Journal of Biomedical Optics

SPIEDigitalLibrary.org/jbo

Biomarkers spectral subspace for cancer detection

Yi Sun
Yang Pu
Yuanlong Yang
Robert R. Alfano

SPIE

Biomarkers spectral subspace for cancer detection

Yi Sun,[a,b] Yang Pu,[b] Yuanlong Yang,[b] and Robert R. Alfano[a,b]
[a]City College of City University of New York, Electrical Engineering Department, Convent Avenue at 138th Street, New York, New York 10031
[b]City College of City University of New York, Institute for Ultrafast Spectroscopy and Lasers and Physics Department, Convent Avenue at 138th Street, New York, New York 10031

Abstract. A novel approach to cancer detection in biomarkers spectral subspace (BSS) is proposed. The basis spectra of the subspace spanned by fluorescence spectra of biomarkers are obtained by the Gram-Schmidt method. A support vector machine classifier (SVM) is trained in the subspace. The spectrum of a sample tissue is projected onto and is classified in the subspace. In addition to sensitivity and specificity, the metrics of positive predictivity, Score1, maximum Score1, and accuracy (AC) are employed for performance evaluation. The proposed BSS using SVM is applied to breast cancer detection using four biomarkers: collagen, NADH, flavin, and elastin, with 340-nm excitation. It is found that the BSS SVM outperforms the approach based on multivariate curve resolution (MCR) using SVM and achieves the best performance of principal component analysis (PCA) using SVM among all combinations of PCs. The descent order of efficacy of the four biomarkers in the breast cancer detection of this experiment is collagen, NADH, elastin, and flavin. The advantage of BSS is twofold. First, all diagnostically useful information of biomarkers for cancer detection is retained while dimensionality of data is significantly reduced to obviate the curse of dimensionality. Second, the efficacy of biomarkers in cancer detection can be determined. [DOI: 10.1117/1.JBO.17.10.107005]

Keywords: fluorescence spectra; cancer detection; biomarkers; multivariate curve resolution; principal component analysis; sufficient statistic; support vector machine; optical biopsy.

Paper 12420 received Jul. 3, 2012; revised manuscript received Sep. 19, 2012; accepted for publication Sep. 20, 2012; published online Oct. 10, 2012.

1 Introduction

Since the first report on the laser-induced fluorescence spectroscopy from biological tissues for cancer detection in the late 1980s,[1,2] fluorescence spectroscopy has been widely used for noninvasive diagnoses of cancer, developed into a field called optical biopsy. Fluorescence spectra of biochemicals play an important role in discrimination of cancerous tissue from normal tissue.[3,4] It is observed that a cancerous tissue contains more key molecules tryptophan and nicotinamide adenine dinucleotide (NADH), and less collagen than a normal tissue.[3,5,6] Quantification of the key biochemical components or biomarkers in a tissue provides a means of cancer detection. Multivariate curve resolution (MCR),[7–9] widely applied in chemometrics, can be employed to quantify the biochemical components by analyzing fluorescence spectrum. The fluorescence spectrum of a tissue is a superposition of spectra of various salient biochemicals and molecules. Thus component quantification is essentially an ill-posed problem. MCR, which is a model-free data fitting method, might result in invalid solutions; moreover, component quantification cannot guarantee to obtain sufficient statistics of the biomarkers, and the diagnostically useful information for cancer detection is inevitably reduced in terms of information theory. Spectral unmixing method (SUM),[10] similar to MCR, is employed in remote sensing to determine the endmembers and quantify their fractions in a hyperspectral imaginary. Compared with MCR, SUM further requires the full additivity of component fractions that might further reduce the useful information for cancer detection. On the other hand, in the cancer detection point of view, without knowing biomarker components, a classifier can be trained directly using training samples of normal and cancerous fluorescence spectra. A fluorescence spectrum is sampled to become a high-dimensional vector. To obviate the effect of curse of dimensionality, principal component analysis (PCA)[11] can be applied to reduce the dimensionality by projecting fluorescence spectra of tissue samples onto a few of selected principal components and then a tissue is classified in the spectral subspace spanned by the selected principal components.[12] PCA can effectively employ most power of sample spectra with dimensionality reduction. The contribution of important biomarkers to the principal components is usually unknown, while many unknown biochemicals may contribute to the principal components. It is of interest to employ spectra of biomarkers in cancer detection and know the efficacy of each biomarker.

In detection of a signal coming from a set of known signals with observed data corrupted by noise and inference, a sufficient statistic can be obtained by projection of the data onto the subspace spanned by the set of signals. The sufficient statistic contains all information about the signal to be detected and removes the redundant dimension and reduces the power brought by the noise and inference. The sufficient statistic can be used in detection of the signal without loss of information. This principle is extensively applied in signal detection and telecommunications.[13] Similar to but slightly different from the signal detection problem, cancer detection needs to determine whether an observed spectrum, corrupted by noise and the spectra of unknown biochemicals, comes from a cancerous or normal tissue based on a set of known spectra of biomarkers. Given a set of observed fluorescence spectra of tissues, a set of sufficient statistics can be obtained by projection of the observed spectra onto the subspace spanned by the spectra of

Address all correspondence to: R. Alfano, Electrical Engineering Department, The City College of City University of New York, Convent Avenue at 138th Street, New York, New York 10031. Tel: 2126505541; Fax: 2126505530; E-mail: ralfano@sci.ccny.cuny.edu

0091-3286/2012/$25.00 © 2012 SPIE

biomarkers. The sufficient statistics contain all diagnostically useful information provided by the biomarkers. Hence cancer can be detected in this subspace without loss of diagnostically useful information while significantly reducing dimensionality.

In this paper, for the first time to our knowledge, a new theoretical approach of spectral analysis to cancer detection in the biomarkers spectral subspace (BSS) is proposed. Specifically, the bases of BSS are obtained by the Gram-Schmidt method using the spectra of biomarkers. The spectra of human cancerous and normal tissue samples are projected onto the BSS. The support vector machine (SVM)[14,15] in the BSS is trained by using training samples. To evaluate the efficacy of a classifier for cancer detection, the metrics of positive predictivity, Score1, maximum Score1, and accuracy (AC) are employed in addition to the commonly used sensitivity, specificity, and receiver operating characteristic (ROC). The BSS approach is applied in this paper to analyze fluorescence spectra for breast cancer detection with 340-nm excitation and using four biomarkers, collagen, NADH, flavin, and elastin.

2 Methods and Materials

2.1 *Spectral Mixture Model*

Consider that a tissue consists of a number of components including the biomarkers of interest. After excitation, each component emits a particular fluorescence spectrum. Since a tissue is usually not particulate, the fluorescence light mostly comes from the surface of the tissue. The mixed spectrum is a linear combination of the spectra of all components. Given K biomarkers each having a fluorescence spectrum represented by an N-dimensional vector $\mathbf{c}_k$, $k = 1, \ldots, K$ where the dimension is indexed by wavelength. Without loss of generality, assume that $\mathbf{c}_k$'s are linearly independent, which is usually true in practice due to the fact of $K \ll N$. The fluorescence spectrum of the tissue can be written as

$$\mathbf{y} = \sum_{k=1}^{K} a_k \mathbf{c}_k + \mathbf{z} = \mathbf{C}\mathbf{a} + \mathbf{z}, \tag{1}$$

where $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K)$ is the matrix of spectra of biomarkers and $\mathbf{a} = (a_1, a_2, \ldots, a_K)^T$. Here, a_k is a constant proportional to the quantity of the k'th biomarker in the tissue. $\mathbf{z}$ is the spectrum due to other biochemicals and noise.

In chemometrics, MCR[7–9] is applied to quantify chemical components of a substance by analyzing its spectrum. In remote sensing, SUM[10] is applied to determine the endmembers and quantify their fractions through the hyperspectral imaginary in an area. In both cases, the linear mixing model Eq. (1) can be employed.[7–10] Since a_k represents the fraction of the k'th component, a_k is nonnegative, and this condition is imposed in both MCR and SUM. Moreover, in remote sensing, if the entire composition is considered in a mixed pixel, the full additivity of $\sum_{k=1}^{K} a_k = 1$ is further constrained in SUM.[10] Hence, in the framework of the linear mixing model Eq. (1), MCR and SUM solve the similar problem but the full additivity is further imposed in SUM. Clearly, quantification of component fractions a_k's that MCR and SUM solve is an ill-posed problem.

Unlike the chemometrics and remote sensing that aim at quantifying component fractions, cancer detection via analyzing fluorescence spectrum of a tissue aims at classification of the tissue by exploiting the information provided by the spectra of biomarkers. Hence cancer detection does not need to quantify the fractions a_k's of biomarkers in the tissue, and the condition of nonnegative a_k's is useless. In fact, all diagnostically useful information for cancer detection provided by the biomarkers spectra is retained in the well-defined spectral subspace of biomarkers.

2.2 *Biomarkers Spectral Subspace*

Since cancerous and normal tissues shall be discriminated by using the biomarkers, all information useful in cancer detection is embedded in the subspace spanned by the fluorescence spectra $\mathbf{c}_k$'s. The basis spectra of the subspace can be obtained by singular value decomposition (SVD) for $\mathbf{C}$. However, to clearly see how each biomarker contributes to the subspace and affects classification, the Gram-Schmidt method is employed to obtain the basis spectra as

$$\mathbf{b}_1 = \mathbf{c}_1 / \|\mathbf{c}_1\| \tag{2}$$

$$\mathbf{b}_k = \frac{\mathbf{c}_k - \sum_{i=1}^{k-1} (\mathbf{c}_k^T \mathbf{b}_i)\mathbf{b}_i}{\|\mathbf{c}_k - \sum_{i=1}^{k-1} (\mathbf{c}_k^T \mathbf{b}_i)\mathbf{b}_i\|}, \qquad k = 2, \ldots, K. \tag{3}$$

The spectrum $\mathbf{y}$ of a tissue sample is projected onto the subspace and forms a K-dimensional vector

$$\mathbf{s} = \mathbf{B}^T \mathbf{y}, \tag{4}$$

where $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_K)$ is the matrix of basis spectra. By the projection, $\mathbf{s}$ retains the component in the spectral subspace of biomarkers but gets rid of all other components orthogonal to the subspace. Statistically speaking, $\mathbf{s}$ is a sufficient statistic and contains all information useful in cancer detection when the biomarkers are used. Meanwhile, since K is usually much smaller than N, the projection of $\mathbf{y}$ onto the subspace also significantly reduces the dimensionality and effectively obviates the curse of dimensionality when training a classifier. A tissue shall be classified in the spectral subspace of biomarkers using $\mathbf{s}$. The order of biomarker spectra presented in the Gram-Schmidt method affects $\mathbf{B}$ only by a unitary transform and therefore does not affect the topological relations of a set of tissue spectra in the subspace; therefore, the order does not affect classification of tissues for a given classifier.

The power of a sample spectrum $\mathbf{y}$ is defined as $\|\mathbf{y}\|^2$ and the total power of a set of spectra $\mathbf{y}_i$ for $i = 1, 2, \ldots, M$ is equal to $\sum_{i=1}^{M} \|\mathbf{y}_i\|^T$. By BSS, the spectral dimensionality is significantly reduced. Consequently, the spectral power of sample tissues in the subspace is usually less than the total spectral power of the original samples. The ratio of spectral power in the spectral subspace of biomarkers to the total spectral power measures the power usage of the biomarkers in the BSS. Denote by $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M)$ the matrix of M sample tissue spectra, the power usage of the biomarkers in the BSS is equal to

$$R = \frac{\mathrm{tr}[(\mathbf{B}^T\mathbf{Y})^T(\mathbf{B}^T\mathbf{Y})]}{\mathrm{tr}(\mathbf{Y}^T\mathbf{Y})}, \tag{5}$$

where tr means the trace of matrix. It is expected that a set of proper biomarkers can attain a high power usage while significantly reducing the dimensionality of spectral space.

SVM is well-known one of the most powerful classifiers.[14,15] By using training samples of cancerous and normal tissues, a classifier that is defined by a hyperplane $\mathbf{w}^T\mathbf{s} = b$ in the subspace is obtained by SVM with a linear kernel. In general, the SVM classifier is determined by and most effectively discriminates a number of samples, so called the support vectors, located at the boundary between cancerous and normal tissues in the spectral space. A schematic diagram of the proposed BSS approach to cancer detection in the spectral subspace of biomarkers is shown in Fig. 1.

MCR and PCA are popularly used in tissue classification. The approach and performance of BSS will be compared with those of the MCR and PCA based classifiers using SVM, which for completeness are presented in the following subsections.

2.3 *Multivariate Curve Resolution*

In the approach of MCR, the fractions a_k's of biomarkers in a tissue are estimated from the tissue spectrum by[7,8]

$$\mathbf{a}^* = \arg\min_{a_k \geq 0} \|\mathbf{C}\mathbf{a} - \mathbf{y}\|. \tag{6}$$

Then the tissue is classified according to $\mathbf{a}^*$. Specifically, an SVM classifier is trained using the estimated fractions of tissue samples.

Although the vector $\mathbf{a}^*$ also has K dimensions, it is usually not located in the spectral subspace where $\mathbf{s}$ in Eq. (4) is located. If there was no constraint of $a_k \geq 0$, a solution to Eq. (6) would be $\mathbf{a}^* = \mathbf{C}^+\mathbf{y}$ where $\mathbf{C}^+$ is the pseudoinverse of $\mathbf{C}$. By SVD, $\mathbf{C} = \mathbf{U}\Lambda\mathbf{V}^T$ where $\mathbf{U}$ and $\mathbf{V}$ are the matrices of column and row singular vectors of $\mathbf{C}$, respectively, and Λ is the diagonal matrix of nonzero singular values of $\mathbf{C}$; and then $\mathbf{C}^+ = \mathbf{V}\Lambda^{-1}\mathbf{U}^T$. Note that the columns of $\mathbf{U}$ are also a set of bases of the spectral subspace of biomarkers (but usually different from the set of bases of $\mathbf{B}$ by a unitary transform). Hence, replacing $\mathbf{B}$ by $\mathbf{U}$ in Eq. (4) produces the same topological relationship of samples in the spectral subspace and therefore does not affect classification in the BSS approach. This demonstrates that the MCR approach is different from the BSS approach in two ways: 1. the constraint of $a_k \geq 0$ is imposed and 2. in the case that $\mathbf{a}^* = \mathbf{C}^+\mathbf{y}$ can satisfy the constraint of $a_k \geq 0$, $\mathbf{C}^+$ further performs the transform of $\mathbf{V}\Lambda^{-1}$ after projecting $\mathbf{y}$ onto the spectral subspace by $\mathbf{U}^T$. Since all useful information for cancer detection is in the spectral subspace of biomarkers, the MCR process of Eq. (6) usually reduces the useful information and makes the classification less reliable in the case when biomarkers, which can provide sufficient discriminability, are used in cancer detection. SUM is similar to MCR but further requires the full additivity, which might further reduce the diagnostically useful information in cancer detection. In fact, though the full additivity can be realistic in remote sensing, it is usually unrealistic in cancer detection as a tissue may contain many other biochemicals in addition to the biomarkers.

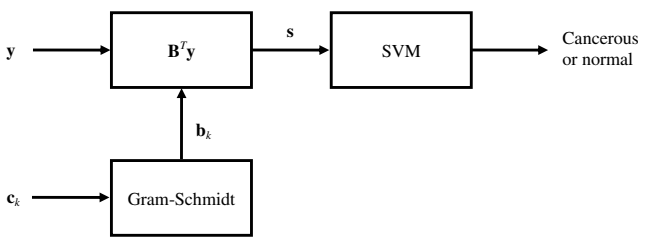

Fig. 1 The BSS approach. Biomarkers spectra $\mathbf{c}_k$'s yield the basis spectra $\mathbf{b}_k$'s of the subspace by the Gram-Schmidt method. The spectrum of a sample tissue $\mathbf{y}$ is projected onto the subspace and yields $\mathbf{s}$, based on which the sample is determined as cancerous or normal tissue by the SVM classifier.

2.4 *Principal Component Analysis*

In the approach of PCA, the spectra of biomarkers are not employed. Instead, the principal components (PCs), along which the most power of sample spectra is located, are obtained by SVD of sample spectra. Let $\mathbf{x}_i = \mathbf{y}_i - \bar{\mathbf{y}}$ be the spectrum of the i'th sample tissue with subtraction of the average spectrum $\bar{\mathbf{y}}$ of $\mathbf{y}_i$'s and let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M)$. By SVD, $\mathbf{X} = \mathbf{PTQ}^T$ where $\mathbf{P}$ and $\mathbf{Q}$ are the matrices of column and row singular vectors of $\mathbf{X}$, respectively, and $\mathbf{T} = \mathrm{diag}(t_1, \ldots, t_M)$ for $M < N$ is the diagonal matrix of singular values in the descent order. Let $\mathbf{P}_L = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_L)$ be the matrix of the first L PCs of $\mathbf{P}$. Then $\mathbf{v}_i = \mathbf{P}_L^T\mathbf{x}_i$ is the component of $\mathbf{x}_i$ in the subspace spanned by the first L PCs. An SVM classifier in the subspace is trained by $\mathbf{v}_i$'s.

By PCA, the spectral dimensionality can be significantly reduced as L can be chosen to be much smaller than N. Consequently, the spectral power of tissue samples in the PC's subspace is usually less than the total spectral power. The power usage of PCA, defined as the ratio of spectral power of tissue samples in the subspace to the total spectral power, is equal to

$$R = \frac{\mathrm{tr}[(P_L^T\mathbf{X})^T(P_L^T\mathbf{X})]}{\mathrm{tr}(\mathbf{X}^T\mathbf{X})} = \frac{\sum_{i=1}^{L} t_i^2}{\sum_{i=1}^{M} t_i^2}. \tag{7}$$

Therefore, PCA selects a subspace, spanned by the first L PCs, such that the power usage attains the maximum among all L-dimensional subspaces. Like BSS, PCA can significantly reduce the spectral dimensionality, while retaining the most power of samples. However, the contribution of biomarkers to the PCs is unknown and many unknown biochemicals can contribute to the PCs.

2.5 *Performance Metrics*

To evaluate performance of a classifier in cancer detection, sensitivity and specificity are commonly used metrics

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{8}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}, \tag{9}$$

where TP, FP, FN, TN are the numbers of true positive, false positive, false negative, and true negative samples, respectively. In cancer detection, sensitivity is much more important than specificity. Moreover, the portion of cancerous samples in all samples in a clinical practice might be small. In this case, increase of the number of samples that are classified as cancer, i.e., TP + FP, can increase the sensitivity, but this does not mean the classifier is more sensitive to the true positive samples, i.e. the cancerous samples. Hence it is worthy to employ the metrics of positive predictivity ${}^+P$ and Score1[16]

$$^{+}P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{10}$$

$$\mathrm{Score1} = \min(\mathrm{Sensitivity}, {}^{+}P). \tag{11}$$

Positive predictivity (^{+}P) is the portion of true positive samples among all the samples classified as positive. As TP + FP increases, positive predictivity (^{+}P) usually decreases, opposite to sensitivity. In order to fairly evaluate the performance of a classifier in sensitivity to true positive samples, both sensitivity and positive predictivity must be considered and therefore their minimum, Score1, fairly evaluates the performance of a classifier[16] in detection of true positive samples. Positive predictivity is also called positive predictive value (PPV) that depends on prevalence, a health-related state of population that is used in epidemiology. To fairly evaluate performance of a classifier, accuracy (AC) can also be employed

$$\mathrm{AC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \tag{12}$$

which is the ratio of the number of all correctly classified samples to the total number of samples.

Given an SVM classifier $\mathbf{w}^T\mathbf{s} = b$, the sensitivity, specificity, positive predictivity, Score1, and AC can be calculated accordingly. Changing the distance b of the hyperplane to the origin but retaining the normal angle determined by $\mathbf{w}$, different pairs of sensitivity and specificity can be obtained and form ROC of the classifier in terms of sensitivity versus 1−specificity. The closer to one the area under curve (AUC) of ROC is, the better the performance of a classifier is. Meanwhile, changing b with fixed $\mathbf{w}$ also yields a curve of positive predictivity versus 1−specificity. In general, as the specificity decreases, sensitivity increases but positive predictivity decreases, and therefore their intersection is the maximum Score1 that all the classifiers, $\mathbf{w}^T\mathbf{s} = b$ with changing b, can attain.

2.6 *Biomarkers and Samples*

In the experiments, the proposed BSS approach is applied to breast cancer detection via fluorescence spectral analysis and is compared with the MCR and PCA approaches in performance. Six main fluorophores—tryptophan, collagen, elastin, NADH, flavin, and tyrosine—have been reported to exist in breast cells and tissues.[5,6] The primary fluorophore in the breast tissue extracellular matrix is type I collagen.[17] According to the Scarff-Bloom-Richardson (SBR) system,[18] breast cancerous cells present the features of higher cell density, uncontrollable cell division, and nonuniform larger cellular nuclei. Correspondingly, increased fluorescence of the main fluorophores—tryptophan, NADH, and flavin—is expected inside the cells. It has been observed that a cancerous tissue contains more key molecules tryptophan and NADH and less collagen than a normal tissue.[3,5,6] The wavelengths of absorption and emission peaks are (275, 303) for tyrosine, (287, 342) for tryptophan, (339, 380) for collagen, (351, 410) for elastin, (340, 460) for NADH, and (375, 525) for flavin, respectively, in the order of increasing wavelength in the unit of nm. In order for a set of biomarkers to have their peak emission wavelength close to the excitation wavelength, four biomarkers—collagen, NADH, flavin, and elastin—with 340-nm wavelength excitation were chosen for breast cancer detection in the experiments. These biomarkers were obtained commercially from Mallinckrodt Baker, Inc. Their emission spectra with the same concentration of about 0.75 mg/cm^3 were measured individually with the excitation at 340 nm using the same experiment method as that for breast tissues.

Cancerous and normal breast tissue samples were provided by the Co-operation Human Tissue Network (CHTN) and National Disease Research Interchange (NDRI) under IRB approvals. The cancerous and normal breast tissue samples were diagnosed by a pathology medical doctor. Samples were neither chemically treated nor were frozen prior to the experiments. The time elapsed between tissue resection and measurement may vary for different sample sources. The elapsed times are about 30 h. Fluorescence spectra of 37 normal tissues and 37 breast cancerous tissues excited by 340-nm light were measured using Perkin-Elmer LS-50 spectrometer. In the spectrometer, an equipped Xenon lamp coupled to the monochromator delivers light to the sample spot at a desired wavelength, and the emission from the sample is collected by emission monochromator connected to a photomultiplier tube. The fluorescence was measured from the front-face of tissue samples to reduce the distortion of absorption and scattering, which is often done for turbid or opaque samples.[19]

3 Experimental Results

The objectives of experiments are to examine performance of the proposed BSS approach, to investigate the efficacy of biomarkers in cancer detection using the proposed BSS, and to compare the performance with those of MCR and PCA using SVM. BSS, MCR, and PCA are applied to breast cancer detection with four biomarkers—collagen, NADH, flavin, and elastin—with 340-nm wavelength excitation. Figure 2 shows the average spectra of the 74 tissues with standard deviations, which, for both cancerous and normal tissues, are large. The fluorescence spectra of the four biomarkers are shown in Fig. 3, and their crosscorrelations are given in Table 1. The crosscorrelation between biomarker spectra $\mathbf{c}_i$, $\mathbf{c}_j$ is defined as $\mathbf{c}_i^T\mathbf{c}_j/(\|\mathbf{c}_i\|\|\mathbf{c}_j\|)$. Flavin has low crosscorrelations with other biomarkers in the fluorescence spectrum. This implies that flavin can provide a significant component additional to the subspace spanned by {collagen, NADH, elastin}.

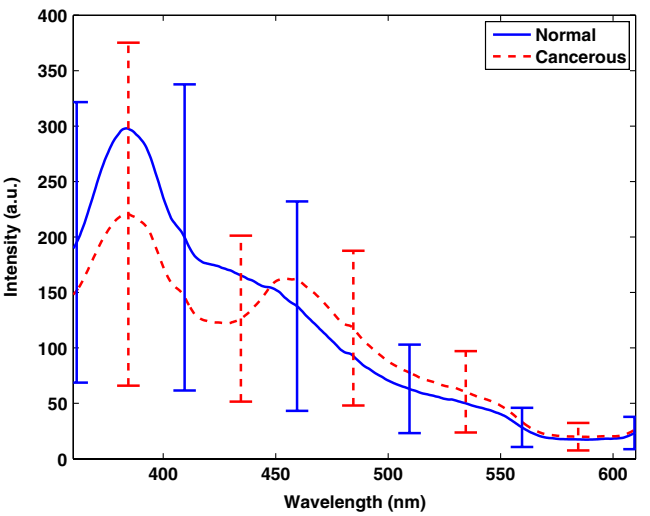

Fig. 2 Average spectra of normal and cancerous samples with standard deviations.

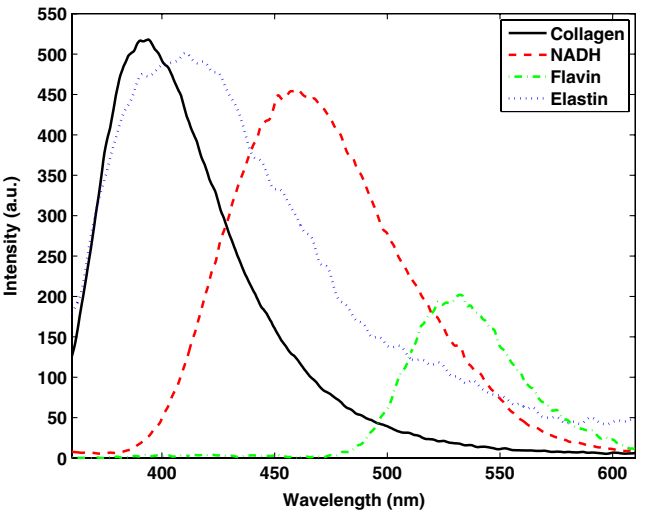

Fig. 3 Fluorescence spectra of four biomarkers.

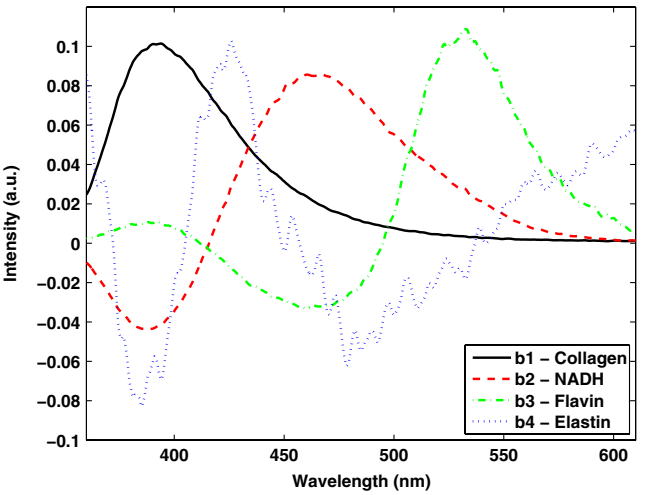

Fig. 4 Basis spectra of four biomarkers obtained by the Gram-Schmidt method in the order of collagen, NADH, flavin, and elastin.

Table 1 Crosscorrelations of biomarkers spectra.

Biomarker	1	2	3	4
1	1	0.423	0.065	0.940
2	0.423	1	0.342	0.691
3	0.065	0.342	1	0.218
4	0.940	0.691	0.218	1

1: Collagen; 2: NADH; 3: Flavin; 4: Elastin

Table 2 Correlations between the biomarker and basis spectra.

Biomarker	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{b}_4$
Collagen	1	0	0	0
NADH	0.423	0.900	0	0
Flavin	0.065	0.347	0.936	0
Elastin	0.940	0.324	0.048	0.099

In all the experiments, the linear kernel is applied in the SVM classifier. To classify the data in an identical space, the sample spectra in the subspace of biomarker spectra $\mathbf{s}$ are identically linearly transformed to locate in the cube $[0,1]^K$ (replacing K by L for the PCA approach). The linear transform retains the topological relationship among sample spectra and therefore does not affect the classification result. Then the parameter of C in the SVM classifier can be identical in all experiments. By a number of tests on the data of fluorescence spectra in this paper and others for Raman, resonant Raman, and Stokes shifts spectra, it is found that $C = 1000$ can achieve reasonably good performances and is then used in all experiments.

Figure 4 illustrates the basis spectra of the subspace spanned by the spectra of all four biomarkers, which are obtained by the Gram-Schmidt method in the order of collagen, NADH, flavin, and elastin. That is, $\mathbf{b}_1$ is obtained by normalizing spectrum of collagen to unit length, $\mathbf{b}_2$ is obtained by taking the component of NADH spectrum that is orthogonal to $\mathbf{b}_1$ and then normalized to unit length, etc. The correlations between the biomarker spectra and the basis spectra are illustrated in Table 2. If all the four biomarkers are used to classify the 74 samples using SVM, the power usage by BSS is 0.949, specificity is 0.973, sensitivity is 0.919, positive predictivity is 0.971, and therefore Score1 is 0.919, and AC is 0.946. By changing b with the fixed $\mathbf{w}$, the ROC AUC is 0.993, close to one, the maximum Score1 is 0.973, and the number of support vectors (SV #) is 18 as listed in the last row of Table 3.

To investigate the efficacy of a biomarker in cancer detection, performance of all possible combinations of four biomarkers is calculated, and the results are presented in Table 3. In each row, a subset of biomarkers is given, the basis spectra of their spectral subspace are obtained by the Gram-Schmidt method, and then the SVM is trained. For example, Fig. 5 illustrates the data samples, support vectors, SVM classifier and NADH in the subspace spanned by the spectra of {collagen, NADH}. It is clear that collagen alone is incapable of properly classifying the normal and cancerous samples as shown in Fig. 5 and the first row of Table 3. However, collagen and NADH jointly can classify very well the samples with specificity of 0.973, sensitivity of 0.919, ${}^+P$ of 0.971, Score1 of 0.919, AC of 0.946, ROC AUC of 0.962, and maximum Score1 of 0.919 shown in Table 3. Figure 6 shows the sensitivity and positive predictivity (${}^+P$) versus 1−specificity. The ROC AUC, the area under the sensitivity curve, is equal to 0.962. As the specificity decreases, the sensitivity increases but ${}^+P$ decreases, and their intersection is the maximum Score1 = 0.919.

From Table 3 the order of efficacy of biomarkers can be obtained. If one biomarker is used, collagen outperforms the other three biomarkers. If two biomarkers are used, {NADH, elastin} outperforms the other five combinations. If three biomarkers are used, {collagen, NADH, elastin} outperforms the other three combinations. If all four biomarkers are used, the performance is the best in terms of the maximum Score1 and ROC AUC. Hence the order of efficacy of biomarkers in breast cancer detection in this experiment is collagen, NADH, elastin, and flavin.

Table 3 Performance of BSS.

Biomarkers	Power usage	Spec.	Sens.	^{+}P	Score1	AC	ROC AUC	Max. Score1	SV #
1	0.846	0.514	0.649	0.571	0.571	0.581	0.589	0.596	74
2	0.417	0.541	0.486	0.514	0.486	0.514	0.530	0.558	74
3	0.057	0.568	0.459	0.515	0.459	0.514	0.583	0.569	74
4	0.908	0.514	0.595	0.550	0.550	0.554	0.557	0.574	74
1, 2	0.934	0.973	0.919	0.971	0.919	0.946	0.962	0.919	30
1, 3	0.880	0.865	0.838	0.861	0.838	0.851	0.907	0.842	48
1, 4	0.923	0.946	0.865	0.941	0.865	0.905	0.927	0.868	42
2, 3	0.420	0.730	0.486	0.643	0.486	0.608	0.678	0.659	73
2, 4	0.919	0.973	0.946	0.972	0.946	0.959	0.971	0.946	30
3, 4	0.912	0.838	0.784	0.829	0.784	0.811	0.884	0.811	54
1, 2, 3	0.942	0.919	0.919	0.919	0.919	0.919	0.969	0.923	28
1, 2, 4	0.936	0.973	0.946	0.972	0.946	0.959	0.993	0.949	18
1, 3, 4	0.929	0.919	0.838	0.912	0.838	0.878	0.952	0.895	35
2, 3, 4	0.923	0.946	0.919	0.944	0.919	0.932	0.970	0.946	29
1, 2, 3, 4	0.949	0.973	0.919	0.971	0.919	0.946	0.993	0.973	18

1: Collagen; 2: NADH; 3: Flavin; 4: Elastin

Table 3 also demonstrates that whenever collagen or elastin is used, about more than 85% power of data spectra is used in the cancer detection, which means a high power usage by the biomarkers. Table 3 also implies that when a combination of biomarkers makes samples easier to be classified in a subspace, the support vectors are fewer.

The performance of the MCR using SVM classifier is presented in Table 4. When one biomarker is employed, the performance of MCR is identical to that of BSS. However, when more biomarkers are employed, overall the BSS with ROC AUC of 0.993 and maximum Score1 of 0.973 outperforms the MCR with ROC AUC of 0.973 and maximum Score1 of

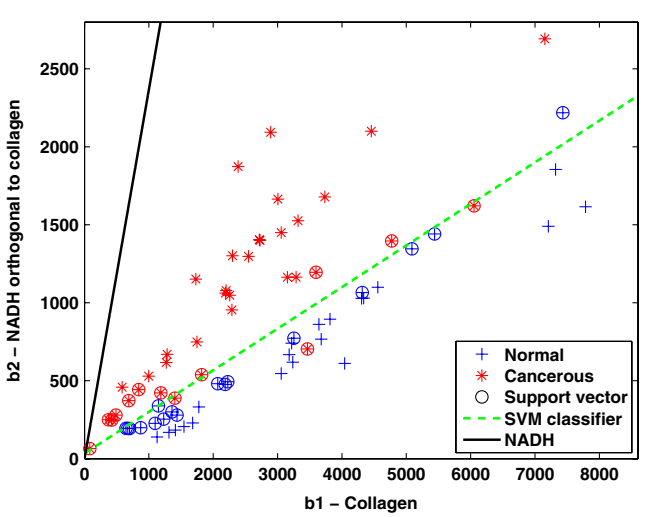

Fig. 5 Normal and cancerous samples, support vectors, SVM classifier and NADH component in the subspace spanned by the spectra of collagen and NADH. Specificity is 0.973, sensitivity is 0.919, ^{+}P is 0.971, Score1 is 0.919, and AC is 0.946.

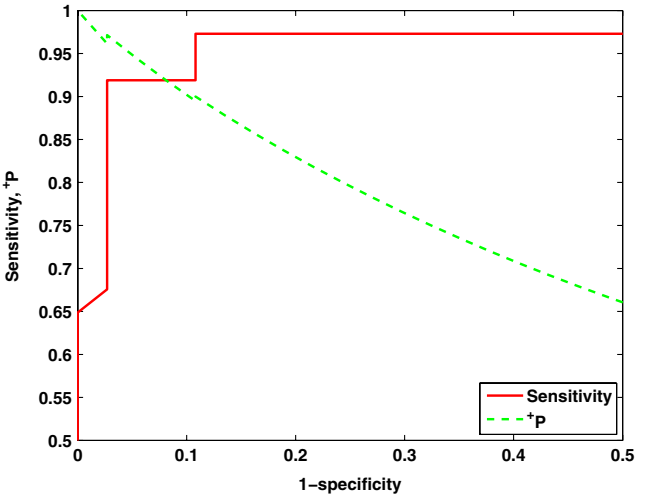

Fig. 6 Sensitivity and positive predictivity (^{+}P) versus 1—specificity when collagen and NADH are used, corresponding to the condition in Fig. 5. The ROC AUC is 0.962. As the specificity decreases, the sensitivity increases but ^{+}P decreases and their intersection is the maximum Score1 = 0.919.

Table 4 Performance of MCR.

Biomarkers	Spec.	Sens.	^{+}P	Score1	AC	ROC AUC	Max. Score1	SV #
1	0.514	0.649	0.571	0.571	0.581	0.589	0.596	74
2	0.541	0.486	0.514	0.486	0.514	0.530	0.558	74
3	0.568	0.459	0.515	0.459	0.514	0.583	0.569	74
4	0.514	0.595	0.550	0.550	0.554	0.557	0.574	74
1, 2	0.946	0.919	0.944	0.919	0.932	0.963	0.919	29
1, 3	0.865	0.838	0.861	0.838	0.851	0.907	0.842	48
1, 4	0.838	0.919	0.850	0.850	0.865	0.922	0.865	44
2, 3	0.730	0.568	0.677	0.568	0.649	0.665	0.632	72
2, 4	1.000	0.865	1.000	0.865	0.932	0.869	0.865	45
3, 4	0.865	0.784	0.853	0.784	0.824	0.885	0.825	53
1, 2, 3	0.946	0.892	0.943	0.892	0.919	0.970	0.947	27
1, 2, 4	0.946	0.946	0.946	0.946	0.946	0.979	0.946	23
1, 3, 4	0.919	0.838	0.912	0.838	0.878	0.949	0.892	35
2, 3, 4	0.946	0.865	0.941	0.865	0.905	0.927	0.865	39
1, 2, 3, 4	0.946	0.946	0.946	0.946	0.946	0.973	0.947	24

1: Collagen; 2: NADH; 3: Flavin; 4: Elastin

Table 5 Performance of PCA.

PCs	Power usage	Spec.	Sens.	^{+}P	Score1	AC	ROC AUC	Max. Score1	SV #
1	0.926	0.514	0.595	0.550	0.550	0.554	0.563	0.574	74
2	0.036	1.000	0.946	1.000	0.946	0.973	0.994	0.947	14
3	0.025	0.703	0.514	0.633	0.514	0.608	0.576	0.541	74
4	0.008	0.432	0.432	0.432	0.432	0.432	0.486	0.507	74
1, 2	0.962	1.000	0.946	1.000	0.946	0.973	0.995	0.973	14
1, 3	0.951	0.459	0.676	0.556	0.556	0.568	0.673	0.622	74
1, 4	0.934	0.459	0.568	0.512	0.512	0.514	0.571	0.566	74
2, 3	0.061	0.973	0.946	0.972	0.946	0.959	0.997	0.973	12
2, 4	0.043	0.973	0.946	0.972	0.946	0.959	0.995	0.947	14
3, 4	0.033	0.649	0.622	0.639	0.622	0.635	0.606	0.622	74
1, 2, 3	0.987	0.973	0.973	0.973	0.973	0.973	0.994	0.973	13
1, 2, 4	0.970	1.000	0.973	1.000	0.973	0.986	0.994	0.973	13
1, 3, 4	0.959	0.459	0.649	0.545	0.545	0.554	0.655	0.622	73
2, 3, 4	0.068	0.919	0.973	0.923	0.923	0.946	0.986	0.923	14
1, 2, 3, 4	0.995	0.865	0.973	0.878	0.878	0.919	0.957	0.921	14

Table 6 Performance in cross validation with leave-one-out.

	BSS			MCR			PCA		
Biomarkers or PCs	Spec.	Sens.	AC	Spec.	Sens.	AC	Spec.	Sens.	AC
1	0.514	0.595	0.554	0.514	0.595	0.554	0.514	0.595	0.554
2	0.514	0.351	0.432	0.514	0.351	0.432	0.973	0.919	0.946
3	0.568	0.460	0.514	0.568	0.460	0.514	0.676	0.487	0.581
4	0.514	0.541	0.527	0.514	0.541	0.527	0.432	0.432	0.432
1, 2	0.919	0.919	0.919	0.946	0.892	0.919	0.946	0.946	0.946
1, 3	0.865	0.838	0.851	0.865	0.838	0.851	0.460	0.676	0.568
1, 4	0.865	0.839	0.851	0.838	0.919	0.878	0.460	0.514	0.487
2, 3	0.649	0.405	0.527	0.730	0.541	0.635	0.973	0.919	0.946
2, 4	0.973	0.919	0.946	1.000	0.865	0.932	0.973	0.919	0.946
3, 4	0.811	0.784	0.797	0.811	0.757	0.784	0.595	0.595	0.595
1, 2, 3	0.917	0.865	0.892	0.919	0.865	0.892	0.946	0.919	0.932
1, 2, 4	0.946	0.865	0.905	0.865	0.946	0.905	0.973	0.892	0.932
1, 3, 4	0.892	0.838	0.865	0.892	0.838	0.865	0.405	0.595	0.500
2, 3, 4	0.919	0.919	0.919	0.892	0.811	0.851	0.919	0.919	0.919
1, 2, 3, 4	0.973	0.892	0.932	0.946	0.892	0.919	0.838	0.892	0.865

For BSS and MCR: 1: Collagen; 2: NADH; 3: Flavin; 4: Elastin

0.947 as some useful information might be lost by the MCR process.

The performance of the PCA using SVM classifier is presented in Table 5. Although taking into account 92.6% power of sample spectra, PC 1 is not the most effective PC in cancer detection; instead, taking only 3.6% power of sample spectra, PC 2 is the most effective PC in cancer detection. If two PCs are used, PCs 1, 2 outperform the other combinations. If three PCs are used, PCs 1, 2, 4 outperform the other combinations. Hence the order of efficacy of PCs in cancer detection is PC 2, PC 1, PC 4, and PC 3. In fact, PCs 1, 2, 4 outperform PCs 1, 2, 3, 4. In other words, PC 3 deteriorates the performance. These results imply that employing the PCs with more power of sample spectra may not result in a better performance, and using more PCs may not necessarily improve the performance in cancer detection. Overall, with properly chosen PCs, the PCA can achieve the same performance of the BSS using all biomarkers. However, it is unknown how each biomarker contributes to the PCs. Moreover, the PCs and their subspace are determined by the training samples, which may change as the number of tissue samples increases while the biomarkers spectral space is completely determined by the biomarkers.

In the above experiments, all 74 samples are used to train the SVM. To further compare the performance, cross validation with leave-one-out is also carried out for BSS, MCR, and PCA using the SVM classifier. The results are presented in Table 6. As expected, BSS outperforms MCR and achieves the highest accuracy of PCA among all combinations of PCs. All the analyses are performed in Matlab and the codes are available at http://www-ee.engr.ccny.cuny.edu/www/web/ysun/NanoscopeLab/Codes/BSS-JBO2012/BSS.htm.

4 Conclusions

We propose a novel theoretical approach to cancer detection from the fluorescence spectral subspace of biomarkers. Projection of sample spectra onto the biomarkers spectra subspace (BSS) retains all useful information for cancer detection while significantly reducing data dimensionality. The efficacy of biomarkers in cancer detection can be determined in the subspace. In comparison, multivariate curve resolution (MCR) and spectral unmixing methods (SUM) may decrease the diagnostically useful information of sample spectra and make cancer detection less reliable. Principal component analysis (PCA) can achieve good performance if the best combination of PCs is employed, but PCA relies on the spectra of all sample tissues, and the contribution of biomarkers to PCs is unknown.

Applied to breast cancer detection with fluorescence spectrum of 340-nm wavelength excitation, the proposed BSS with support vector machine (SVM) outperforms the MCR-SVM and achieves the best performance of the PCA-SVM among all combinations of PCs. It is found that the efficacy of biomarkers in the breast cancer detection is in the order of collagen, NADH, elastin, and flavin. In principle, the BSS approach can be applied extensively to detection of other types of cancers using fluorescence and other spectra.

In addition to the commonly used sensitivity and specificity, the metrics of positive predictivity, Score1, maximum Score1, and accuracy (AC) can jointly provide a proper measurement of the sensitivity of a classifier to the true positive cancerous samples.

Finally, it needs to be pointed out that the efficacy of the BSS approach in cancer detection ultimately depends on the diagnostic power provided by the used biomarkers; moreover, the BSS approach does not attempt to quantify the fractions of biomarkers in a tissue. However, the BSS approach can retain all the diagnostic information of biomarkers, while significantly reducing the dimensionality, and provide a means to determine the efficacy of a biomarker in cancer detection.

Acknowledgments

This research is supported in part by the U.S. Army Medical Research and Material Command grants of W81XWH-08-1-0717 (CUNY RF 47170-00-01), W81XWH-11-1-0335 (CUNY RF # 47204-00-01) and Army Research Office (ARO). The authors acknowledge the help of CHTN and NDRI for providing normal and cancerous breast tissue samples for the measurements.

References

1. R. R. Alfano et al., "Laser induced fluorescence spectroscopy from native cancerous and normal tissue," *IEEE J. Quantum Electron.* **20** (12), 1507–1511 (1984).
2. R. R. Alfano et al., "Fluorescence spectra from cancerous and normal human breast and lung tissues," *IEEE J. of Quant. Electron QE* **23**(10), 1806–1811 (1987).
3. Y. Pu et al., "Changes of collagen and nicotinamide adenine dinucleotide in human cancerous and normal prostate tissues studied using fluorescence spectroscopy with selective excitation wavelength," *J. Biomed. Opt.* **15**, 047008 (2010).
4. Y. Pu et al., "Native fluorescence spectroscopic evaluation of chemotherapeutic effects on malignant cells using nonnegative matrix factorization analysis," *Technol. Cancer Res. Treat. (TCRT)* **10**(2), 113–120 (2011).
5. I. Georgakoudi et al., "NAD(P)H and collagen as in vivo quantitative fluorescent biomarkers of epithelial precancerous changes," *Cancer Res.* **62**, 682–687 (2002).
6. R. Drezek et al., "Understanding the contributions of NADH and collagen to cervical tissue fluorescence spectra: modeling, measurements, and implications," *J. Biomed. Opt.* **6**(4), 385–396 (2001).
7. R. Tauler, A. Smilde, and R. Kovalski, "Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution," *J. Chemom.* **9**, 31–58 (1995).
8. A. De Juan and R. Tauler, "Multivariate curve resolution (MCR) from 2000: progress in concepts and applications," *Crit. Rev. Analyt. Chem.* **36**, 163–176 (2006).
9. A. Kandelbauer, W. Kessler, and R. W. Kessler, "Online UV-visible spectroscopy and multivariate curve resolution as powerful tool for model-free investigation of laccase-catalysed oxidation," *Ana. Bioanal. Chem.* **390**, 1303–1315 (2008).
10. N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE Signal Proc. Mag.* **19**(1), 44–57 (2002).
11. I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., Springer, New York (2002).
12. J. Mo et al., "High wavenumber Raman spectroscopy for *in vivo* detection of cervical dysplasia," *Anal. Chem.* **81**, 8908–8915 (2009).
13. J. G. Proakis, *Digital Communications*, 4th ed., McGraw Hill, New York, NY (2000).
14. C. Cortes and V. N. Vapnik, "Support-vector networks," *Mach. Learn.* **20**, 273–297 (1995).
15. W. H. Press et al., "Section 16.5. support vector machines," *Numerical Recipes: The Art of Scientific Computing*, 3rd ed., Cambridge University Press, New York (2007).
16. PhysioNet, "Predicting mortality of ICU patients: the PhysioNet/computing in cardiology challenge 2012," http://www.physionet.org/challenge/2012.
17. G. Fenhalls, D. M. Dent, and M. I. Parker, "Breast tumour cell-induced down-regulation of type I collagen mRNA in fibroblasts," *Br. J. Cancer* **81**, 1142–1149 (1999).
18. H. J. G. Bloom and W. W. Richardson, "Histological grading and prognosis in breast cancer: a study of 1409 cases of which 359 have been followed for 15 years," *Br. J. Cancer* **11**(3), 359–377 (1957).
19. J. Eisinger and J. Flores, "Front-face fluorometry of liquid samples," *Anal. Biochem.* **94**(1), 15–21 (1979).