for comparison to the EOSC method. Two standard kernels were used to evaluate the SVM for the effectiveness of the signal correction methods, the RBF kernel and a simple linear kernel. To explain the feasibility and effectiveness of EOSC, a comparison of EOSC and PC-OSC using synthetic data and real NIR spectra of endometrial cancer samples is presented.

2. Theory

2.1. Complementary orthogonal subspaces

Any N-dimensional space can be partitioned into mutually orthogonal subspaces. In Fig. 1, a 3D space is defined by plane A ($x+y+z=0$) and straight line $\boldsymbol{l}$ ($x=y=z$). In this space, a basis for plane A (dimensions of 2) is $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$ and a basis for the straight line $\boldsymbol{l}$

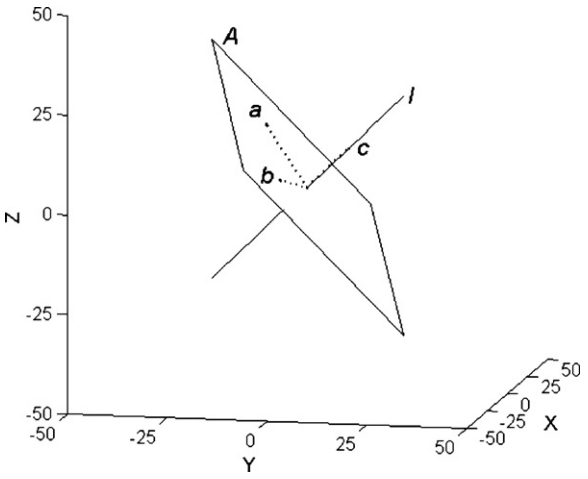

Fig. 1. Two complementary orthogonal subspaces in 3D space. The plane is subsapce N_1 and the line $\boldsymbol{l}$ is subspace N_2.

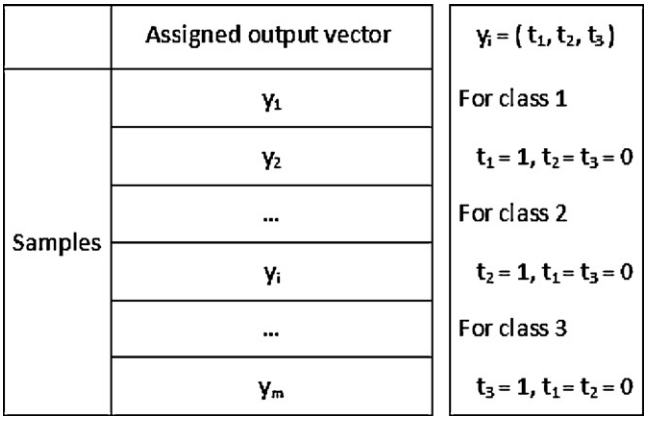

Fig. 2. Construction of the binary encoded $\boldsymbol{Y}$-matrix.

(dimension of 1) is $\boldsymbol{c}$. An N-dimensional space can be separated into two mutually complementary orthogonal subspaces of dimensions N_1 and N_2 as defined by:

$$N = N_1 + N_2 \tag{1}$$

for which N is the dimension of the full space; N_1 and N_2 are dimensions of the two complementary orthogonal subspaces.

2.2. PC-OSC and EOSC methods

The data matrix $\boldsymbol{X}$ comprises m rows of spectra. A binary matrix $\boldsymbol{Y}$ encodes the classes by having a single value of unity in each row and the columns designate the class membership. The other terms in the $\boldsymbol{Y}$ matrix are zero. $\boldsymbol{X}$ is an $m \times p$ matrix for which p are the number of spectral variables or measurements. $\boldsymbol{Y}$ is an $m \times k$ matrix for k is the number of classes. The binary encoded $\boldsymbol{Y}$-matrix is given in Fig. 2.

Principal component orthogonal signal correction methods work by defining a subspace that is orthogonal to the property matrix $\boldsymbol{Y}$. The orthogonal subspace $\boldsymbol{X}_0$ to the property matrix $\boldsymbol{Y}$ is obtained by:

$$\boldsymbol{X}_0 = (\boldsymbol{I} - \boldsymbol{Y}(\boldsymbol{Y}^T\boldsymbol{Y})^{-1}\boldsymbol{Y}^T)\boldsymbol{X} \tag{2}$$

Typically control of the subspace dimensionality determines the amount of background correction of the data set $\boldsymbol{X}$. Too few components will underfit the data and leave some background components in the spectra and too many components will overfit the data and remove signal from the spectra. The PC-OSC [20,22] procedure is given in Table 1.

Table 1
Procedure for PC-OSC.

Inverseleast squares model for $\boldsymbol{X}$ and $\boldsymbol{Y}$	$\boldsymbol{X} = \boldsymbol{YB} + \boldsymbol{E}$ $\boldsymbol{B} = (\boldsymbol{Y}^T\boldsymbol{Y})^{-1}\boldsymbol{Y}^T\boldsymbol{X} + \boldsymbol{E}$
1. Correct $\boldsymbol{X}$ and $\boldsymbol{Y}$ by subtracting their means	$\boldsymbol{X}_1 = \boldsymbol{X} - \bar{\boldsymbol{X}}$ $\boldsymbol{Y}_1 = \boldsymbol{Y} - \bar{\boldsymbol{Y}}$
2. Calculate the background from the least squares model	$\boldsymbol{X}_0 = \boldsymbol{X}_1 - \hat{\boldsymbol{X}}_1 = \boldsymbol{X}_1 - \boldsymbol{Y}_1({\boldsymbol{Y}_1}^T\boldsymbol{Y}_1)^{-1}\boldsymbol{Y}_1^T\boldsymbol{X}_1$
3. Calculate a basis from the background using the row-space eigenvectors from SVD	$\boldsymbol{X}_0 = \boldsymbol{USV}^T$
4. Define the basis by selecting n components	$\boldsymbol{V}_n = [\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_n]$
5. Calculate corrected spectrum $\boldsymbol{x}_{pc}$ from new spectrum $\boldsymbol{x}_p$	$\boldsymbol{x}_{pc} = (\boldsymbol{x}_p - \bar{\boldsymbol{X}}) - [(\boldsymbol{x}_p - \bar{\boldsymbol{X}})\boldsymbol{V}_n]\boldsymbol{V}_n^T$

The EOSC method is based on the following relationships:

$$\boldsymbol{M} = (\boldsymbol{Y} - \bar{\boldsymbol{Y}})^T(\boldsymbol{X} - \bar{\boldsymbol{X}}) \quad (3)$$

$$\boldsymbol{B} = \mathbf{null}(\boldsymbol{M}) \quad (4)$$

$$\boldsymbol{Q} = (\boldsymbol{X} - \bar{\boldsymbol{X}})\boldsymbol{B} \quad (5)$$

for which $\boldsymbol{B}$ is a $p \times (p-k)$ null basis formed from the $p-k$ eigenvectors of the smallest eigenvalues (secondary components) of the covariance matrix $\boldsymbol{M}$. The subspace defined by $\boldsymbol{Q}$ is a complementary orthogonal subspace of the subspace defined by $\boldsymbol{M}$ and as a result $\boldsymbol{Y}$. The null space spans the full residual rank and multiplication by $\boldsymbol{X}$ is required to remove additional dimensions that may not be found in $\boldsymbol{X}$. Thus, $\boldsymbol{Q}$ is the intersection of the null space $\boldsymbol{B}$ and the mean corrected spectral data space $(\boldsymbol{X} - \bar{\boldsymbol{X}})$. However, background variations in the prediction set may not be the same as those found in $\boldsymbol{X}$ and may reside in the null space $\boldsymbol{B}$.

Singular value decomposition is used to decompose $\boldsymbol{Q}$ as given below:

$$\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T \quad (6)$$

for which $\boldsymbol{Q}$ is decomposed into row $\boldsymbol{U}$ and column $\boldsymbol{V}$ eigenvectors and a diagonal matrix of the singular values $\boldsymbol{S}$. These matrices are used to invert $\boldsymbol{Q}$ and the number of components used to define the pseudoinverse controls the fit of the correction. The goal is to calculate a transformation matrix by using singular value decomposition to divide $\boldsymbol{Q}$ by itself and the ratio is subtracted from the identity matrix $\boldsymbol{I}_p$ to yield the correction matrix $\boldsymbol{D}$.

$$\boldsymbol{D} = \boldsymbol{I}_p - \boldsymbol{B}\boldsymbol{V}\boldsymbol{S}^{-1}\boldsymbol{U}^T(\boldsymbol{X} - \bar{\boldsymbol{X}}) \quad (7)$$

$$\hat{\boldsymbol{X}} = (\boldsymbol{X} - \bar{\boldsymbol{X}})\boldsymbol{D} \quad (8)$$

for which $\hat{\boldsymbol{X}}$ is the orthogonal signal corrected data, and $\boldsymbol{D}$ is the conversion matrix. For the prediction set:

$$\hat{\boldsymbol{X}}_{\text{prediction}} = (\boldsymbol{X}_{\text{prediction}} - \bar{\boldsymbol{X}}_{\text{training}})\boldsymbol{D}_{\text{training}} \quad (9)$$

for which $\hat{\boldsymbol{X}}_{\text{prediction}}$ is the orthogonal signal corrected prediction data matrix $\boldsymbol{X}_{\text{prediction}}$. The EOSC procedure is given in Table 2.

3. Experimental

3.1. Instrumentation

A Nicolet 6700 extended Fourier transform near infrared (FT-NIR) spectrometer (Thermo Electron, USA) equipped with InGaAs detector was used for the NIR measurement. The spectrometer was controlled by OMNIC service software of version 7.3. The measurement was performed on the top of the glass plate, and each section was measured at 5 different locations and the average spectra of tissue sections were used as the spectra of cases in the following analysis. Data analysis was done using MATLAB software (The MathWorks Inc., South Natick, MA, USA).

Table 2
Procedure for EOSC.

1. Correct $\boldsymbol{X}$ and $\boldsymbol{Y}$ by subtracting their means.	$\boldsymbol{X}_1 = \boldsymbol{X} - \bar{\boldsymbol{X}}$ $\boldsymbol{Y}_1 = \boldsymbol{Y} - \bar{\boldsymbol{Y}}$
2. Calculate the covariance between $\boldsymbol{X}_1$ and $\boldsymbol{Y}_1$	$\boldsymbol{M} = \boldsymbol{Y}_1^T\boldsymbol{X}_1$
3. Calculate the null space of the covariance $\boldsymbol{M}$	$\boldsymbol{B} = \mathbf{null}\,(\boldsymbol{M})$
4. Calculate the intersection Q of the null space and $\boldsymbol{X}_1$	$\boldsymbol{Q} = \boldsymbol{X}_1\boldsymbol{B}$
5. Use SVD to decompose $\boldsymbol{Q}$	$\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$
6. Use n components to calculate the pseudoinverse of $\boldsymbol{Q}_n^+$	$\boldsymbol{Q}_n^+ = \boldsymbol{V}_n\boldsymbol{S}_n^{-1}\boldsymbol{U}_n^T$
7. Construct the transformation matrix $\boldsymbol{D}$ by dividing $\boldsymbol{Q}$ by its pseudoinverse $\boldsymbol{Q}^+$ and subtracting from the identity matrix $\boldsymbol{I}$	$\boldsymbol{D} = \boldsymbol{I} - \boldsymbol{B}\boldsymbol{Q}^+\boldsymbol{X}_1$
8. Calculate corrected spectrum $\boldsymbol{x}_{\text{pc}}$ from new spectrum $\boldsymbol{x}_{\text{p}}$	$\boldsymbol{x}_{\text{pc}} = (\boldsymbol{x}_{\text{p}} - \bar{\boldsymbol{X}})\boldsymbol{D}$

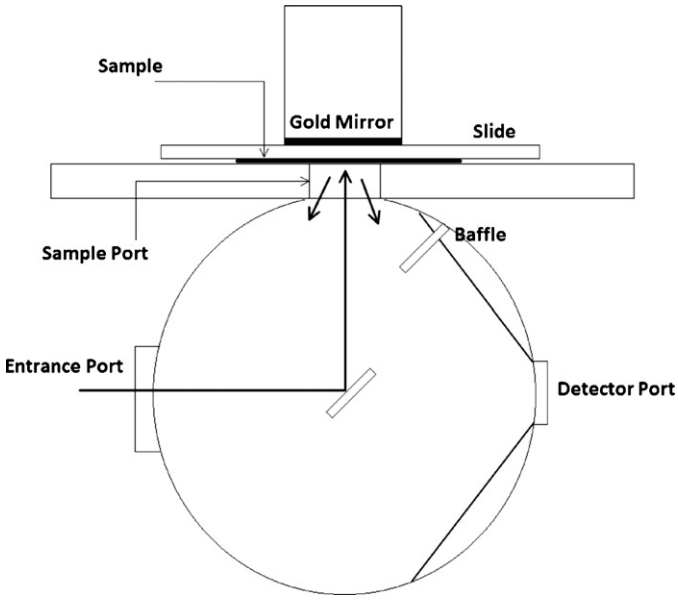

Fig. 3. Sample arrangement for NIR measurement.

3.2. Samples

Endometrial tissues sections (18 normal, 30 hyperplasic, and 29 cancerous) were obtained from Beijing Red Cross ChaoYang Hospital. Seventy-seven paraffin sections of endometrial tissues were supplied by Beijing Obstetrics and Gynecology Hospital, attached to the Capital Medical University. The mean age was 46 with the oldest and youngest patients having respective ages of 71 and 19 years. All the endometrial tissues were put in a 4% formaldehyde solution to be stabilized, and then were washed with a series of increasing concentrations of ethanol solutions (30%, 50%, 70%, 85%, and 95%, respectively) for dehydration. The samples were put into xylene for 2 h, embedded in paraffin wax, and then sliced into 4 μm thick sections. The sections were put on glass slides and dried at 45 °C, and then fixed with a neural gum mounting. Upon completion of all the above procedures, the samples were ready for measurement by NIRS. The sample arrangement for NIR measurement is given in Fig. 3.

The paraffin sections were placed in the integrating sphere of the NIR spectrometer. The NIR diffuse reflection spectra were collected with a nominal optical resolution of 4 cm^{-1} across the spectral range of 4000 and 10,000 cm^{-1} by using Thermo Fisher Omnic software version 7.3. A background spectrum was recorded using an air reference at 25 °C. Each sample section was scanned five times at different positions, and the average of five scans was used as the measured spectrum. The NIR spectra of the paraffin sections are given in Fig. 4. Each spectrum was obtained from a tissue section of a different patient.

3.3. Data treatment and computation

The aim of this work is to develop a new method for diagnosis of endometrial cancer as well as other cancers, therefore, some synthetic data with various properties and characteristics were generated for method evaluation. A set of underdetermined and overdetermined synthetic data were constructed in MATLAB. For both sets, the number of variables was 200. For the underdetermined set, the number of objects was 100 and for the overdetermined set the number of objects was 400. All other following parameters were the same for the two sets of synthetic data. Normally distributed random deviates with a mean of zero

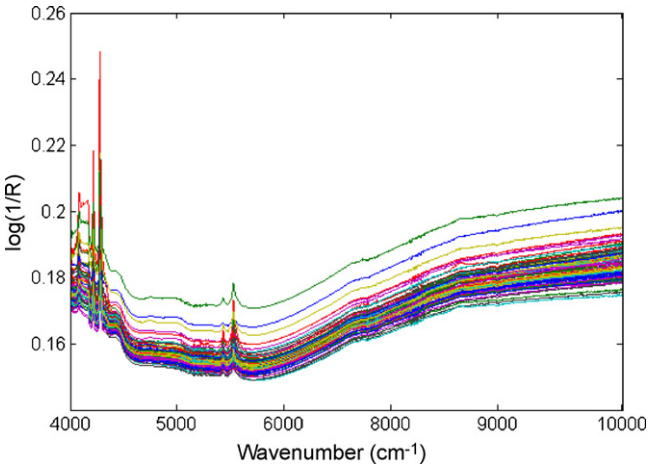

Fig. 4. Average NIR reflectance spectra of the 77 tissue thin sections.

and standard deviation of 0.1 were added to all the points of the data set to simulate noise. The background variation was obtained from a Gaussian peak with amplitude of 1, standard deviation of 10 points, and a randomly selected position. Eighty simulated peaks were generated and randomly added to half of the objects in the data set, so that the simulated backgrounds would be independent of the signal. A signal vector was formed using 4 Gaussian peaks with amplitudes of 0.3, standard deviations of 5 points, and centered at 40, 80, 120, and 160 point number. The signal was added to 50% of the objects and these objects were designated as class A. The other objects were designated as class B. This set represents a case for which a trace quantity of analyte is to be detected from a blank matrix with complex background variations.

For illuminating the differences between OSC calibration and prediction data sets, a Venetian blind was used to partition the data into two equally sized sets. The spectra in the even numbered rows of the data matrix were used for constructing the OSC model and the spectra in the odd numbered rows were used as the prediction set.

Four Latin partitions were used to split the data into four training-prediction set pairs. The Latin partition method maintains the same class distributions in the training and prediction sets

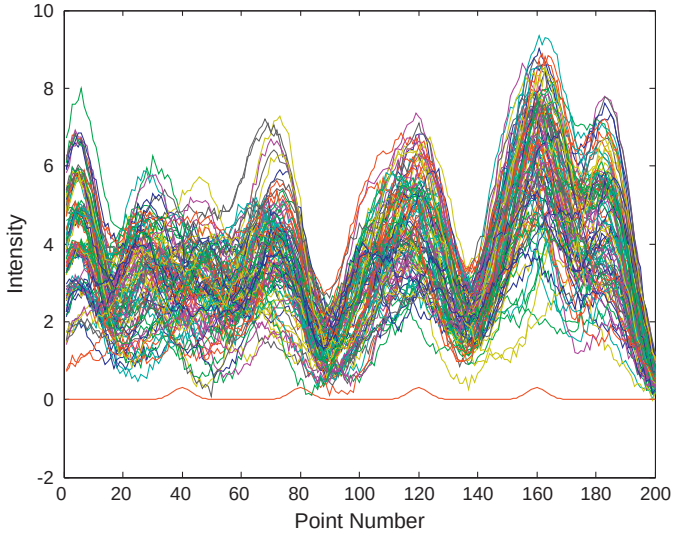

Fig. 5. Underdetermined synthetic data set with pure signal component in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

Table 3
Factors and levels in the orthogonal experimental design.

Levels	Factors			
	OSC method	*C*	*g*	nc
1	EOSC	11	−15	5
2	PC-OSC	12	−14	20
3	No OSC	13	−13	30

Note: The second column is the OSC methods used in this work, three methods were taken as three levels of OSC factor. EOSC: emphatic orthogonal signal correction; PC-OSC: principal component orthogonal signal correction; no OSC: with no OSC. *C*: RBF kernel pennalty parameter; *g*: RBF kernel spread parameter; nc: number of components.

while randomly selecting objects for these data sets [25,26]. Each of the four training sets comprised 75% of the objects and the prediction set comprised the other 25%. The results for the four prediction sets were pooled so every object was used once for prediction and three times for model-building. This approach was used for all the SVM evaluations and measures of prediction rates. For these evaluations 100 bootstraps with 4 Latin partitions were used except for the screening study that used 25 bootstraps.

Based on our preliminary work on derivative spectra for the tissue sections evaluations, the spectra were converted to their second derivatives after multiplicative scatter correction (MSC) and OSC by using a home built Savitzky–Golay (SG) filter. The SG filter used a 9 point window and a cubic polynomial. The MSC was applied first to the training set and the prediction set objects were fit to the mean of the train set. The signal correction models obtained from the training set were used for correcting the prediction sets.

Four factors, the OSC method, RBF kernel spread parameter (g), RBF kernel penalty parameter (C), and number of components for the OSC model were screened. Two interactions terms, the number of components for the OSC models and the two SVM parameters, g with C were evaluated. Factors at various levels and the interactions were arranged in an orthogonal experimental design table according to Taguchi's orthogonal arrays $L_{81}3^{40}$ [27]. Each factor and the interactions comprised columns of the Taguchi's orthogonal array. The factors and their levels used for the experimental design are given in Table 3. The experiment was bootstrapped 25 times to calculate the average prediction rates and to determine their confidence intervals.

4. Results and discussion

4.1. Synthetic data evaluations

Support vector machines are powerful classifiers so that the data sets were designed with a low-signal to noise ratio (i.e., 3) and a very large background variation that is typical of many biological systems. Fig. 5 is the underdetermined data set with the pure analytical signal visible along the bottom of the figure. Fig. 6 gives the object scores on the first two principal components.

The data set was divided by rows into a training and a prediction set by selecting the even row objects for the training set and the odd row objects for the prediction set. The training set was used to construct the OSC models. Fig. 7 gives a comparison of the principal component scores for models built with 40 components. Note the difference in the range of component scores between PC-OSC and the EOSC corrected data. The EOSC data contains a larger amount of signal variance.

Fig. 8 gives a comparison among the OSC objects. Again the range is larger for the EOSC data and the EOSC has retained prominent features that match up to the underlying signal peaks in both the training and prediction sets at positions 120 and 180. When normally distributed noise is removed from the synthetic data all

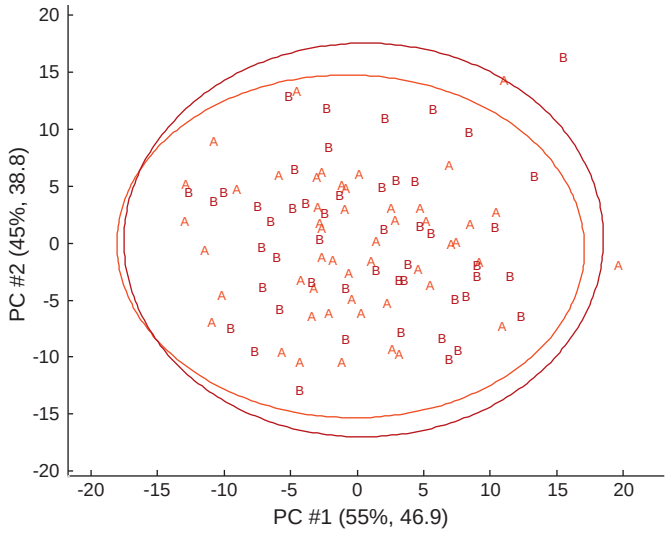

Fig. 6. Principal component scores of the underdetermined synthetic data without signal correction. Positive (signal and background) score is designated by A and negative score (background only) is designated by B.

four signal peaks are visible in the EOSC data. The PC-OSC data retains the signal information but it has become distributed across all the variables so that the original signal peaks are no longer discernable.

The prediction effectiveness was evaluated with respect to signal correction component numbers for the underdetermined and overdetermined data sets. The evaluations used 100 bootstraps to gain statistical power and control the variability of the SVM models. All non-linear optimization algorithms are susceptible to model variability that arises from local or degenerate optima that may result in different models for the same training set.

The underdetermined synthetic data set had 100 objects and 200 variables. Fig. 9 gives the average SVM prediction rates across the 100 bootstraps for this data set after the application of the two signal correction methods and without signal correction as a reference. Without signal correction, the SVM predicted $81.5 \pm 0.8\%$ of the objections correctly. For 100 bootstraps, the average prediction rate is constant across the 30 components, which demonstrates the power of the bootstrap Latin partition approach. PC-OSC is more efficient at achieving its maximum prediction rate of $89.9 \pm 0.6\%$ which is achieved with 15 components. EOSC achieves an equivalent average prediction rate of $89.8 \pm 0.7\%$ with 25 components and remains constant, while the PC-OSC average prediction rate declines as it begins to overfit the data after 35 components. There are 50 objects in the training set so 45 components is beginning to span the full data space of the training set; remarkably no decrease of prediction accuracy is observed with the EOSC data.

Fig. 10 is the same experiment but this time with the overdetermined synthetic data set of 400 objects and 200 variables. For the overdetermined data sets both OSC methods significantly improved the average prediction rates compared with not correcting the data at all which was $91.4 \pm 0.2\%$ effective. As with most classifiers increasing the number of objects in the training set improves both the accuracy and the precision for the SVM. The PC-OSC data is improved to $94.7 \pm 0.2\%$ with 9 components, while the EOSC date yielded a $95.4 \pm 0.2\%$ average prediction rate. EOSC allows SVM classifiers to reach a significantly better classification rate and at the same time resists overfitting the data. The PC-OSC method is very sensitive to overfitting with overdetermined data. Although these differences in performance are small, they can be detected with the high statistical power afforded by the 100 Latin partition bootstraps.

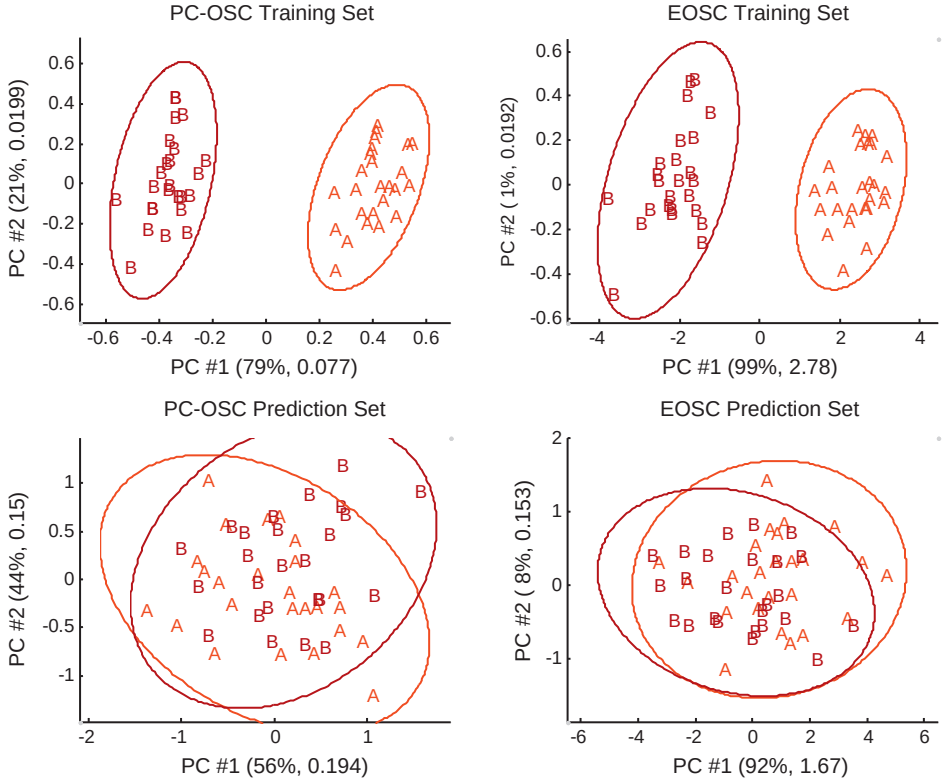

Fig. 7. Comparison of principal component scores for PC-OSC and EOSC synthetic underdetermined data. The signal corrections each had 30 components. Positive (signal and background) score is designated by A and negative score (background only) is designated by B.

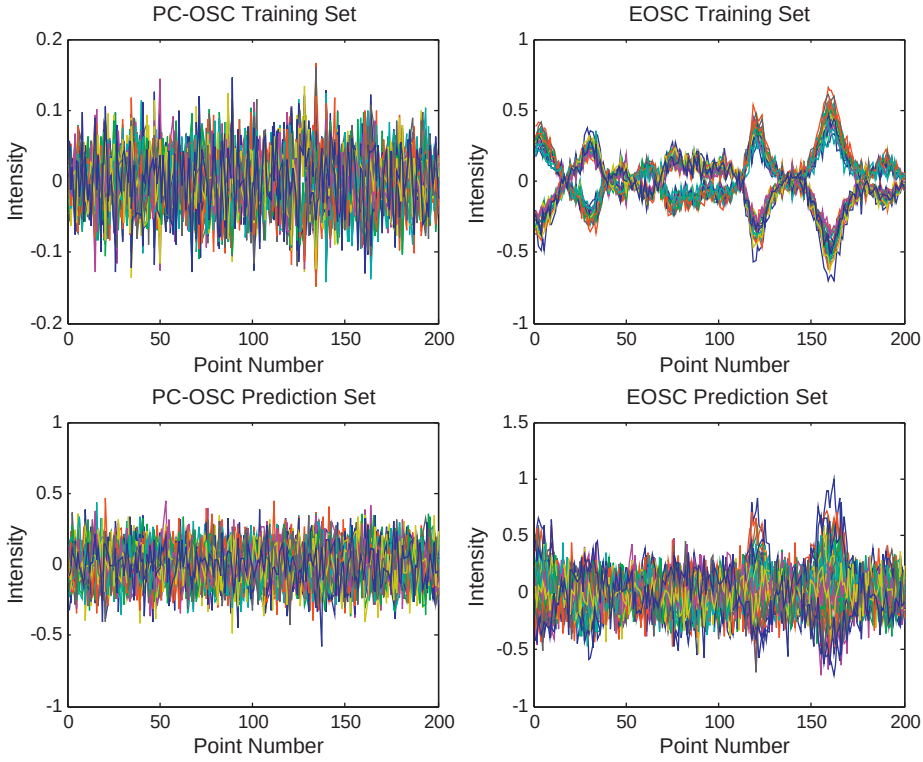

Fig. 8. Comparison of PC-OSC and EOSC data objects for training and prediction sets. The signal corrections each had 30 components. Pure signal located at point numbers 40, 80, 120, and 160.

4.2. Screening of the model parameters

ANOVA was applied to the results of an orthogonal experimental design to screen the main factors of the models. The results of the ANOVA are reported in Table 4. The significant factors were model component number and the OSC method. The SVM kernel function parameters were not influential on the prediction accuracy. The trends obtained by linear kernel and RBF kernel functions are almost the same. This finding was confirmed when the data was evaluated with an SVM using a linear kernel function and nearly identical prediction results were obtained as to the SVM with the RBF kernel function. Results show that the OSC method is the most significant factor in the experiment.

4.3. Comparison of the two OSC methods for classifying cancerous tissue

The RBF kernel had a C parameter of 11 and a g parameter of −15, but the following results did not differ significantly when a linear kernel was used (data are not shown). The OSC methods were evaluated using 100 bootstrap Latin partitions, with 4 partitions of each bootstrap from 1 to 30 components. The average prediction accuracies with 95% confidence intervals are given

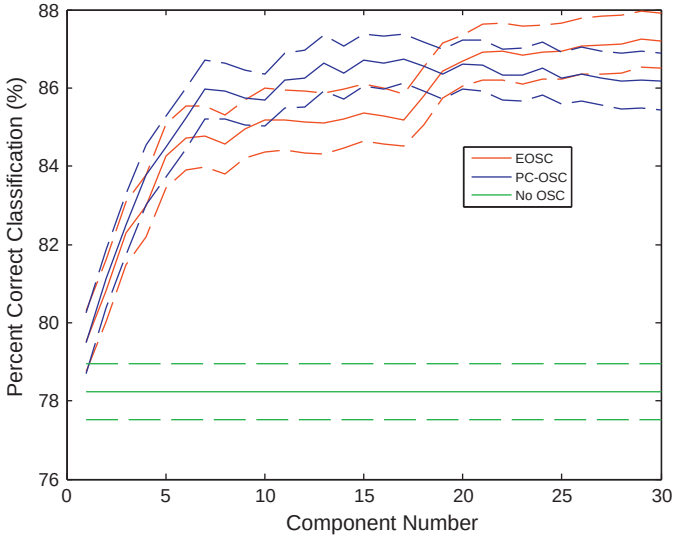

Fig. 9. Prediction rate averages with 95% confidence intervals for the underdetermined synthetic data set with respect to OSC component number obtained with two Latin partitions and 100 bootstraps.

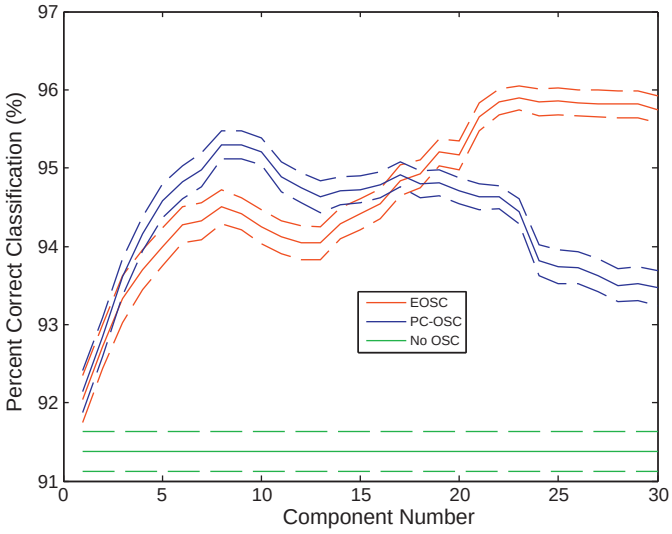

Fig. 10. Average prediction results with 95% confidence intervals for the overdetermined synthetic data set obtained with two Latin partitions and 100 bootstraps.

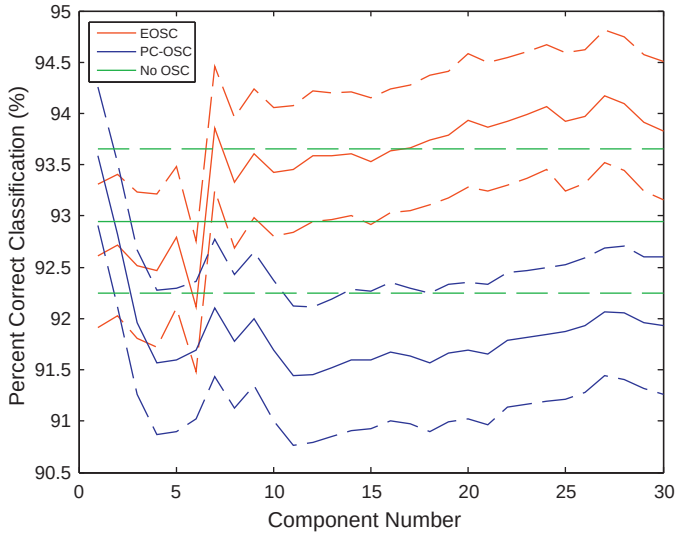

Fig. 11. Comparison of signal correction methods for the optimized RBF SVM classifier for the classification of the tissue thin sections. The average classification rate with respect to OSC component number is plotted with 95% confidence intervals for 4 Latin partitions and 100 bootstraps.

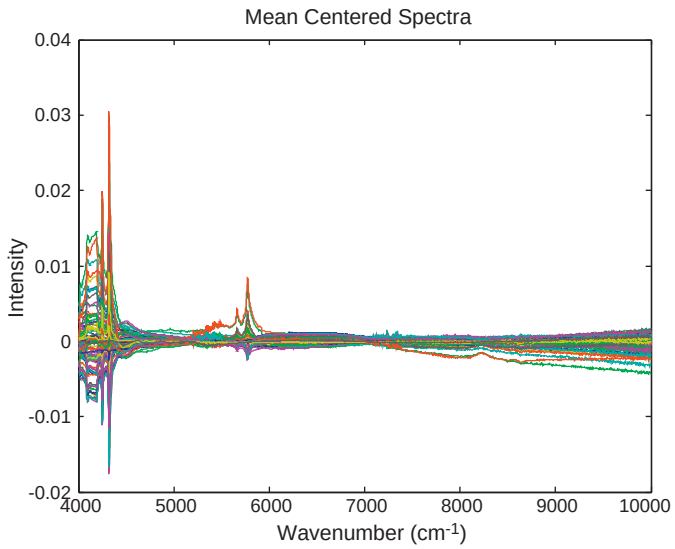

Fig. 12. NIR reflectance spectra after MSC and mean centering to compare with the OSC data.

Table 4
ANOVA results for the screening experiment.

Source	Sum sq.	d.f.	Mean sq.	*F*	*p*-Value
OSC	0.064	2	0.032	35.239	0.0
nc	0.003	2	0.002	1.795	0.2
C	0.000	2	0.000	0.000	1
g	0.000	2	0.000	0.121	0.9
OSC × nc	0.016	4	0.004	4.302	0.002
C × *g*	0.000	4	0.000	0.000	1
Error	1.828	2008	0.001		
Total	1.912	2024			

in Fig. 11. Without using signal correction, good performance of $93.0 \pm 0.7\%$ was obtained. The PC-OSC had improved the prediction accuracy with a result of $93.6 \pm 0.7\%$ that used only a single component. EOSC gave a maximum prediction accuracy at $94.5 \pm 0.6\%$ with 27 components. A matched sample *t*-test revealed that this EOSC gave a significantly better result than without using signal correction. The *p*-value was 2×10^{-9} for the matched sample *t*-test so that the prediction accuracy had statistically improved over no signal correction. A similar *p*-value of 3×10^{-5} was obtained for the significance test with respect to EOSC compared to the PC-OSC method. The confidence intervals in Fig. 11 characterize the

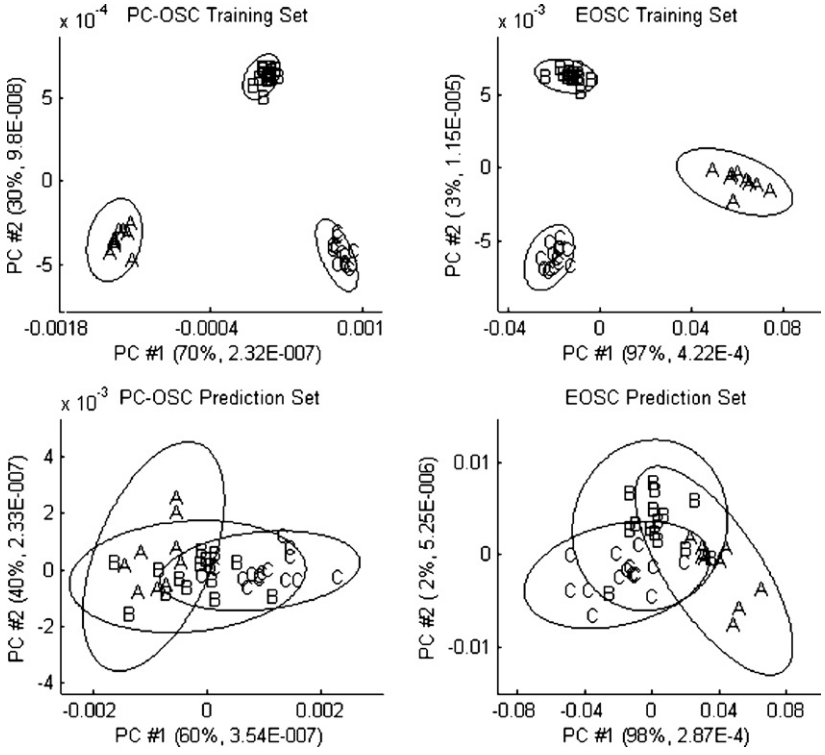

Fig. 13. Comparison of principal component scores for PC-OSC and EOSC of the NIR measurements of A: normal, B: hyperplasic, and C: cancerous tissue sections. The number of components for each model was 27 which was the optimal number with respect to the EOSC SVM classifier. The optimal number of components for PC-OSC SVM classifier was 1, so for this case some overfitting may have occurred.

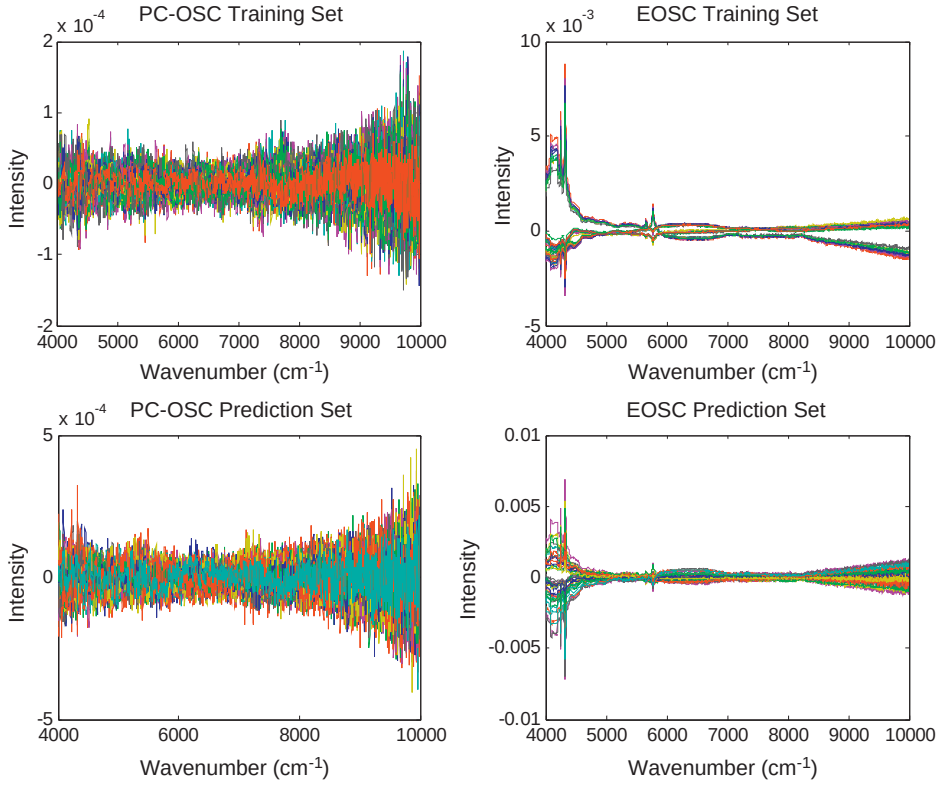

Fig. 14. The OSC spectra (left panels PC-OSC) and (right panels EOSC) using 27 components in the model. EOSC does not overfit the data and retains many of the spectra features after the correction.

variation in prediction accuracy with respect to the composition of the training and prediction data sets that vary among the bootstraps. The large variability indicates that the number of samples are not adequate to solve this problem, but the study was constrained by the number of samples that were available from the hospital. The matched sample *t*-test removes this additional source of variation for comparing the prediction results.

Two-way ANOVA of bootstrapped prediction results comparing EOSC and without signal correction also yielded a statistical difference of 1×10^{-6}, while the component number and interaction were insignificant at a 95% confidence level.

EOSC with 27 components was compared to the PC-OSC corrected data with the same number of components. The NIR spectra are in Fig. 4. The sharp peaks from 4300 to 5800 cm^{-1} are wax peaks, and the background shift was found because of the variation in pathlength and scatter from the sample matrices and sampling issues such as geometry of the sample. The background shift can be corrected by data OSC. The data used the same pretreatment as the previous SVM evaluations in that MSC was used first to correct the data. Fig. 12 gives the spectra after the MSC and mean-centering steps. This preprocessing step was done so the unprocessed spectra would be comparable to the signal corrected spectra.

The spectra were divided systematically into training and prediction sets by placing spectra in the even numbered rows of the data matrix into the training set and the odd numbered rows into the prediction set (i.e., Venetian blind partition). Fig. 13 compares the principal component object scores in a similar fashion as the synthetic data. The same trends are observed. The relative positions of the scores appear similar between the two OSC methods, but the range of the scores for EOSC is respectively 40 and 20 times greater than PC-OSC for the training and prediction sets, respectively. A similar tendency appears in the spectra given in Fig. 14. For EOSC recognizable spectral features are apparent, while for PC-OSC although spectral information is still present and no distinguishing spectral features are visible. EOSC spectra also have the same increase in the range of intensities as seen with the synthetic data. EOSC appears to be even softer than the soft method PC-OSC which is helpful for fine-tuning background correction and preventing overfitting of the spectra during signal correction.

5. Conclusions

NIR spectra of tissue samples often have severe background variation because of scattering and the unavailability of suitable reference materials. Therefore background corrections may be required and OSC methods are effective. The EOSC method is introduced for the first time in the chemical literature. Coupled with support vector machines the EOSC method provides a statistical improvement over the case without any background correction. Both simulated and real studies demonstrate the inherent advantages of EOSC over PC-OSC when coupled with an SVM classifier. EOSC performed as well or better than PC-OSC although it is softer in that it requires more components to correct the data. This lower efficiency is an advantage because it allows the background correction to be fine-tuned. Another advantage of EOSC is its resistance to overfitting or overcorrecting the data, especially for overdetermined data sets. The most important advantage of the EOSC method is that it preserves characteristic information so that the signal corrected spectra are still amenable to interpretation. Therefore, EOSC has promise for differential diagnosis of endometrial cancer based on NIR detection. This technique may provide a reliable method for correcting NIR spectra that may develop into a novel and robust approach to the non-invasive diagnosis for tumors.

Acknowledgements

This work was supported by the Natural Science Foundation of China (Grant Nos. 20875065 and 30772322), Beijing Natural

Science Foundation (2102010), and Beijing Municipal Knowledge Innovation Team Project (PHR20100718). Xiaobo Sun and Zhanfeng Xu are thanked for their helpful comments.

References

[1] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray, M.J. Thun, Cancer Statistics 58 (2008) 71–96.
[2] M.E. Ossewaarde, M.L. Bots, A.L.M. Verbeek, P.H.M. Peeters, Y. van der Graff, D.E. Grobbee, Y.T. van der Schouw, Epidemiology 16 (2005) 556–562.
[3] M. Lambe, J. Wuu, E. Weiderpass, C.-C. Hsieh, Cancer Causes and Control 10 (1999) 43–49.
[4] A. Newcomb Polly, A. Trentham-Dietz, Cancer Causes and Control 14 (2003) 195–201.
[5] K.E. Anderson, E. Anderson, P.J. Mink, C.P. Hong, L.H. Kushi, T.A. Sellers, D. Lazovich, A.R. Folsom, Cancer Epidemiology Biomarkers and Prevention 10 (2001) 611–616.
[6] H. Xu Wang, W. Zheng, Y.B. Xiang, Z.M. Ruan, J.R. Cheng, Q. Dai, Y.T. Gao, X.O. Shu, British Medical Journal 328 (2004) 1285.
[7] P. Gujral, M. Amrhein, D. Bonvin, J.P. Vallee, X. Montet, N. Michoux, Chemometrics and Intelligent Laborary Systems 98 (2009) 173–181.
[8] M.J. Bruwer, J.F. MacGregor, M.D. Noseworthy, Journal of Chemistry 22 (2008) 708–716.
[9] J. Luts, F. Ojeda, R. Van de Plas, B. De Moor, S. Van Huffel, J.A.K. Suykens, Analytica Chimica Acta 665 (2010) 129–145.
[10] T. Mu, A.K. Nandi, Journal of The Franklin Institute 344 (2007) 285–311.
[11] E.D. Übeyli, Expert Systems with Applications 35 (2008) 1733–1740.
[12] C.-L. Huang, H.-C. Liao, M.-C. Chen, Expert Systems with Applications 34 (2008) 578–587.
[13] M.S. Neofytou, M.S. Pattichis, C.S. Pattichis, V. Tanos, E.C. Kyriacou, D.D. Koutsouris, Conference Proceedings of IEEE Engineering in Medicine and Biology Society 1 (2006) 3005–3008.
[14] Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, N. Jent, Journal of Pharmaceutical and Biomedical Analysis 44 (2007) 683–700.
[15] B. Schrader, B. Dippel, S. Fendel, S. Keller, T. Lochte, M. Riedl, R. Schulte, E. Tatsch, Journal of Molecular Structure 408–409 (1997) 23–31.
[16] P.C. Williams, S.G. Stevensen, Trends in Food and Science Technology 1 (1990) 44–48.
[17] J. Moros, S. Armenta, S. Garrigues, M.d.l. Guardia, Analytica Chimica Acta 579 (2006) 17–24.
[18] V.R. Kondepati, M. Keese, R. Mueller, B.C. Manegold, J. Backhaus, Vibrational Spectroscopy 44 (2007) 236–242.
[19] S.B. Kim, C. Temiyasathit, K. Bensalah, A. Tuncel, J. Cadeddu, W. Kabbani, A.V. Mathker, H. Liu, Expert Systems with Applications 37 (2010) 3863–3869.
[20] P.B. Harrington, J. Kister, J. Artaud, N. Dupuy, Analytical Chemistry 81 (2009) 7160–7169.
[21] S. Wold, H. Antti, F. Lindgren, J. Öhman, Chemometrics and Intelligent Laborary Systems 44 (1998) 175–185.
[22] C.A. Andersson, Chemometrics and Intelligent Laborary Systems 47 (1999) 51–63.
[23] B. Cheng, X. Wu, Advances in Mathematics 28 (1999) 365.
[24] H. Wang, Z. Wu, J. Meng, Partial Least-Squares Regression-Linear and Nonlinear Methods, first ed., National Defense Industrial Press, Beijing, 2006.
[25] C.H. Wan, P.B. Harrington, Analytica Chimica Acta 408 (2000) 1–12.
[26] P.B. Harrington, Trends in Analytical Chemistry 25 (2006) 1112–1124.
[27] G. Taguchi, S. Konishi, Orthogonal Arrays Linear Graphs: Tools for Quality Engineering, first ed., American Supplier Institute, Dearborn, 1987.