Hyper-spectral microscopic discrimination between normal and cancerous colon biopsies

Franco Woolfe*, Mauro Maggioni, Gustave Davis, Frederick Warner, Ronald Coifman, and Steven Zucker

***Abstract*— The spectral study of cancer dates back 50 years, but it is still not known whether spectral measurements suffice to distinguish cancerous from normal tissue. An objective approach to that question is designing automatic classifiers for discrimination between these two classes and then estimating generalization error rates. Previous studies have not estimated errors adequately: it is not a priori clear whether unseen spectra from patients in the algorithm's test set are sufficiently independent of the training data to provide a fair evaluation. We show experimentally that to obtain accurate error estimations, spectra from unseen patients are necessary. Our results suggest that although spectra are not sufficient to distinguish fully between cancerous and normal tissue, some high degree of discrimination is possible. This leads us to ask how discriminatory spectral features should be selected. The features in previous work on cancer spectroscopy have been chosen according to heuristics. We use the "best basis" algorithm to select a Haar wavelet packet basis which is optimal for the discrimination task at hand. These provide interpretable spectral features consisting of contiguous wavelength bands. However they are outperformed by features which use information from all parts of the spectrum, combined linearly at random.**

I. Introduction

HYPER-SPECTRAL imaging for the characterization of cancer dates back more than 50 years [1]. A natural question is whether information in the spectrum is sufficient to distinguish cancerous from normal tissue. To answer it we can design automatic classifiers that are blind to all other information. The error rate of the classifier with respect to the entire population then quantifies the spectral information which this classifier is able to access and which is useful for discriminating between cancerous and normal tissue. The two problems are (1) that there might be information available that a particular classifier misses and (2) that we can never evaluate a classifier on the entire population to find its true error rate. The first problem is one of building an optimal classifier, which is the subject of statistical learning (see for example [2]) and has not been solved in general. Certainly however, using a particular classifier will yield a lower bound on the available information provided that its error rate is estimated reliably.

The second problem, estimating the error rate of a classifier given only a finite data sample, has been well studied. The standard solution is cross-validation, introduced by Stone in [3]: one partitions the data at random into a training set (used to build the classifier) and a testing set (used to obtain one estimate of its error rate). By partitioning many times one can calculate reliable estimates for the true error rate. Unfortunately, for the general problem of statistical model performance estimation "analytical results are difficult, if not impossible" according to [4]. On the other hand, extensive simulation studies [4], [5] have shown the reliability of cross validation empirically.

The underlying rationale behind testing a classifier on unseen data is that the unseen data should be independent of that used for training. For example, suppose a classifier is trained on spectra from a certain patient and evaluated on different spectra from the same patient. We call that approach *weak cross validation*. By contrast we use the term *strong cross validation* to refer to training on data from some patients and testing on different patients. Molckovsky et al. in [6] introduce their use of weak cross validation saying that "although multiple spectra could be obtained from a large polyp, each Raman spectrum was considered independent". Other uses of weak cross validation in the literature on cancer recognition algorithms include [7]–[11]. On the other hand some works [12]–[14] make use of strong cross validation and yet others do not specify [15]–[18]. This suggests some researchers may be unaware that the distinction between weak and strong cross validation can be an issue. The question is: are the success rates reported by weak cross validation studies believable estimates for the error rate on the total population? In this paper we answer that question by evaluating our algorithms using both the strong and weak cross validation frameworks, for comparison. Our results indicate that weak cross validation is not sufficient to reliably estimate the out of sample error rate of a classification algorithm: it tends to over estimate success rates.

Thus in order to obtain a true lower bound on the useful spectral information content of tissue, for the task of cancer recognition, we must use strong cross validation. Our classification results using that framework do suggest that some information relevant to discrimination between normal and cancerous colon tissue is available in visible light spectra. We proceed to ask whether that information is confined to certain wavelength bands. Is the entire spectrum needed for classification or do a smaller number of spectral features suffice to extract the available information? Identifying a smaller number of relevant features has the additional advantages that we can decrease computational and image acquisition times since data volume will be lower. What is more, the discrimination will be more straightforward with only a few features due to lower dimensionality of the problem (alleviating the "curse of dimensionality" [2]). Previous studies on how to select

Manuscript received January 99, 9999; revised November 99, 9999.

Mauro Maggioni is with Duke University Mathematics Department. The other authors are with the Yale University Program in Applied Mathematics. *Franco Woolfe is the corresponding author; email: Franco.Woolfe@Yale.edu.

relevant spectral features for cancer discrimination include [15] which uses linear regression to select spectral bands as features. On the other hand [12] uses AdaBoost, the machine learning technique, and [13] prefers a genetic algorithm. Other approaches use wavelets for selecting relevant spectral features, for example [16], [19]. A principled way to select the best wavelet basis for a particular discrimination task is given by the best basis algorithm [20]–[22]. It has the advantage of finding the best set of features available for a specific discrimination task by searching a wavelet packet tree, and has been used in the hyper-spectral geosensing literature [23], [24]. It provides a principled way to select relevant features. We adopt that method in this paper, with the Haar basis.

The features provided by the "best basis" algorithm with Haar wavelets, on our data set, consist of contiguous bands of wavelengths. Thus they have the potential to be give interpretable information about what parts of the spectrum are important. Other features do not have this property, in particular those obtained from the random projections method [25]–[27]. The random projections features consist of randomly chosen wavelengths of light. Each feature is spread randomly across the entire spectrum, so it may be counterintuitive that they should preserve any relevant information at all. However results in [28] suggest that separation between classes is indeed preserved under such a random mixing procedure. In our experiments we find that the random projection method outperforms the best Haar wavelet packet basis, as a feature selection technique. Thus there is a trade off between interpretable results about important parts of the spectrum ("best basis") and higher classification accuracy (random projections).

Having acquired spectral measurements, a nonlinear technique for classification is employed. This is Laplacian eigenmaps [29]–[34], which takes into consideration the curved geometry of the subspace occupied by the spectral measurements. By contrast to linear techniques for similar tasks [15], Laplacian eigenmaps parameterize non-planar manifolds of data so are more general and more powerful.

Our experimental set up is geared towards ease of use by practicing pathologists. For that reason we use a hyper-spectral light source that can easily be attached to a visible light microscope [35]. Other studies of discrimination between normal and cancerous colon tissue use ultraviolet [36] or infrared [6], [16], [37], [38] micro-spectroscopy: these types of techniques are not in general use by pathologists. In addition, we use biopsies stained by hematoxylin and eosin (H&E), the standard pathology staining method. On the other hand [15], [16], [39] use unstained samples for non-visible microscopic techniques, which would not usually be available to a pathologist.

This paragraph summarizes the issues addressed by this paper. We explicitly compare weak and strong cross validation to determine which is more appropriate for estimating classification error rates. Strong cross validation allows us to bound from below the information content of visible light spectra for discriminating between normal and cancerous colon biopsies. We evaluate two approaches for spectral feature selection: Haar wavelet packet best bases (active sensing) and random projections. Throughout we ensure that the methods we develop can be used by a practicing pathologist, with no more than access to a desktop computer and a hyper-spectral light source which can be attached to a microscope.

Novel contributions of this paper include an explicit comparison of the strong and weak cross-validation frameworks and use of the best basis algorithm for cancer classification. In addition this is the first published use of random projections for hyper-spectral colon cancer classification. It is the first study of hyper-spectral imaging of the colon using visible light spectra, H&E stained biopsies and strong cross validation. Finally this is the first study of hyper-spectral cancer analysis which makes use of Laplacian eigenmaps to take into account the nonlinear geometry in the design of learning algorithms.

II. Summary of our approach

This section summarizes section IV.

We photograph 20 normal and 20 cancerous (adenocarcinoma) human colon biopsies from over 200 such biopsies on a tissue micro-array, obtained from the Yale Tissue Microarray facility [40]. Different biopsies come from different patients and the preparation (H&E staining) may vary from patient to patient. The prototype tuned light source [35] generates combinations of visible light at 128 different wavelengths. These transilluminate the biopsies, passing through a Nikon Biophot microscope.

We design algorithms to automatically discriminate between the normal and the cancerous biopsies. In order to evaluate the error rates of our algorithms, we use cross validation. This consists of selecting a random subset of the data which is used for training and evaluating the algorithm on the remaining data. This is repeated many times to obtain an accurate error estimate. *Weak cross validation* consists in having at least some spectra in the testing and training sets taken from the same patient. It is only an acceptable approach if one assumes independence between spectra from the same patient. On the other hand *strong cross-validation* consists of testing the algorithms on spectra from unseen patients. In order to test the independence assumption on different spectra from the same patient, we try both frameworks. We find that weak cross-validation artificially inflates the algorithm's success rate. This is because of differences in biopsy preparation between different patients, leading to differences between biopsies that do not correlate with diagnosis. For example total spectral energy varies from patient to patient but contains no information about being normal versus cancerous (see Figure 1). Testing an algorithm on the same patients it was trained on (weak cross-validation) means it does not have to normalize for these irrelevant differences. Therefore it has an artificially easier task and appears to perform better.

A. Acquisition modes

Initially biopsies are imaged at all available wavelengths; call this approach *passive sensing*. Since classification accuracies with strong cross validation using entire spectra are encouraging we conclude that there is at least some spectral information relevant to cancer detection. This leads us naturally to consider where and how this information is contained in the spectra. To study that problem we consider spectral feature

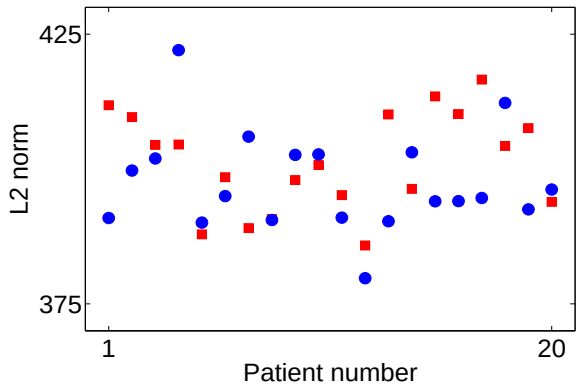

Fig. 1. Mean $L2$ norm of nuclei spectra for each biopsy. There are significant differences between patients, due in part to differences in sample preparation. But these differences do not correlate with being cancerous (red squares) or normal (blue circles). To obtain reliable success rate estimates, classifiers must be evaluated on unseen patients (strong cross validation) rather than unseen spectra from known patients (weak cross validation).

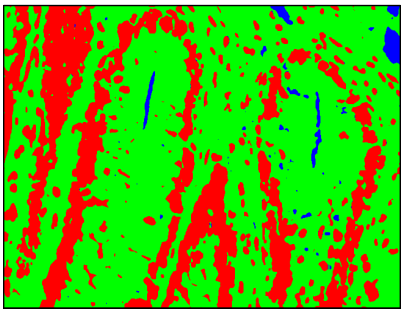

Fig. 2. Passive sensing tissue classification of a normal biopsy (all available wavelengths were used). Cell nuclei are in red, glandular cell cytoplasm is in green and lamina propria is in blue.

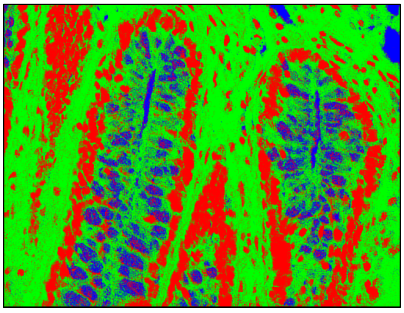

Fig. 3. Active sensing tissue classification of a normal biopsy (a carefully chosen subset of relevant wavelengths was used). Cell nuclei are in red, glandular cell cytoplasm is in green and lamina propria is in blue.

selection. In particular we evaluate two classes of spectral features. The first consists of bands of wavelengths, chosen by local discriminant bases to be relevant to our discrimination task. These features can be measured directly by our instrument [35], which can be programmed to shine exactly those bands in question, a procedure we call *active sensing*. We go on to consider features consisting of randomly chosen subsets of all available wavelengths. Surprisingly, theoretical results show that such measurements not only preserve separation between separable classes [28] and approximate distances between points [26], [41] but also can be used to reconstruct the full spectra [27]. Again these features can be measured directly by our instrument which can be programmed to shine only those wavelengths of light which have been selected. We call that approach *random sensing*.

B. Algorithmic stages

- *Tissue segmentation* Since histologic changes in the nucleus are a hallmark of cancer, our recognition algorithms work on nucleic spectra. To identify the locations of nuclei, the biopsies are initially segmented into regions corresponding to three biologically significant tissue classes. These are cell nuclei, glandular cell cytoplasm and lamina propria cell cytoplasm. For examples see Figures 2, 3 and 4.
- *Nucleus classification* Having performed tissue segmentation, spectral measurements belonging cell nuclei are automatically extracted. A classification algorithm is trained on nucleus features which are labeled as normal or cancerous. Nuclear classification can be either weakly or strongly cross validated.
- *Biopsy classification* A biopsy is classified as normal if at least some fraction of its nuclei are classified as normal (above). Biopsy classification can be either weakly or strongly cross validated.

Identifying small sets of relevant features has many engineering advantages beyond revealing how the discriminatory information is arranged inside spectra. These include making classification tasks "easier" and smaller numbers of features lead to lower imaging and computational times. See table II for some typical imaging time savings.

To achieve a more objective basis for comparison of results, we use the same set of passive measurements for all our computational experiments. Active and random sensing are simulated by taking averages of sets of wavelengths as needed, before inputting to the classification algorithms.

III. Mathematical methods used

A. Hadamard spectroscopy

To obtain low noise hyper-spectral images in a short period of time, imaging at all wavelengths, we exploit Hadamard multiplexing [42]. Our hyper-spectral device is capable of shining N different wavelengths $\{\nu_i\}_{i=1}^N$ of light and combinations thereof. To measure the biopsy's response when transilluminated by each wavelength the naive approach (called a raster scan) consists in shining one wavelength at a time through the biopsy. However this means that for a given intensity I of the light source the energy shone per pattern is only I/N. Thus the photographic exposure time must be very long for a reasonable signal to noise ratio (SNR): about 1024s per biopsy would be typical. Instead we use multiplexing to shine patterns consisting of many wavelengths at once. Let $\psi_i^j \in \{0,1\}$ denote whether or not light is shone at frequency ν_j in the

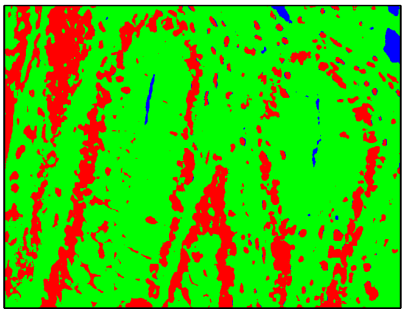

Fig. 4. Random sensing tissue classification of a normal biopsy (randomly chosen spectral features were used). Cell nuclei are in red, glandular cell cytoplasm is in green and lamina propria is in blue.

i^{th} pattern. In order to reconstruct physical spectra from our measurements, it is important that the vectors ψ_i span the whole of $\mathbb{R}^N$. Hadamard basis vectors (see [42]) suffice and have the following carefully designed properties:

- For each i there are $N/2$ non zero $(\psi_H)_i^j$. So the intensity of the light shining through the biopsy is about $I/2$ for each of the patterns.
- The set $\{(\psi_H)_1, \ldots, (\psi_H)_N\}$ of vectors is a complete (spanning) set for $\mathbb{R}^N$.
- The index i parametrizes rapidity of oscillation: for small i, $(\psi_H)_i^j$ is slowly varying as a function of j whereas for larger i, $(\psi_H)_i^j$ oscillates rapidly as a function of j.

Here more light is shone per pattern and we can expect less noisy measurements, while still spanning $\mathbb{R}^N$ (that is, the transform is still invertible). However, because naturally occurring spectra are gradual functions of frequency, the response to gradually varying patterns (ψ_i for small i) will be higher than for rapidly varying patterns; i.e. the signal to noise ratio decreases as a function of i. To ensure a minimum SNR for all measurements, an exposure time long enough to accommodate the most rapidly varying pattern must be used. In that case the SNR for the lowest frequency pattern will be higher than we need. To remedy this inefficiency we use permuted Hadamard vectors, which come from applying the same random permutation to the elements of each basis vector ψ_i. These allow us to maximize the SNR uniformly across patterns shone. In more detail we build a single random bijection $m : \{1, \ldots, N\} \to \{1, \ldots, N\}$ and define $(\psi_{RH})_i^j = (\psi_H)_i^{m(j)}$. We compute the permutation once and use that shuffling in all our measurements. After shuffling, all Hadamard basis vectors have about the same frequency of oscillation. Of course, the change of variable m just induces an orthogonal transformation between $\{(\psi_H)_i\}$ and $\{(\psi_{RH})_i\}$, so the permuted Hadamard transform is also still invertible. Also the size of all the collected coefficients is roughly constant in i. So the exposure time can be kept constant for all the patterns. We choose the exposure time just under the saturation level of the CCD. (In our experiments the exposure time was 250 ms and $N = 128$).

B. Laplacian eigenmaps

In order to classify spectra as to their class of biological tissue, we need to take into consideration the geometric relationships between spectral measurements, thought of as points in space. These points lie on or near a curved subspace (see Figure 5).

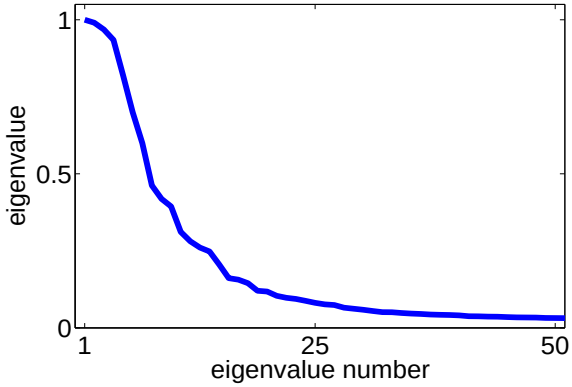

Fig. 5. The Laplacian eigenvalues of a graph whose nodes are constructed from tissue spectra (passive sensing). Note how rapidly the eigenvalues decay: the top 20 eigenvalues account for 98% of the total energy. This justifies the assumption that the spectra lie on or near a curved subspace of $\mathbb{R}^D$ and our use of only the top eigenvectors for embedding.

Laplacian eigenmaps [29]–[34] allow us to study the intrinsic geometry of curved subspaces such as those in question (see Figure 9). In particular we measure vectors $\{v_i\}_{i=1}^M \subset \mathbb{R}^D$ which represent spectra and lie on a low dimensional ($d << D$) manifold $\mathcal{M}$. We can set up a coordinate system for $\mathcal{M}$ using vectors with only d elements. Thus we can represent the position of each data point v_i by only d coordinates $\{w_i\}_{i=1}^M \subset \mathbb{R}^d$ rather than D, which can lead to massive savings in storage, computation and data acquisition time.

To achieve such a parametrization, consider the Laplacian operator on the manifold $\mathcal{M}$. The idea is that the d coordinates of point i should be the values of the top d Laplacian eigenvectors at point i. One way of understanding this approach is that it optimally preserves distances between nearby points: if v_i and v_j are close together in the D dimensional space it is desirable that their images in d dimensions remain close. In a sense which can be made precise, the Laplacian embedding preserves local distances as well as possible [29]. Another way of understanding the Laplacian embedding is via heat diffusion. Assume that the manifold conducts heat, but is insulated from the remainder of the high dimensional space. Consider a point source of heat located at one of the points on the manifold. The time taken for that heat to diffuse to other points on the manifold is related to the distances between those points in the low d dimensional embedding space. This diffusion metric is robust to noisy data since heat will not travel quickly between points on the manifold unless there are many paths connecting them. So the creation or deletion of some small number of paths, due to measurement noise in the $\{v_i\}_{i=1}^M$, will not strongly affect the distances between points in the low d dimensional embedding space.

In order to compute the low dimensional embedding coordinates of data, it is necessary to approximate the continuous

Laplacian operator by a matrix. Define the Laplacian matrix as $L = I - A$ where $A = (a_{i,j})_1^M$ and $a_{i,j} = \exp(-||v_i - v_j||/\epsilon)$, before normalization. Here A is called the adjacency matrix. In fact, we used a Laplace-Beltrami normalization [43] for L. The advantage of that normalization is invariance to the sampling density of the points on the manifold. Now diagonalize L and let $w_i(m)$ be the i^{th} coordinate of the m^{th} eigenvector (for $m = 1, \ldots, d$). For any particular data set, the values of d and ϵ are not known a priori and need to be determined experimentally.

C. Local discriminant bases

Despite the considerable savings in imaging time obtained through use of Hadamard spectroscopy (section III-A), it still takes 32s to acquire a hyper-spectral image of a single biopsy. That is because Hadamard spectroscopy is intended to capture images at all available frequencies. We show that faster acquisition times can be obtained by only measuring the response at those frequencies which are relevant to the discrimination task at hand. We call this approach active sensing which we implement using local discriminant bases (LDB) of Coifman and Saito [20]–[22].

Given labeled sets of high dimensional training data points the method finds a small number of directions such that projecting the data onto those directions preserves the separation between the classes. The high dimensional training points in question are physical spectra in this situation. Projecting the data onto a small number of directions corresponds to shining a small subset of the available wavelengths which have been chosen specifically with the goal preserving separation between classes. A classifier is then learned in the projected lower dimensional space corresponding to the small number of measurements made. This leads to faster image acquisition and data-analysis times.

The search for features in high dimensional spaces is notoriously difficult. LDB performs dimensionality reduction by searching sub-optimal projections among hierarchically well-organized dictionaries of wavelet or Fourier packets. There are fast algorithms with to perform such a search and efficiently compute the projections onto ensembles of these patterns. We use a version of LDB that searches arbitrary Haar packet decompositions. Note the LDB method is fundamentally limited by the constraint that the projections available in its search space come from Haar packet decompositions. Nonetheless we find it can reduce imaging time with little subsequent loss of classification accuracy.

IV. Algorithms

A. Tissue segmentation

Since histologic changes in cell nuclei occur with the onset of cancer, our spectral cancer discrimination algorithms work on nuclei. In order to locate nuclei patches they begin by segmenting the spectral measurements into three classes corresponding to cell nuclei, glandular cell cytoplasm and lamina propria. The H&E stain used in slide preparation differentiates between nuclei and other tissue components, which makes the task tractable. The major difficulty is to achieve consistency across different biopsies. There are significant differences between biopsies which should be ignored, for example due to different uptake of the stain, slightly different focus and lighting settings in the microscope (see Figure 1).

The tissue segmentation algorithms are supervised: we label a subset of the locations from three biopsies according to tissue class. Given these labeled points, the algorithm assigns a class to each unlabeled location. Measurements are made at the labeled points in either passive, active or random sensing mode. The measurements made are given to the algorithm in section IV-A.2. For the results, see Figures 2, 3 and 4.

1) Measurements:

- *Passive sensing* The input to the tissue segmentation algorithm in the next section is the full spectra - that is the response to N different wavelengths of light at each location in the biopsies (see Figure 6).

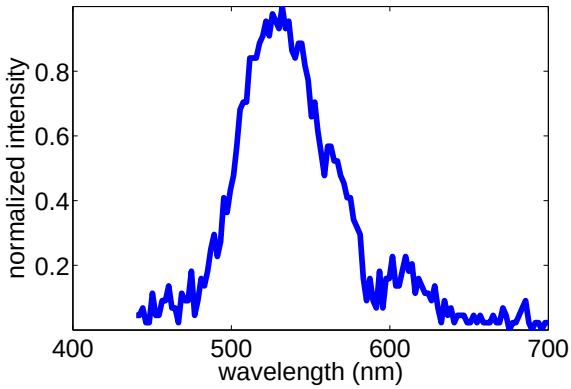

Fig. 6. A colon transmittance spectrum between 440nm and 700nm. These measurements are used by the "passive" algorithms.

- *Active sensing* We use the LDB algorithm (see section III-C and Figure 7) to find d features that discriminate among the full spectra of different tissue classes in the training set. The features are optimal subject to being the first d vectors of a Haar packet basis for $\mathbb{R}^N$. These responses of the tissue to these features (d numbers per location in the biopsies) are subsequently classified using the tissue segmentation algorithm in the next section.
- *Random sensing* We also make only d spectral measurements, instead of N. However the measurements made each consist of shining a random subset of the available wavelengths (see Figure 8). Again, these d measurements per location are input into the following algorithm.

2) Algorithm: We use a nonlinear classifier that takes advantage of the curved geometry of the measurement space using Laplacian eigenmaps (see Figure 9 and section III-B) as follows. Suppose $I(x, y) \in \mathbb{R}^D$ is the measurement made at a location (x, y), by either passive, active or random sensing. Let $S = \{s_i\}_{i=1}^{25}$ be the measurements in a 5 by 5 neighborhood about (x, y). S is used to calculate k local statistics which capture variation near (x, y), assuming that S contains realizations of the same random variable. The local statistics are the top k eigenvalues $\{\sigma_i\}_{i=1}^k$ of the local covariance matrix

$$C = \sum (s_i - \mu)^\top (s_i - \mu),$$

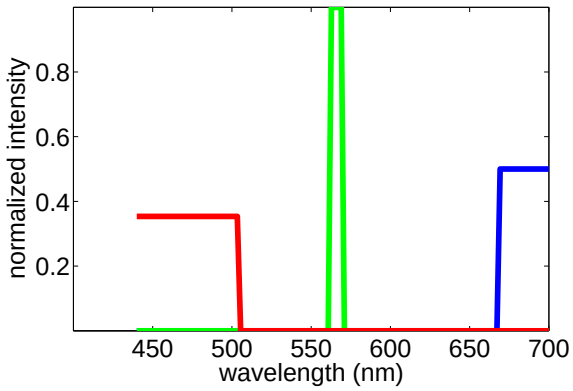

Fig. 7. Three wavelength bands carefully selected to be relevant to the task of tissue classification (active sensing). Feature vectors used by the active sensing algorithms consist of averages of transmittance spectra (Figure 6) over bands such as those shown. The vectors plotted are from the LDB algorithm (see section III-C).

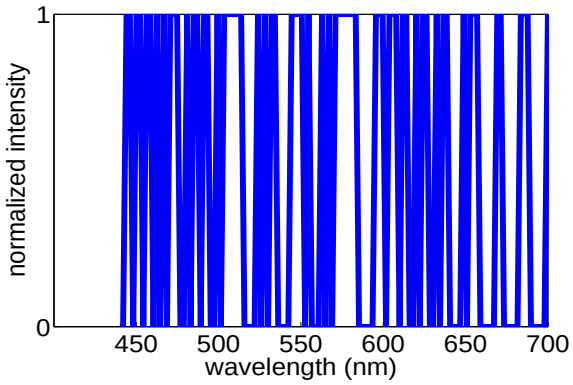

Fig. 8. The measurements used by the random sensing learning algorithms are sums of transmittances at randomly chosen wavelengths.

where μ is the mean spectrum of S. Then we form the feature vector at (x, y): $f(x, y) = [\mu^\top, \sigma_1, \ldots, \sigma_k]^\top \in \mathbb{R}^{D+k}$. The feature vector is normalized by linearly mapping each coordinate into the interval $[0, 1]$.

The choice of a 5 by 5 neighborhood is motivated by this being small enough to fit inside most cell nuclei. The spatial characteristics considered are at such a low scale as not to capture information about histological structures. Essentially they provide a very local measure of spectral variation about a point.

The physical and biological spectral compositions can be modeled by constraining the set of feature vectors to lie on or near a manifold $\mathcal{M} \subsetneq \mathbb{R}^{D+k}$ whose intrinsic dimensionality is less than $D+k$. So we construct an empirical parametrization of the point cloud by using Laplacian eigenmaps (see section III-B). This nonlinear map from $\mathbb{R}^{D+k}$ to $\mathbb{R}^n$, which we learn on the training spectra, is extended to all other spectra (in all other biopsies) by the Nyström extension technique [44]–[46]. To classify a new spectrum, we compute the local spatial statistics, assemble the feature vector, apply Nyström extension to deduce its low dimensional embedding coordinates in $\mathbb{R}^n$ and use a 10-nearest neighbor classifier [2] in the n dimensional diffusion space.

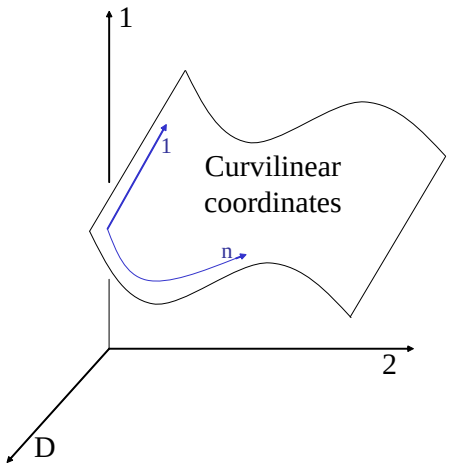

Fig. 9. Laplacian eigenmaps allow us to perform classification of data on a curved manifold. Distances between unseen points and labeled training points are measured in curvilinear coordinates along the manifold rather than in the D dimensional ambient space.

B. Classification of Nuclei

Nuclear classification can be either weakly or strongly cross validated (see section II). That is the nuclear classifier can be trained on all biopsies (weak) or trained on some and evaluated on others (strong).

1) Passive and random sensing: In passive and random sensing, once we have detected where nuclei are located, they are classified by partial least squares regression (PLSR, see [2], [47]–[51]). The PLSR algorithm is given full spectra for passive sensing (128 measurements at each location) and only 16 random measurements per location for random sensing. We use 15 latent vectors in the PLSR algorithm.

2) Active sensing: In active sensing, once we have detected where nuclei are located, we make use of a second set of measurements which have been chosen by the LDB algorithm (section III-C) to discriminate between normal and cancerous nuclei. Given the measurements we run a $10-$nearest neighbors classifier [2] to assign one of the two classes.

C. Classification of biopsies

To classify a biopsy, we collect between 40 and 60 nuclei. We find that selecting spectra from near the centers of nuclei increases confidence that the spectra really are nucleic, thus improving diagnostic efficiency (when cross validating strongly). The number of nuclei selected varies from biopsy to biopsy according to availability and confidence the algorithm has in the locations really being nucleic. Each nucleus is classified as in section IV-B. Each biopsy is classified as cancerous if at least some fraction of its nuclei spectra are classified as cancerous. It is possible to vary the sensitivity

and specificity of our algorithms by altering that policy, for example conservatively classifying a biopsy as cancerous if at least 10% of nuclei spectra are deemed cancerous (Figures 11 and 10).

V. Experiments

A micro-array was obtained from the Yale Tissue micro-array facility, containing normal and cancerous human colon biopsies stained with H&E. Each biopsy comes from a different patient and biopsy preparation may vary from patient to patient. We consider 40 such biopsies: 20 normal biopsies and 20 cancerous carcinomas. One of us (Dr Gustave L Davis, M.D., a board-certified pathologist) photographed the biopsies after confirming the diagnoses and evaluating adequacy of each biopsy for selection.

The prototype tuned light source [35] can generate 128 light frequencies, between 440 nm and 700 nm, with a wavelength resolution of about 6 nm. A fiber optic cable connects the light source to a Nikon Biophot microscope. We work at 400X magnification. The instrument is flexible in that it can transilluminate a sample with a combination of wavelengths at the same time. It is a prototype obtained from Plain Sights Systems Inc., Hamden, CT. Related hyper-spectral light sources are available commercially. For example Tidal Photonics Inc., Vancouver, Canada [52] produces a hyper-spectral light source that can also be configured to shine combinations of light of different frequencies. It has been used to construct a hyper-spectral bronchoscope in [53]. Cambridge Research and Instrumentation Inc., Woburn, MA [54] sells a liquid crystal tunable filter which has been used for cytological analysis of bladder cancer in [55]. In addition Spectral Dimensions, Inc., Olney, MD, [56] produces microscopes equipped with hyper-spectral light sources.

We take advantage of the flexibility to shine combinations of light wavelengths in three different ways. In passive sensing we acquire the image at all available wavelengths by shining randomized Hadamard patterns which minimize the data acquisition time subject to the signal to noise ratio being below a level we choose (see section III-A for more details). In random sensing we shine randomly chosen wavelengths (see section IV-A.1). In active sensing we shine wavelengths which have been chosen to be particularly relevant to the discrimination task at hand (see section III-C) out of a Haar wavelet packet tree.

A single data cube is collected for each biopsy. It is a set of D images, $\{I_i\}_{i=1}^{D}$. Each image has size 491 by 652 pixels which are the response of the biopsy when transilluminated by a particular combination of wavelengths of light.

The number of measurements made (D) and number of spatial features calculated (k) for the tissue segmentation algorithm (section IV-A.2) are given in table I. For the tissue segmentation stage, the active sensing algorithm is given 8 measurements which is fewer than the random sensing algorithm which has access to 16 measurements. This is in order to make the final comparison between success rates fair. The reason is that the random sensing method re-uses the same random measurements twice: once for tissue segmentation and again for nuclear classification. On the other hand the active sensing method relies on being able to choose relevant measurements for each task separately. For the task of biopsy classification as a whole the random and active methods have 16 measurements each.

TABLE I
Parameters used in the algorithms (see IV-A.2).

acquisition mode	D	k	n
passive	128	20	20
active	8	8	4
random	16	16	8

VI. Results

Tables II and III show the strongly cross validated results of the cancer recognition and tissue segmentation stages of the algorithm, respectively. In particular, table III shows the percentage of the locations automatically recognized as nuclei which are indeed nuclei, as verified by a pathologist.

The algorithms described above classify a biopsy as cancerous if at least some fraction of its nuclei are classified as cancerous. That fraction can be varied, to produce Figures 10 and 11, for the case of random sensing. Analogous Figures for active and passive sensing look almost identical. These Figures show robustness of the schemes proposed when the threshold parameter is varied. In addition Figure 11 shows the trade off between sensitivity and specificity. The area under the receiver operator characteristic (ROC) curve is a measure of robustness of a test, and must be between 0 and 1. For our ROC curves this area is about 0.9.

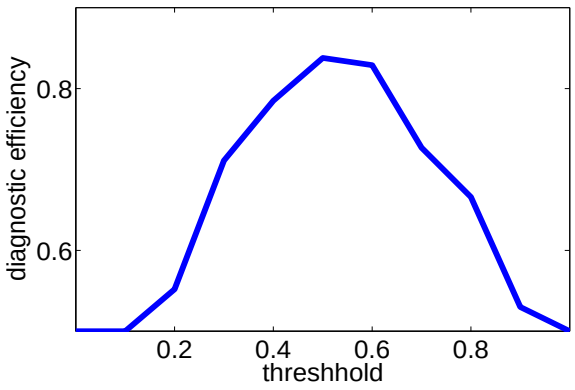

Fig. 10. Diagnostic efficiency as a function of the threshold fraction of spectra. This is the fraction of nuclei from a biopsy which need to be classified as malignant in order for the whole biopsy to be classified as malignant. Each point on the graph is the average diagnostic efficiency over 100 runs of strong cross-validation on the data set of 40 patients. The optimal threshold is 0.5.

For the weakly cross validated task (testing on the same patients as the algorithm trains on) we achieve almost perfect results as in [6], [8] which also use weak cross validation. Our diagnostic efficiency is 99%. For the harder task of strong cross-validation (testing on unseen patients), our tissue classification results are near 100%. For strongly cross validated cancer recognition, we achieve 85% diagnostic efficiency in passive sensing mode, that is when the algorithm has access

TABLE II

CANCER DETECTION SUCCESS RATES

sensing mode	cross validation type	diagnostic efficiency (%)	sensitivity (%)	specificity (%)	time per biopsy (s)
passive	weak	99	99	100	32
passive	strong	85	77	94	32
active	strong	82	85	78	4
random	strong	85	78	92	4

TABLE III

NUCLEI DETECTION SUCCESS RATES, STRONG CROSS VALIDATION

sensing mode	success rate (%)
passive	100
active	99
random	100

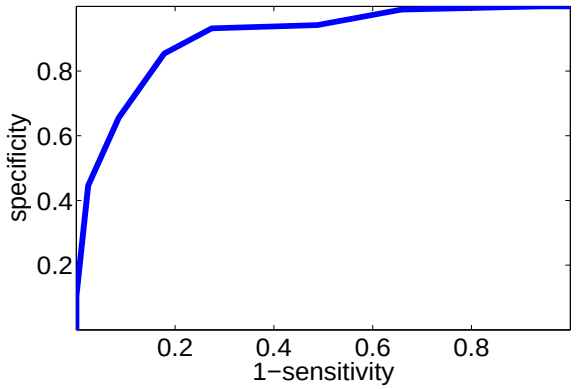

Fig. 11. Receiver operator characteristic (ROC) curve for the random sensing algorithm. This shows the trade-off between too many false alarms versus catching all true positives. True positive rate, or specificity, is plotted on the y-axis as a function of $1-$sensitivity, or false alarm rate, on the x-axis. The area under the ROC curve is a standard measure of robustness of a test, and is always between 0 and 1. In this case it is 0.9.

to all spectral bands. This requires an acquisition time of 32s per biopsy. For active and random sensing, the measurement time per data cube falls to 4s. The diagnostic efficiency only decreases to 82% for active and remains at 85% for random sensing with strong cross validation. The features chosen at random slightly, but consistently, outperform those chosen by LDB (active sensing).

VII. CONCLUSIONS AND FURTHER WORK

An example of a lower bound on the information content is that in at least 85% of cases spectral information alone is sufficient to discriminate between normal and cancerous biopsies. Thus the question arises as to what form this spectral information takes. Are the entire spectra needed in order to obtain this classification accuracy? The answer is no: using only 16 spectral measurements are sufficient in order to obtain that success rate.

Of the two classes of features we use, random measurements outperform spectral band features selected by LDB (active sensing). One might have expected the reverse since in active sensing the algorithm makes a choice about which patterns will be particularly helpful. On the other hand, the set of candidate patterns that LDB (active sensing) can choose from is limited to start with: they must be basis vectors from Haar wavelet packets. Random sensing has no such restriction. In addition active sensing is handicapped in that it cannot use the same measurements twice for different tasks. The measurements made are specific to either tissue segmentation or nuclear classification. To make the comparison between random and active fair, both are given access to the same number of total measurements. Thus for each sub-task (tissue segmentation and nuclear classification) the active algorithms have access to fewer measurements. A priori it is not obvious which of these methods will perform best for a particular application. In this case the benefits of random sensing outweigh those of active.

We have successfully designed algorithms for the discrimination of normal and cancerous colon biopsies given little more than materials in general use by pathologists. In particular we have used H&E stained biopsies and a visible light microscope with a hyper-spectral light source attached to it as described in [35]. We recognize the importance of keeping imaging times low. These relatively standard materials and ambitious goals have necessitated the use of sophisticated algorithms which nonetheless run on a standard desktop computer. The algorithms are specifically tailored to the tasks at hand. In this way we have achieved strongly cross validated diagnostic efficiencies of 85% which have only previously been available with sophisticated imaging equipment, for colon cancer [15], [39].

One problem with our method is that cancer only develops in the *glandular* cells. However we use both glandular and lamina propria cell nuclei in training the cancer recognition algorithms (local discriminant bases and partial least squares). The reason for this is that we have thus far been unable to tell the difference between these two populations of cell nuclei, automatically. In the future we hope to achieve this discrimination using immunoperoxidase markers for epithelium and leukocytes. These stains will facilitate the task of automatically differentiating between lamina propria nuclei and glandular nuclei.

ACKNOWLEDGMENTS

We thank Dr. Boaz Nadler for interesting discussions on the application of the PLSR method.

Partially funded by National Science Foundation: NSF DMS-01399; Defense Advanced Research Projects Agency

(DARPA) and Air Force Office of Scientific Research (AFOSR). MM is grateful for partial support from NSF DMS Grant 0512050.

Results presented in part at Optical Biopsy VI, January 2006, San Jose, CA and the Fifth Inter-Institute Workshop on Optical Diagnostic Imaging from Bench to Bedside at the National Institutes Of Health, September 2006, Bethesda, MD.

REFERENCES

[1] D. Woernley, "Ir absorption curves for normal and neoplastic tissues and related biological substances," *Cancer Res.*, vol. 12, pp. 516–523, 1952.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2001.

[3] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 2, pp. 111–147, 1974.

[4] L. Breiman and P. Spector, "Submodel Selection and Evaluation in Regression. The X-Random Case," *International Statistical Review/Revue Internationale de Statistique*, vol. 60, no. 3, pp. 291–319, 1992.

[5] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, vol. 2, pp. 1137–1145, 1995.

[6] A. Molckovsky, L. Song, M. Shim, N. Marcon, and B. Wilson, "Diagnostic potential of near-infrared raman spectroscopy in the colon: Differentiating adenomatous from hyperplastic polyps," *Gastrointest. Endosc.*, vol. 57, no. 3, pp. 396–402, 2003.

[7] K. Rajpoot, N. Rajpoot, and M. Turner, "Hyperspectral colon tissue cell classification," in *Proc. SPIE Medical Imaging (MI'04)*, San Diego, USA, Feb. 2004.

[8] K. Rajpoot and N. Rajpoot, "Svm optimization for hyperspectral colon tissue cell classification," in *Medical Image Computing and Computer Assisted Intervention (MICCAI'04)*, St-Malo, France, Sept. 2004.

[9] G. Davis, M. Maggioni, F. Warner, and F. Geshwind, "Hyper-spectral analysis of normal and malignant colon tissue microarray sections using a novel dmd system," in *Proc. Fourth inter-institute workshop on optical diagnosis from bench to bedside at the national institutes of health*, Bethesda, MD, Sept. 2004.

[10] I. Bigio, S. Brown, M. Briggs, C. Kelley, S. Lakhani, D. Pickard, P. Ripley, I. Rose, and C. Saunders, "Diagnosis of breast cancer using elastic-scattering spectroscopy: preliminary clinical results," *J. Biomedical Optics*, vol. 5, pp. 221–228.

[11] X. Qi, M. Sivak, G. Isenberg, J. Willis, and A. Rollins, "Computer-aided diagnosis of dysplasia in barretts esophagus using endoscopic optical coherence tomography," *Journal of biomedical optics*, vol. 11, 2006.

[12] Y. Qu, B. Adam, Y. Yasui, M. Ward, L. Cazares, P. Schellhammer, Z. Feng, O. Semmes, and G. Wright, "Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients," *clinical chemistry*, vol. 48, pp. 1835–1843, 2002.

[13] E. Petricoin, A. Ardekani, B. Hitt, P. Levine, V. Fusaro, S. Steinberg, G. Mills, C. Simone, D. Fishman, E. Kohn, and L. Liotta, "Use of proteomic patterns in serum to identify ovarian cancer," *The Lancet*, vol. 359, no. 9306, pp. 572–577, 2002.

[14] K. Masood, N. Rajpoot, H. Qureshi, and K. Rajpoot, "Co-occurrence and morphological analysis for colon tissue biopsy classification," in *Proceedings 4th International Workshop on Frontiers of Information Technology (FIT'06)*, Islamabad, Pakistan, Dec. 2006.

[15] Z. Ge, K. Schonmacker, and N. Nishioka, "Identification of colonic dysplasia and neoplasia by diffuse reflectance spectroscopy and pattern recognition techniques," *Applied spectroscopy*, vol. 52, no. 6, pp. 833–839, 1998.

[16] S. Argov, J. Ramesh, A. Salman, I. Sinelnikov, J. Goldstein, H. Guterman, and S. Mordechai, "Diagnostic potential of Fourier-transform infrared microspectroscopy and advanced computational methods in colon cancer patients," *Journal of Biomedical Optics*, vol. 7, no. 2, pp. 248–254, 2002.

[17] D. Dicker, J. Lerner, P. Vanbelle, S. Barth, , D. Guerry, M. Herlyn, D. Elder, and W. Eldeiry, "Differentiation of normal skin and melanoma using high resolution hyperspectral imaging," *cancer biology and therapy*, vol. 5, 2006.

[18] M. Astion and P. Wilding, "Application of neural networks to the interpretation of laboratory data in cancer diagnosis," *Clinical Chemistry*, vol. 38, no. 1, pp. 34–38, 1992.

[19] M. Mehrubeoglu, N. Kehtarnavaz, G. Marquez, M. Duvic, and L. Wang, "Skin lesion classification using oblique-incidence diffuse reflectance spectroscopic imaging," *applied optics*, vol. 41, pp. 182–192, 2002.

[20] R. R. Coifman, N. Saito, F. B. Geshwind, and F. Warner, "Discriminant feature extraction using empirical probability density estimation and a local basis library," *Pattern Recognition*, 2002.

[21] R. R. Coifman and N. Saito, "Constructions of local orthonormal bases for classification and regression," *C. R. Acad. Sci. Paris*, vol. 319 Série I, pp. 191–196, 1994.

[22] N. Saito, "Local feature extraction and its applications using a library of bases," Ph.D. dissertation, Yale Mathematics Department, May 1994.

[23] P. Hsu and Y. Tseng, "Feature extraction of hyperspectral data using the best wavelet packet basis," *Geoscience and Remote Sensing Symposium, 2002. IGARSS'02. 2002 IEEE International*, vol. 3, pp. 1667– 1669, 2002.

[24] S. Kumar, J. Ghosh, and M. Crawford, "Best-bases feature extraction algorithms for classification ofhyperspectral data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 39, no. 7, pp. 1368–1379, 2001.

[25] D. Fradkin and D. Madigan, "Experiments with random projections for machine learning," *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 517–522, 2003.

[26] S. Vempala, *The Random Projection Method*. American Mathematical Society, 2004.

[27] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[28] A. Blum, "Random projection, margins, kernels and feature selection," in *Proc. SLSFS 2005*, Bohinj, Slovenia, Feb. 2005.

[29] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 6, no. 15, pp. 1373–1396, June 2003.

[30] R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, pp. 5–30, 2006.

[31] B. Nadler, S. Lafon, R. Coifman, and I. Kevrekidis, "Diffusion maps, spectral clustering and reaction coordinates of dynamical systems," *Applied and Computational Harmonic Analysis*, vol. 21, pp. 113–127, 2006.

[32] R. R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data. Part I: Diffusion maps," *Proc. of Nat. Acad. Sci.*, no. 102, pp. 7426–7431, May 2005.

[33] ——, "Geometric diffusions as a tool for harmonic analysis and structure definition of data. Part II: Multiscale methods," *Proc. of Nat. Acad. Sci.*, no. 102, pp. 7432–7438, May 2005.

[34] M. Maggioni, F. J. Warner, G. L. Davis, R. R. Coifman, F. B. Geshwind, A. C. Coppi, and R. A. DeVerse, "Algorithms from signal and data processing applied to hyperspectral analysis: Application to discriminating normal and malignant microarray colon tissue sections," Yale University, Dept. Comp. Sci., Tech. Rep. 1311, Feb 2004.

[35] R. DeVerse, R. Coifman, A. Coppi, W. Fateley, F. Geshwind, R. Hammaker, S. Valenti, F. Warner, and G. Davis, "Application of spatial light modulators for new modalities in spectrometry and imaging," *Proc. SPIE*, vol. 4959, pp. 12–22, 2003.

[36] N. Boustany, J. Crawford, R. Manoharan, R. Dasari, and M. Feld, "Analysis of nucleotides and aromatic amino acids in normal and neoplastic colon mucosa by ultraviolet resonance raman spectroscopy," *Lab Invest*, vol. 79, pp. 1201–1214, 1999.

[37] B. Rigas, S. Morgello, I. Goldman, and P. Wong, "Human colorectal cancers display abnormal fourier-transform infrared spectra," *Proc. Natl. Acad. Sci.*, vol. 87, no. 20, pp. 8140–8144, 1990.

[38] D. Fernandez, R. Bhargava, S. Hewitt, and I. Levin, "Infrared spectroscopic imaging for histopathologic recognition," *Nature biotechnology*, vol. 23, pp. 469–474, 2005.

[39] X. Li, X. Li, M. Lei, D. Wang, and J. Lin, "Detection of colon cancer by laser induced fluorescence and raman spectroscopy," in *Proc. IEEE Engineering in medicine and biology annual conference*, Shanghai, China, Sept. 2005, pp. 6961–6964.

[40] http://tissuearray.org/yale/index.html.

[41] S. Dasgupta and A. Gupta, "An elementary proof of a theorem of Johnson and Lindenstrauss," *Random Structures and Algorithms*, vol. 22, no. 1, pp. 60–65, 2003.

[42] Harwit, *Hadamard transform optics*. Ed. New York: Academic, 1979.

[43] S. Lafon, "Diffusion maps and geometric harmonics," Ph.D. dissertation, Yale Univ., New Haven, May 2004. [Online]. Available: http://www.math.yale.edu/~sl349/publications/publications.htm

[44] C. K. I. Williams and M. Seeger, "Using the nyström method to speed up kernel machines." in *NIPS*, 2000, pp. 682–688.
[45] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the nyström method," *IEEE PAMI*, Feb 2004.
[46] J. C. Platt, "FastMap, MetricMap, and Landmark MDS are all Nyström algorithms," Microsoft Research, Tech. Rep. MSR-TR-2004-26, Sep 2004.
[47] H. Martens and T. Naes, *Multivariate Calibration*. Wiley, 1988.
[48] T. Naes, T. Isaksson, T. Fearn, and T. Davies, *User-friendly Guide to Multivariate Calibration and Classification*. NIR Publications, 2002.
[49] S. Wold, M. Sjostrom, and L. Eriksson, "Pls-regression: a basic tool of chemometrics," *Chem. Int. Lab. Sys.*, vol. 58, pp. 109–130, 2001.
[50] D. Haaland and E. Thomas, "Partial least-squares methods for spectral analysis. 1.relation to other quantitative calibration methods and the extraction of qualitative information," *Anal. Chem.*, vol. 60, pp. 211–228, 1988.
[51] A. Hoskuldsson, "Pls regression methods," *J. Chem.*, vol. 2, pp. 211–228, 1988.
[52] http://www.tidalphotonics.com.
[53] N. MacKinnon, U. Stange, P. Lane, C. MacAulay, and M. Quatrevalet, "Spectrally programmable light engine for in vitro or in vivo molecular imaging and spectroscopy," *Applied Optics*, vol. 44, 2005.
[54] http://www.cri-inc.com.
[55] C. Angeletti, N. Harvey, V. Khomitch, A. Fischer, R. Levenson, and D. Rimm, "Detection of malignancy in cytology specimens using spectral-spatial analysis," *Laboratory Investigation*, vol. 85, pp. 1555–1564, 2005.
[56] http://www.spectraldimensions.com.