Algorithm for Classification of Fluorescence Spectra of Caner and Normal Tissue

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Philosophy in Electrical and Electronic Engineering

The Hong Kong University of Science and Technology

Hong Kong

by Chang, Han Peng

Department of Electrical and Electronic Engineering B. Eng. in Electronic Engineering

The Hong Kong University of Science and Technology

August 2001
Algorithm for Classification of Fluorescence Spectra of

Caner and Normal Tissue

by

Chang, Han Peng

Approved by:

Dr. Jianan Qu
Thesis Supervisor

Dr. Ho-chi Huang
Thesis Examination Committee Member (Chairman)

Dr. Li-xin Wang
Thesis Examination Committee Member

Prof. Philip C. H. Chan
Head of Department

Department of Electrical and Electronic Engineering

The Hong Kong University of Science and Technology

August 2001
Acknowledgements

Firstly, I wish to express my gratitude to my research supervisor Dr Jianan Qu. I would like to thank him for his invaluable guidance, enormous support, and encouragement throughout this project. He has shared his valuable experience and his intellectual maturity with me. His advice has been particularly helpful in stimulating and guiding this research project in the right direction.

Secondly, I would like to sincerely thank all the research group members who have participated in this project. I would like to extend my appreciation to Mr Bruce Lee and Dr Huang Zhi-Jian, who collected the patient light induced autofluorescence tissue spectral data over three months from patients at the Queen Mary Hospital. Without their help, I could not have completed this analysis smoothly.

Thirdly, I warmly thank all my colleagues at the Biomedical Research Laboratory of the Hong Kong University of Science and Technology. Their kindly help and cooperation made the working environment peaceful, enjoyable and memorable.

Last, but not least, I would like to thank my parents for their love and support.
Algorithm for Classification of Fluorescence Spectra

of Caner and Normal Tissue

by
Chang, Han Peng
for the degree of
Master of Philosophy in Electrical and Electronic Engineering
at The Hong Kong University of Science and Technology
in August 2001

Abstract

To detect the small lesion and identify the margin of observable tumors, in vivo, the potential of light-induced fluorescence (LIF) spectroscopic imaging was evaluated to improve the accuracy of conventional white light endoscopy. In this study, further investigations were carried out using a general multivariate spectral classification algorithm.

A conventional endoscopic system with a multiple channel spectrometer was used to measure the autofluorescence of nasopharyngeal tissue in vivo. Classification was based on the spectral difference between the carcinoma and normal tissue. A sophisticated algorithm based on Principal Component Analysis (PCA) was developed to differentiate between the nasopharyngeal carcinoma from the normal tissue. Firstly, preprocessing was done to reduce noise and to calibrate the different measurement distances and geometry about which there was no prior information. Secondly, processing using Principal Component Analysis was done to effectively reduce the variable dimensions while maintaining useful information for analysis. Thirdly, various post-processing techniques were investigated and the classification performance was compared. Algorithms using the ratio of auto-fluorescence spectra
intensity at two wavelength and three wavelength bands were applied for diagnostic performance comparisons[1]. In addition, the robustness of the PCA based algorithm in noisy environment was investigated. Finally, the possibility of applying of the optical processing techniques based on the PCA-based algorithm using optical filter in real time diagnosis were analyzed.

Without prior knowledge of tissue optics and blood absorption characteristics, the application of the PCA-based method gives significantly better diagnostic performance than the previous two-wavelength and three-wavelength algorithms. Based on the entire spectra, the two-wavelength ratio algorithm gives a sensitivity of 88% and a specificity of 92% in respect to the detection of nasopharyngeal carcinomas. A sensitivity of 92% and specificity of 96% are achieved by the PCA-based algorithm. For the three-wavelength algorithm, a sensitivity of 88% and specificity of 95% are achieved. Also, the PCA-based algorithm shows less than 1% degradation in specificity with 5% additional random noise on the collected spectra. In addition, the investigation into the application of the algorithm using optical filters for real-time diagnosis showed that the diagnostic performance degradation is insignificant compared to the theoretical results.

In conclusion, the PCA based multivariate statistical algorithm is an efficient way to improve the spectral classification performance of nasopharyngeal carcinoma. The results of which can be read by non-experts. The algorithm is also robust in respect to spectra collected in a noisy environment and it is feasible to implement the algorithm using optical processing techniques for real time diagnosis.
# Table of Contents

Title Page  
Signature Page  
Acknowledgements  
Abstract  
Table of Contents  
List of Figures  
List of Table  

Chapter 1: Introduction  
1.1 Motivation  
1.2 Problems  

Chapter 2: Light Induced Autofluorescence Spectroscopy  
2.1 Principles of Fluorescence  
2.1.1 Phenomenon of Fluorescence  
2.1.2 Characteristics of Fluorescence Emission Spectrum  
2.2 Biological Fluorophores  
2.3 Application to Characterization of Tissue Pathology  
2.3.1 Motivation  
2.3.3 Limitations  

Chapter 3: Instrumentation and Method of Measurement  
3.1 Instrumentation  
3.1.1 Light Source  
3.1.2 Detector  
3.1.3 Spectral and Spatial Information  
3.2 Method of Measurement  

Chapter 4: Application of Two-wavelength and Three-wavelength Detection Algorithm based on the observed Nasopharyngeal Carcinoma (NPC) Spectra  
4.1 Introduction  
4.2 Nasopharyngeal Carcinoma Tissue Spectra at 390nm excitation  
4.3 Two-wavelength and Three-wavelength Detection Algorithms  
4.3.1 Methods  
4.3.2 Prospective Evaluation of Diagnostic Performance  
4.3.3 Results and Discussions  

Chapter 5: Development of Multivariate Statistical Algorithm based on Principal Component Analysis for in vivo NPC Tissue Diagnosis  
5.1 Introduction  
5.2 Multivariate Statistical Algorithms Development – PCA  
5.2.1 Motivation  
5.2.2 Preprocessing  
5.2.3 Processing  
5.2.4 Post-Processing  
5.2.5 Choice of Different Principal Components  
5.2.6 Prospective Evaluation of Diagnostic Performance
5.3 Results and Discussions ........................................ 32
5.4 Overall Comparison on Different Algorithms .......... 45
5.5 Robustness of the PCA based Detection Algorithm on Noisy Data .... 48

Chapter 6: Investigation on the Possibility of applying the Algorithm using Optical Processing Technique for real time clinical diagnosis
6.1 Introduction .................................................. 51
6.2 Investigation of the Optical PC Filters ................. 52
6.3 Diagnostic Performance Analysis of the Optical Processing System ...... 54
6.4 Noise Robustness of the Optical Processing System ........ 55
6.5 Conclusions ................................................. 58

Chapter 7: Summary, Conclusions and Future Directions 59
References ................................................................... 61
Appendix A .......................................................... 65
Appendix B .......................................................... 67
List of Figures

Figure 1. Jablonski diagram

Figure 2. *In vivo* nasopharyngeal tissue autofluorescence measuring system

Figure 3(a). Typical LIF emission spectra of five different subjects of normal tissue

Figure 3(b). Typical LIF emission spectra of five different subjects of carcinoma tissue

Figure 4. Two-wavelength and three-wavelength algorithm diagnostic performance

Figure 5 (a). Preprocessed LIF emission spectra of five different subjects of normal tissue correspond to Figure 1(a)

Figure 5(b). Preprocessed LIF emission spectra of five different subjects of Carcinoma tissue corresponding to Figure 1(b)

Figure 6. SCREE Plot to illustrate the variation explained by first five principal components

Figure 7. The first four principal component loading values

Figure 8. Scatter plot of principal component score 2 verse score 1

Figure 9. Principal component score threshold based algorithm diagnostic accuracy

Figure 10. Principal component 1,2 score distribution for normal and carcinoma tissue using Silverman density estimation

Figure 11. Principal component 1-4 score distribution approximates to normality

Figure 12. Posterior probability of a given tissue belong to carcinoma category based on principal component 1,2

Figure 13. PCA logistic discrimination diagnosis accuracy

Figure 14. Performance comparison of PCA based algorithms and two-wavelength ratio algorithm

Figure 15. Typical spectral data with different amount of gaussian random noise

Figure 16. PC 1-2 diagnostic line algorithm diagnostic accuracy on noisy spectra

Figure 17. PC 1-2 logistic discrimination algorithm diagnostic accuracy on noisy spectra
Figure 18. Transmissions of PC1 and PC2 Filters

Figure 19. Diagnostic performance of PCA diagnostic line algorithm using optical processing on noisy fluorescence spectra

Figure 20. Diagnostic performance of PCA logistic discrimination algorithm using optical processing on noisy fluorescence spectra
List of Table

<table>
<thead>
<tr>
<th>Table 1.</th>
<th>Performance of PCA diagnostic line algorithms using optical processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.</td>
<td>Performance of PCA logistic discrimination algorithms using optical processing</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Motivation

The biomedical application of fluorescence-based techniques is increasing. The main reason for this is that non-invasive fluorescence imaging has the potential to provide *in vivo* diagnostic information in many clinical specialities, especially in oncological applications.

The current gold standard for early cancer detection is a biopsy following a white-light endoscopic examination [2,3]. However, the random nature of the tissue removal process often leads to misdiagnosis. In order to improve the diagnostic accuracy and to reduce the unnecessary suffering of patients from the random biopsy, a non-invasive light induced fluorescence (LIF) spectroscopic remote imaging technique, which measures LIF spectra at multiple tissue sites simultaneously, was developed for the collection of LIF spectra *in vivo*. The measured LIF spectra provide valuable optical characteristics of human tissue including absorption, scattering, and fluorescence. These optical characteristics shown in the spectra can reveal the biochemical and morphological changes that occur as tissue becomes neoplastic. Although a complete quantitative analysis of the biochemical composition of various kinds of tissue has not been achieved, this *in vivo* fluorescence spectral information can be correlated to tissue histopathology. Thus, LIF has the potential to improve the differentiation between diseased tissue and non-diseased tissue. Comprehensive reviews of the characterization of tissue studied using LIF spectroscopy have been
provided by Wagnieres et al.[2] and Ramanujam[3]. The review by Richards-Kortum and Sevick-Muraca presented the biomedical application of fluorescence; the optical interactions of tissue was extensively described[4].

There are several major advantages of using this LIF spectroscopy for clinical screening and diagnosis that are worth emphasizing. The remote imaging technique requires no expert knowledge of the characteristics of the endoscopic images of tumors for immediate treatment because the collected spectral data can be processed by computer pre-programmed mathematical algorithm for tissue identification. Also, as the LIF spectra contain more diagnostic information than white light endoscopic images, higher diagnostic and screening accuracy is expected. The neoplastic tissue is sometimes invisible to the conventional endoscopy. Further, the high dimension collected spectral data allows multivariate-based statistical algorithms to be applied to the calibration data set and prediction data set so as to obtain optimal decision criteria and perform related performance analysis. Moreover, it is a non-invasive approach that can effectively reduce the likelihood of unnecessary trauma caused by random biopsies. In addition, this remote sensing technique with multivariate statistical algorithms could be applicable to the classification of other kinds of tissues.

1.2 Problems

LIF imaging systems for the early detection of malignant tumors based on the differences in fluorescence spectral characteristics between diseased and non-diseased tissue have been developed by several other groups [5-8]. The diseased sites of various human organs have been evaluated by other groups using various spectral classification algorithms, in which multivariate statistical algorithms were common
discrimination strategies with high sensitivity and specificity [9-21]. These results show that measuring fluorescence spectra with statistical algorithm is an effective way to classify normal and abnormal tissue.

This research focused on the detection of nasopharyngeal carcinomas (NPC) and the surrounding normal tissue based on LIF spectra. This research is significant because it is difficult to localize and detect small NPC lesions and to identify the margin of observable tumors when using conventional white light endoscopy. NPCs have a high occurrence in Asia but they are often misdiagnosed because of the difficulty of examining the nasopharynx. This study involved the application of classification algorithms to process the spectra, and to carry out performance analysis and performance comparisons. In previous study, the algorithms that were developed were based on an observation of LIF characteristics. There is a prominent difference in the ratio of fluorescence signal at two wavelength bands which result in high sensitivity and specificity [1]. With the compensation of tissue blood absorption characteristics that reflect the fluorescence signals, the algorithm was extended using a three wavelength bands ratio. In this study, the possibility of applying multivariate statistical analysis to the measured tissue spectra was investigated. In vivo LIF spectra from 85 carcinoma tissue sites and 131 normal tissue sites from a total of 59 patients were collected and used for analysis. After the known normal and abnormal tissue sample spectra were collected from the patient, various pre-processing, processing, and post-processing procedures were carried out in order to retrieve the meaningful diagnostic information. Preprocessing involved normalization to its area to calibrate the errors caused by different measuring distances and geometry. This, thus, reduced the intra-patient and inter-patient variation within a tissue category. Also, the random
noise in the raw signals was smoothed out by adjacent averaging. Processing included the application of a general multivariate statistical algorithm to the tissue spectra. Principal component analysis was used to concentrate the high dimensional spectral data into a few dimensions that captured the most diagnostic information for further analysis. Post-processing included the use of the resulting processed scores to optimize the diagnostic accuracy of normal and carcinoma tissue. The detailed procedures and the results of the algorithms are presented in the next few chapters.

The primary goal of this study was to investigate the algorithms using Principal Component Analysis (PCA) on the in vivo autofluorescence spectra for the detection of nasopharyngeal carcinoma so as to achieve high sensitivity and specificity. This study also evaluated the improvement in the diagnostic accuracy of the algorithm based on wavelength bands ratio to differentiate between nasopharyngeal carcinoma tissue and surrounding normal tissue. Various types of statistical analysis were carried out and the results were compared. In addition, the robustness of the PCA based algorithm in noisy environment was investigated. Finally, the possibility of applying of the processing technique implemented using optical filters for real time diagnosis was analyzed.

This study focused on the application of tissue detection technique on LIF spectra of NPC. This thesis is organized as follows. In Chapter 2, the background of light induced autofluoroscence is presented. The spectroscopic measurement instrument and method are given in Chapter 3. Based on the observed spectra, two-wavelength and three-wavelength algorithms are presented in Chapter 4, together with the diagnostic performance evaluation. In Chapter 5, the application of spectral
multivariate statistical classification algorithm on the spectra is presented. The detailed procedures, the various processing stages to develop the whole diagnostic algorithm, and the prospective evaluation result are discussed. The robustness of the algorithm over random noisy spectral signals is also reported. In Chapter 6, the possibility of applying the algorithm in practical implementations using optical filters is presented. The Conclusion and future directions are discussed in Chapter 7.
Chapter 2

Light Induced Autofluorescence Spectroscopy

The aim of this chapter is to present an overview of the fluorescence background information in order to give an understanding of the potential of light induced autofluorescence spectroscopy for neoplastic tissue diagnosis. An excellent review can be found in [22].

2.1 Principles of Fluorescence

2.1.1 Phenomenon of Fluorescence

The phenomenon of fluorescence was first successfully observed by Stokes in 1852. The issue that fluorescence has the potential to be used for diagnostic proposes was recognized much later by Stuebel in 1911. Since then, much effort has been put on investigating the origin of this tissue fluorescent property.

Luminescence is the emission of light from any substance, and it occurs in electronically excited states. Fluorescence is the luminescence that results from the return of the lower orbital from the singlet excited orbital of the electron. The electron in the singlet excited orbital is paired (of opposite spin) to the second electron in the ground-state orbital. The emission rates of fluorescence are typically $10^8 \text{s}^{-1}$, so that a typical fluorescence lifetime is near $10 \text{ns}$ ($10 \times 10^{-9} \text{s}$). The absorption and emission of light by a molecule can be illustrated by the Jablonski diagram as shown in Figure 1 [22]. The singlet ground, first and second electronic states are depicted by $S_0$, $S_1$ and $S_2$, respectively. At each of these electronic energy levels the fluorophores exist in a
number of vibrational energy levels.

![Jablonski Diagram](image)

**Figure 1.** Jablonski diagram

Absorption typically occurs from fluorophores with the lowest vibrational energy in which the electrons are activated to some higher excited vibrational level of either $S_1$ or $S_2$. As the excited states are unstable, the excited electrons tend to return to the ground state ($S_0$). This involves internal conversion where the electron rapidly relaxes to the lowest vibration level of $S_1$ and reaches thermal equilibrium. The fluorescence emission occurs where the electrons make a transition to a vibrational level of the ground state. Radiationless transition can also exist.

### 2.1.2 Characteristics of Fluorescence Emission Spectrum

Fluorescence spectral data are generally presented as emission spectra. A fluorescence emission spectrum is a plot of the fluorescence intensity versus wavelength (nanometers) or wavenumber ($cm^{-1}$). Emission spectra vary widely and are dependent upon the chemical structure of the fluorophore and the solvent in which it is
dissolved.

In general, the phenomenon of fluorescence displays several characteristics in the emission spectrum of a particular biological molecule. Firstly, as there is a loss of energy between emission and absorption (see the Jablonski diagram), the fluorescence typically shows a red-shift, i.e., lower energies or longer wavelengths. Secondly, the emission spectra are typically independent of the excitation wavelength because there is a rapid relaxation of the fluorophore upon excitation to higher electronic and vibrational levels. This relaxation results in the dissipation of thermal energy, leaving the fluorophore in the lowest vibrational level of the electronically excited state. Thirdly, the fluorescence emission spectrum is generally a mirror image of its absorption spectrum. The biological fluorescence is characterized by its quantum yield and its lifetime. The quantum yield is the ratio of the number of photons emitted to the number absorbed, while the lifetime is the average time a molecule spends in the excited state. In addition, the biological fluorescence emission spectrum signal is a function of fluorophore concentration, its extinction coefficient (absorption power) at the excitation wavelength, and its quantum yield at the emission wavelength [23].

2.2 Biological Fluorophores

In general, fluorescence can be induced by endogenous fluorophores (intrinsic fluorescence) or exogenous fluorophores (extrinsic fluorescence). Autofluorescence is the natural intrinsic fluorescence from biological tissue fluorophores.

There are several fluorophore characteristics that are important in respect to clinical studies. First, each fluorophore has a distinct excitation and emission spectrum.
Second, a particular type of tissue contains a mixture of many fluorophores of different concentrations. Third, the fluorophores are not uniformly distributed in tissue; they may vary markedly according to the depth below the tissue surface [2].

There is a wide variety of biological endogenous natural fluorophores that display significant fluorescence. [22] Common intrinsic fluorophores includes tryptophan and tyrosine in proteins, the cofactors, NADH, Riboflavin, FMN (flavin mononucleotide), and FAD (flavin adenine dinucleotide). Proteins absorb light near 280nm, and fluorescence emission maxima range from 320 to 350nm. NADH has an absorption and an emission maxima of 340 and 450nm, respectively. Riboflavin and flavin absorb light at around 450nm and emit at around 515nm.

For extrinsic fluorescence where fluorophores are externally induced to the body, HpD, pheophorbide-a, mTHPC and benzoporphyrin derivative (BPD) have been used for clinical applications [2].

In this study, light induced autofluorescence emission spectra from 390nm excitation were used.

2.3 Application to Characterization of Tissue Pathology

2.3.1 Motivation

Recently, the change in the concentration of the tissue fluorophores (e.g., NADH) or the redox state of flavin co-factors (e.g., FAD) have been used as the fluorescence spectral signatures that correlate with tissue pathologies such as cancer. Classification can be based on the difference of the above parameters that are revealed in the
fluorescence spectrum. It should be noted that it is difficult to measure biomechemical molecules quantitatively using the spectra information (that contain emissions from different kinds of fluorophores, tissue scattering and absorption). Therefore, a qualitative analysis is more suitable for tissue diagnosis.

The different general diagnostic strategies applied to classify tissue obtained from the LIF spectroscopy at various tissue sites have been summarized in the recent review by Ramanujam [3]. The techniques developed in past researches that made use of the LIF for tissue diagnosis have been successfully applied in various human organs tissue site, both in vivo or in vitro. Human organs tissue sites including the brain, the head and neck, the esophagus, the bronchus, the breast, the bile duct, the stomach, the colon, the bladder, the cervix and the skin have been investigated [10-21]. Past ENT analysis has focused on the oral cavity and larynx site. The nasopharynx, which is significant, has received minimal attention, and the use of multivariate statistical analysis has not been applied. Most diagnostic performance results show high sensitivity and specificity with significant diagnostic accuracy improvement over the current white light endoscopy. The non-invasive tissue diagnosis reduces the need of biopsy, the current gold standard pathology, and the trauma experienced by patient.

Concerning the classification scheme, methods based on several wavelength band ratios of the observed tissue emission spectra has been commonly used. Several groups have performed statistical analysis using binary or multiple classification schemes. More complicated non-linear multivariate statistical algorithm have also been investigated [3].
2.3.2 Limitations

There are, however, some limitations to the biomedical application of fluorescence spectral signals for extracting diagnostic information. One challenging issue is the tissue optical interaction upon illumination. According to the review by Richards-Kortum et al. [4], the tissue random scattering and absorption of the excitation light sources complicate the emission that make the extraction of the biochemical or diagnostic information more difficult. It should be noted, though, that the scattering and absorption characteristics sometimes provide diagnostic information.

In addition, the differences in tissue collection distance and geometry between samples alter the measured spectra intensity. Although spectral normalization can uncouple part of the influenced signals, this remains an error that cannot be easily resolved through signal processing.

Another limitation inherent in the application of LIF to tissue pathology is the limited sample size of both the non-diseased and the diseased tissues. It is clear that great effort is needed to collect a statistically significant tissue sample size from patients and the available number of patients is usually limited to reveal the true population. Quite often, there are fixed statistical inaccuracies because of the limited sample size.
Chapter 3

Instrumentation and Method of Measurement

3.1 Instrumentation

The *in vivo* autofluorescence spectra measuring device consists of a conventional white light nasal endoscope with a multiple channel spectrometer. The schematics of the arrangement are illustrated in Figure 2.

![Diagram of instrument setup](image)

**Figure 2.** *In vivo* nasopharyngeal tissue autofluorescence measuring system

The important parts of the system are the light source, detector, the spectral and spatial information.
3.1.1 LIGHT SOURCES

Excitation source:

Tissue sites were illuminated by a 100W mercury arc lamp.

Band pass filter:

The excitation lamp consists of a filter with a band pass wavelength 390-450nm.

3.1.2 DETECTORS

Endoscope:

A commercial white light nasal endoscope was used. At the endoscope tip, the excitation power is approximately 50mW.

Dichroic mirror:

This was used to filter out signal wavelengths below 470nm as these contain no diagnostic information. It also divided the optical signal from the endoscope into the reflection and fluorescence.

Fiber bundle:

This consists of seven optical fibers (200um in diameter) that deliver the fluorescence signal to the image spectrograph. The fibers were evenly distributed in fluorescence image plane of the endoscope as shown in Figure 2.

CCD/ICCD:

The charge coupled device (CCD) camera was used to record the reflection of excitation from the tissue surface, and the intensified CCD camera was used to image
the fluorescence.

### 3.1.3 SPECTRAL AND SPATIAL INFORMATION

**Image spectrograph:**

This was used to receive the fluorescence signal from the fiber and to convert the signals to spectra information.

**Frame grabber:**

This obtained the real time ICCD images at a rate of 25 frame per second.

**Computer:**

This stored the spectra data permanently.

### 3.2 Method of Measurement

Using the above apparatus, tissue spectral signals from multiple points can be collected each time. In general, there are two types of measuring systems, point and image spectrofluorometry. In the point measurement, the entire tissue fluorescence spectrum is recorded for different desired excitation wavelengths. In general, the restriction to a single wavelength seriously limits the reliability of fluorescence diagnosis. When using the imaging measurement technique, the multiple excitation-emission wavelength pairs are usually limited. When the tissue surface is imaged, usually only a small number of combinations of excitation and emission wavelengths from a relatively large tissue surface are recorded. Their major diagnostic information extraction strategies are quite different. Point measurement provides more
spectroscopic information so diagnostic techniques based on spectral different are more important. Also, it is more efficient if the tissue surface under investigation is small. Another advantage of point measurement is that it often possesses higher signal sensitivity. In image measurement, more spatial information can be collected. The image analysis techniques, which are based on pixel-by-pixel comparisons or other spatial domain analyses, are more useful for extracting suitable diagnostic information, although a combination of spatial and spectral domain analysis is most desirable. The reviews presented in [23] and [24] provide good accounts of fluorescence spectroscopy in that they were focused on point and image measurements, respectively.

*In vivo* autofluorescence spectra were collected from nasopharyngeal tissue sites through a nasal endoscopic imaging system at the Queen Mary Hospital, University of Hong Kong. The excitation light source wavelength was 390nm and samples of the emission spectra at wavelength range from approximately 430nm to 680nm were measured with the multi-channel imaging spectrograph. Each emission spectra signal is separated from its neighbors by a 0.3-0.4nm wavelength. A long pass filter was used to cut off the wavelength below 470nm. The objective of the filter is to minimize the detection of the back-scattered excitation light, which is much stronger than the weaker emitted light. The long-pass filtered fluorescence was then passed to the image spectrograph through an optical fiber bundle with seven optical fibers. The ICCD was used to image the spectra. The images captured by the frame grabber were eventually stored in the computer.

Concerning the illumination and detection optics, as a remote imaging method was
used, and because the tissue surface was irregular, the source-tissue measuring
distance and geometry were not fixed for different measurements. The advantage of
this non-contact type measurement is that it can reduce the distortion in the
fluorescence spectrum due to the changes in the local blood content caused by
pressure on tissue. However, the variable source-tissue distance and geometry affect
the detected signal strength, and the effect, which needs to be ameliorated, can be
partially corrected using spectral normalization.

In the study undertaken, *in vivo* LIF spectra from 85 carcinoma tissue sites and 131
normal tissue sites from a total of 59 patients were collected for analysis, which lasted
approximately 6 months. The raw spectral data was then processed to extract
necessary diagnostic information. This is discussed in Chapter 4 and 5.
Chapter 4

Application of Two-wavelength and Three-wavelength Detection Algorithm based on the observed Nasopharyngeal Carcinoma (NPC) Spectra

4.1 Introduction

In the simplest case in which fluorescence imaging performed at a single excitation wavelength $\lambda_{ex}$ and a single emission wavelength $\lambda_{em}$, the extraction of diagnostic information is simply achieved using the univariate spectral intensity as the diagnostic threshold. In this study, in contrast, the emission spectra contained a series of different wavelengths. This allowed multivariate analysis in order to extract maximal information.

As discussed in Chapter 2, a simple, yet commonly used classification scheme by many previous researchers is the wavelength bands ratio based algorithm. In many cases, significant differences between neoplastic and normal tissue spectra in two or three wavelength bands are observed. These two or three wavelength ratio scores can be used to construct the diagnostic algorithm. An obvious advantage of this method is that it is simple to use, sometimes results excellent sensitivity and specificity. It is also an efficient method if the diagnostic accuracy is already very high or when the spectral signal-to-noise ratio is relatively low, because other algorithms that make use of the fine detailed spectral correlation may not bring about improvement. Therefore, this method is useful as a first step analysis for preliminary study.

However, since only two or three wavelength bands are used, the out-of-band
information, which may be useful for diagnosis, was not captured. Some tissue scattering and absorption characteristics between normal and neoplastic tissue are not obvious in spectra intensity over a large bandwidth. Instead, they are revealed in the fine detailed spectral shape. In addition, the wavelength ratio score may not follow a particular density distribution and this complicates the analysis.

4.2 Nasopharyngeal Carcinoma Tissue Spectra at 390nm Excitation

TYPICAL LIF EMISSION SPECTRA

The spectra was successfully measured and collected. Typical raw LIF spectra collected from the normal tissue and carcinoma are shown in Figure 3.

![Figure 3 (a). Typical LIF Emission Spectra of Five different Normal Tissue Subjects](image-url)
**Figure 3 (b).** Typical LIF Emission Spectra of five different Carcinoma Tissue Subjects

**OBSERVATIONS AND DISCUSSION:**

In general, the normal tissue and carcinoma tissue spectra have their own particular characteristics. As shown in Figure 3, for each observation there is a high intensity peak at around 510nm, the green region, and there is a low intensity trough at around 580nm, the yellow region. There is another peak at around 610nm but the amplitude is smaller than the one at 510nm. However, between the normal and carcinoma tissue, there are not any obvious differentiable differences in intensity and characteristics at a particular wavelength. This results from the fact that for different subjects, the measurement geometry and measurement distances were different. This is the reason that calibration is required (as described in the preprocessing part) to enhance the variation between normal and carcinoma tissue.
Another observation is that the peak at 610nm is much smaller than the peak at the 510nm region for normal tissue spectra, while the peak at 610nm is not so obviously smaller than the peak at the 510nm region for carcinoma tissue spectra in general. Hence, it is natural to use the ratio of these two wavelength bands for classification.

4.3 Two-wavelength and Three-wavelength Detection Algorithms

4.3.1 Methods

In the two-wavelength algorithm, the discrimination between normal and carcinoma tissue was based on the difference in two wavelength bands ratio of the observed spectra. An exhaustive search was used to find the optimal wavelength bands with the highest statistical difference between the normal and the abnormal two-wavelength bands score, using an unpaired one-sided Students’ T-test [26]. (See Appendix A) The exhaustive search was done by adjusting the two wavelength bands, including the center wavelengths and the wavelength bandwidths, and search for the set such that highest t-statistics results. In the three-wavelength algorithm, variations in blood content were taken into account from the fact that the blood appears to have very strong absorption in the wavelength region 530-590nm. The result score was obtained using

\[ R = \frac{I(C1 \pm B1)}{I(C3 \pm B3)} \left[ \frac{I(C1 \pm B1)}{I(C2 \pm B2)} \right]^k \]  \hspace{1cm} (4.1)

where \( C1 \pm B1, C2 \pm B2, C3 \pm B3 \) are the three wavelength band center wavelength and bandwidth. Again, an exhaustive search was done on the normal and abnormal R-score to find the optimal wavelength bands using the Students’ T-test.
4.3.2 Prospective Evaluation of Diagnostic Performance

The usual criteria for diagnostic performance evaluation in clinical studies of tissue fluorescence spectroscopy, i.e. sensitivity, specificity, were used [27].

Sensitivity is defined by

\[
Sensitivity = \frac{True \ Positive}{True \ Positive + False \ Negative} \tag{4.2}
\]

Specificity is defined by

\[
Specificity = \frac{True \ Negative}{True \ Negative + False \ Positive} \tag{4.3}
\]

Accuracy is defined by

\[
Accuracy = \frac{True \ Positive + True \ Negative}{True \ Negative + True \ Negative + False \ Positive + False \ Negative} \tag{4.4}
\]

In clinical applications, sensitivity is the relatively more important parameter because misclassifying a diseased patient as non-diseased is most undesirable.

To compare the diagnostic accuracy of the different algorithms, the highest possible specificities, given the particular sensitivities, were used. For the two-wavelength and three-wavelength algorithm, a calibration set was used to calculate the different diagnostic score thresholds of the desired sensitivities. The thresholds were then used for prediction set evaluation.

Normally, data is randomly divided into roughly half for calibration and half for prediction, as was the case in [9]. However, this method requires a large sample size.
In the absence of a large sample size (less than 100 carcinoma tissue samples in our case), cross-validation was suggested in which one observation is held out from the data set each time [28]. The remaining data is used to construct and optimize the algorithm. The held-out observation is then used to evaluate the algorithm. This process is repeated until all the held-out observations are used.

4.3.3 Results and Discussions

RESULTS

After the exhaustive search, the optimal wavelength bands were found to be 505 ± 20nm and 605 ± 40nm for the two-wavelength algorithm. For the three-wavelength algorithm, the R-function was obtained as:

\[ R = \frac{I(510 \pm 20)}{I(610 \pm 40)} \left[ \frac{I(510 \pm 20)}{I(550 \pm 20)} \right]^{0.35} \]  \hspace{1cm} (4.5)

The resulting diagnostic accuracy of the two-wavelength and three-wavelength algorithms are shown in Figure 4.
**Figure 4.** Two-wavelength and Three-wavelength algorithm Diagnostic Performance

**OBSERVATIONS AND DISCUSSION**

In general, the three-wavelength algorithm has a significantly higher specificity than the two-wavelength algorithm when its sensitivity is between 75\% and 90\%. In the other range of sensitivity, both the three-wavelength algorithm and two-wavelength algorithm give similar diagnostic accuracy.
Chapter 5

Development of Multivariate Statistical Algorithm based on Principal Component Analysis for \textit{in vivo} NPC tissue diagnosis

5.1 Introduction

As pointed out in the previous chapter, although the use of the method based on the wavelength-band ratio gives a satisfactory accuracy in regard to the detection of NPC tissue, it also results in the disposal of out-of-band information. This may result in the loss of important diagnostic information including tissue scattering, hemoglobin absorption, and characteristics of the fluorophores.

As discussed in a later part of this chapter, the wavelength ratio scores do not fit the normal distribution well, so using the students’ T-test to search for the wavelength bands is not optimal. An algorithm based on normally distributed scores is desirable.

In this chapter, sophisticated multivariate statistical based algorithms used for tissue detection are presented together with evaluation results. The first goal of the analysis presented in this chapter is to investigate a systematic multivariate algorithm, which utilizes the whole spectra information. The objective of the algorithm is to improve diagnostic accuracy and, at the same time, to be simple for real time diagnosis. The second goal is to seek the possibility of the algorithm to detect the presence of carcinoma tissue as well as its level of certainty in regard to the normal or abnormal tissue. The third goal is to investigate the diagnostic performance degradation under noisy signal environments.
5.2 Multivariate Statistical Algorithms Development – PCA

5.2.1 Motivation

If the tissue spectral wavelength is represented as a dimension of the tissue samples, multivariate statistical analysis can be utilized for diagnosis. The reason is clear that the spectra can, qualitatively, reveal the biochemical composition. The multivariate statistical techniques were applied successfully in the past, as described in Chapter 2. According to O’Brien et al. [9], seven spectral classification algorithms are commonly used. These include four linear algorithms, including multivariate linear regression, stepwise multivariate linear regression, principal component analysis (PCA), and decision plane analysis. Other non-linear classification algorithms include the one using Bayes decision theory, and two based on simple features of the spectral shape, that is, the principal peak ratio and spectral width. In the analysis of [9], the principal component analysis based algorithm, of all the suggested methods, gave the highest validation set accuracy of all the suggested methods. It is suitable in the absence of a large sample size, in contrast to linear regression where only the training set is optimized and gives high training set accuracy, but it requires a large sample size to achieve a high prediction set accuracy. Also, PCA is flexible in choosing the number of reduced sets of variables for further analysis. Further, Ramanujam N. et al. [20] have successfully applied PCA in cervical tissue classification. Hence, PCA was the natural choice for our statistical analysis. A good account of applied multivariate analysis can be found in [28]. A comprehensive review, with extensive case studies in which PCA was used, is given by J. Edward Jackson [29] and a good account of the use of PCA based on multivariate images is given by P. Geladi and H. Grahn [30].
The development of the PCA based screen and diagnostic algorithm included that of PC score-based algorithms. These are in comparison to the previous wavelength bands ratio based algorithm. The algorithms involved preprocessing, processing and postprocessing parts.

5.2.2 Preprocessing

Preprocessing includes normalization to calibrate for the normal and abnormal tissue fluorescence spectral so that the greatest diagnostic information can be obtained for further processing and analysis. Normalization can be done based on the maximum fluorescence intensity or based on the area. Normalization based on the maximum fluorescence intensity involves dividing the intensity of the tissue spectra by the maximal intensity of the spectra. Normalization by area involves scaling the spectral intensity while keeping the total area of the resulting tissue spectra the same.

Both methods have their advantages and disadvantages. For instance, normalization by area demonstrates a better noise resistance when there is sudden high amplitude noise in the collected spectrum. Normalization by maximum amplitude gives a better observed spectral intensity amplitude comparison because all the normal and abnormal spectral are relative to the maximum fixed intensity. Normalization by area was done in this study. It should be noted that both methods give exactly the same result for the two-wavelength algorithm.

Another preprocessing technique is smoothing the spectral shape. It is obvious that random noise and sudden noise are caused by changes in the environment and by errors during measurement, as shown in the observed tissue spectral shown in Figure
3. The noise is more significant when the collected spectral signal is weak, as seen in the typical carcinoma tissue spectra in which the signal-to-noise ratio is low. Smoothing can reduce this random noise and the uncorrelated variation caused by the noise. The noise is significant because the objective of the PCA is to capture the variation between observations. The random noise in the raw signals was smoothed out using the adjacent averaging spectral intensity.

In order to simulate the actual diagnostic performance, before the data is processed, we had to consider how the limited tissue spectral data samples were to be categorized as calibration set and prediction set. The calibration set data was trained to form the models or thresholds, while the prediction set data was used to prospectively evaluate the diagnostic accuracy. As discussed in the previous chapter, a cross-validation method was used to prospectively evaluate the diagnostic performance.

5.2.3 Processing

Principal component analysis (PCA) was applied to transform the original spectral variables into new sets of linear combinations of the orthogonal components that explained most of the data variation. The objective was to make use of the reduced set of components, called principal components, to extract the principal difference between normal and carcinoma tissue for diagnostic purposes. A brief overview of the mathematical, data analytical aspects, and how they apply to the data analysis, is given below. Details of the mathematical formulation and the model can be found in [29] and [30].

In contrast to the previous chapter that made use of two wavelength bands or three
wavelength bands with the compensation of blood absorption factor, this study made use of all the wavelengths in the spectra as variables (a total of 211 dimensions), and extracted the maximum data variation, and reduced this to a few dimensions for analysis.

With the preprocessed spectral data, $X_{nxp}$, arranged as follows:

$$
X = \begin{pmatrix}
    x_{11} & x_{12} & \cdots & x_{1p} \\
    x_{21} & x_{22} & \cdots & x_{2p} \\
    \vdots & \vdots & \ddots & \vdots \\
    x_{n1} & x_{n2} & \cdots & x_{np}
\end{pmatrix}
$$

(5.1)

where $n$ is the number of tissue samples and $p$ is the emission wavelength variable, $x_{ij}$ is the intensity of the data at $i$-th subject and $j$-th emission wavelength.

At all the emission wavelengths, the intensities were measured with the same scale. So, it is natural to use covariance matrix to find the principal component without the need of standardization.

The unbiased covariance matrix is obtained by:

$$
Z_{\text{cov}} = \frac{1}{n-1} (X - \overline{X})(X - \overline{X}^\top)
$$

(5.2)

where $\overline{X}$ is the mean of $X$

Next, singular value decomposition is performed on $Z_{\text{cov}}$:

$$
Z_{\text{cov}} = EDE^\top
$$

(5.3)
where \[ E = \begin{pmatrix} \mathbf{e}_{11} & \cdots & \mathbf{e}_{1p} \\ \mathbf{e}_{21} & \cdots & \mathbf{e}_{2p} \\ \vdots & & \vdots \\ \mathbf{e}_{p1} & \cdots & \mathbf{e}_{pp} \end{pmatrix} \] whose \( j \)-th column is the \( j \)-th eigenvector and

\[ D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix} \] with \( \lambda_1 \geq \lambda_2 \geq \cdots \lambda_p \geq 0 \) are the eigenvalues of \( Z_{\text{cov}} \).

Then, the principal component score (PC) is projection of \( X \) on \( E \):

\[ Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = XE \quad (5.4) \]

where \( Y \) is the PC score correspond to the original data which is useful for diagnosis.

The eigenvalue \( \lambda_i \) of \( Z_{\text{cov}} \) account for the variation explained by the transformation \( y_i \). Since \( \lambda_1 \geq \lambda_2 \geq \cdots \lambda_p \geq 0 \), the first PC accounts for the greatest variation among the PCs and the second PC accounts for the greatest variation of the remaining PCs. Quite often, the first two to three PCs account for more than 90% of the total data variation and the remaining PCs mostly contribute to noise. This will be further discussed in the next section.

The proportion of variation explained by the \( i \)-th PC is calculated by:

\[ \frac{\text{Var}(y_i)}{\text{Var}(y_1) + \text{Var}(y_2) + \cdots + \text{Var}(y_p)} = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \quad (5.5) \]

which is useful as one of the criteria to choose the retained PC for analysis.

### 5.2.4 Post-Processing

In the development of the algorithm, the focus was on comparing three algorithms.

1. **PCA with score diagnostic threshold**
As there is a significant difference between the PC score of normal and carcinoma tissue, the classification can be based on the score value. For a single PC score only, the objective was to find the suitable threshold points for diagnosis. For analyzes involving two or three PC scores, the decision lines or planes were estimated, respectively. An exhaustive search was done to find the optimum decision threshold so that the highest sensitivity and specificity could be achieved.

2. PCA with logistic discrimination

The classification criterion is based on the posterior probability of the given observation being normal or carcinoma tissue [28]. The diagnostic accuracy of the algorithm relies on accurate PC score density estimation [31]. The posterior probability of a given observation, \( x_i \), being carcinoma, \( C \), can be estimated using Bayes theory as follows:

\[
P(C | x_i) = \frac{P(x_i | C)P(C)}{P(x_i | C)P(C) + P(x_i | \bar{C})P(\bar{C})}
\]

(5.6)

where \( P(x_i | C) \) is the conditional probability that the carcinoma tissue sample of will have principal component score \( x_i \). \( P(C) \) and \( P(\bar{C}) \) is the prior probability of the tissue being carcinoma and normal respectively in the sample population. The prior probability is an estimation of the likelihood that a sample belongs to a particular group when no information about it is available. If the sample size is considered as the representative of the population, then the prior probability can be estimated using the observed tissue proportion.

One obvious advantage of applying logistic discrimination is that it can increase the number of PCs for analysis without much increase in computation as in the score
based diagnostic line method. Also, the resulting posterior probability, ranging from 0 to 1, can be conveniently used as indicator of how likely the tissue is carcinoma. However, an accurate multivariate PC score joint density estimation is usually difficult to achieve. As the number of PC used increases, the variable dimension increases, and the density estimation accuracy drops [31]. Further, in the first stage of the analysis, the PC score conditional density was assumed to be multivariate normally distributed using the density model in Equation (5.7):

$$f(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} e^{-\frac{(x-u)\Sigma^{-1}(x-u)}{2}}$$

(5.7)

where $u$ and $\Sigma$ are the mean and covariance matrix of the multivariate $p$-dimension variable $x$, respectively. However, this density model may deviate from reality.

### 5.2.5 Choice of Different Principal Components

Before the application of the PCA based algorithm for processing, we had to consider which PCs should be used and how many PCs would be necessary in order to give reasonable diagnostic accuracy. As many of the PCs account mainly for noise, and do not provide diagnostic information, it is reasonable to use only those informative PCs to reduce the diagnostic algorithms complexity. An unpaired Students’ T-test was used to evaluate the difference between normal and carcinoma tissue in the PCs. Also, higher order PCs often account for less than 1% of the total data variation and are very noisy. Therefore, the eigenvalue of the normalized data, which explains the proportion of data variation, was used as the choice for the PCs retained for investigation, together with the Students’ T-test.
5.2.6 Prospective Evaluation of Diagnostic Performance

As stated in Chapter 4, sensitivity and specificity were used as the criteria for evaluation [27]. The highest possible specificities, given the particular sensitivities, were used to compare the diagnostic accuracy of different algorithms. Cross-validation was done to calibrate and to prospectively evaluate the data. For the PCA diagnostic line algorithm, the calibration set was used to calculate the principal loadings and to search for the optimal diagnostic threshold that returned the highest specificities for given sensitivities. Then the prediction set preprocessed data was applied to the principal loading and scores were returned for diagnosis. For the PCA with logistic discrimination, the calibration set was used to estimate the PC score distribution of given tissue types using Equation (5.7), and the prediction set score was applied to the Bayes theorem based Equation (5.6) using the calibrated density function.

5.3 Results and Discussions

PREPROCESSING

RESULTS

As discussed in the previous section, smoothing can remove random noise from the tissue spectra, and normalization can reduce the intra-category variation and, thus, enhance the difference between normal and carcinoma tissue spectra. Typical preprocessed spectra corresponding to Figure 3 are shown in Figure 5.
Figure 5(a). Preprocessed LIF Emission Spectra of Five different Normal Tissue Subjects correspond to Figure 3(a)

Figure 5(b). Preprocessed LIF Emission Spectra of Five different Carcinoma Tissue Subjects correspond to Figure 3(b)
OBSERVATIONS AND DISCUSSION

From Figure 5, it can be seen that the variation within each tissue category (normal and abnormal) is much reduced after preprocessing and the general spectral shape, that contains the information needed for the diagnosis of normal and abnormal tissue, is maintained.

PROCESSING

The proportion of variation explained by the first five principal components are shown in the scree plot in Figure 6. A scree plot is a plot of eigenvalues against the component number [30]. It was introduced by Cattell in [32] also well described by Jackson in [29]. From Figure 6, it can be observed that the first two PCs account for 90% of the total data variation, and the first three PCs account for more than 95% of the total data variation. This result shows that using two or three PCs are enough to extract significant data variation. Higher order PCs mainly account for the random noise that does not contain diagnostic information.
Figure 6. SCREE Plot to illustrate the variation explained by first five Principal Components.

RESULT
The first four principal component loadings obtained from all the spectra data are shown in Figure 7.
Figure 7. The first four Principal Component Loading values.

OBSERVATIONS AND DISCUSSION

The PC loading is significant as the ‘model’ trained by the calibration set, and is commonly used by new observations to calculate the score for diagnosis. Theoretically, the larger the number of training samples, the more accurate the PC loading value and, thus, the more realistic the diagnostic performance. From the PC1 loading shown in Figure 7, it can be observed that there is a relatively high negative
loading value at around the 500-550nm region. The opposite (positive) loading value is shown in the 600-650nm region. This implies that the first principal component mainly captured the difference between the peak of the data spectra between the 500-550nm and the 600-650nm regions. From the PC2 loading, a large positive loading and a negative loading of 500nm and 550nm, respectively, implies that the PC2 captured the difference between the 500nm peak and 550nm trough of the original spectra. Higher order PCs are not significant as they contribute mostly noise. They do not have an obvious loading value to explain.

RESULT

The analysis of the PC score was mainly based on PC1-2 as higher order PCs are not significant for diagnosis. The scatter plot of the PC1 and PC2 scores from a run of cross-validation with the diagnostic line at sensitivity=94% and specificity=97% is shown in Figure 8.
Figure 8. Scatter plot of Principal Component Score 2 verse Score 1.

OBSERVATIONS AND DISCUSSION:

Several observations can be made from the above score plot. Firstly, it can be observed that there is overlapping in a small number of the observations of normal and abnormal tissue. This shows that, if only the PC1 and PC2 scores are used, the two different kinds of tissue cannot be completely classified, although the majority can be distinguished. Secondly, it is observed that the optimal threshold line which gives the highest sensitivity and specificity is not vertical. This implies that PC2 can help diagnosis as well as PC1. Note that the higher order PCs, which mainly account for the random noise, are not shown here.
POSTPROCESSING

1. PCA with score diagnostic threshold

For the PC score based method, the proportion of variations explained by the principal components determine the number of PCs suitable for analysis. The SCREE plot in Figure 6 illustrates the fact that the first two PCs are enough to account for 90% of the data variation, and that the higher order PCs account for less than 5% of the PCs. To choose between different PCs for diagnosis, an unpaired students’ t-test (on the hypothesis that there is an equal mean PC score between normal and carcinoma tissue) was conducted. The p-value of the first two PCs are extremely small (P<<0.001), showing that there is significant statistical difference between normal and carcinoma tissue PC1,2 scores.(See Appendix A) The relatively large p-value for higher order PC (PC3=0.2814, PC4=0.4932) shows that the difference between normal and carcinoma tissue is not significant and, thus, that higher order PCs are not useful for tissue classification. Hence, it is reasonable to use only first two principal components for tissue classification.

The overall diagnostic accuracy of the PC score threshold based algorithms are shown in Figure 9.
Figure 9: Principal Component score threshold based algorithm diagnostic accuracy

OBSERVATIONS AND DISCUSSION:

From Figure 9, it is obvious that there is a significant improvement in diagnostic accuracy when using PC1 and PC2 rather than just PC1. In general, the specificity of the algorithm based on PC1 and PC2 is 20-25% higher than that of PC1 for the same sensitivity. Also, the calibration set diagnostic accuracy, which is based on the calibration set optimal threshold, gives 1% higher specificity than the prediction set specificity on average. For the PC1,2 and score-based algorithm, a prediction set sensitivity of 92% is achieved when the specificity is 96%.

2. PCA with logistic discrimination

The conditional probability of a given tissue being normal or carcinoma relies on the
PC score distribution of different principal components. The commercial statistical software, EViews, was used to estimate the PC score marginal distribution. In Figure 10, the marginal score density distributions, estimated using Silverman's Method [31] with the PC1 and PC2 estimated score density of each tissue category, are shown for illustration.

**Figure 10.** Principal Component 1,2 score distribution for Normal and Carcinoma tissue using Silverman density estimation

The posterior probability calculation was based on the estimated multivariate normal density model. The marginal densities of which are shown in Figure 11.
**Figure 11.** Principal Component 1-4 score distribution approximates to Normality

**OBSERVATIONS AND DISCUSSION**

If the marginal densities of the joint distribution are not normally distributed, then the variable is not jointly normal. Using the Q-Q plot correlation coefficient normality test on the PC1-5 score [26], it was found that, at the 5% significance level, the normality assumption held for almost all the PC scores, except for the PC1 abnormal and PC4 normal tissue score which was slightly less than the critical point. (See Appendix B) In this study, the normal density model for the PC score given tissue type was assumed, and Equation (5.7) was used.
The resulting posterior probability that the given sample belongs to a carcinoma tissue category, based on PC1 and PC2, is shown in Figure 12.

**Figure 12.** Posterior Probability of a given Tissue belongs to Carcinoma Category based on Principal Component 1,2

**OBSERVATIONS AND DISCUSSION**

In Figure 12, it can be observed that in general, the normal tissue samples tend to concentrate at a probability of zero, while carcinoma tissue samples tend to concentrate at a probability of one. A small number of observations deviate from the majority that makes no diagnostic threshold is able to completely separate these two kinds of tissue. Adjusting the decision threshold can achieve the desired sensitivity or
specificity. Further, those tissue samples posterior probabilities reveal their certainty to be normal or carcinoma. Thus, the score can be used as an indicator for further analysis, especially the scores with a posterior probability close to 0.5.

The overall diagnostic accuracy of the PC posterior probability based algorithm, when different numbers of PCs are used for analysis, is compared in Figure 13.

![Figure 13. PCA Logistic Discrimination diagnosis accuracy](image)

**OBSERVATIONS AND DISCUSSION**

There are several important observations that can be made from Figure 13. Firstly, the
overall prediction set diagnostic accuracy is slightly worse than the calibration set diagnostic accuracy when the same number of PC are used for analysis, as in the case of other algorithms. Secondly, calibration set diagnostic accuracy, when more than two principal components are used, is more-or-less the same. Thirdly, as the number of PCs used for analysis increases, the degradation from the calibration set to prediction set diagnostic accuracy becomes more severe. Only a 1-2% specificity degradation for same sensitivity is shown when PC1 and PC2 are used, while the degradation is around 10% specificity when PC1-10 are used for analysis. One possible reason for the degradation is that the assumed normal score distribution model does not exactly hold and, therefore, introduces error. Also, higher order PCs mainly account for the uncorrelated random noise and do not contain diagnostic information. The Students' T-test on higher order PCs give relatively high p-value. Moreover, analyzing more PCs dimensions generally gives less accurate multivariate joint density estimations because of the limited sample size [31]. Therefore, algorithms based on optimizing calibration sets may not show optimal results for prediction sets for PCA with logistic discrimination when many PCs are used.

5.4 Overall Comparison of Different Algorithms

RESULTS

The PC based algorithms were compared with the previous wavelength ratio based algorithm. The prediction set diagnostic accuracy of different algorithms is presented in Figure 14.
Figure 14. Performance Comparison of PCA based algorithms and the Two-wavelength ratio algorithm

OBSERVATIONS AND DISCUSSION

In general, the PC1,2 score diagnostic threshold based algorithm gives the best diagnostic accuracy of all algorithms. The result shown in Figure 14 shows that highest overall sensitivity and specificity can be achieved when using the PC1,2 score threshold based algorithm. Especially when the sensitivity is in the range of 90-94%, the specificity is higher than 95%. The overall diagnostic accuracy is slightly less (within a 2% specificity on average) for PCA with logistic discrimination based on
the posterior probability as the score, when using the PC1-2. For PC with logistic discrimination, using higher order PC than the first two principal components did not help discrimination. As shown in Figure 14, using logistic discrimination with PC1-5 slightly worsens the diagnostic performance. Analysis using PC1-10, as expected, makes the diagnostic performance even worse. The diagnostic performance of the two-wavelength algorithm is between PC1-5 and PC1-10 with logistic discrimination. PCA using only the first principal component gives the worst diagnostic accuracy. Clearly, a lot of important diagnostic information is not captured by the first principal component.

The diagnostic accuracy of two-wavelength and three-wavelength algorithm, in general, is higher than PCA using only the first principal component and PC1-10 with logistic discrimination. However, the wavelength band ratio based algorithm has a slightly worse diagnostic performance than the PCA using PC1-2.

From the above discussion, it can be concluded that the PC1,2 score based algorithm is the most reasonable choice to achieve the best diagnostic accuracy and that the algorithm is feasible for implementation with real time diagnosis. However, this PC score threshold based algorithm is difficult to further analyze when more PCs are used as this requires an exhaustive search of higher dimension variables, which is computationally intensive. However, the PCA with the logistic discrimination method has the advantage that it is flexible in that it can be combined with other algorithms to further improve diagnostic accuracy, provided that the PC score distribution can be accurately estimated.
5.5 Robustness of the PCA based Detection Algorithms on Noisy Data

As the PCA based algorithm is shown to be sophisticated in regard to diagnostic performance (sensitivity, specificity), and because the data is likely collected in noisy environments in most cases while different amounts of noise are possible, it worths the effort to evaluate the performance of the algorithm under noise. In this section, investigations of the noise robustness of the PCA based detection algorithms were performed.

THE METHODS

To simulate the robustness of the PCA based algorithm on noisy data, different amounts of zero-mean Gaussian random noise were intentionally added to the spectral data. The noises range from 0% to 30% were added and the proportion of the added noise was obtained as

\[
SNR = \frac{Average\ Signal\ Amplitude}{RMS\ Noise}
\]  \hspace{1cm} (5.8)

Sensitivity and specificity degradation were used as the criteria of the noise robustness evaluation. Also, the robustness of PCA with score diagnostic threshold algorithm was compared with the PCA with logistic discrimination.

1. **PCA with score diagnostic threshold**

In each run of cross-validation, firstly, principal component analysis was performed on the calibration set smoothed and normalized (preprocessed) spectra to calculate the PC scores. Secondly, PC1-2 scores were used to calculate the optimal diagnostic threshold line by exhaustive search. The choice of principal component was explained
in Section 5.3. The optimal diagnostic line, together with the PC loading, were trained as the ‘model’ that were validated by the prediction set. Thirdly, different amount of noise was added to the prediction set preprocessed spectrum. Finally, the resulting sensitivity and specificity of this ‘noisy’ spectrum was calculated according to the calibrated PC loading and the diagnostic threshold, and the overall sensitivity and specificity is their average values over all runs of cross-validation, and the degradation of diagnostic accuracy was evaluated.

2. PCA with logistic discrimination

In each run of cross-validation, firstly, principal component analysis was performed on the calibration set smoothed and normalized (preprocessed) spectra to calculate the PC scores. Secondly, PC1-2 scores were used to construct the conditional multivariate Gaussian density function that was given in Equation (5.6) and (5.7). Again, the choice of principal component was explained in Section 5.3. The density function, together with the prior probability of each tissue category, was used to calculate the posterior probability that a particular tissue being carcinoma according to the Bayesian theory based equations in Equation (5.6). Here, the PC loading, the estimated density function and the posterior probability threshold were trained as the ‘model’ that was validated by the prediction set. Thirdly, different amount of noise was added to the prediction set preprocessed spectra. Finally, different resulting sensitivity and specificity of the ‘noisy’ spectra were evaluated according to the calibrated PC loading, the estimated density function, and different threshold of posterior probability. The overall sensitivity and specificity is their average values over all runs of cross-validation. The degradation of diagnostic accuracy at different noise level was compared.
RESULTS AND DISCUSSION

Typical pre-processed spectral data with different amounts of Gaussian noise are illustrated in Figure 15. The effect of the different amounts of noise on the cross-validation prediction set diagnostic accuracy of the PC1-2 score based diagnostic line algorithm is shown in Figure 16. The result shows that there is only within 1% average diagnostic accuracy degradation with 10% addition of random noise. With 20% additional Gaussian noise, the degradation is within 3%. As a comparison, the effect of different amount of noise on the prediction set diagnostic accuracy of PCA with logistic discrimination is shown in Figure 17. From Figure 17, it is clear that less than 1% average diagnostic accuracy degradation was resulted when there is 10% additive Gaussian noise. The degradation upon 20% additional Gaussian noise is less than 2%. The result also demonstrates the noise robustness of PCA algorithm based on posterior probability is slightly better than the PCA algorithm based on score diagnostic threshold.

In the actual noisy environment where the tissue spectra are collected, the random noise is most likely less than 10% of the spectral data, and the pre-processing procedure helps to remove part of the noise. Even with 10% of random noise on the pre-processed spectral data, the degradation of specificity is within 1%. Hence, the PC Score based algorithm is robust to random noise that is likely to be encountered in clinical applications.
Figure 15. Typical LIF Spectra with different amount of Gaussian Random Noise
Figure 16. PC 1-2 Diagnostic Line algorithm Diagnostic Accuracy on Noisy Spectra
Figure 17. PC 1-2 Logistic Discrimination algorithm Diagnostic Accuracy on Noisy Spectra
Chapter 6

Investigation into the Possibility of applying the Algorithm using Optical Processing Technique for real time clinical diagnosis

6.1 Introduction

In the collection of fluorescence spectrum of small tissue surface, the use of an ordinary spectrograph to record a single pixel spectrum on the tissue surface in one measurement is suitable. However, when taking measurements in a large area of tissue surfaces, the pixel-by-pixel examination may not be practical for clinical purposes. An imaging approach that can rapidly process the spectral signal of every pixel over the entire imaged tissue surface is more desirable.

In this chapter, an optical imaging processing method, that can extract the clinically useful information for characterizing tissue pathology, based on the principal component analysis of light-induced autofluorescence developed in previous chapter, is presented. An optical processing system is designed for the in vivo autofluorescence spectra recorded from the nasopharyngeal tissue. The processing involves the application of a set of optical spectral filters, which are related to the principal component loading vectors, that process the autofluorescence signal optically and generate principal component scores from the autofluorescence spectra. The scores are then correlated with the tissue pathology.

The objective of this analysis is to investigate the diagnostic performance of this PCA based optical processing system. This chapter is outlined as follow. First, the
relationship among the optical processing filter characteristics, the principal component analysis, and the tissue diagnostic information is presented. Second, the diagnostic accuracy degradation analysis of this system, compared to that of the previous non-optical processing system, is reported. Third, the spectral noise robustness of the optical processing system is discussed.

6.2 Investigation of the Optical PC Filters

A set of optical filters with the desired spectral transmission characteristics is needed. The relationship between the filter transmission characteristics and the principal component scores is described. A simple color system consists of a set of spectral filters for color recognition. The emission spectrum $S(\lambda)$ is projected on to this set of spectral filters as

$$C_i = \int F_i(\lambda)S(\lambda)d\lambda$$

(6.1)

where $F_i(\lambda)$ is the transmission of the $i$-th spectral filter. This projection yields a set of color responses, $C_i$. Similarly, if a spectral analysis system is equipped with a filter set, $F_i(\lambda)$, related to the loading vectors of the PCs, the scores of the incident spectrum projected on to the PCs can be extracted from the system responses. A 'principal component filter' can be designed as $F_i(\lambda) = \alpha_i P_i(\lambda) + \beta_i$ such that $0 \leq F_i(\lambda) \leq 1$. Here, $P_i(\lambda)$ is the spectrum of the $i$-th PC loading vector in the wavelength space. The $\alpha_i$ and $\beta_i$ are two parameters to ensure the value of $F_i(\lambda)$ in the range of 0 to 100% because the value of $P_i(\lambda)$ at different wavelengths can be positive or negative. The $i$-th PC score then can be calculated as

$$I_i = \frac{1}{\alpha_i} \int F_i(\lambda)S(\lambda)d\lambda - \frac{\beta_i}{\alpha_i} \int S(\lambda)d\lambda .$$

(6.2)
where \( \int F_i(\lambda)S(\lambda)d\lambda \) is the response of spectrum \( S(\lambda) \) to filter \( F_i(\lambda) \) and \( \int S(\lambda)d\lambda \) is the system response of \( S(\lambda) \) without any filter. Both terms are measurable. This means that a spectral analyzer with \( i \) PC filters needs, in total, \( i+1 \) measurements to generate all the PC scores. In an imaging system with \( i+1 \) channels of \( i \) PC filters, the spectral signal of each pixel can be processed in parallel, optically, and the score images of all the PCs can be produced simultaneously. An image with diagnostic information can then be created using the algorithm built on the PC scores.

The details of the instrument for recording the in vivo LIF spectral data was described in Chapter 3 (see Figure 2). An additional spectral analyzer with a set of PC filters based on the autofluorescence spectra collected from nasopharyngeal tissue in vivo was designed. According to the results of Chapter 5, the first two PC account for 90% of the total spectral data variation. Therefore, the higher order PCs are much less significant. Also, significant statistical difference between carcinoma and normal tissue, using unpaired students' t-test on the PC score as the differentiation criterion, only shows in the first two PCs. As a result, it is reasonable to choose two filters corresponding to PC1 and PC2 to develop the diagnostic algorithm.

The optical filters were designed by our collaborator with a commercial software for optical thin films (Film Wizard, Scientific computing, CA). The transmissions of ideal PC filters and designed filters are shown in Figure 18. Here, the values of the scaling and shifting parameters \( a_i \) and \( \beta_i \) in Equation (6.2) were selected so as to keep the transmission of the calculated PC1 and PC2 filters in the range from 10% to 90%. The transmission of the designed optimal thin film filters obtained from the software was found to be quite closed to the transmission calculated from PC1 and PC2.
average difference between the software fitted filters transmission and the calculated filter transmission were less than 1% (PC1) and 2% (PC2).

Figure 18. Transmissions of PC1 and PC2 Filters
6.3 Diagnostic Performance Analysis of the Optical Processing System

As explained in Chapter 4 and 5, because of the limited sample size, cross-validation was used to evaluate the diagnostic performance of the optimal processing method for differentiating carcinoma tissue from normal tissue. In vivo LIF spectra from 85 carcinoma tissue sites and 131 normal tissue sites from a total of 59 patients, were collected and used for analysis (see Chapter 4 and 5). In the validation procedure, one spectrum was held out and the PC optical filters were designed using the PC1 and PC2 calculated from the remaining 215 spectra. Based on Equation (6.2), the $i$-th PC scores of the normalized spectra were calculated from the measurements with and without filter, using the Equation (6.3).

$$I_i^N = \frac{1}{\alpha_i} \left[ \frac{\int F_i(\lambda)S(\lambda)d\lambda}{\int S(\lambda)d\lambda} - \beta_i \right]$$  \hspace{1cm} (6.3)

Both the PCA score based diagnostic line algorithm and the PCA logistic discrimination algorithm in Chapter 5 can be used to differentiate carcinoma from normal tissue using PC filters.

1. PC filters with score diagnostic threshold

The optimal diagnostic line in the PC2 versus PC1 filtered score plot from exhaustive search, defined as $Y = aX + b$ was calibrated from the training data set in each round of the cross-validation. The PC1 and PC2 scores in Equation (6.3) were the $X$ and $Y$ respectively. The diagnostic accuracy of the algorithm can be controlled by choosing the parameters $a$ and $b$. The results of the optical process system diagnostic performance are shown in Table 1. The calibration set sensitivity was set at 92%, 95%
and 98%.

<table>
<thead>
<tr>
<th>SE</th>
<th>SP</th>
<th>SE</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>98%</td>
<td>83%</td>
<td>96%</td>
<td>82%</td>
</tr>
<tr>
<td>95%</td>
<td>93%</td>
<td>94%</td>
<td>90%</td>
</tr>
<tr>
<td>92%</td>
<td>97%</td>
<td>92%</td>
<td>96%</td>
</tr>
</tbody>
</table>

SE and SP refer to sensitivity and specificity.

Table 1. Performance of PCA Diagnostic Line Algorithm using Optical Processing

From Table 1, it can be seen that a high sensitivity of 92% and a specificity of 96% were achieved. A comparison of the optical PC filters diagnostic performance and the PCA algorithm based on the calculated PCs in Chapter 5 reveals that the diagnostic accuracies were more or less, the same. This is because the fitting of the transmissions of designed optical filters to those of the optical transmission directly calculated from the PCs is precise, with less than 2%, error as described in the previous section.

2. PC filters with logistic discrimination

The multivariate density function model was fitted in the PC1 and PC2 filtered score using the normality assumption and resulting posterior probability was calculated, as indicated in Chapter 5, from the training set in each round of the cross-validation. The estimated density function, together with different posterior probability thresholds, were then validated by the prediction data set. The diagnostic accuracy of the algorithm can be controlled by choosing the posterior probability thresholds. The results of the optical process system diagnostic performance are shown in Table 2.
The calibration set sensitivity was set at 90%, 94% and 96%.

<table>
<thead>
<tr>
<th>Training set</th>
<th>Cross-validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE</td>
<td>SP</td>
</tr>
<tr>
<td>96%</td>
<td>82%</td>
</tr>
<tr>
<td>94%</td>
<td>93%</td>
</tr>
<tr>
<td>90%</td>
<td>94%</td>
</tr>
</tbody>
</table>

SE and SP refer to sensitivity and specificity.

**Table 2.** Performance of PCA Logistic Discrimination Algorithm using Optical Processing

From Table 2, it is clear that a sensitivity of 93% and a specificity of 93% were achieved with the PC filters with logistic discrimination. A comparison of the optical PC filters diagnostic performance and the PCA algorithm based on the calculated PCs in Chapter 5 reveals that the diagnostic accuracies were more or less, the same. This is because the fitting of the transmissions of designed optical filters to those of the optical transmission directly calculated from the PCs is precise, with less than 2%, error as described in the previous section. The comparison between Table 1 and Table 2 reveals that PC filters with diagnostic threshold algorithm is a better classification scheme than the PC filters with logistic discrimination algorithm in term of diagnostic performance because logistic discrimination algorithm requires the normality assumption of PC score distribution.
6.4 Noise Robustness of the Optical Processing System

As discussed in the previous chapter, in many clinical measurements, the spectra are collected in noisy environments. This is particularly the case in the remote imaging method where the source-tissue distances are relatively large and not fixed. Also, the carcinoma tissue spectra, generally, contain more noise than the normal tissue spectra because the emission spectral signals of carcinoma are weaker. Furthermore, the application of the optical filters as a part of the signal processing does not involve the noise smoothing as a preprocessing step. Thus, the noise problem is more significant in this optical processing system than in the system discussed in Chapter 3. Hence, it is important to evaluate the performance of the optical processing system at different noise levels. The noise robustness analysis in this study involved the investigation into the diagnostic performance degradation when additional spectral noise was applied. This is presented in this section.

THE METHODS
As in the analysis of Chapter 5, zero mean Gaussian random noise was assumed. Different amounts of noise are added to the \textit{in vivo} autofluorescence spectra. The noises ranging from 0\% to 30\% of the spectral signals were added to the validation set data, and the signal-to-noise ratio was the same as that described in Section 5.5 (see Figure 15). The two optical PC filters with a transmission range from 10\% to 90\% (presented in Section 6.2) were used to filter the incident noisy spectral signals and generate the 'noisy' PC1 and PC2 score. Again, sensitivity and specificity degradations were used as the criteria of the noise robustness evaluation. Also, the robustness of PC filters system with score diagnostic threshold algorithm was compared with the PC filters with logistic discrimination.
1. **PC filters with score diagnostic threshold**

The pre-calibrated and exhaustive searched optimal PC diagnostic lines of desired sensitivity were used to discriminate the carcinoma tissue from the normal tissue, using the PC filter scores. The degradation of the diagnostic accuracy was calculated.

2. **PC filters with logistic discrimination**

The pre-calibrated multivariate normality density model and the posterior probability thresholds were used to discriminate the carcinoma tissue from the normal tissue, using the PC filter scores. The degradation of the diagnostic accuracy was calculated.

**RESULTS AND DISCUSSION**

1. **PCA with diagnostic threshold**

In Figure 19, a comparison of the effect of different amounts of noise on the cross-validation prediction set diagnostic accuracy of the optical processing system is given. From Figure 19, it is clear that the diagnostic accuracy is only degraded by less than 1%, on average, with 10% of additional Gaussian random noise. With 20% additional noise, the degradation is less than 3%.
Figure 19. Diagnostic Performance of PCA Diagnostic Line Algorithm using Optical Processing on Noisy Fluorescence Spectra

2. PCA with logistic discrimination

The effect of different amounts of noise on the prediction set diagnostic accuracy of the optical processing system is compared in Figure 20. The diagnostic accuracy is only degraded by less than 1%, on average, with 10% of additional Gaussian random noise. With 20% additional noise, the degradation is less than 2%.
Figure 20. Diagnostic Performance of PCA Logistic Discrimination Algorithm using Optical Processing on Noisy Fluorescence Spectra

From the comparison of the diagnostic accuracy degradation of PC filters with diagnostic threshold algorithm and the PC filters with logistic discrimination under noisy optical processing environment, the result demonstrates a slightly better noise robustness for the PC filters based on logistic discrimination algorithm under the same amount of random noise, especially when the signal-to-noise ratio is low.

In clinical practice, the collected \textit{in vivo} autofluorescence tissue spectra signals are
noisy, especially when the fluorescence emission spectral signals are weak, as in the case of nasopharyngeal carcinoma tissue. But the noise is likely to be less than 10% of the spectral signals. Even with 10% random noise on the fluorescence emission spectral signals, the degradation of specificity is less than 1% for the same sensitivity. Hence, the PCA based optical processing technique is promising for fast and reliable tissue classification.

6.5 Conclusions

The possibility of applying optical filtering techniques to process the in vivo autofluorescence spectra that uses the PCA based algorithm was presented. To evaluate the optical filtering technique, an optical processing system was proposed. The system was based on the previous system used to develop the PCA based algorithm. A set of optical PC filters with designed transmission-wavelength characteristics was applied to process the fluorescence emission signals. Simple mathematical calculations were adequate for the tissue classification. This system is particularly useful when collecting the in vivo autofluorescence spectra of human organs from a large tissue surface area, since all the pixels of any two-dimensional tissue site image can be processed simultaneously. Because this tremendously increases the processing efficiency, the use of the multivariate spectral information for real-time tissue diagnosis for image measurement is feasible. The optical processing technique is reliable for tissue diagnosis even when the spectra are recorded in noisy environments. The diagnostic performance degradation under noise, when the PCA based algorithm was applied in the optical processing system, was shown to be small. To conclude, the proposed optical processing technique simplifies the instrumentation and the computer processing needed for tissue diagnosis. It is both possible, and
feasible, to utilize for practical clinical applications.
Chapter 7

Summary, Conclusions and Future Directions

In this study, our major goals were achieved. Multivariate statistical algorithms based on PCA were successfully developed, and the result showed significant improvement in the diagnostic accuracy in respect to the previously reported two-wavelength ratio algorithm. A sensitivity of 92% is achieved when specificity is set to 96% for the PCA based algorithms. However, the sensitivity is only about 88% when the specificity is set to 95% for the wavelength ratio based algorithm. In addition, the algorithm is robust with spectra collected in a noisy environment. Further, it is feasible to apply the algorithm implemented by optical filter for real time diagnosis.

In conclusion, there are several advantages in using the PCA based algorithm for tissue diagnosis that are worth mentioning. Firstly, the algorithms can be easily applied without prior knowledge of the optical properties of tissue such as blood absorption. The diagnostic algorithms can be easily applicable to other kinds of tumor tissue once their tissue spectra are obtained. Secondly, the diagnostic algorithm involves simple mathematical computations that make it both possible and feasible to implement practically for real time diagnosis using a simple optical filtering technique and personal computer setup. Hence, this method could be used for non-invasive early cancer detection without the need of expert hand.

In the future, several further improvements are possible. In the present study, analysis was based on the available total of 216 tissue samples. This limited the score density
estimation accuracy. Therefore, in the future, as more samples are available for calibration, more sophisticated statistical analysis would be possible to give higher confident measure and diagnostic accuracy. Also, other multivariate statistical algorithms could be further investigated and compared to show that PCA is the most suitable method for detecting nasopharyngeal carcinoma. In addition, a deeper knowledge and understanding of the nasopharyngeal carcinoma spectral characteristics could be further established by suitable modeling of the carcinoma and the surrounding normal tissue spectra that takes account of the tissue optical properties such as blood absorption and scattering. Furthermore, the possibility for implementation could be further analyzed.
References


"Laser induced autofluorescence diagnosis of bladder cancer," J Urol 156 (5),

N. S., and Deutsch T. F., "Ultraviolet laser induced fluorescence of colonic tissue:

statistical algorithm to analyze human cervical tissue fluorescence spectra

[21] Chwirot B.W., Chwirot S., Redzinski J., and Michiniewicz Z., "Detection of
melanomas by digital imaging of spectrally resolved ultraviolet light induced


[23] Campbell I.D., and Dwek R.A., Biological Spectroscopy, Benjamin/Cummings

[24] Bigio I.J., and Mourant J. R., "Ultraviolet and visible spectroscopies for tissue
diagnostics fluorescence spectroscopy and elastic-scattering spectroscopy," Phys.

(1997).

[26] Richard A.J., Dean W.W., Applied multivariate statistical analysis, Prentice-Hall,


Appendix A  The Students’ $t$-test

A good reference of the students’ $t$-test was given in [26] and [34]. A brief but concise review that was applied in this analysis is presented in this Appendix.

A.1 The $t$-distribution

For a random sample from a normal distribution with known population mean $\mu$ and $\sigma^2$ variance, the quantity

$$ Z = \frac{(\bar{Y} - \mu)}{\sigma / \sqrt{n}} \quad (A1) $$

is a pivotal quantity which has a normal distribution $N(0,1)$. If the variance is unknown, the $t$-distribution, which was introduced by the statistician W. S. Gossett in 1908 who published the result under the pseudonym “Student”, can be modeled as the sample distribution.

$$ t_{n-1} = \frac{\bar{Y} - \mu}{s / \sqrt{n}} \quad (A2) $$

where the index $(n-1)$ is the degree of freedom and $n$ is the sample size. The distribution of this variable is now called Student’s $t$-distribution. The distribution of $t$ is similar to that of the normal distribution, but somewhat more “heavy-tailed”, and that for each sample size there is a different distribution [34].

A.2 One-sided Unpaired students’ $t$-test on Two-population

To perform mean comparison that involves normally distributed samples from multiple populations, the one-side unpaired Students’ $t$-test is commonly used. In our analysis, the samples were the principal component scores of the normal and
carcinoma tissue spectra. Univariate two samples analysis were used to compare the
difference in each principal component score between normal and abnormal tissue.

Suppose \( \bar{X}_1 \) and \( \bar{X}_2 \) are the means and \( s_1 \) and \( s_2 \) are the standard derivation from two
random samples of size \( n_1 \) and \( n_2 \), respectively. If \( X_1 \) and \( X_2 \) are normally
distributed with population means \( \mu_1 \) and \( \mu_2 \), and variances \( \sigma_1^2 \) and \( \sigma_2^2 \), then

\[
\bar{X} - \bar{Y} \sim N \left( \mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right)
\]  

(A3)

so that

\[
\frac{(X - Y) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}} = t_w = Z \text{(Standard Normal Distribution)}
\]  

(A4)

If their population means and variances are unknown and unequal, and the samples of
different categories are not paired, then, under the one-sided hypothesis

\[ H_0 : \mu_x \leq \mu_y \quad \quad \quad H_1 : \mu_x > \mu_y \]  

(A5)

The test statistic is given by

\[
T = \frac{\bar{X} - \bar{Y}}{\sqrt{s_x^2 / n_1 + s_y^2 / n_2}}
\]  

(A6)

The null hypothesis, \( H_0 \), is rejected when the value \( T \) is high.

The p-value is given by the probability \( P[T > t_\alpha] \). Hence, a high p-value means the
hypothesis \( H_0 \) is not rejected, which implies that the difference between the two
samples under investigated is insignificant.
Appendix B  Q-Q plots and Normality Test

A good reference of the Q-Q plots and normality test is given in [26]. A brief but concise review is presented this Appendix.

B. 1 Q-Q plots

To assess the assumption of normality, special plots called Q-Q plots are commonly used. These plots can be made for the marginal distributions of the sample observations on each variable. Briefly, Q-Q plots are the plots of the sample quantile versus the quantile one would expect to observe if the observations actually were normally distributed.

Let \( x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \) represent the ordered \( n \) observations on \( X_i \) according to magnitude. For a standard normal distribution, the quantiles of \( j \)-th observations, \( q_{(j)} \), are defined by the relation

\[
P[Z \leq q(j)] = \int_{-\infty}^{q(j)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \, dz = p_{(j)} = \frac{j - \frac{1}{2}}{n}
\]

(B1)

where \( p_{(j)} \) is the probability of getting a value less than or equal to \( q_{(j)} \) from a standard normal population. A Q-Q plot is the plot of the ordered data \( x_{(j)} \) against the normal quantiles \( q_{(j)} \). When the observations arise from a normal population, the pairs \( (q_{(j)}, x_{(j)}) \), will be approximately linearly related, since \( \sigma \cdot q(j) + \mu \), where \( \sigma \) and \( \mu \) are the population standard derivation and mean respectively, is nearly the expected sample quantile and the pairs of points \( (q_{(j)}, x_{(j)}) \) lie very nearly along a straight line.
B.2 Normality Test

The straightness of the Q-Q plot can be measured by calculating the correlation coefficient of the points in the plot. The correlation coefficient for the Q-Q plot is defined by

\[
 r_q = \frac{\sum_{j=1}^{n} (x_{(j)} - \bar{x})(q_{(j)} - \bar{q})}{\sqrt{\sum_{j=1}^{n} (x_{(j)} - \bar{x})^2} \sqrt{\sum_{j=1}^{n} (q_{(j)} - \bar{q})^2}} \tag{B2}
\]

and the critical points for these Q-Q plot correlation coefficient can be used as a test of normality. The critical points shown in Table B1 were used in this analysis where the hypothesis of normality at level of significance \( \alpha \) is rejected if \( r_q \) fall below the appropriate values [26].

<table>
<thead>
<tr>
<th>Sample Size ( n )</th>
<th>Significance levels ( \alpha )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
</tr>
<tr>
<td>150</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
</tr>
<tr>
<td>300</td>
<td></td>
</tr>
</tbody>
</table>

**Table B1.** Critical points for the Q-Q plot correlation coefficient test for normality