STATISTICAL ASPECTS OF ELASTIC SCATTERING SPECTROSCOPY WITH APPLICATIONS TO CANCER DIAGNOSIS

A thesis submitted for the degree of Doctor of Philosophy (PhD) of University College London

By

Ying Zhu

Department of Statistical Science
National Medical Laser Centre
University College London

March 2009

“The quiet statisticians have changed our world; not by discovering new facts or technical developments, but by changing the ways that we reason, experiment and form our opinions….”

Ian Hacking, Contemporary Philosopher

ABSTRACT

Elastic scattering spectroscopy (ESS), is a non-invasive and real-time *in vivo* optical diagnosis technique sensitive to changes in the physical properties of human tissue, and thus able to detect early cancer and precancerous changes. This thesis focuses on the statistical issue on how to eliminate irrelevant variations in the high-dimensional ESS spectra and extract the most useful information to enable the classification of tissue as normal or abnormal.

Multivariate statistical methods have been used to tackle the problems, among which principal component discriminant analysis and partial least squares discriminant analysis are the most explored throughout the thesis as general tools for supervised dimension reduction and classification. Customized multivariate methods are proposed in the specific context of ESS.

When ESS spectra are measured *in vivo* by a hand-held optical probe, differences in the angle and pressure of the probe are a major source of variability between the spectra from replicate measurements. A customized spectral pre-treatment called error removal by orthogonal subtraction (EROS) is designed to ameliorate the effect of this variability. This pre-treatment reduces the complexity and increases both the accuracy and interpretability of the subsequent classification models when applied to early detection of cancer risk in Barrett’s oesophagus.

For the application of ESS to diagnosis of sentinel lymph node metastases in breast cancer, an automated ESS scanner was developed to take measurements from a larger area of tissue to produce ESS images for cancer diagnosis. Problems arise due to the existence of background area in the image with considerable between-node variation and no training data available. A partially supervised Bayesian multivariate finite mixture classification model with a Markov random field spatial prior in a reduced dimensional space is proposed to recognise the background area automatically at the same time as distinguishing normal from metastatic tissue.

ACKNOWLEDGEMENTS

The achievement of this work would not have been possible without the immense help of my principal supervisor, Professor Tom Fearn, to whom I cannot express my appreciation enough for bringing me to this most interesting statistical subject. I feel very fortunate to have such a knowledgeable and inspiring supervisor to share his invaluable ideas and time throughout the whole project. I am most grateful to my project supervisor Professor Stephen Bown and to Dr Laurence Lovat who gave me the opportunity to work on this very exciting real-world statistical application project and supported me throughout.

Special thanks are due to Dr Ben Clark for his innovative and practical suggestion during the first year of my PhD, which brought a smooth transition and refreshing challenge to my research. I would also like to thank Dr Martin Austwick and Dr Sandy Mosse for their patient explanation on the interaction between physics and medicine, which has been crucial. Dr Gary Mackenzie, Dr Wayne Chicken and Dr Santosh Somasundaram are greatly acknowledged for providing me with the clinical data from their diligent clinical work, and for their medical knowledge in these fields. This work was only made possible by the generosity of my scholarship sponsors, the Peacock Trust, and the ORS Awards Scheme, to whom I am deeply indebted.

The friendship, support and encouragement from my fellow researchers and all the members from National Medical Laser Centre and Statistical Science department of University College London have been of enormous value to me and have made my PhD study in London an extraordinarily enjoyable experience.

Lastly I would like to thank my family. The constant love and encouragement from my fiancé, Kang, enliven my research life and have brought me smiles even during the hardest times. My deepest gratitude and love have to go to my dear parents who gave me the impetus to begin and the constant encouragement to pursue my academic research, and were always ready to share all my happiness and difficulties throughout these years. Nobody else has sacrificed more during my research, and been more enthusiastic about my progress, than them. My sanity owes a lot to their dedication.

STATEMENT OF ORIGINALITY

I, Ying Zhu, certify that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

This chapter gives an overview of the logical structure of the thesis by introducing the background to this work, describing the major problems studied, and presenting the statistical methodologies used and the proposed solutions to the problems.

1.1 Context

This thesis is based on a joint project by National Medical Laser Centre and Statistical Science Department of UCL to examine the problem of medical diagnosis using high-dimensional elastic scattering spectroscopic (ESS) data. The aim of the project is to use elastic scattering spectroscopy to detect pre-cancerous and early cancerous changes in human tissue, with current application in Barrett's oesophagus, colon lesions and sentinel lymph nodes. In ESS a lamp delivers to the tissue light spanning a broad range of wavelengths. There is evidence that normal and abnormal tissue have different patterns of scattering and absorbance. It should be possible to use measured scattering and absorbance spectra to classify an area of tissue as normal or abnormal, or even to grade the level of abnormality.

Although the ESS measurement is simple to perform, ESS spectra often contain major sources of variation that are of little or no predictive value for the property of interest and thus can hinder the recognition of normal or abnormal tissue. The main statistical issue of this project is how to eliminate the irrelevant variations and extract the most useful information from high-dimensional spectroscopic data for classification and clinical diagnostic purposes. This is a challenging problem.

The overall aims of this research are twofold: a) to find approaches to tackle the problems in ESS clinical applications described above and contribute to patient care, and (b) to make some methodological contributions to the area of classification with high-dimensional data as a whole, in the following areas: spectral pre-treatment, dimension reduction, classification and discrimination, and image analysis.

1.2 Optical diagnosis and elastic scattering spectroscopy

There is little doubt that detection of early cancer or pre-cancerous change is of crucial importance for achieving a positive outcome for cancer therapy. The conventional clinical diagnosis procedure of diagnosing certain types of cancer by taking biopsies from the patient is time consuming, labour intensive and has a low detection rate for some high risk cancers. The challenge for clinicians and scientists is to develop new technologies for detecting patients at high risk of progression to cancer. Ideally this would be accurate, easy to use, inexpensive, and provide results rapidly, preferably without the need to remove tissue.

The application of optical technology to problems in clinical diagnosis is a rapidly growing field. In a diagnostic application, the goal is to learn something about the physiology or pathology of the tissue through its interaction with light. For example, the light reflected by tissue after absorption and scattering is characteristic of its chemical and morphological composition which, in turn, can depend on the nature and stage of certain diseases such as cancer. Absorption and scattering are two of the fundamental optical processes that can be exploited for diagnostic information.

This thesis focuses on elastic scattering spectroscopy (ESS), also known as diffuse reflectance spectroscopy, measured in the visible and near-infrared ranges of the electromagnetic spectrum, which is a cheap, fast and simple-to-use tool compared with other spectroscopic techniques. ESS, a novel optical diagnostic technique, is sensitive to the changes in the physical properties of tissue such as nuclear size, density and changes in the nucleus. Since these physical properties change with dysplasia, this technique could be used as a noninvasive and real-time *in vivo* test for early cancer or for precancerous detection. More details about the technical aspects of elastic scattering spectroscopy and its interaction with human tissues are presented in Chapter 2.

1.3 Statistical aspects of ESS

1.3.1 Multivariate methods for ESS

From physical and chemical points of view, the ESS spectrum, as a visible and NIR spectrum from a diffuse measurement, is a mixture of several physical effects (the strong dependence of reflectance on the scattering properties of the tissue sample, particularly on the particle size) and chemical effects (absorbances by molecular bonds) resulting in many overlapping and often broad peaks which are the product of complex patterns of scattering and absorption by numerous structural and biochemical components. These spectra are difficult or impossible to interpret using univariate methods. In most applications of ESS technology the number of variables (wavelengths) p of the spectral matrix is usually much larger than the number of observations (samples or subjects) n. A typical study includes variables from 1800 dimensions for only 300 spectra. These variables are collinear since there are high correlations and near linear relations among them. Investigating the association of spectral absorption and scattering properties with specific clinical outcomes from such high-dimensional data is a difficult and challenging task.

This thesis focuses on the analysis of ESS spectra, applying dimension reduction techniques as well as supervised and partially supervised classification. Most of the topics of multivariate statistics relevant to this thesis, regarding spectral pretreatment, dimension reduction, classification and discrimination, feature extraction, wavelength selection, etc., are reviewed in Chapter 3.

Among all these multivariate techniques, principal component discriminant analysis (PCDA) and partial least squares discriminant analysis (PLSDA) are the most explored as tools for dimension reduction and supervised classification. Both methods project the high-dimensional spectral variables onto a low-dimensional space spanned by components constructed as linear combinations of the original variables. The linear coefficients (loadings) are potentially interpretable for the intensity contribution to the classification at each wavelength. Both methods are described in Chapter 3 and are exploited in Chapter 4 with applications to detection of cancer or pre-cancerous change in Barrett's esophagus and colon lesions.

1.3.2 Spectral pretreatments for ESS

When ESS spectra are used for quantitative or qualitative analysis, data pre-treatment can be an important step in the construction of an effective regression or classification model.

When ESS spectra are measured *in vivo* by an optical probe in physical contact with human tissue, although the taking of each measurement is very simple to perform, ESS spectral acquisition under accurate control is very difficult in the clinical setting, even for experienced endoscopists. Therefore ESS spectra often contain major sources of variations caused by small changes in the angle and pressure of the probe used to acquire the spectra. These factors cause considerable measurement variability between replicated spectra taken at the same site, and can hinder the development of a diagnostic model for cancer risk. A customised pretreatment called error removal by orthogonal subtraction (EROS) is proposed in Chapter 4 to model the measurement variability and ameliorate the effects of it on the spectra as a pre-treatment step. Two applications of this method to the clinical diagnosis of colon lesions and high grade dysplasia or cancer in Barrett's oesophagus are presented in Chapter 4. For the purpose of model interpretation, an experiment was designed to find how pressure and angle of the probe influence the variations in measured spectral data collected under controlled conditions and to explain the measurement variability removed by EROS.

1.3.3 ESS images

For the application of ESS to the diagnosis of sentinel lymph node metastases in breast cancer, an automated two-dimensional ESS scanning device was developed to take measurements from the entire cut surface of the excised nodes. Instead of determining whether an individual point of tissue is abnormal or not, the ESS scanner examines a larger area of tissue by taking measurements at 400 points (pixels) in a 20×20 grid, and has the ability to produce diagnostic images to assess the cancer risk.

Although using the ESS scanner avoids experimental variation caused by angle or pressure of a hand-held probe and produces much more information, additional problems arise in the analysis and interpretation of the ESS scanning data. There is no reference pathology available for individual pixels in the image, which causes problems in deriving an algorithm to classify pixels. We do have manual measurements on other nodes taken at an earlier stage of this research, where we do have reference pathology for spectra from normal and metastatic tissue. The obvious approach is to use the manual measurement data to train a classification algorithm for use with the scanning data, though it is possible that the manually measured points are not completely representative of the data obtained in the scans. A further difficulty comes from the existence of a third group in the scanning data, which is a non-nodal group from a background area, possibly contaminated by blood or lipid. No training data are available for this non-nodal group and these background areas are very variable from node to node.

One opportunity also arises in the ESS scanning data analysis. The scanner generates a spectral image, and it should therefore be possible to use smoothness assumptions to improve the classification of individual pixels.

In Chapter 5, two different options for dimension reduction (for both manual and scanning data) are explored before constructing a classification model. One uses PCA on individual nodes, the other includes a common dimension, the canonical variable from a linear discriminant analysis on the manual data. A partially supervised image classification algorithm employing a Bayesian multivariate finite mixture model is then developed on the low-dimensional data to model three unknown groups (normal, metastatic and non-nodal) in the images from scanned measurements. Prior distributions for the parameters of the normal and metastatic groups are derived from the manually measured data.

In Chapter 6 we take into consideration the spatial correlation between adjacent pixels of image and a Markov random field (MRF) spatial prior is then incorporated into the model proposed in Chapter 5 in order to represent the continuity of the image. An implementation of the whole model system for a rapid intraoperative sentinel lymph node diagnosis is presented and discussed in Chapter 7.

CHAPTER 2

ELASTIC SCATTERING SPECTROSCOPY

This chapter gives a brief description of the basics of elastic scattering spectroscopy (ESS). It describes the light absorption and scattering properties of materials and in particular biological tissues, and the practical aspects of ESS. In the applications studied in this thesis, ESS is employed as an *in vivo* point measurement and an *ex vivo* imaging system for detection of early cancer and pre-cancerous changes in human tissues. The background provided in this chapter can be used for further physical understanding and interpretation of spectral features in the rest of the thesis.

2.1 Visible and NIR radiation

The discovery of invisible radiation was made by Herschel in 1800 when he observed that the relative heating effect of different portions of the spectrum of sunlight generated by a prism continued to increase when he placed a thermometer beyond the red end of the visible spectrum. The existence of near infrared (NIR) energy was proved by this classic experiment and subsequent discoveries led to the idea of the continuous electromagnetic spectrum. The NIR region spanning from 780 to 2500nm (nanometers) has become divided into two sub-regions mainly due to instrumental considerations: 780~1100nm and 1100~2500nm. The second region has been widely used in a variety of applications. However, there is increasing interest in the first subregion which is sometimes called the Herschel region in honour of its discover.

The spectra in this thesis range from 300 to 920 nm covering some of the ultraviolet (UV), the visible (Vis) and some of the near infrared (NIR) regions, the positions of which in the electromagnetic spectrum are shown in Figure 2.1. Among these regions, only a narrow band from about 400 to 750nm is visible to the human eye. The NIR region of the electromagnetic spectrum was little used in spectroscopy until the late 1950's when visible region spectrometers extended their range into the NIR region.

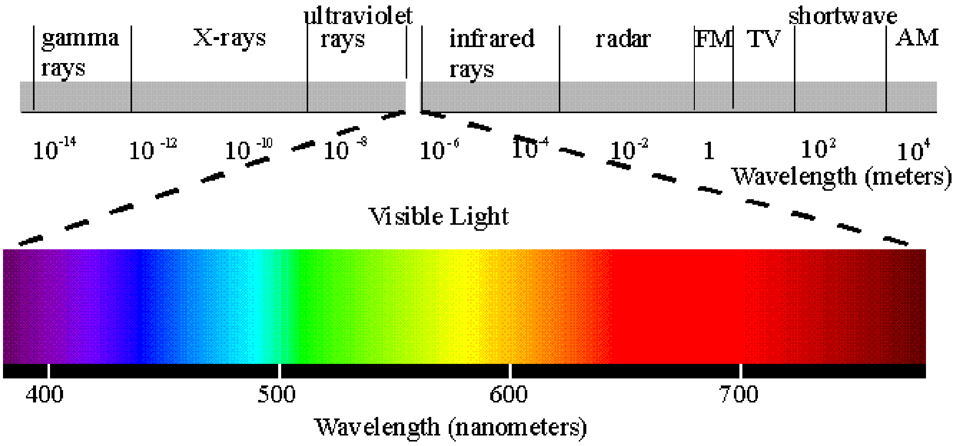


Figure 2. 1: Electromagnetic spectrum. (1meter =10^9 nanometer)

2.2 Elastic scattering of light and its interaction with tissue

2.2.1 Mie theory and elastic scattering

Scattering is the process by which small particles suspended in a medium of a different refractive index diffuse incident radiation in all directions.

Mie theory is a mathematical theory of the scattering of electromagnetic radiation by homogeneous spherical particles, developed by Gustav Mie in 1908. Mie theory applies to scattering by particles of a size comparable to the wavelength of the light being scattered. A particle, for the purposes of elastic scattering, can be defined as an aggregation of material that has a different refractive index to its surroundings. This is the appropriate theory for the scattering of light by the organelles of living cells. The shorter the wavelength of light, the greater the degree of scatter. For example blue light is scattered more than red by particles in the earth's stratosphere. Therefore, red light (especially when the sun is low in the sky) is seen to be coming directly from the sun, whereas blue light is scattered in all directions and make the whole sky blue on a sunny day.

In biomedical optics, scattering of photons is an important event and potentially has diagnostic value. The light scattered by a tissue has interacted with its ultrastructure which encompasses membranes and collagen fibers through to nuclei and cellular components.

Elastic scattering spectroscopy is a measure of both the elastic scattering of light via Mie theory and the absorptive components of tissue. In elastic scattering, there is no energy transfer and no change in the wavelength of the light; there is only a change in the spatial distribution of the radiation. The total intensity of the scattered light remains comparable to that of the incident light, reduced only by light that has been absorbed. In contrast, with the various forms of inelastic scattering, such as fluorescence and Raman spectroscopy (Bigio and Mourant 1997), light is absorbed at the incident wavelength and re-emitted at a different wavelength. Thus, the use of elastic rather than inelastic scattering allows for a large optical signal, approximately 100 times stronger than fluorescence and 1000 probably more like 10,000 times greater than Raman scattering.

The beauty of elastic scattering is its simplicity compared with other optical spectroscopic techniques. A pulse of white light is shone into the tissue, and this light is scattered by the tissue components (see Figure 2.2). The elastic scatter signal that is detected is due to light that has entered the tissue, been scattered one or more times, and exited nearby on the same side of the tissue. The degree of light scatter at each wavelength is dependent on the size of cellular organelles and the density of cellular packing. Only light that has undergone a 180^{o} turn due to multiple scattering events is detected by the detection fiber.

Elastic scattering spectroscopy uses a white light source, so spectral analysis of the light scattered back from tissue can include all wavelengths from the blue to the near infrared part of the spectrum. This gives a great deal of information about the tissue up to about 1 mm from where the probe is touching the surface of the tissue being interrogated. Elastic scattering spectroscopy (ESS) (Lovat and Bown 2004) is also known as diffuse reflectance spectroscopy (Georgakoudi *et al*. 2001). For simplicity these phenomena will be referred to collectively as ESS from now on.

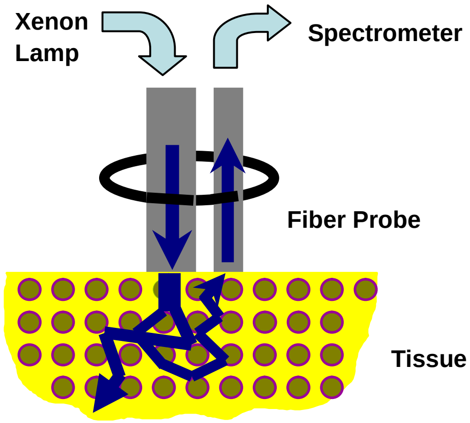


Figure 2.2: Schematic representation of elastic scattering spectroscopy probe. (Lovat and Bown 2004)

2.2.2 ESS instrumentation system and operations in practice

The ESS system used to make the measurements studied in this thesis consists of a pulsed xenon arc lamp, an optical probe, a spectrometer, and a computer to control various components and record the spectra (Figure 2.3). The arc lamp, spectrometer and their power supply are mounted in a briefcase size unit which is portable and can be easily connected to a laptop computer (Figure 2.4).

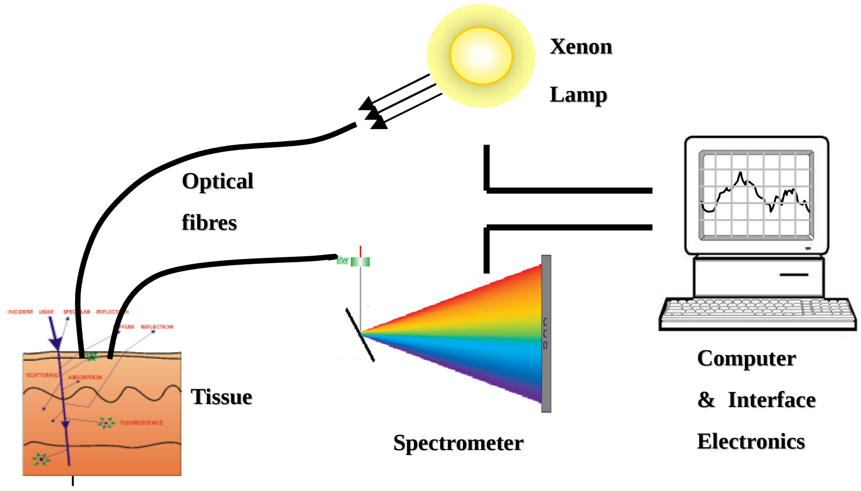


Figure 2.3: Schematic diagram of elastic scattering spectroscopy (ESS) system.

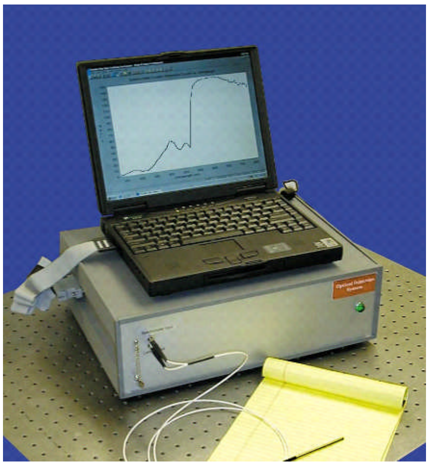


Figure 2.4: Picture of the ESS System.

In operation, short pulses of white light (320-920 nm) from the xenon arc lamp are directed through a flexible optical fibre (400µm diameter) which is gently placed in contact with the tissue under examination. For safety reasons ultraviolet B & C (100-315nm) light is filtered out to avoid the risk to patients. A collection fibre (200µm diameter) collects the light back-scattered from the upper layers of the tissue. These two fibres are encased in an outer sheath (outer diameter 2.0mm) and the fibres are separated by a distance of 50µm with a fixed centre-centre separation distance of 350 µm (Figure 2.5). This makes it convenient to place the optic fibre down the biopsy channel of a standard endoscope if the technique is to be used during endoscopy. It has been shown that the smaller the optical fibre separation the greater the sensitivity of ESS to the size of scatterers in the tissue (Mourant *et al*. 1997). The light is propagated into the spectrometer by the collection optical fibre and the output spectrum of light is recorded in a laptop computer for further analysis. In total each reading (collection and recording of a single spectrum) takes less than 200 milliseconds to perform. The fibre can be cleaned with the endoscope.

Before any tissue spectra are taken, a white reference spectrum is recorded from the flat surface of a Spectralon standard, whose elastic scattering is spectrally flat between 250 and 1000nm, allowing the system to account for spectral variations in the light source, spectrometer, fibre transmission, and fibre coupling. Each tissue spectrum is divided by this standard reference to give the system-independent spectrum of the site being investigated prior to any statistical processing.

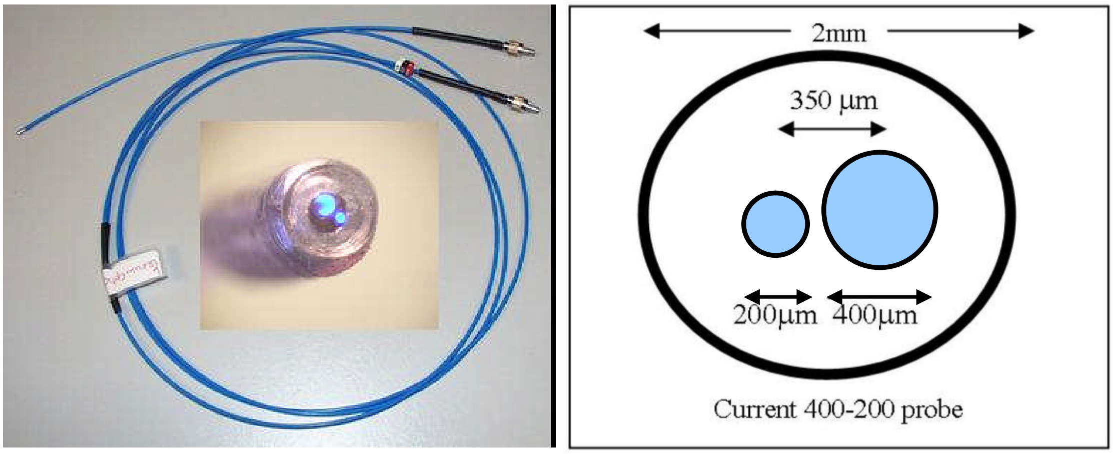


Figure 2.5: Picture of optical fibre (400-200µm) with calibration pot and schematic cross-section.

Immediately prior to (100 milliseconds before) any spectral measurement (Spectralon or tissue), the automated system records a “dark” spectrum (without triggering the lamp), which is subtracted from the spectrum (with lamp) that follows. Thus, the tissue spectrum that is stored and displayed is determined by the expression

$$\frac{S_{tissue} - D_{tissue}}{S_{ref} - D_{ref}},$$

where *ref* indicates a measurement with the Spectralon reference material, *S* indicates a spectrum recorded with the lamp triggered and *D* indicates a dark recording without the lamp. In this manner, the site-specific ambient light at the moment of measurement and detector dark current are accounted for.

What the reference measurement cannot correct for are changes in the angle and pressure of the probe in contact with the tissue. This problem will be discussed and solved from statistical point of view in Chapter 4.

2.2.3 Light interaction with tissue

In ESS light interaction with tissue involves scattering and absorption. Incident light is delivered to the tissue surface, part of which is absorbed in the tissue, whereas the non-absorbed part is subject to multiple elastic scattering and eventually emerges from the surface as diffuse reflectance carrying quantitative information about tissue structure and composition. A certain fraction of this emerging light is collected by the probe, whereas the remaining part escapes undetected. The amount of the light collected depends on the optical properties of scattering and absorption coefficients, as well as on the probe radius.

The absorption coefficient is directly related to the concentration of physiologically relevant absorbers in the tissue, which include oxygenated and deoxygenated haemoglobin. The scattering coefficient reflects information on the size and density of scattering centers in tissue, such as cells and nuclei, which is a very important factor that changes when cells become malignant (Webb and Jones 2004).

At wavelengths shorter than 600nm (in the UV and visible region) light is strongly attenuated in tissue due to absorption and therefore fails to penetrate more than approximately 1mm of tissue. However in the 600~800 nm range (in the NIR region) the absorption of light is significantly lower and scattering plays a key role in the spectral features. Below 370nm the xenon arc lamp has a low light output, causing very low signal to noise ratio of the spectra, thus only the window between 370 and 890 nm is explored in the ESS analysis here.

2.2.3.1 ESS and light absorption in tissues

Light from 350nm-450nm (in the UV region) has been shown to be mainly absorbed by haemoglobin. This absorption feature, which dominates the spectrum of light at these wavelengths, has been called the Soret band, with peak absorption at 414nm.

In the visible spectrum region (450nm-700nm) oxy- and deoxyhaemoglobin have been described as the biggest absorbers of light in biological tissue. This absorption occurs between 500-630nm with the peak for oxyhaemoglobin (HbO) and deoxyhaemoglobin (Hb) at 542nm, and a second peak for oxyhaemoglobin at 577nm. These absorption features have been termed the Q Bands of haemoglobin. Again, up to 630nm, haemoglobin has a smaller but significant effect on the ESS spectra. The relative absorption strengths of these chromophores and other principal tissue chromophores have been compiled in Figure 2.6. In the near infrared region, when wavelength is above 900nm, the absorption by water is the strongest contributor to tissue absorption with an increase in its absorption exceeding haemoglobin above 930nm with a small peak at 976nm (VandenBerg and Spekreijse 1997) as shown in Figure 2.6. Realistically, the ESS system is limited to a maximum wavelength of 1100nm because of very intense water absorption above this wavelength.

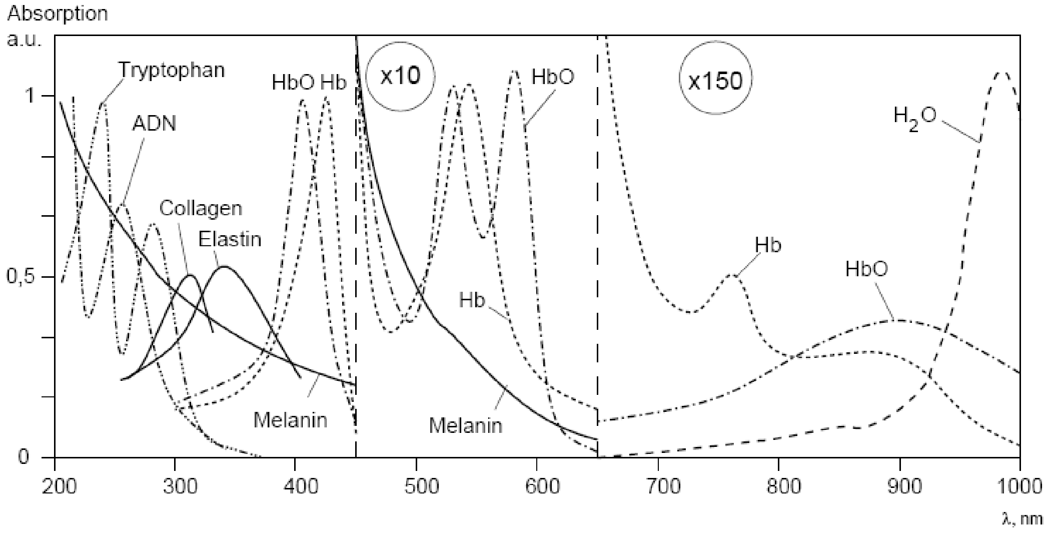


Figure 2.6: Compilation of absorption spectra of principle chromophores in tissue. (Hale and Querry 1973, VandenBerg and Spekreijse 1997, Zijlistra et al. 2000)

An example ESS spectrum measured *in vivo* using our ESS system from a patient with Barrett's esophagus is given in Figure 2.7. It shows that haemoglobin has a marked effect on the ESS spectrum with strong presence of the Soret band and Q bands of oxy- and deoxyhaemoglobin. The strong absorption of haemoglobin in the visible wavelengths of blue and green light is the reason that blood appears as a vivid red colour. More interpretations on absorption are addressed in Chapter 4 and 5 with applications to Barrett's oesophagus and sentinel lymph node for diagnostic purposes.

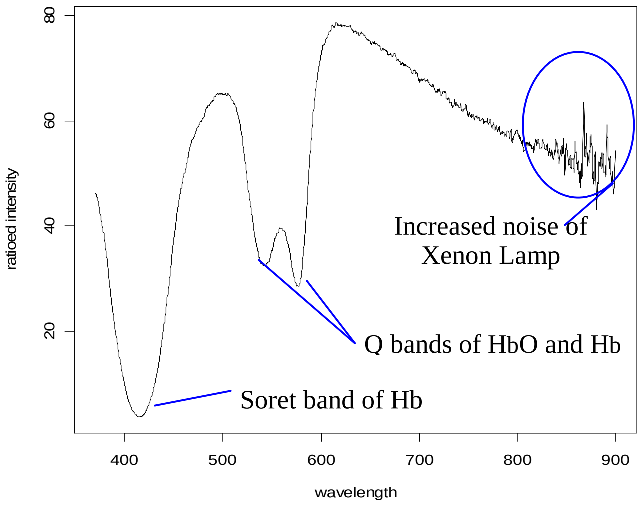


Figure 2.7: An example ESS Spectrum with the Soret and Q bands marked.

2.2.3.2 *ESS and light scattering in tissues*

One of the principal cellular scatterers for ESS is the nucleus. Several properties of the nucleus are known to change the pattern of elastic scattering. The first and potentially simplest to measure is nuclear size which particularly affects high angle scattering (Mourant *et al.* 2000). Secondly, nuclear density has also been shown to change the ESS spectra (Zonios *et al.* 1999) and nuclear chromatin content has been shown to affect the spectra of scattered light (Backman *et al.* 2000, Mourant *et al.* 2000). Finally, the nucleus:cytoplasmic ratio, along with other intracellular changes, has been shown to have a significant effect on scattering in Monte Carlo modeling (Drezek *et al.* 1999).

Another group of principle scatterers in cells, the mitochondria, and the cellular packing density have also been shown to change the elastic scattering of light in ESS (Mourant *et al*. 2000, Wallace *et al*. 2000).

2.3 ESS and early-cancerous and pre-cancerous changes in tissue

As described above ESS is sensitive to the changes in the physical properties of biological tissue. Some clinical pilot studies show that these physical properties do change with premalignant or malignant conditions, which suggests that ESS can be used as an non-invasive and real-time *in vivo* test to detect early cancer or precancerous change in human tissues.

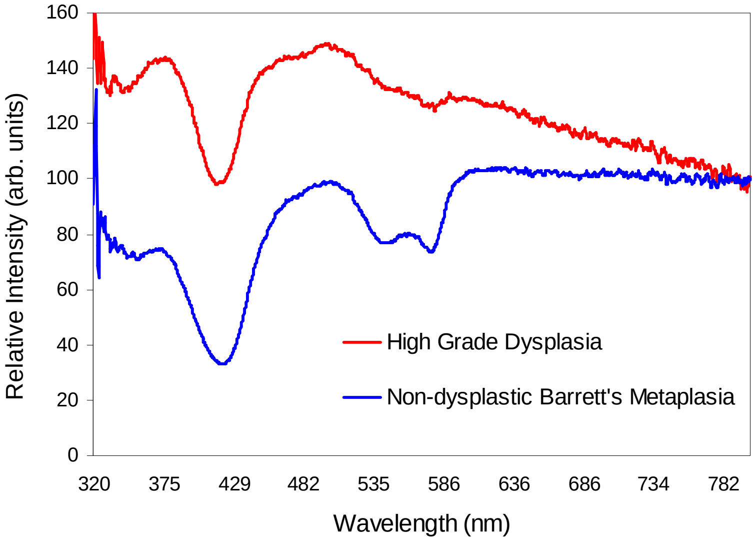


Figure 2.8: Representative spectra from patients with Barrett’s oesophagus.

In diagnosing high grade dysplasia (HGD) in Barrett’s oesophagus (BE) (which is the current most robust predictor of future cancer risk in patients with BE and is described in Chapter 4), nuclear size has been shown to change by up to 50% in HGD compared to low grade dysplasia (LGD) nuclei (Polkowski *et al*. 1998). Cellular packing and nucleus:cytoplasmic ratio are also used as criteria by histologists for the definition of HGD. Representative spectra from BE data are shown in Figure 2.8. The absolute values in the ultraviolet (reflecting haemoglobin absorption peaks) and the relative gradient (reflecting some scattering effect) in the red and near-infra-red parts of the spectrum discriminate between nondysplastic and HGD Barrett’s mucosa. Histological findings are correlated with the spectral patterns. A comprehensive study on diagnosis of HGD in BE using ESS is explored in Chapter 4.

2.4 ESS as a point or imaging measurement for cancer detection

With the advantages of being cheap, simple-to-use, robust, with minimal and uncomplicated calibration, and fast in measurement acquisition, ESS is currently usable in the endoscopy room. For any optical diagnostic test used endoscopically, one of the important issues is whether this technique looks at a single point of tissue (an “optical biopsy”) or whether it can survey the whole lining of the gastrointestinal tract, as an endoscope does. ESS, in its present form, is only a point measurement as an *in vivo* optical biopsy technique, but preliminary work shows that it may be possible to develop this concept to a field imaging detection technique. Although simple in principle, this requires the capability to image at many wavelengths simultaneously (or rapidly in succession) over a fairly wide spectral range and to do this fast enough and at low enough cost to be clinically acceptable. If this proves possible in practice, the technique may become an invaluable tool to endoscopists.

In this thesis, ESS is first employed as an *in vivo* manual point measurement in Chapter 4 for clinical diagnosis in Barrett’s esophagus. ESS is then studied as an *ex vivo* imaging system using an ESS scanner in Chapters 5-7 for sentinel lymph node metastases diagnosis in breast cancer. Although the ESS scanner still uses point measurement technology in a somewhat crude fashion, the imaging analysis method developed for that may become a very useful statistical software solution for later use for ESS *in vivo* field detection device technology.

CHAPTER 3

MULTIVARIATE METHODS FOR ESS

This chapter gives a review of the multivariate statistical methods used in this thesis for ESS spectral data analysis. The methods of smoothing, scattering correction and normalization are applied in Chapter 4 and Chapter 7 as standard tools for spectral pretreatments. The methods of principal component discriminant analysis and partial least squares discriminant analysis are used in Chapter 4 as general tools for supervised dimension reduction and classification, and also provide a basis for the partially supervised dimension reduction method developed in Chapters 5, 6 and 7. The methods of data validation and assessment are applied throughout the thesis as criteria for choosing models and assessing classification performance. This chapter is intended not only to provide background knowledge for the following chapters, but also to give a self-contained if brief overview of these interesting fields of research.

3.1 Need for multivariate methods for ESS

As described in Chapter 2, elastic scattering spectroscopy (ESS) measures a mixture of several physical and chemical effects. These effects result in spectra with many overlapping and often broad peaks which are the product of complex patterns of scattering and absorption by numerous structural and biochemical components. It is difficult or impossible to interpret these spectra using univariate methods. Use of multivariate methods is therefore necessary to reveal specific and useful information from ESS spectra. Investigating the association of spectral absorption and scattering properties with specific clinical outcomes from such high-dimensional data for qualitative and quantitative assessment of human tissue is a difficult and challenging task.

3.2 Spectral pre-treatment

Spectral pre-treatment is an essential step for success before constructing an effective regression or classification model. This is especially true in biomedical studies where the material, e.g. human tissue, under study with very complex structures often gives spectra containing much variability that is of little or no predictive value for the property of interest. This section reviews some widely used spectral pre-treatment methods, with a focus on scattering correction and smoothing, aimed at removing noise and other non-useful spectral variation.

3.2.1 Scatter corrections

Diffuse reflectance spectra typically show the sort of variability seen in Figure 3.1 (top panel). There are small differences in shape between the spectra and large differences in level and slope. In most applications the latter variations carry little useful information, and a range of so-called scatter corrections have been designed to remove them. Of course there is always the risk of removing signal rather than noise, but even in this application, where the general level of scattering may be informative, their use appears to be beneficial.

3.2.1.1 Multiplicative scatter correction (MSC)

The multiplicative scatter correction (MSC) method (Geladi *et al.* 1985) corrects differences between samples due to multiplicative and additive effect of light scatter by fitting a linear regression to the relationship between the spectrum of interest and an average spectrum. For different samples different regression lines are interpreted as differences due to scatter effects while the deviations from the regression lines are interpreted as the chemical information in the spectra. The MSC model for each individual spectrum is

$$x_i = a_i + b_i \bar{x} + \varepsilon_i \tag{3.1}$$

where x_i is the $p \times 1$ spectral vector of sample *i*, the scalars a_i and b_i represent the additive effect and multiplicative effect for sample *i*, $\bar{x}$, $p \times 1$, is the mean of the set of spectra, and the error vector, ε_i , corresponds to all spectral effects in the sample *i* that cannot be modeled by an additive and multiplicative constant.

By using MSC transform, the corrected spectral vector x_i^* of sample *i* is expressed as below:

$$x_i^* = (x_i - \hat{a}_i) / \hat{b}_i \tag{3.2}$$

where $\hat{a}_i$, $\hat{b}_i$ are estimated by least squares, and most of the dominating additive and multiplicative effects have been eliminated.

3.2.1.2 Standard normal variate (SNV)

The standard normal variate (SNV) method, described by Barnes *et al*. (1989), is a mathematical transformation method of the spectra used to put spectra on a common scale. Each spectrum is corrected individually by first centring the spectral values (i.e. subtracting the mean intensity value of the whole spectrum from intensity value at each wavelength) and then scaling the centred spectra by dividing by the standard deviation calculated from the individual spectral values. The transformed value is

$$x_i^* = (x_i - m_i) / s_i \tag{3.3}$$

where x_i is the *p*-dimensional spectral vector of sample *i*, m_i is the mean of the *p* spectral measurements for sample *i* and s_i is the standard deviation of the same *p* measurements. The effect of the SNV transform is that, on the vertical scale, each spectrum is centered on zero and varies roughly from -2 to +2.

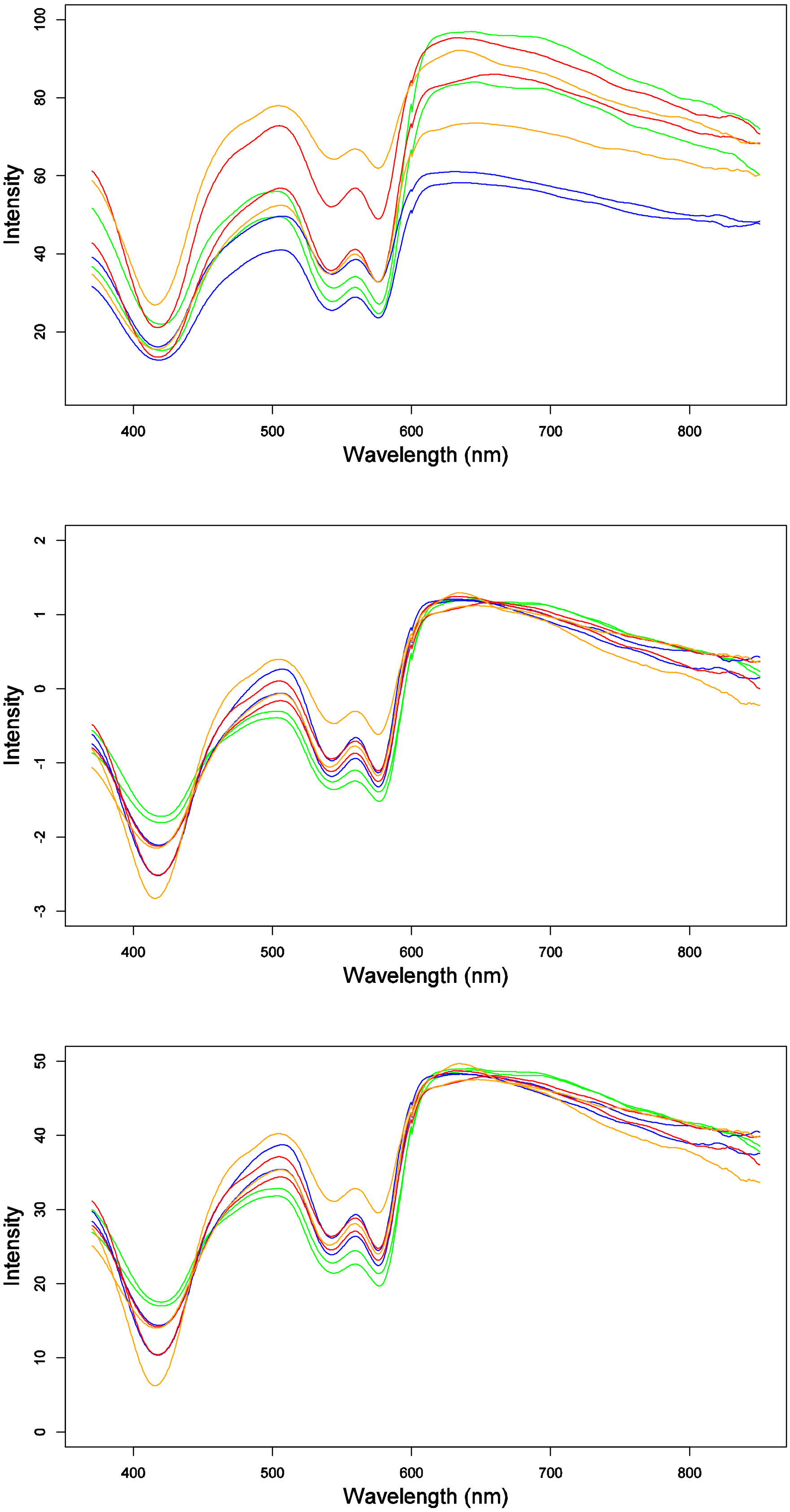


Figure 3.1: Top: the colon spectra before scatter correction; middle: the colon spectra after SNV transform; bottom: the colon spectra after MSC transform.

To compare the effects of these different scatter correction methods, some typical spectra are selected from the colon data that will be described later in Chapter 4. As presented in Figure 3.1 (top), the blue and green spectra are from two different sites of hyperplastic polyps; the red and orange spectra are from adenomatous polyps. The corresponding effects of SNV transform and MSC transform are illustrated in Figure 3.1 (middle and bottom). As can be seen SNV has an effect very much like that of MSC apart from the different scaling. Most of the variation among the spectra is eliminated after transform, i.e. the dominating additive and multiplicative effects have been removed. The main practical difference is that SNV standardizes each spectrum using only the data from that spectrum while MSC uses the mean of a set of spectra. As already noted, there is a risk in using such pre-treatments that we will remove useful information. However, inspection of spectra shows that they contain considerable variability of the sort removed by these pre-treatments, and that this variability appears to be unrelated to the diagnosis of the tissue. For the reasons of normalizing the scale, throughout the thesis we use SNV as a spectral normalization step.

3.2.2 Derivatives

Spectral derivatives are widely used to remove variable baselines and slopes from spectra. The first derivative spectrum is the slope at each point of the original spectrum. It has peaks where the original has maximum slope, and crosses zero at peaks in the original. The second derivative is the slope of the first derivative. It is more similar to the original in some ways, having peaks in roughly the same places, although they are inverted in direction. It is a measure of the curvature in the original spectrum at each point. Taking the first derivative removes an additive baseline. The spectra parallel to each other but shifted upwards or downwards would have the same first derivative spectrum since the slope would be the same everywhere. Taking a second derivative removes a linear baseline.

As the measured spectrum is not a continuous mathematical curve, but a series of measurements at equally spaced discrete points, the obvious way to calculate derivatives with such data would be to use differences between the values at adjacent points. In practice it is almost always necessary to incorporate some smoothing (for instance, Savitzky Golay smoothing described below) into the derivative calculation.

3.2.3 Smoothing

Savitzky Golay smoothing, described by Savitzky and Golay (1964), essentially performs a local polynomial regression of degree k on at least k+1 equally spaced points to determine the smoothed value for each point. It is also used to compute smoothed first and second derivatives.

A narrow window (say 7 to 20 points) is taken centred at the wavelength of interest and a low-order polynomial is fitted to the data points in the window using least squares. The choice of window size (or fitted length) involves a trade-off between noise reduction or smoothing (for which we want as wide a window as possible) and distortion of the curve (which will happen if the window is too wide). The effect of using too wide a window is to round off the peaks and troughs. One way to find a suitable width is to start with three points and increase the window width until the smoothed (fitted) spectra are not visibly noisy. The smoothed value or derivative is then computed from the polynomial.

The Savitzky-Golay approach is preferable to the use of simple moving averages, tending to preserve features of the spectra such as relative maxima, minima and width. If the two methods are compared on artificial examples it is clear that Savitzky-Golay gives a less distorted estimate of the original spectra.

3.3 Dimension reduction

Dimension reduction plays a key role in constructing a multivariate classification/dicrimination model by reducing a large number of spectroscopic variables, say 1800, to, a small number of variables, say 10, with little or no loss of information. Using linear discriminant analysis (LDA), p+1 samples can be classified by p variables perfectly into two groups. However, for ESS spectra the number of variables (wavelengths) of the spectral matrix X often exceeds the number of samples. Directly applying LDA to the spectral data causes an exact collinearity problem. Even if the number of samples exceeds the number of variables there is still a serious problem: the same or similar spectral information may be reflected at different wavelengths, which means some of the variables can be written approximately as linear functions of other variables. All these lead to exact and near collinearity among the variables in the matrix X and LDA runs into the same problem as multiple regression in this situation. Directly employing the unreduced spectral data in LDA allows too much scope for discrimination to be achieved by chance, in directions that mainly represent noise. Overfitting thus happens easily and prediction performance may be unstable and poor.

Two kinds of approaches of solving the collinearity problem are commonly used to reduce the dimensions of the spectral data. They are the methods using regression or classification for a few selected variables, and the methods constructing new components and factors by a few linear combinations of the original variables in the matrix X. Throughout the thesis the latter kind of dimension reduction methods is most explored for tackling the collinear problem in high-dimensional spectral data. The former kind of dimension reduction method by wavelength selection is mentioned with an example in Section 4.8.4 of Chapter 4.

3.3.1 Principal component analysis (PCA)

Principal component analysis (PCA) is the most common dimension reduction technique. From a large number of variables measured on a given set of samples, it extracts a small to moderate number of new variables which account for most of the variability between samples. The new variables, called principal components, are linear combinations of all the original spectral measurements and are uncorrelated to each other. The first principal component captures as much as possible of the variability in all the original ones, and each successive principal component accounts for as much of the remaining variability as possible. A very large of proportion of the variability in high-dimensional data may often be described with a modest number, *a*, of these new variables. Mathematically, PCA relies on an eigenvector decomposition of the covariance matrix of the original data matrix. The eigenvectors then become the weight vectors for construction of the new variables, and the corresponding eigenvalues tell how much of the original variance has been captured in each new variable. The decomposition of spectral matrix X by PCA can be written as

$$X = \hat{T}\hat{P}^T + \hat{E}_X \qquad (3.5)$$

where $\hat{T}$ is score matrix whose columns are the *a* most dominating principal component scores of *X*, $\hat{P}$ is the corresponding matrix of loadings and residual matrix $\hat{E}_X$ is the unexplained part of X. $\hat{P}$ is an orthogonal matrix whose columns are the unit length eigenvectors of the matrix $X^T X$, ordered according to the magnitude (value) of the corresponding eigenvalues. The scores matrix $\hat{T}$ can easily be found by $\hat{T} = X\hat{P}$. More details on PCA can be found in Mardia *et al*. (1979).

PCA often works well in spectral analysis. However there is no guarantee that the scores with large variance are necessarily the best new variables for classification. In ESS spectroscopy it might well be that the large variation is due to scatter that reflects measurement artefacts rather than the variation of variable of interest. Large variation due to measurement variability may lead to principal component scores that are poor classifiers. This is studied in Chapter 4 with applications to early cancer and precancerous detection in Barrett’s esophagus and colon lesions.

3.3.2 Partial least squares (PLS)

Partial least squares (PLS) (Martens and Næs 1998) is a dimension reduction technique that seeks to find a small to modest number of latent variables that maximize the covariance between these new spectral variables and the response variable *y*. Each latent variable is obtained by maximizing the covariance between *y* and all possible linear functions of X , which leads to variables more directly related to variability in y than the principal components.

PLS decomposes X and y into the form

$$X = \hat{T}\hat{P}^T + \hat{E}_X \tag{3.6}$$

$$y = \hat{T}\hat{q} + \hat{E}_Y \tag{3.7}$$

where $\hat{T}$ is a matrix of the extracted *a* score vectors, $\hat{P}$ is a loading matrix, $\hat{q}$ is a loading vector, and $\hat{E}_X$ and $\hat{E}_Y$ are residual matrices. The direction of the first PLS component denoted by $\hat{w}_1$ (with unit length and called the first loading weight vector) is obtained by maximizing the covariance criterion. The scores along this axis are computed as $\hat{t}_1 = X\hat{w}_1$. By regressing X on $\hat{t}_1$ and y on $\hat{t}_1$ the loading vector $\hat{p}_1$ and the regression coefficient $\hat{q}_1$ can be obtained. The product of $\hat{t}_1$ and $\hat{p}_1$ is then subtracted from X , and $\hat{t}_1\hat{q}_1$ is subtracted from *y*. The second direction is found in the same way as the first, but using the residuals $X - \hat{t}_1\hat{p}_1^T$ instead of the original data X . The process is continued in the same way until the desired number of components, *a*, has been extracted.

The Equations (3.6)-(3.7) are sometimes called bilinear since they are linear in both loadings and scores. The estimated parameters in both (3.6) and (3.7) can be combined into the regression vector to be used in the PLS prediction equation

$$\hat{y} = \hat{b}_0 + X\hat{b} \tag{3.8}$$

where the intercept $\hat{b}_0$ becomes equal to $\bar{y}$ when using mean-centred *X* variables, $\hat{b}$ represents the regression coefficient vector, which is given by

$$\hat{b} = \hat{W}(\hat{P}^T\hat{W})^{-1}\hat{q} \tag{3.9}$$

where the $\hat{W}$ is the matrix of loading weights.

Using factors (PLS components) determined by employing both *X* and *y* in the estimation directly, PLS avoids the dilemma in PCA of deciding which components to later use in regression or classification.

3.4 Multivariate classification

As described in Section 2.4, ESS is sensitive to the changes in the physical properties of tissue such as scatter that reflects nuclear size, density and changes in the nucleus and to the chemistry of hemoglobin. Since these properties change with dysplasia, and there is evidence that normal and abnormal tissue show different spectral patterns of scattering and absorbance, reference pathologies (for normal or abnormal tissues) are found to correlate with the spectra. Appropriate multivariate analysis enables ESS to be a useful tool for early cancer or precancerous detection, by investigating the association of spectral absorption and scattering properties with specific clinical outcomes. Qualitative classification is explored in this thesis since the quantity (reference pathology) to be predicted is not a continuous measurement value, but a categorical group variable. The problem studied in this thesis is to assign samples to classes or groups on the basis of spectral measurements taken on the samples. It means deciding whether a particular sample tissue is normal or abnormal using a spectral measurement. Two different types of classification are addressed in this chapter: supervised and unsupervised classification. Supervised classification is applied in Chapter 4 for clinical diagnosis of cancer or pre-cancerous change in Barrett's esophagus and colon lesions. A partially supervised method developed from both supervised and unsupervised classification models is introduced in Chapter 5-6 and are exploited in Chapter 7 with an application to sentinel lymph node diagnosis. General knowledge on multivariate classification can be found in textbooks by Næs, Isaksson, Fearn and Davies (2002), McLachlan (1992), Ripley (1996) and Mardia *et al.* (1979).

3.4.1 Supervised classification/ Discrimination

Supervised classification, also known under the name of discriminant analysis or supervised pattern recognition, is a class of methods primarily used to learn classification rules from training data belonging to known groups. These rules are later used on test data to assign new and unknown samples to the most probable groups.

3.4.1.1 Linear discriminant analysis (LDA)

The classical method of linear discriminant analysis (LDA), described by Fisher for two classes and extended to more by Rao, is called Fisher's linear discriminant analysis or canonical variate analysis (CVA). It seeks a linear combination $X\hat{a}$ of the spectral variables which maximizes the ratio of between-group variance to within-group variance, which is equivalent to finding a direction defined by the vector $\hat{a}$ in multivariate space, that maximises the quantity

$$\hat{a}^T B\hat{a} / \hat{a}^T W\hat{a} \tag{3.10}$$

where W is the within-group sum of squares and products (SSP) matrix, and B is the between-group SSP matrix, defined by

$$W = \sum_{j=1}^{G} (X_j - 1\bar{x}_j^T)^T \ (X_j - 1\bar{x}_j^T \ ) \tag{3.11}$$

$$B = \sum_{j=1}^{G} N_j (\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})^T \tag{3.12}$$

where X_j is the $n \times p$ spectral matrix for samples from group *j*, $\bar{x}_j$ is the average spectrum for samples from group *j*, $\bar{x}$ is the average spectrum for the whole dataset, N_j is the sample size of group *j* ($N = \sum_{j=1}^{G} N_j$), and 1 is an $n \times 1$ vector of 1's. The *W* and *B* matrices are proportional to the within-group and between-group variance matrices, the proportionality constant having no effect on the solution of Equation (3.10).

The definition of *B* in Equation (3.12) weights the groups by their size N_j in the training set. An alternative is to weight the samples in group *j* by $N\pi_j$, where π_j is a prior probability of group *j*, in forming B, W and $\bar{x}$.

The vector *a* can be found as the eigenvector of $W^{-1}B$ which corresponds to the largest eigenvalue, called the first canonical variate (CV). When there are more than two groups, we need more than one discriminant axis (canonical variate) to discriminate between them. In the general case, to discriminate between *G* groups, we normally need *G*-1 canonical variates which correspond to the successive eigenvectors of $W^{-1}B$.

The value of the new variable, the canonical variate, $z = X\hat{a}$ can then be calculated and used for assigning samples to groups. A cutoff point is needed by the discriminant rule. The obvious choice is the value of *z* midway between the two mean *z* values for the two groups case. Other choices of cutoff can be set to favour either higher specificity or higher sensitivity in some specific clinical diagnosis situations, as is the case in the application of Barrett’s data and colon data in Chapter 4.

An alternative approach to develop a discriminant rule is by use of Bayes’ rule. Based on the normal distribution assumption with identical covariance matrices of the two (or more) groups, the Bayes’ rule derived from the posterior probability of the groups is to allocate a new sample with measured vector *x* to the group *j* with the smallest value of

$$L_j = -2\log p(j\,|\,x) = (x - \bar{x}_j)^T \hat{\Sigma}^{-1}(x - \bar{x}_j) - 2\log \pi_j \tag{3.13}$$

where $\bar{x}_j$ is the sample mean vector for group *j*, π_j is the prior probability for group *j*, and $\hat{\Sigma}$ is the pooled within-groups sample covariance matrix. The first term on the right of Equation (3.13) is known as the Mahalanobis distance from *x* to the group mean. It can be seen that the difference $L_j - L_k$ (for groups *j* and *k*) can be reduced to a linear function of *x*, that is, in the two groups case the allocation rule is indicated by a straight line separating the two groups, which gives the name LDA.

When the covariance matrices $\hat{\Sigma}_j$ are different for different groups, Bayes’ rule is to allocate a new sample *x* to the group for which

$$L_j = -2\log p(j\,|\,x) = (x - \bar{x}_j)^T \hat{\Sigma}_j^{-1}(x - \bar{x}_j) + \log\left|\hat{\Sigma}_j\right| - 2\log \pi_j \tag{3.14}$$

is the smallest. This is called quadratic discriminant analysis (QDA), because we now have a quadratic function of *x* for the difference $L_j - L_k$. Thus curved lines in the multivariate space are used to separate the groups of samples.

It can be shown that using LDA based on Bayes' rule by L_j in (3.13) to assign to groups is equivalent (when the prior probabilities are equal) to using the canonical variable *z* with a cutoff midway between the scores of the two group means when they are projected onto the CV. Thus LDA based on Bayes' rule gives another derivation of what is the same way of assigning to groups as Fisher's LDA. Using the pooled covariance matrix, LDA often gives robust and reliable, though not always optimal, classification results, even for training data sets with moderate sample size and covariance matrices that may not be equal. In this thesis the discriminant analyses are based on Fisher's LDA which is available in R.

3.4.1.2 Principal component discriminant analysis (PCDA)

Principal component discriminant analysis (PCDA) is used for classification on the spectroscopic data with large numbers of variables. This involves a PCA on the spectral data as a data reduction step followed by a LDA as a classification step (McLachlan 1992). Using PCA the variance in the original high-dimensional spectral variables can be effectively captured in 10 or 20 PCs with little or no loss of information. Using the scores on these 10 or 20 PCs as the raw data for input to a discriminant analysis, we find canonical variates to display the data and use discriminant functions to assign new samples. PCA projects the raw data onto a low dimensional space composed of 10-20 principal components capturing a mixture of between- and within-group variance. The data separation along individual PCs might be very poor unless this combined variation is dominated by the variation between groups. LDA projects the data onto the axis of canonical variate, along which the separation reaches its optimum by maximizing the ratio of between-group to within-group variance.

The PC scores are linear combinations of the original spectral variables via the loading matrix $\hat{P}$, and the canonical variate scores are, in turn, linear combinations of the PC scores defined by the vector $\hat{a}$ in Equation (3.10), so that by combining the loadings from the two steps we can find the effective loading, $\hat{P}\hat{a}$, called the LDA loading here, of each of the original spectral variables in constructing a canonical variable.

3.4.1.3 Partial least squares discriminant analysis (PLSDA)

An alternative approach is partial least squares discriminant analysis (PLSDA) (Barker and Rayens 2003). This involves a PLS regression on the spectral data using as y a dummy variable, typically coded as 0/1 in the two-group case. This preliminary data reduction step is followed by LDA, taking the scores on the PLS factors as input to the LDA.

In the same way as for the PCDA loading, combining the loadings from the PLS regression and LDA gives the PLSDA loading of the original variables in constructing a canonical variate. Both PLSDA and PCDA loadings show the contribution of each wavelength to the classification and may enable the interpretation of the canonical variate.

Though PCDA is better understood than PLSDA, PLSDA sometimes appears to do a better job than PCDA (though the difference is not usually large), giving comparable prediction results but using fewer optimal number of components in constructing the canonical variate. However, when the difference between two groups is subtle and there are large variabilities existing within the groups, PLSDA is not necessarily better than PCDA. Examples and comparisons will be given in Chapter 4.

3.4.2 Unsupervised classification

Unsupervised classification, also known as cluster analysis, is a class of methods used to identify groups from spectral patterns of the samples without the knowledge of group assignments of samples and possibly even without knowing the number of groups.

The aim of this technique is to find or identify tendencies of samples to cluster in subgroups, which may be very useful at an early stage of an investigation. It uses distances between the objects to identify samples that are close to each other. Before the actual computation of distances PCA is often performed for visual inspection and understanding of the group structure. A more comprehensive and general description of cluster analysis can be found in Kaufman and Rousseeuw (1990) and Mardia *et al*. (1979).

An alternative approach to unsupervised classification is based on parametric mixture models (Titterington *et al*. 1985, McLachlan and Peel 2000, Fraley and Raftery 2002). The idea is to model the distribution of the diagnostic variables by a mixture of parametric distributions, simultaneously estimating with the parameters of the distributions and the group membership of the samples. More details on finite mixture model are addressed in Chapter 5 with applications to sentinel lymph node diagnosis given in Chapter 7.

3.5 Validation and assessment

When doing classification and prediction, the performance of a rule or formula is typically estimated using subsampling (sample-reuse) techniques when the sample sizes are not large enough to use a separate withheld test set. Such methods include *k*-fold cross-validation, leave-one-out cross-validation, bootstrap methods and repeated holdout (Dudoit *et al.* 2002, Braga-Neto and Dougherty 2004, Molinaro 2005). The estimates of the performance are then used for model selection, parameter tuning, feature selection, or performance evaluation in empirical comparisons of discrimination methods.

3.5.1 Repeated holdout

Repeat holdout, also known as random splitting, isolates a proportion *a* of the data into the test set and uses the remaining proportion 1-*a* of the data for training. The commonly used technique is to choose 1/3 as proportion for *a*. To ensure statistical robustness, that is to say, to ensure that the results obtained are not training-set specific and to lower the variance of area under the curve (AUC) described in Section

3.5.3, this process is repeated a large number of times with different randomly selected training and test sets. The final estimate is calculated by averaging the estimates from different runs. To reduce the bias, stratified repeated holdout is suggested to sample independently from each of the classes in the test sets using the proportion as close as possible to the overall class proportions. In our implementation, the holdout method used in Chapter 4 divides the data into two sets: 80% of the total amount of data is used for training and the remaining 20% is used for testing, with 50 replications.

3.5.2 Cross-Validation

Two kinds of cross-validation are commonly used: they are k-fold cross-validation and leave-one-out cross-validation (Stone 1974).

In *k*-fold cross-validation the whole set of dataset *X* is randomly or systematically divided into *k* subsets $X_1,\ X_2,\ldots,\ X_k$. After the division each subset in turn is used for testing and the remaining subsets are combined and used for training. This means that in *k*-fold cross-validation the training and testing is repeated *k* times. Final estimation, in our case a measurement of classification accuracy, is calculated by averaging the estimates from each of the training and testing runs. To guarantee that each class is properly represented in both training and testing sets, we use stratified *k*-fold cross-validation which means that the division into *k* subsets takes the classes of samples into account. In our case, each cross-validation subset contains the same proportion of samples from different classes as the original data set.

Leave-one-out cross-validation (LOOCV), also known as complete cross-validation, can be represented as *n*-fold cross-validation where *n* is the total number of all samples in the original data set. Leave-one-out cross-validation does not use random splitting of the whole dataset because all of the samples are used alternately for testing and the remaining for training. A problem with the method is its time consumption. The computational costs may become large, if complete cross-validation is applied to large datasets.

In this thesis LOOCV are used in Chapter 4.

3.5.3 ROC curve and AUC

For comparing and assessing the performance of classification algorithms, the receiver operating characteristic curve (ROC) and the associated area under the ROC curve (AUC) are frequently used measures of performance.

The ROC shows the trade-off between sensitivity and specificity for a two-class classifier or diagnostic system by generating a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for the different possible cutoffs of a diagnostic test. It has long been used in medical diagnosis (Swets 1988) and has become widely used in evaluating machine learning algorithms.

The AUC (Bradley 1997, Hanley and McNeil 1982) summarises the ROC and provides a single measure of the performance of a classifier and the discriminability of a model: a random model has an AUC of 0.5 and a perfectly discriminable model has AUC of 1.0. The AUC has an important statistical property: the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. This is equivalent to the Wilcoxon test of ranks (Hanley and McNeil 1982).

When combining the results of the replicates of a validation procedure to compute an overall estimate of the ROC curve or AUC, there are two main approaches: the pooling and averaging strategies (Swets and Pickets 1982, Bradley 1997, Fawcett 2004, Witten and Frank 2005).

The pooling strategy involves collecting the classifier scoring outputs determined on each test set， and a series of sensitivities and specificities are found with varying cut off scores to generate a ROC curve and then calculate the AUC. Here the pooling strategy for the ROC curve and the AUC calculation assumes that the classifier outputs across all the replicates are comparable and thus can be globally ordered, this is often not the case and can lead to test samples in different replicates being ranked in the wrong order. The pooling strategy is often used in cross validation, and particularly for LOOCV it is the only method of measuring the AUC and calculating the ROC curve.

In the averaging strategy, for each replicate a series of sensitivities and specificities for each cut off canonical score are used to trace a ROC curve. A separate AUC is computed for each test set, and the mean of these AUCs over all the replicates is calculated with its standard deviation; for each replicate the cut off score discriminating two classes from test set can be adjusted to meet the same criteria (either giving a balanced sensitivity and specificity, or favoring a fixed value of sensitivity or specificity). Averaging sensitivity and specificity from all the replicates gives the mean sensitivity and the mean specificity with its standard deviation. The mean sensitivity, mean specificity and the average AUC are then used as measurements to assess the performance of the classification. The problem with this averaging approach is that it does not give an average ROC curve. The averaging strategy is often used in repeated holdout validation.

Both strategies are used in practice and the current literature is equivocal about which approach is to be recommended. Both approaches are described as valid estimates of AUC and ROC curves (Witten and Frank 2005, Fawcett 2004, Flach 2004). Witten and Frank (2005) noted that the pooling strategy has the advantage that it is easier to implement. Also, it is expected to have a lower variance assuming that the results of each replicate are samples from the same population (Swets and Pickets 1982). However in practice, this assumption is generally not valid for cross-validation and can lead to large pessimistic biases.

CHAPTER 4

SPECTRAL PRETREATMENT

4.1 Introduction

This Chapter is based on the paper "Error removal by orthogonal subtraction (EROS): a customised pre-treatment for spectroscopic data" (Zhu, Fearn *et al.* 2008).

When reflectance spectra are used for quantitative analysis, data pre-treatment can be an important step in the construction of an effective regression or classification model. The ESS spectra often contain major sources of variation that are of little or no predictive value for the property of interest and may thus hinder the development of a diagnostic model. An appropriate pre-treatment can help to reduce this variability. For example, when diffuse reflectance spectra are measured on powdered or ground samples, the variability in the spectra between samples is typically dominated by scatter effects caused by variations in particle size. Pre-treatments such as multiplicative scatter correction (MSC) (Geladi *et al.* 1985, Næs *et al.* 2002) are commonly applied to the spectra to reduce this variability. This allows the spectral variability due to differences in the chemical composition of the samples, which is usually of primary interest, to be seen more easily, both by eye and by the chemometric approaches commonly used for calibration in these situations.

Although off-the-shelf pre-treatments such as MSC are very useful, there may be situations where it is desirable to construct a pre-treatment that is tailored to remove variability from a specific identified source. The work described here was motivated by the variability observed in ESS spectra measured using a probe in physical contact with human tissue. When the measurements were replicated by removing and then replacing the probe in an attempt to take a spectrum at exactly the same point there was considerable variability between the replicate spectra. Small changes in angle and pressure are amongst the possible causes of this variability.

In this chapter a customized pre-treatment called EROS, for Error Removal by Orthogonal Subtraction, was constructed to remove this measurement variability before the development of a diagnostic model. The approach was explored in two applications involving *in vivo* measurement of dysplasia in Barrett's oesophagus and *in vivo* measurement of polyps in the human colon. To understand the measurement variability removed by EROS, an experiment was carried out to find how pressure and angle of the probe influence the variations in measured spectral data collected under controlled conditions. An independent dataset was also used to test the model prediction performance.

4.2 Measurement variability

Two major potential sources of measurement variability when collecting data *in vivo* include differences in pressure of the probe on the tissue and the angle at which the probe is held to the tissue. The importance of pressure applied and the angle of approach to the *in vivo* BE have been noted before (Gonzalez-Correa *et al.* 2000). As pressure and angle of the probe are difficult to control by hand and even small changes of these factors may have a big influence on the spectra, it may be beneficial to remove or at least reduce this influence with a pre-treatment.

4.3 Spectral pretreatment method

To ameliorate the effects of measurement variability, a customised spectral pre-treatment method called error removal by orthogonal subtraction (EROS) is proposed here. EROS projects the spectra onto a subspace orthogonal to the subspace in which most of the measurement variability occurs, so that the subsequent classification step is robust to this measurement variability.

4.3.1 Orthogonal projection

The orthogonal projection used by EROS to subtract interfering variability works as follows. Let X be an $n \times p$ matrix whose rows are the spectra of n samples, and let P be a $p \times k$ matrix whose columns are mutually orthogonal vectors of unit length. Then

$$\tilde{X} = X\,(I - PP^{\mathrm{T}})\,, \qquad (4.1)$$

where I is the identity matrix, is the projection of X onto the $p - k$ dimensional subspace orthogonal to the columns of P. Writing this as $\tilde{X} = X - X_P$ where $X_P = XPP^{\mathrm{T}}$ emphasizes that this is a subtraction from X of the part X_P that lies in the subspace defined by P.

A number of existing approaches employ this simple method of subtracting variability, the differences between them being in the way P is chosen. It is useful to distinguish between methods that use the dependent variable y in the construction of P and those that do not. The former are an integral part of the calibration process, with important implications for validation, while the latter can be regarded as pre-treatments.

4.3.1.1 Approaches that use y

The original method of this type is orthogonal signal correction (OSC), which was introduced by Wold *et al*. (1998) and developed by Sjöblom *et al*. (1998), Fearn (2000), Westerhuis *et al*. (2001) and Trygg and Wold (2002) among others. The idea of OSC and related methods is to construct factors that account for a large amount of spectral variation whilst being uncorrelated with the dependent variable y. The loadings for these factors form the columns of P, and this variability is subtracted from X. The details vary, and some of the procedures involve the sequential subtraction of single dimensions, but the basic idea of subtracting directions uncorrelated with *y* is common to all these approaches. The involvement of *y* means that OSC, in all its versions, is part of the calibration process, as indeed it is explicitly presented in the OPLS method of Trygg and Wold (2002). One important consequence is that it is vital that the construction and subtraction of OSC factors needs to be included in any cross-validation procedure, and not done outside the loop, or there is a risk of overfitting. The main advantage usually claimed for OSC is that its use with PLS improves the interpretability of the PLS model, because the PLS factors do not need to account for the interfering variability that has been subtracted.

4.3.1.2 Approaches that do not use y

A very different way of selecting directions to subtract from *X* is used by independent interference reduction (IIR), Hansen (2001), and external parameter orthogonalization (EPO), Roger *et al*. (2003). Both IIR and EPO use spectra from a set of dedicated experiments to identify structured variations resulting from interfering factors, and choose *P* to be an orthogonal basis for the space in which these variations lie. They do not use *y*, and therefore are genuine pre-treatments.

In IIR, which builds on work by Sun (1997), spectra are measured on samples that do not contain the analyte of interest, but are otherwise variable. PCA is applied to the matrix *Z* of these spectra to find the principal dimensions in which they vary, and the loading vectors of the first *k* principal components become the columns of *P*. IIR is an attempt to identify and remove all sources of interfering variability from the spectra. The success of this is clearly dependent on the extent to which *Z* represents these sources of variability, and implies that it must include a substantial number of samples. It will, of course, not always be possible to find such samples.

In contrast, EPO is more focused on removing interfering variability from one particular source. In the EPO experiment a small number of samples are taken and each is measured under a range of conditions. For example Roger *et al*. (2003) measured the samples at different temperatures, and Andrew and Fearn (2004) measured the samples on different instruments. A matrix *Z* of difference spectra may then be constructed in which between-sample variability has been eliminated and which represents variability due to the changing conditions. The columns of *P* are then the loadings for the first *k* principal components of *Z*. In the applications described the results were calibrations robust to sample temperature (Roger *et al*. 2003) and calibrations transferable between instruments (Andrew and Fearn 2004), but the idea could be applied in any situation where the required experiment is feasible.

4.3.2 Error Removal by Orthogonal Subtraction (EROS)

A method called error reduction by orthogonal subtraction (EROS) is proposed to ameliorate the effects of measurement variability in *in-vivo* ESS spectra. EROS also uses the projection in Equation (4.1) to subtract variability from the spectral matrix *X*, but introduces yet another way of constructing *P*. In some situations there can be considerable variability between the spectra obtained in replicate measurements of what is nominally the same sample. A common way of dealing with this is to take several replicates and average them to reduce this variability. If the variability behaves like white noise then this is probably the best that can be done. If, however, it has some structure, it may be beneficial to identify this structure and subtract appropriate dimensions by projection.

Suppose we have replicate spectral measurements on each of *m* samples. Let R_i be the $r_i \times p$ matrix whose rows are the r_i replicate spectra for the *i*th sample, and let Z_i be this same matrix after centring by subtraction of the column means, i.e. after subtracting from each row the mean spectrum for the sample. Then

$$W = \sum_{i=1}^{m} Z_i^{\mathrm{T}} Z_i \,/ (r - m), \qquad (4.2)$$

with $r = \sum_{i=1}^{m} r_i$, is the pooled within-sample covariance matrix of the spectra. *W* describes the variability between replicate measurements, the variability between samples having been removed by the centring.

The basic idea of EROS is to use the first k eigenvectors of W, for suitably chosen k, as the columns of P for the projection of Equation (4.1). If the replicate variability contains structure, i.e. if it is concentrated in particular directions in the spectral space, these eigenvectors will lie in those directions, and subtracting them will remove this variability.

4.4 Applications and results

Two applications of EROS pre-treatment to the clinical diagnosis of colon lesions and high grade dysplasia or cancer in Barrett's oesophagus are presented here to illustrate the use of EROS for the construction of an effective classification model.

4.4.1 Detection of cancer risk in Barrett's oesophagus (Barrett's data)

4.4.1.1 Background and introduction

The incidence of oesophageal adenocarcinoma has increased dramatically since the 1970s, and it is now the fifth commonest cause of cancer death in the UK. The five year survival rate for this cancer is less than 10% (Newnham *et al*. 2003, Toms 2004). Barrett's oesophagus (BE) is a pre-malignant condition in which the normal squamous epithelium of the oesophagus is replaced by metaplastic columnar epithelium (Phillips and Wong 1991), increasing the risk of developing adenocarcinoma by 30-125 times (Hameeteman 1989, Williamson *et al.* 1991) when compared to the general population. Systematic endoscopic surveillance of BE has been shown to detect oesophageal adenocarcinoma at an early and curable stage (Levine *et al*. 1993). High grade dysplasia (HGD) is the current most robust predictor of future cancer risk in patients with BE, with around 50% progressing to adenocarcinoma at five years if it is not treated (Rabinovitch *et al.* 2001, Sharma *et al.* 2007)

Endoscopic surveillance relies on regularly spaced, but essentially random, biopsies being taken from the four quadrants of the Barrett’s segment every 2cm. It is time consuming, labour intensive and has a low detection rate for HGD even when abnormalities exist (Sandick *et al.* 1998, Messmann *et al.* 1999). The challenge for clinicians and scientists is to develop new technologies for detecting patients at high risk of progression to cancer. Ideally this would be accurate, easy to use, inexpensive, and provide results rapidly, preferably without the need to remove tissue.

As described in Chapter 2 elastic scattering spectroscopy (ESS) is an *in vivo* optical point measurement which, using an appropriate optical geometry, is sensitive to changes in properties of tissue (Lovat and Bown 2004). The optical probe is passed through the working channel of an endoscope and is placed in direct contact with tissue. A flash of light interrogates a cylinder of tissue approximately 0.5mm in diameter and 1 mm deep. Results are available within milliseconds. Since the technology uses white light and produces a strong backscattered signal, components are inexpensive and the system is simple to manufacture. It is also safe because only visible light is used for illumination with ultraviolet light being filtered out. Many of the features that pathologists look for in diagnosing HGD have also been shown to affect light scattering including the nuclear: cytoplasmic ratio (Drezek *et al.* 1999); the cellular packing density (Wallance *et al.* 2000); and the nuclear size (Zonios *et al.*1999). The nuclear chromatin content has also been shown to affect the spectra of both singly scattered light (Gurjar *et al.* 2001, Beckman *et al.* 2000) and high angle scatter in ESS (Mourant *et al.* 2000).

The problem is how to maximise the discrimination between ESS spectra taken from high and low risk sites in order to accurately detect the patients at high future cancer risk. The difficulty is that the spectral differences between normal and abnormal tissue are very subtle. At the same time, ESS spectra often contain major sources of variation that are of little or no predictive value for the detection of cancer risk. In the clinical setting it is extremely difficult even for experienced endoscopists to accurately control all aspects of ESS spectral acquisition especially with respect to the angle and pressure of the optical probe in relation to the tissue with which they are in contact.

4.4.1.2 Spectral acquisition

This study was approved by the joint University College London/University College London Hospitals (UCLH) ethics committee. During the routine endoscopy, optical measurements were taken, followed immediately by biopsy from the same site. A total of 152 matched optical and histological biopsy sites were collected from 81 patients referred to our tertiary referral centre between 2000-2003 for the management of HGD (high grade dysplasia) or early cancer in BE. Informed consent was obtained from all patients prior to their participation in the study.

Before any tissue spectra were taken, a white reference spectrum was recorded from the flat surface of a Spectralon reference. The ESS spectral data used in our analysis is the ratio of the spectral intensity of backscattered light from the tissue to that of the standard reference spectrum from Spectralon. Each spectrum was made up of 1801 points spanning the wavelength range 320-920nm, which lies in the visible and near infrared (NIR) regions.

Measurements were replicated by removing and replacing the probe in an attempt to take a spectrum at exactly the same point. Repeated measurements were taken from 1-3 biopsy sites per patient with a median of four spectra from each site (mean 3.3). Routine quadrantic biopsies were taken every 2 cm as standard practice for management of patients with suspected HGD in BE. Of the 152 matched optical and histological biopsies (corresponding to 506 spectra), 122 (corresponding to 413 spectra) were from low grade dysplasia or no dysplasia (low risk) and 30 (corresponding to 93 spectra) were from high grade dysplasia or cancer (high risk). Each biopsy was assessed and assigned as either high or low risk by three pathologists who met to agree a consensus in cases of disagreement.

All raw spectra were visually examined for any obvious outliers caused by acquisition errors, poor contact of the optical probe with the tissue, or other artefacts; such spectra were excluded from subsequent analysis.

4.4.1.3 Statistical analysis

Standard data pre-processing was carried out on the spectra to improve signal quality (Næs *et al.* 2002, Lovat *et al.* 2006). The spectra were first smoothed using a Savitzky-Golay filter (Savitzky and Golay 1964), using a 7-point window below 620 nm and a 20-point window above 620 nm where noise was greater. To speed subsequent manipulation, the smoothed data were then reduced from the 1801 points, corresponding to the spectrometer resolution, by taking alternate points only. To remove the regions of the spectra with low signal-to-noise ratios arising from the lower light intensity emitted by the Xenon arc lamp at the extremes of its output spectrum, only the wavelengths between 370 and 800nm, with 637 points, were used in the analysis. Using the standard normal variate (SNV) (Barnes *et al.* 1989) method, the spectra were then normalized by setting the mean intensity of each spectrum to zero and the variance to one. The mean spectral patterns from high risk and low risk sites after standard pre-processing are shown in the left panel of Figure 4.1. It illustrates that both high risk and low risk sites spectra are of very similar shape with the mean spectra showing small between-class differences.

EROS, as described in Section 4.3, was then applied to the spectra, using the replication to derive the projection. Various choices of k, the dimension of P in Equation (4.1), were investigated.

Classification rules were derived from both the original spectra and the EROS pre-treated spectra using principal component discriminant analysis (PCDA), and partial least squares discriminant analysis (PLSDA). The PCDA involves an initial PCA on the pre-treated spectra followed by linear discriminant analysis (LDA) on the first q PC scores. The PLSDA involves a PLSR on the pre-treated spectra followed by LDA on the first q PLS latent variable scores. EROS and PCDA or PLSDA were carried out for a grid of values of k and q, with k ranging from 0 (no dimensions subtracted) to 7 and q from 2 to 30. A method called stratified repeated holdout , described in Section 3.5.1, trained the algorithm on stratified randomly sampled 80%

of the data and tested it on the remaining 20% of the data with 50 repetitions of the split. In each split all the replicated spectra from a particular biopsy site were allocated to either the training set or the testing set (to prevent bias being introduced through non-independence of data) and there were the same proportion of sites belonging to each class in the training set as in the whole data set.

For each test set a separate AUC was calculated, and a specific cut off canonical score discriminating the two classes in the test set was chosen to give the same value of sensitivity in each set. The mean sensitivity, mean specificity and the average AUC with its standard deviation were then calculated from the 50 replications to assess the performance of the classification. As mentioned in Section 3.5.3, this approach does not generate an average ROC curve.

All the above analysis was carried out on a per-site-base since this usually gives more reliable results than a per-spectra analysis. In per-site analysis, if any one of the spectra from a particular biopsy site is classified as a spectrum from a high risk site (i.e. the maximum canonical score of the spectra from a particular biopsy site is higher than the cut off score), the whole biopsy site may be regarded as a high risk one. An alternative criterion is that if the average canonical score of the spectra from a particular biopsy site is higher than the cut off score, the whole biopsy site may be classified as one with high risk. Our experience is that for repeated holdout validation choosing mean canonical score rather than maximum canonical score on a per-site-base can give better classification results. So we chose this mean-score-per-site criteria for the analysis of the Barrett’s data.

All the computation and analysis for EROS, PCDA and PLSDA was performed using the R statistical language.

4.4.1.4 Principal component discriminant analysis (PCDA) model

Table 4.1 shows the repeated holdout validation results for detection of HGD or cancer for various combinations of *k*, the number of dimensions removed by the EROS pre-treatment, and *q*, the number of principal components used in the construction of the PCDA diagnostic rule. The specificity and AUC results are plotted in Figure 4.2. In deriving the rule, the cut off canonical score between the high risk and low risk sites was adjusted to give a sensitivity of at least 90%. This high sensitivity comes at the expense of specificity but it was felt to be a greater omission to potentially miss patients at high risk than to have to collect a few additional biopsies. Conventional biopsies would only be taken if the ESS spectrum was indicative of dysplasia or cancer.

Trying to avoid the temptation to over interpret small differences in performance, it seems reasonable to draw the following conclusions:

- Pre-treatment with EROS substantially reduces the number of PC scores *q* needed to attain the best levels of accuracy in the classification, and this reduction in number is greater than *k*, the number of dimensions removed by EROS. Thus the total number of factors involved, even if one counts those in the EROS pre-treatment, is reduced.
- For this Barrett’s dataset, *k* should be at least 3, with the choices 3, 5 and 7 all being reasonable ones.

Before pre-treatment with EROS (k = 0), the choice q = 20 gives the best results with sensitivity, specificity and AUC of 92%, 63% and 85%, respectively. After pre-treatment with EROS, the combinations of k = 3, 5, 7 with *q* from 5 to 7, all give the better results with fewer factors. Amongst these choices, the combination k = 5, q = 5 gives the best results with the fewest factors, with sensitivity, specificity and AUC of 92%, 75% and 87%, respectively, and subsequent discussion is based on this choice. The two panels of Figure 4.1 show the mean spectra for the high risk and low risk biopsy sites, before and after pre-treatment with EROS, in which k = 5 dimensions were removed. The differences between the means are much more evident in the right hand panel after pre-treatment than they are before pre-treatment.

Figures 4.3-4.4 compare the loadings for the PCDA discriminant functions using $k = 0$, $q = 20$ (no EROS) in the left panel and $k = 5$, $q = 5$ in the right panel. These loadings show the contribution at each wavelength to the linear diagnostic rule, and thus permit interpretation of its spectral basis. After using EROS, and thus being able to attain acceptable levels of accuracy with far fewer factors in the classification model, the loading vector is much less noisy and can be related much more easily to features in the spectra. The first few PCs' loadings of PCDA model after EROS pre-treatment in Figure 4.5 show that the first five to seven PCs are most responsible for both the EROS pre-treated spectral difference between two types and the shape of the spectra. After EROS employing large number of PCs in construction of PCDA model may include interference spectral variation which can be seen from the noisy structure of 8th, 9th PC. This gives some evidence that PCDA model with $q = 5$ or 7 can achieve optimal accuracy after EROS with $k = 5$.

As can be seen in Figure 4.4, the most obvious feature is a large positive PCDA loading in the region of 650~800nm, corresponding to clear differences between the mean spectra, which means the measurements in this range contribute strongly to the classification. Physically this might be explained as the effect of scattering differences between the spectra of normal and abnormal tissues in this region. Absorption also seems to play a key role in the spectral differences and the classification between two types of tissue in some specific region of spectra. The region around the 760nm stands out as the most prominent spectral feature with highest PCDA loading. This is consistent with the finding by Jöbsis (1977) that there exists a broad peak at 760 nm characteristic of deoxyhaemoglobin absorption. Two peaks in the PCDA loading around 540 and 580nm might be due to absorption dips of HbO_2 at 542 and 577nm in the spectra of high risk cancer due to increased Hb presence (as described in Section 2.3.3.1). It is known that cancers and pre-cancerous tissues are characterized by increased microvascular volume, hence increased blood content (Jain 1988).

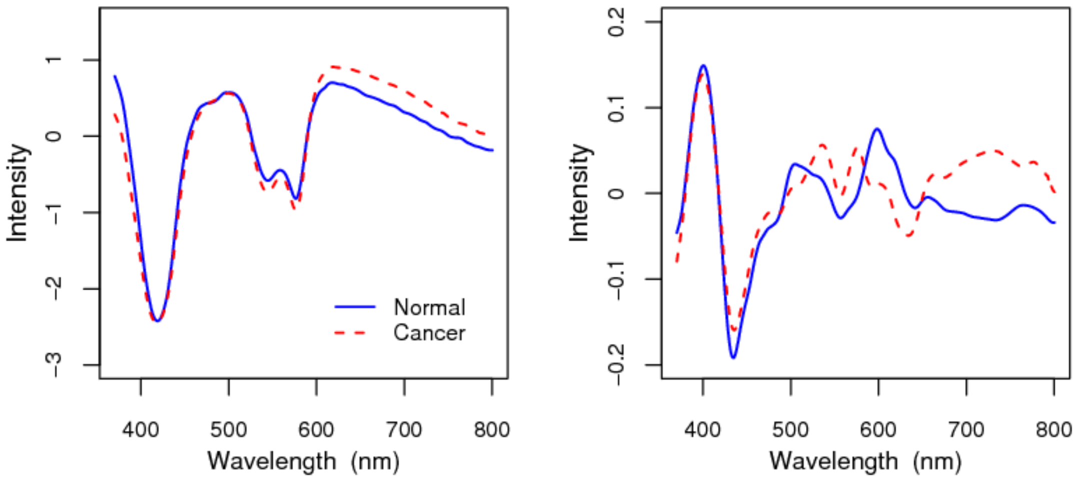


Figure 4.1: Mean spectra from low risk (solid line) and high risk (dotted line) sites of Barrett's data. Left: with standard pre-processing only. Right: with standard pre-processing and EROS ($k = 5$).

The gain in accuracy through the use of EROS is fairly modest one, but the gain in interpretability, and very probably in robustness, of the classification rule is substantial. That the changes are so dramatic is not surprising when one considers that the pre-treatment removes 96.6% of the variability in the spectra. If the information for the classification is in the remaining 3.4%, as it appears to be, it will be much more prominent after pre-treatment. From these result it appears that the EROS pre-treatment has removed much of the undesirable measurement variation, and that the variation left in the spectra is more representative of the spectral features relevant to the detection of high risk cancer. The nature of this measurement variation was explored in a designed experiment described in a later Section 4.5.

Table 4.1: Results of repeated holdout PCDA classification on Barrett's dataset using various combinations of *k* (number of dimensions removed by EROS pre-treatment) and *q* (number of PC scores used in the LDA).

***k* (EROS)**	***q* (PCDA)**	**Prediction Accuracy**			
		Sensitivity (%)	**Specificity (%)**	**AUC.mean**	**AUC.sd**
0	2	91	19	0.61	0.08
0	3	91	16	0.62	0.09
0	5	91	39	0.72	0.11
0	7	91	41	0.72	0.11
0	10	92	34	0.71	0.11
0	15	92	55	0.83	0.10
0	**20**	***92***	***63***	***0.85***	**0.07**
0	30	90	59	0.80	0.09
2	2	91	14	0.67	0.11
2	3	91	38	0.79	0.12
2	5	90	57	0.85	0.09
2	7	92	49	0.83	0.09
2	10	90	57	0.82	0.09
2	15	91	53	0.81	0.09
2	20	91	58	0.82	0.08
2	30	92	59	0.80	0.09
3	2	90	34	0.79	0.11
3	3	90	39	0.83	0.09
3	**5**	***90***	***66***	***0.85***	0.10
3	**7**	***91***	***69***	***0.83***	0.08
3	10	91	59	0.80	0.09
3	15	91	60	0.83	0.10
3	20	92	56	0.82	0.08
3	30	92	52	0.82	0.11
5	2	92	56	0.81	0.09
5	**3**	***91***	***67***	***0.86***	0.08
5	**5**	***92***	***75***	***0.87***	0.08
5	**7**	***90***	***74***	***0.86***	0.08
5	10	93	60	0.84	0.09
5	15	92	62	0.82	0.09
5	20	91	65	0.83	0.10
5	30	92	55	0.79	0.10
7	2	90	48	0.83	0.11
7	**3**	***91***	***66***	***0.85***	0.08
7	**5**	***90***	***67***	***0.85***	0.09
7	**7**	***91***	***66***	***0.85***	0.07
7	10	92	55	0.80	0.10
7	15	92	48	0.76	0.12
7	20	92	43	0.76	0.11
7	30	91	43	0.78	0.09

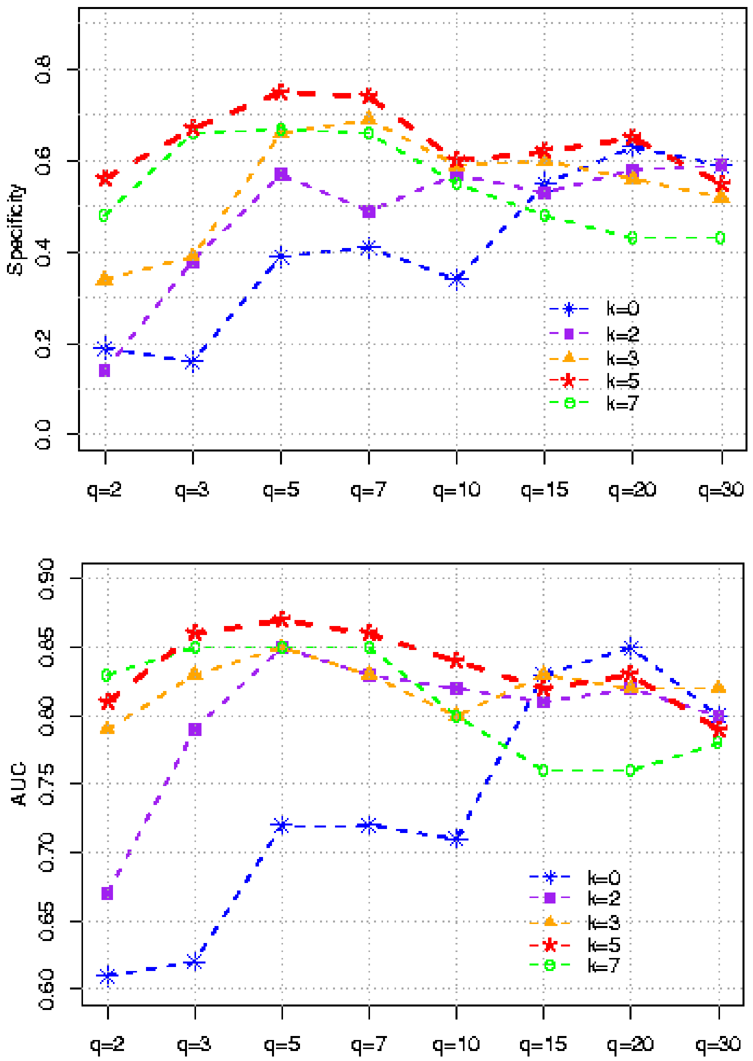


Figure 4.2: Repeated holdout PCDA classification accuracy of Barrett's data as measured by specificity and AUC. The five lines correspond to the choices of k, the x-axis to the choice of q.

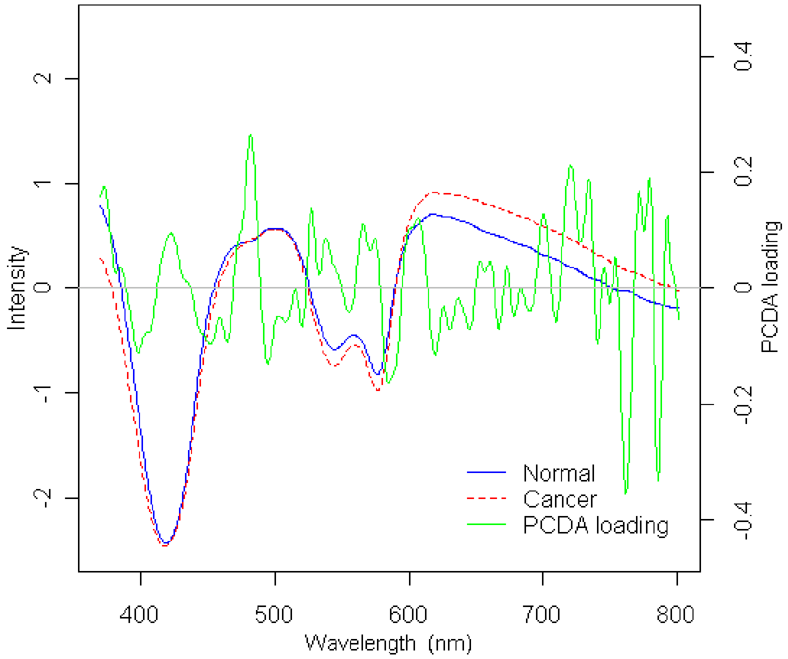


Figure 4.3: PCDA loadings for discrimination between low risk and high risk sites of Barrett's data using model with *k* = 0, *q* = 20 (no EROS). The PCDA loading is shown in green, with the mean spectra for the two types superimposed (low risk solid blue line, high risk dotted red line).

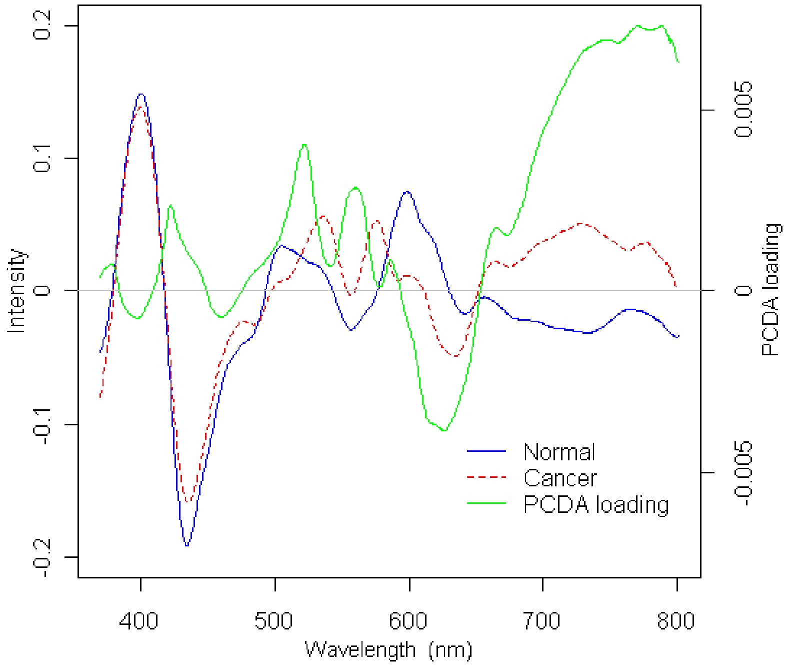


Figure 4.4: PCDA loadings for discrimination between low risk and high risk sites of Barrett's data using model with *k* = 5, *q* = 5. The PCDA loading is shown in green, with the mean spectra for the two types superimposed (low risk solid blue line, high risk dotted red line).

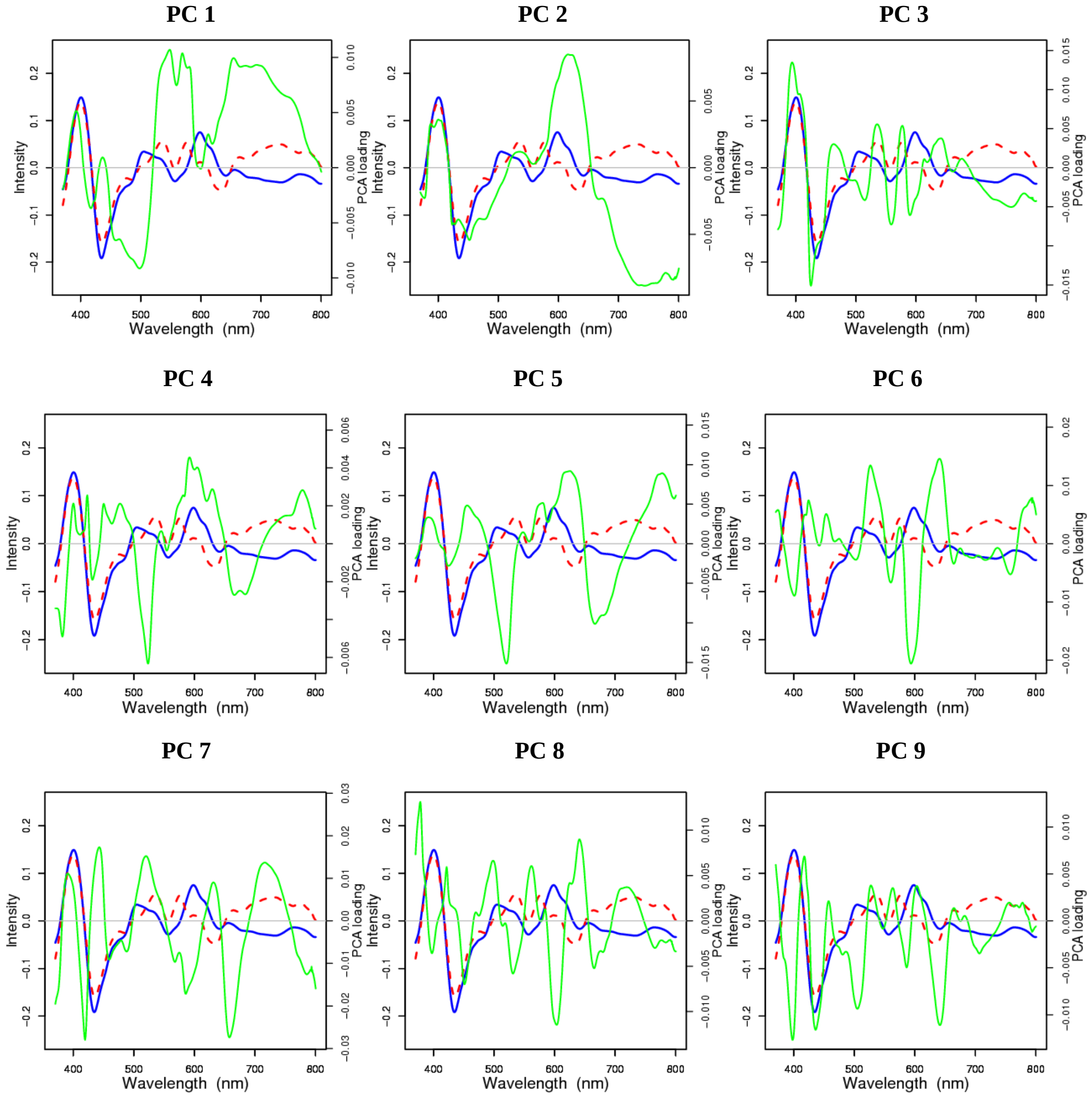


Figure 4.5: The first nine PC loadings based on the EROS pretreated spectra from Barrett's data with $k = 5$. PC loadings are shown in green line, with the mean spectra for the two types superimposed (low risk in solid blue line, high risk in dotted red line).

4.4.1.5 Partial least squares discriminant analysis (PLSDA) model

Table 4.2 shows the repeated holdout validation results on the per-site base for various combinations of k, the number of dimensions removed by EROS pre-treatment, and q, the number of PLS latent variable scores used in the construction of the PLSDA diagnostic rule. The specificity and AUC results are plotted in Figure 4.6. As in the PCDA model, a sensitivity higher than 90% was chosen for all the combinations.

Table 4.2: Results of repeated holdout PLSDA classification on Barrett's dataset using various combinations of *k* (number of dimensions removed by EROS pre-treatment) and *q* (number of PLS scores used in the LDA).

k **(EROS)**	***q*** **(PLSDA)**	**Prediction Accuracy**			
		Sensitivity (%)	**Specificity (%)**	**AUC.mean**	**AUC.sd**
0	2	91	30	0.68	0.08
0	3	91	39	0.68	0.09
0	5	90	42	0.73	0.10
0	**7**	***91***	***53***	***0.79***	0.10
0	10	90	44	0.79	0.11
0	15	91	40	0.78	0.10
0	20	92	41	0.77	0.11
0	30	90	30	0.73	0.11
2	2	90	30	0.73	0.13
2	3	91	44	0.78	0.11
2	**5**	***90***	***61***	***0.84***	0.09
2	7	92	56	0.83	0.09
2	10	92	51	0.77	0.11
2	15	94	51	0.80	0.08
2	20	91	56	0.79	0.10
2	30	91	54	0.78	0.10
3	2	93	43	0.75	0.12
3	**3**	***90***	***64***	***0.83***	0.08
3	**5**	***91***	***66***	***0.84***	0.12
3	7	91	60	0.80	0.08
3	10	91	46	0.80	0.10
3	15	91	50	0.80	0.10
3	20	91	49	0.80	0.09
3	30	90	54	0.81	0.09
5	2	91	60	0.83	0.11
5	**3**	***90***	***72***	***0.87***	0.09
5	**5**	***93***	***65***	***0.81***	0.09
5	7	92	54	0.81	0.09
5	10	92	51	0.81	0.10
5	15	92	47	0.82	0.11
5	20	91	51	0.80	0.10
5	30	91	55	0.79	0.10
7	**2**	***92***	***63***	***0.85***	0.09
7	3	91	60	0.82	0.10
7	5	90	58	0.78	0.09
7	7	92	54	0.79	0.09
7	10	91	45	0.80	0.11
7	15	91	44	0.77	0.11
7	20	91	42	0.77	0.12
7	30	90	50	0.79	0.09

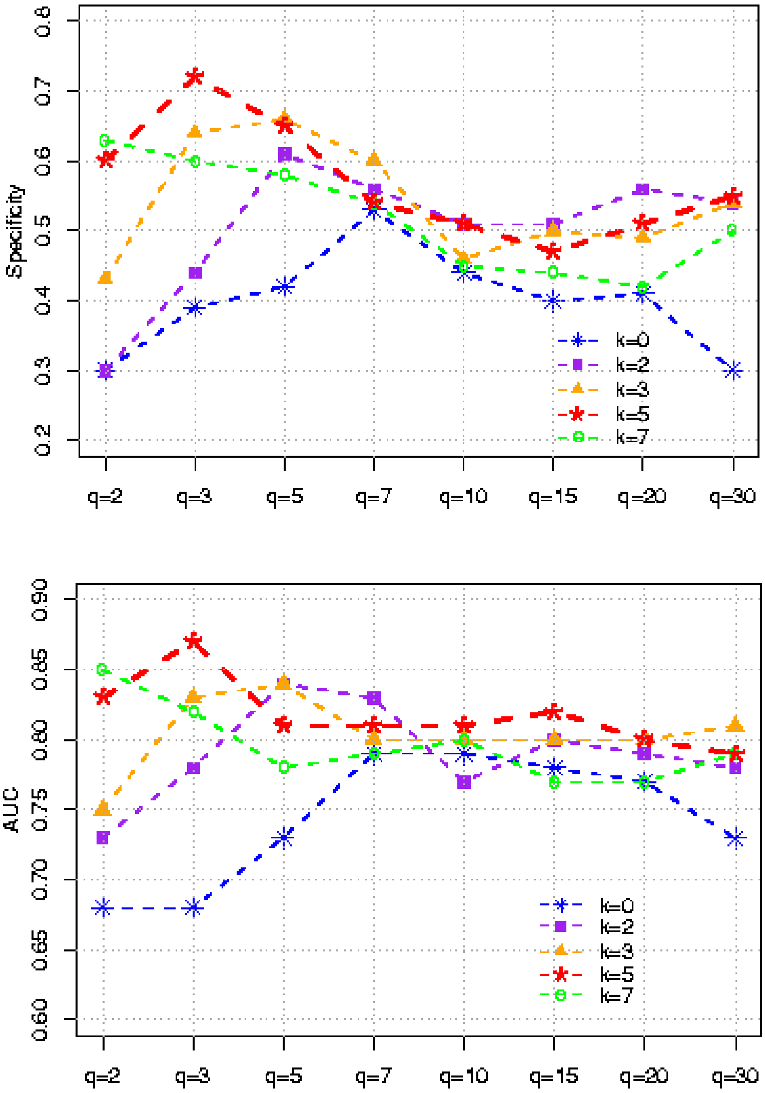


Figure 4.6: Repeated holdout PLSDA classification accuracy of Barrett's data as measured by specificity and AUC. The five lines correspond to the choices of k, the x-axis to the choice of q.

Compared with PCDA, PLSDA uses fewer latent variable components (factors) for classification. Before pre-treatment with EROS (k = 0), the choice q = 7 gives the best results with sensitivity and AUC of 91% and 79%, respectively. But the specificity of 53% is not acceptable. After pre-treatment with EROS, the combination k = 5, q = 3 gives best results using fewer factors, with sensitivity, specificity and AUC of 90%, 72% and 87%, respectively, and subsequent discussion is based on this choice. Other models with combination of k = 3, 5 and q = 3, 5 also give acceptable results. The PLSDA classification accuracy has improved after EROS pretreatment although on the whole the accuracy is not as high as that of PCDA. The right hand panel of Figure 4.7 shows the mean spectra for the high risk and low risk biopsy sites after pre-treatment with EROS. The differences are much more evident than they were before pre-treatment. Figure 4.7 shows the loadings for the PLSDA discriminant functions using k = 0, q = 7 (no EROS) in the left panel and k = 5, q = 3 in the right panel. After using EROS, and thus being able to attain acceptable levels of accuracy with fewer factors in the classification model, the loading vector is much less noisy and can be related much more easily to features in the spectra. Again, the most prominent spectral feature lies in the region around the 760nm with highest PLSDA loading, together with another two peaks of the PLSDA loading around 540 and 580nm. These features are similar to that from the PCDA model. A comparison between the PCDA loading and the PLSDA loading for discrimination between low risk and high risk sites can be seen in Figure 4.8. The PCDA loading vector shows simpler features than the PLSDA loading vector.

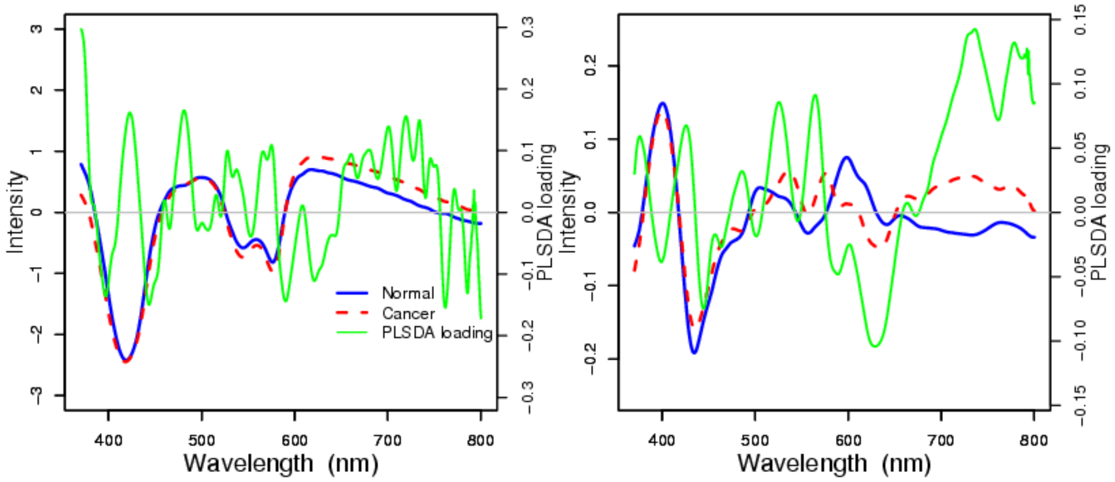


Figure 4.7: PLSDA loadings for discrimination between low risk and high risk sites of Barrett’s data. The PLSDA loading is shown in green, with the mean spectra for the two types superimposed (low risk blue solid line, high risk red dotted line). The left panel is for k = 0, q = 7 (no EROS) and the right panel is for k = 5, q = 3.

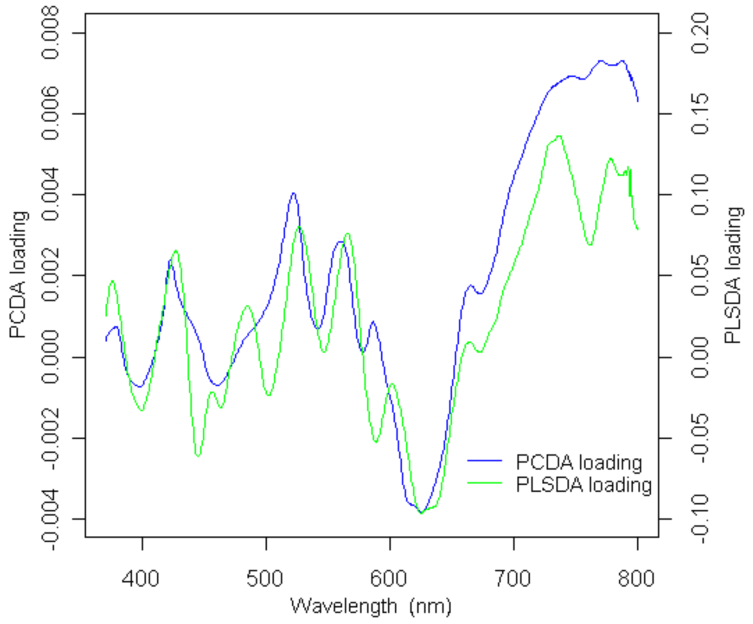


Figure 4.8: Comparison of PCDA loading (blue line) and PLSDA loading (green line) for discrimination between low risk and high risk sites of Barrett's data.

Further exploration on how PLS components contribute to the classification in this dataset can be done by looking at the PLS loading weights as described in Section 3.3.2. The PLS loading weights define the contributions of each wavelength to the constructed components.

In this dataset, the loading weights of the first two PLS components in the top row of Figure 4.9 show many of the same features as the PLSDA loading in Figure 4.7 which is a combination of the loading weights. Comparing these loading weights with the PCA components' loadings in the bottom row of Figure 4.9, PLS is choosing different factors to PCA although more or less similar features can still be found in some spectral ranges. The first three principal components seem to be modeling some mixture of common spectral features and spectral differences between the two types of tissue, though the first two PCs focus more on the latter feature after EROS pretreatment. The first three PLS components seem to have specific focus on between-type features. The comparison of PLS loading weight, PC loading, PLSDA loading and PCDA loading might explain some connection between PLS and LDA and why fewer factors can be chosen for PLSDA model. Further comparison between

PLSDA and PCDA when they are working with EROS will be discussed in Section 4.7.2.

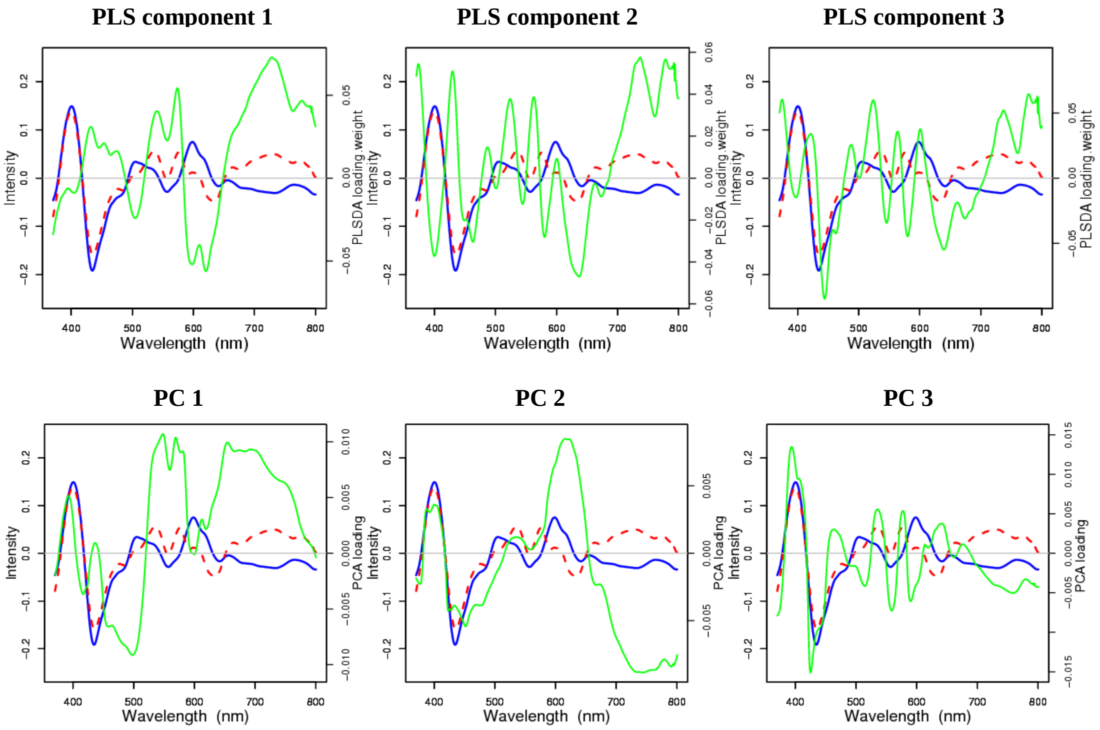


Figure 4.9: The loading weights (top row) of the first three PLS components and the first three PC loadings (bottom row) based on the EROS pretreated spectra from Barrett's data with k = 5. The loadings are shown in green, with the mean spectra for the two types superimposed (low risk in solid blue line, high risk in dotted red line).

4.4.2 Diagnosis of colon lesions (Colon data)

4.4.2.1 Data description

A colon dataset is presented here to illustrate the use of EROS with another application relating to the discrimination between hyperplastic and adenomatous polyps in the colon. Visible-NIR spectra covering the range from 320 to 920nm were collected from 73 human colonic biopsy sites using an optical probe *in vivo* during colonoscopy which was undertaken to assess patients for precancerous adenomatous polyps (Dhar *et al.* 2006). Repeated measurements, on average 5 spectra per site, were taken from each biopsy site. In total 402 spectra were analyzed, including 63 spectra from 11 hyperplastic polyps and 339 spectra from 62 adenomatous polyps.

In clinical diagnosis of colon lesions by use of ESS, measurement variability is a major source of variation. The procedure was done in lightly sedated patients and it was difficult to place the optical probe in constant contact with the polyps. Movement, pressure and other artefacts are therefore very likely to occur and these will influence the experimental result but are irrelevant to classification. These problems are reflected in considerable variability between the measured replicate spectra at a given site. Ideally these measurement variations should be modelled and then removed from the spectra, and this was attempted using EROS.

4.4.2.2 Data analysis

Before applying EROS pre-treatment and constructing a classification model, standard data pre-processing as described in Section 4.4.1.3 was carried out on the spectra to improve signal quality. This involved spectral smoothing using Savitzky-Golay method (Savitzky and Golay 1964), cropping the noisy ends of the spectra, and normalizing using the standard normal variate (SNV) method (Barnes *et al.* 1989).

Spectra from the two types of polyps are of very similar overall shape, with the mean spectra showing small between-class differences (left panel of Figure 4.10).

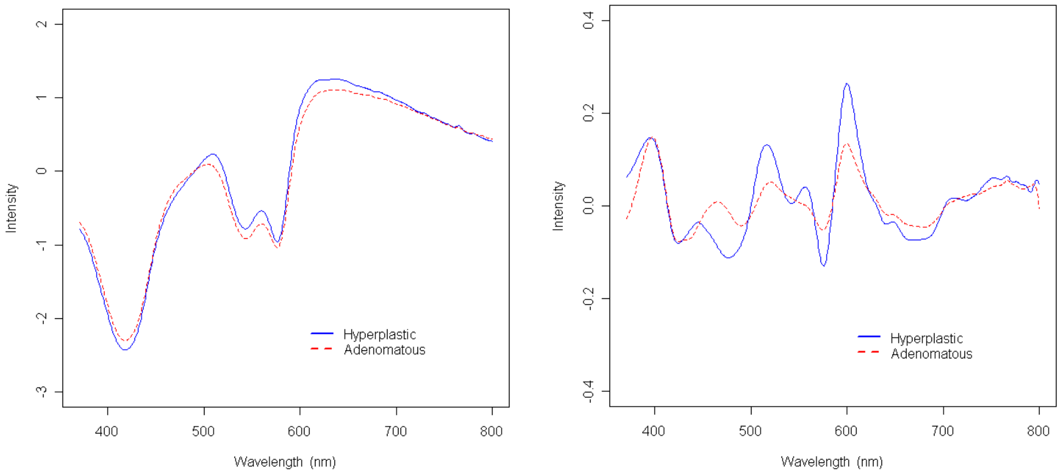


Figure 4.10: Mean spectra from hyperplastic (solid blue line) and adenomatous (dashed red line) polyps of colon data. Left: after standard pre-processing but before application of EROS. Right: after standard pre-processing followed by EROS.

The pooled within-sample covariance matrix was calculated using the replicated spectra measured at each biopsy site, and the first *k* eigenvectors of this matrix were used to construct *P* in Equation (4.1). EROS was applied by projecting the spectra onto the subspace orthogonal to the main sources of measurement variability using Equation (4.1).

Classification rules were derived using both principal component discriminant analysis (PCDA) and partial least squares discriminant analysis (PLSDA).

EROS with PCDA or PLSDA was carried out for a grid of values of *k* and *q*, with *k* ranging from 0 (no dimensions subtracted) to 7 and *q* from 2 to 30. Leave-out-one-site cross-validation (LOOCV), which trains the algorithm on all the data except one site which is then tested, was used to assess performance, the classification accuracy being measured by sensitivity, specificity and AUC, the area under the ROC curve. The construction of *P* was carried out inside the loop, i.e. with the replicates for the omitted site not included in the procedure. The reason for not choosing the repeated holdout validation used in the Barrett's data is that the colon data is a small set, even if we used repeated holdout validation the data randomly split off for testing would not be sufficient to generate a AUC for the held-out set.

The same per-site analysis as described in Section 4.4.1.3 was used for the colon data to classify each biopsy as a high risk site or a low risk site. Also since for LOOCV mean value of the canonical scores on per site base gives better classification results than maximum value (as shown in Figure 4.13 and Figure 4.17), we chose a mean-score-per-site criterion for the analysis of the colon data. The sensitivity, specificity, AUC, and the ROC curves were then calculated by pooling the mean canonical score outputs determined on each biopsy site as described in Section 3.5.3.

4.4.2.3 PCDA model

The leave-out-one-site cross-validation results for various combinations of *k* and *q* are shown in Table 4.3, and the specificity and AUC results are plotted in Figure 4.11. The ROC curves are given in Figure 4.13. Instead of choosing a balanced sensitivity and specificity the cut off canonical score between the normal and abnormal sites was adjusted in order to load towards a higher sensitivity (normally above 90% is preferred) because in practice we don't want to miss any high risk (abnormal) site. Here a sensitivity of 85% was chosen considering the small size of the colon data.

Before pre-treatment with EROS ($k = 0$), the choice $q = 15$ gives the best results with sensitivity, specificity and AUC of 85%, 64% and 86%, respectively. After pre-treatment with EROS, the combination $k = 5$, $q = 7$ gives better results with fewer factors, with sensitivity, specificity and AUC of 85%, 82% and 88%, respectively, and subsequent discussion is based on this choice. The right hand panel of Figure 4.12 shows the mean spectra for the two types of polyp after pre-treatment with EROS. The differences are much more evident than they were before pre-treatment. Figure 4.12 shows the loadings for the LDA discriminant functions using $k = 0$, $q = 15$ (no EROS) in the left panel and $k = 5$, $q = 7$ in the right panel. After using EROS, the loading vector is much less noisy and can be related much more easily to features in the spectra.

What cannot be seen from the figure, because the spectra have rescaled, is that the pretreatment has removed 94% of the variability in the original spectra, and that the information for the classification appears to be in the remaining 6%. This gives another example showing the most gain from EROS probably is to make the subtle difference between two groups become much more prominent after the pre-treatment.

Table 4.3: Results of leave-out-one-site cross-validation on colon dataset using various combinations of *k* (number of dimensions removed by EROS pre-treatment) and *q* (number of PCA scores used in the LDA).

***k* (EROS)**	***q* (PCDA)**	**Prediction Accuracy**		
		Sensitivity (%)	**Specificity (%)**	**AUC**
0	2	85	27	0.70
0	3	85	18	0.68
0	5	85	36	0.79
0	7	85	27	0.76
0	10	85	27	0.68
0	**15**	***85***	***64***	***0.86***
0	20	85	45	0.78
0	30	85	55	0.77
2	2	85	18	0.52
2	3	85	36	0.69
2	5	85	18	0.67
2	7	85	36	0.69
2	10	85	9	0.71
2	**15**	***85***	***73***	***0.84***
2	20	85	55	0.82
2	30	85	64	0.83
3	2	85	27	0.71
3	3	85	27	0.69
3	5	85	45	0.70
3	7	85	64	0.85
3	10	85	64	0.83
3	**15**	***85***	***73***	***0.84***
3	20	85	64	0.84
3	30	85	45	0.79
5	2	85	45	0.71
5	3	85	36	0.71
5	5	85	55	0.76
5	**7**	***85***	***82***	***0.88***
5	10	85	73	0.86
5	15	85	73	0.89
5	20	85	55	0.84
5	30	85	55	0.84
7	2	85	18	0.66
7	**3**	***85***	***64***	***0.86***
7	5	85	64	0.85
7	7	85	64	0.84
7	10	85	64	0.86
7	15	85	45	0.84
7	20	85	45	0.81
7	30	85	27	0.77

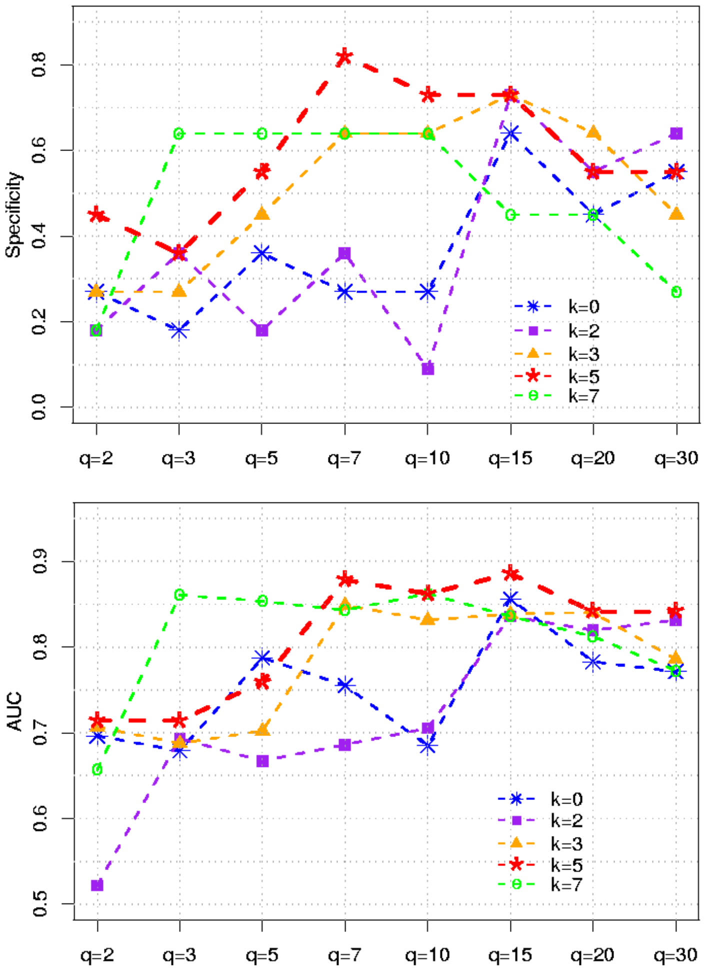


Figure 4.11: Leave-one-site-out cross-validation PCDA classification accuracy of colon data as measured by specificity and AUC. The five lines correspond to the choices of *k*, the x-axis to the choice of *q*.

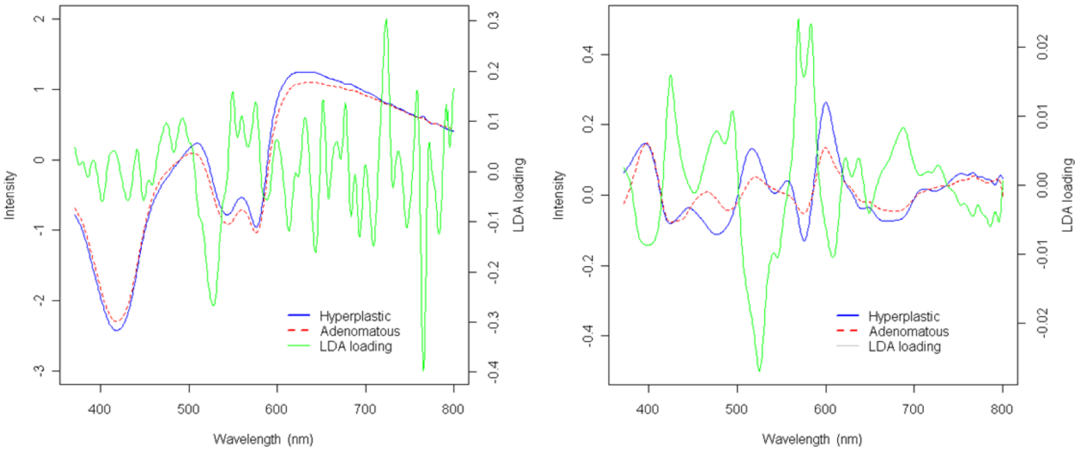


Figure 4.12: LDA loadings for discrimination between hyperplastic and adenomatous polyps of colon data. The LDA loading is shown in green, with the mean spectra for the two types superimposed (hyperplastic solid blue line, adenomatous dashed red line). The left panel is for $k = 0$, $q = 15$ (no EROS) and the right panel is for $k = 5$, $q = 7$.

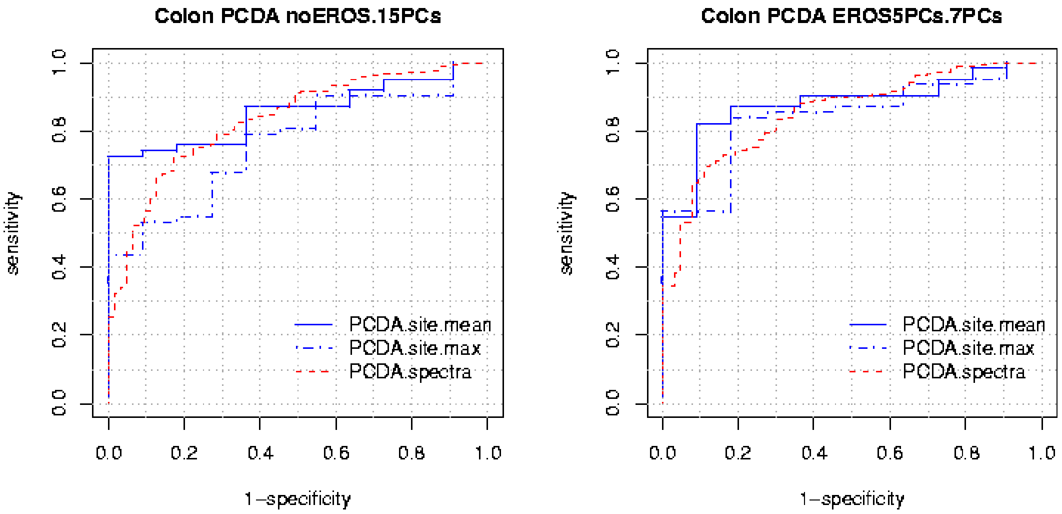


Figure 4.13: ROC curves from leave-one-out cross-validation on colon dataset by using PCDA for discrimination between hyperplastic and adenomatous polyps. The left panel is for *k* = 0, *q* = 15 (no EROS) and the right panel is for *k* = 5, *q* = 7.

4.4.2.4 PLSDA model

The leave-out-one-site cross-validation results on the per-site base for various combinations of *k* and *q* from PLSDA model are shown in Table 4.4, and the specificity, AUC results are plotted in Figure 4.14. The ROC curve is given in Figure 4.17. A sensitivity of 85% was chosen for all the combinations.

Compared with PCDA, PLSDA uses fewer parameters for EROS pretreatment. Before pre-treatment with EROS (*k* = 0), the choice *q* = 7 gives the best results with sensitivity, specificity and AUC of 85%, 73% and 83%, respectively. After pre-treatment with EROS, the combination *k* = 2, *q* = 7 gives best results with fewer factors, with sensitivity, specificity and AUC of 85%, 82% and 85%, respectively, and subsequent discussion is based on this choice. Other models with combination of *k* = 5 and *q* = 5, and combination of *k* = 7 and *q* = 3 also give acceptable results with sensitivity, specificity and AUC of 85% 73% and 84%; 85%, 73% and 86%, respectively.

Table 4.4: Results of leave-out-one-site cross-validation on colon dataset using various combinations of *k* (number of dimensions removed by EROS pre-treatment) and *q* (number of PLS scores used in the LDA).

***k* (EROS)**	***q* (PLSDA)**	**Prediction Accuracy**		
		Sensitivity (%)	**Specificity (%)**	**AUC**
0	2	85	36	0.76
0	3	85	27	0.74
0	5	85	45	0.74
0	**7**	***85***	***73***	***0.83***
0	10	85	55	0.80
0	15	85	45	0.74
0	20	85	36	0.72
0	30	85	36	0.72
2	2	85	36	0.72
2	3	85	36	0.68
2	5	85	45	0.80
2	**7**	***85***	***82***	***0.85***
2	10	85	55	0.80
2	15	85	64	0.85
2	20	85	64	0.84
2	30	85	64	0.83
3	2	85	36	0.71
3	3	85	36	0.72
3	5	85	64	0.81
3	7	85	64	0.82
3	10	85	45	0.82
3	15	85	36	0.74
3	20	85	36	0.78
3	30	85	36	0.77
5	2	85	55	0.78
5	3	85	55	0.78
5	**5**	***85***	***73***	***0.84***
5	7	85	64	0.82
5	10	85	45	0.82
5	15	85	45	0.78
5	20	85	27	0.77
5	30	85	27	0.76
7	2	85	64	0.79
7	**3**	***85***	***73***	***0.86***
7	5	85	36	0.82
7	7	85	55	0.80
7	10	85	36	0.76
7	15	85	45	0.77
7	20	85	45	0.72
7	30	85	36	0.69

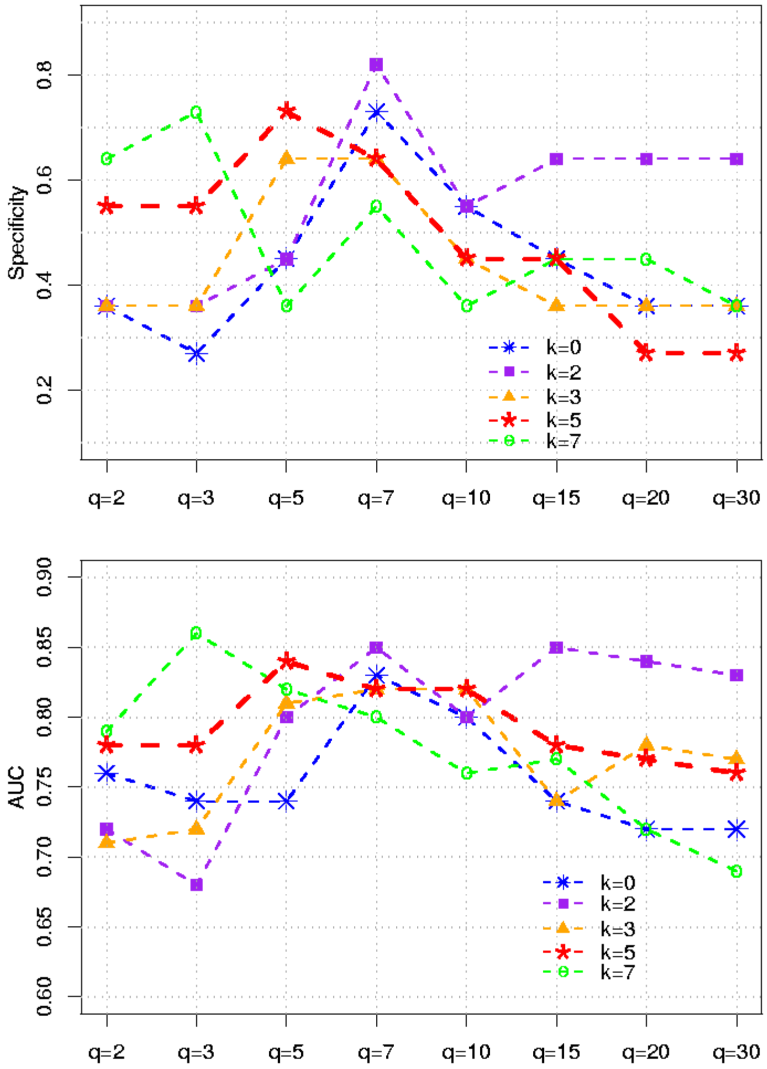


Figure 4.14: Leave-one-site-out cross-validation PLSDA classification accuracy of colon data as measured by specificity and AUC. The five lines correspond to the choices of k, the x-axis to the choice of q.

Panels (b), (c) and (d) of Figure 4.15 show the mean spectra for the two types of polyp after pre-treatment with EROS when k = 2, 5 and 7, respectively. The loadings for the PLSDA discriminant functions of (b) (c) (d) models all show high peaks in the region between 450 and 600nm. After using EROS, the (b) model with best accuracy shows a loading vector with simpler features than others. Also in Figure 4.16 we found in the case of the colon data that the loading vector of optimal models from both PCDA and PLSDA showed similar features, especially in the region between 540 to 580 nm. The PLSDA loadings have more structure than the PCDA loadings. The important regions for classification indicated by the loadings are consistent with the tissue absorption knowledge described in Chapter 2. It was found in a pilot study that by comparing the normalized backscattered signal between 540 and 580 nm with that found between 400 and 440 nm, it was possible to differentiate between adenomatous and dysplastic areas from other non-neoplastic areas of the colon. In these regions of spectra the major spectral differences between the tissues are caused by the changes in blood volume, and oxygen saturation of haemoglobin. Most of these variables are due to absorption rather than scattering, which underlies the complexity of interpreting these data.

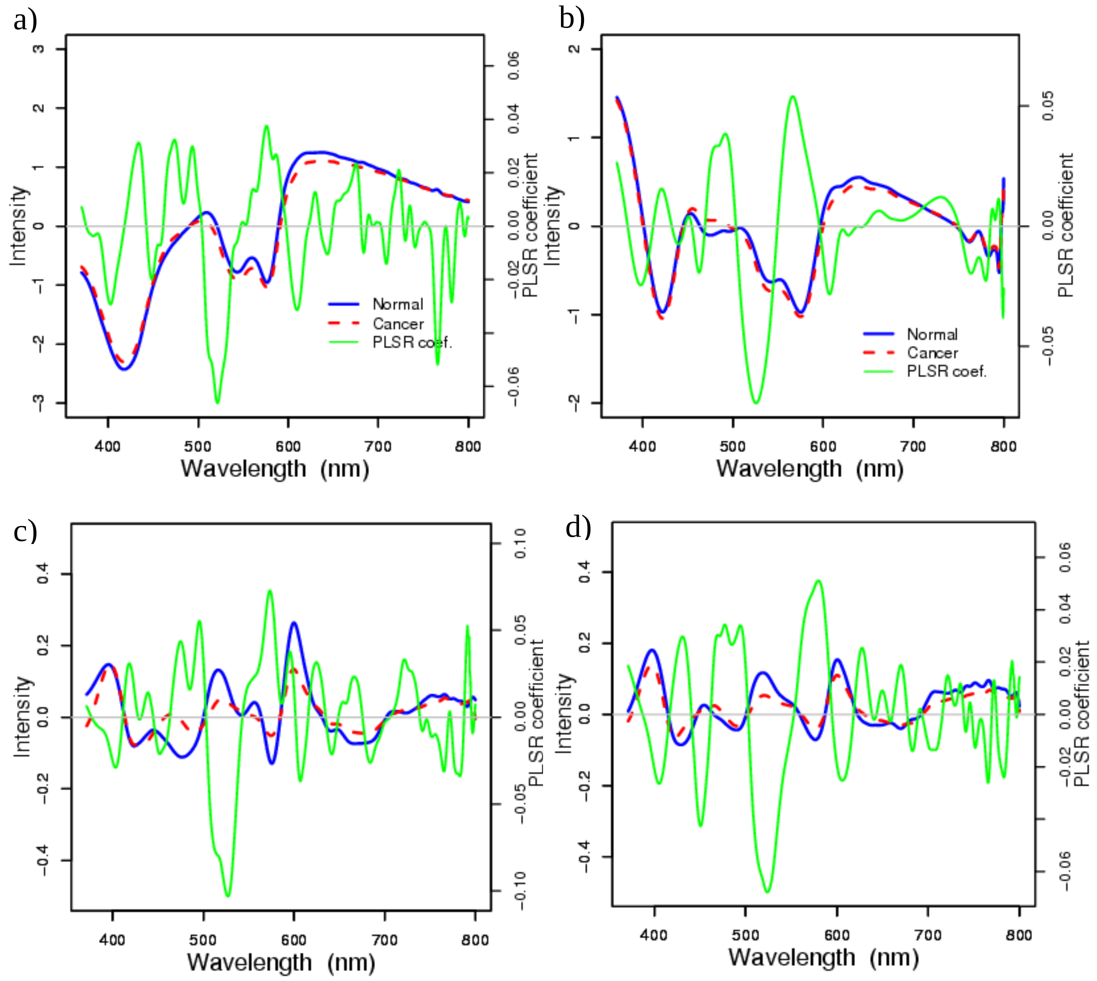


Figure 4.15: PLSDA loadings for discrimination between hyperplastic and adenomatous polyps of colon data. The LDA loading is shown in green, with the mean spectra for the two types superimposed (hyperplastic solid blue line, adenomatous dashed red line). a) The top left panel is for $k = 0$, $q = 7$ (no EROS); b) the top right panel is for $k = 2$, $q = 7$; c) the bottom left panel is for $k = 5$, $q = 5$; d) the bottom right panel is for $k = 7$, $q = 3$.

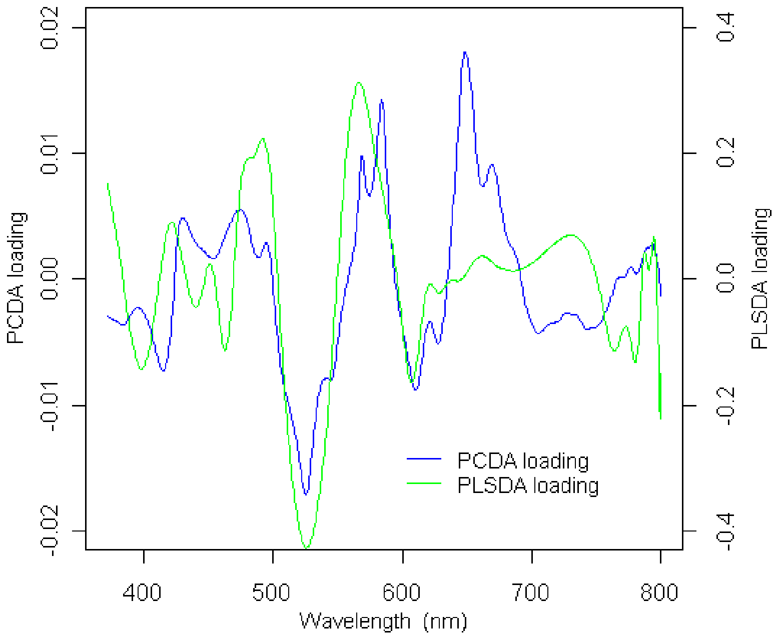


Figure 4.16: Comparison of PCDA loadings and PLSDA loadings for discrimination between hyperplastic and adenomatous polyps of colon data.

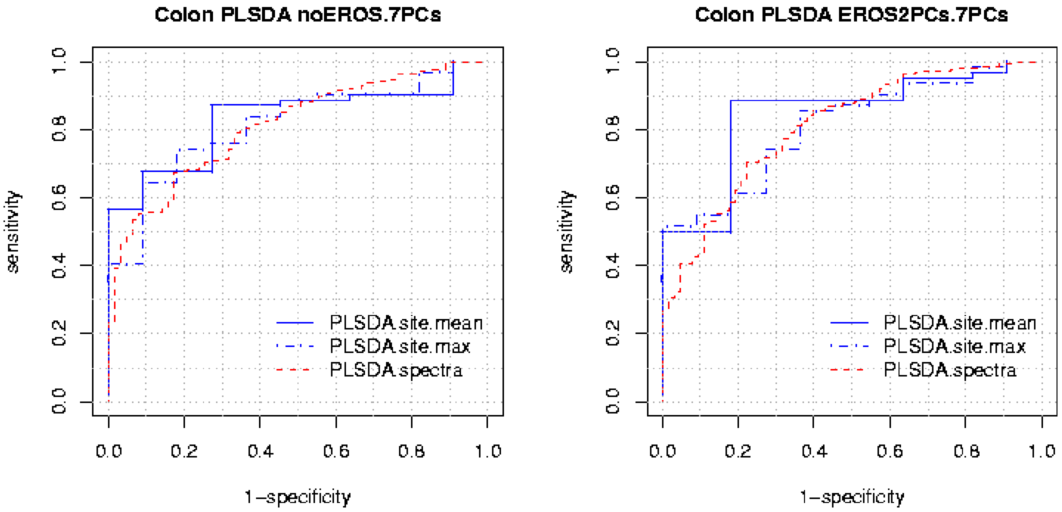


Figure 4.17: ROC curves from leave-one-out cross-validation on colon dataset by using PLSDA for discrimination between hyperplastic and adenomatous polyps. The left panel is for $k = 0$, $q = 7$ (no EROS) and the right panel is for $k = 2$, $q = 7$.

4.5 Experiment

4.5.1 Purpose of experiment

In order to understand the measurement variability modelled and removed by EROS, an experiment was designed to determine how experimental variation affects optical spectra collected under controlled conditions.

Two major potential sources of measurement variability when collecting data *in vivo* include differences in pressure of the probe on the tissue and the angle at which the probe is held to the tissue. These factors are difficult to control during measurement. In the experiment they were deliberately varied in a controlled fashion in order to investigate their effect on the spectra.

4.5.2 Materials, instruments and methods

Two different types of tissues, squamous lined pig oesophagus and columnar lined pig stomach, were resected (we used a portion of approximately 4cm^2 of each tissue), extended on a small piece of cork and fixed with pins. This preparation was placed on an electronic balance as illustrated in Figure 4.18. All measurements were carried out with a 2.5 mm (outer diameter) optical biopsy probe. Data were collected at 10 random sites for each tissue. At each site 10 replicate measurements were taken under the conditions of all possible combinations of four pressures (0kPa, 10kPa, 20kPa, 30kPa) and four angles (0°, 15°, 30°, 45°). The total number of spectra measured for each tissue was 1600. The probe was fixed to a micro-manipulator and pushed downwards until the balance gave a reading close to the desired pressure level. The data for each type of tissue were pooled to mimic the *in vivo* situation. The same spectral pretreatment procedures (smoothing, reduction, cropping and normalization by SNV) as described in Section 4.4.1.3 were carried out this data. We then extracted the first few principal components based on the pooled within-site covariance matrix (Equation (4.2)) using all 160 spectra at each site. By looking at these PC loadings, we can see how pressure and angle affect the experimental spectra. These loadings were then compared to the loadings obtained by EROS from the *in vivo* Barrett's data with the aim of gaining some understanding of the causes of within-site variability in spectra taken at endoscopy.

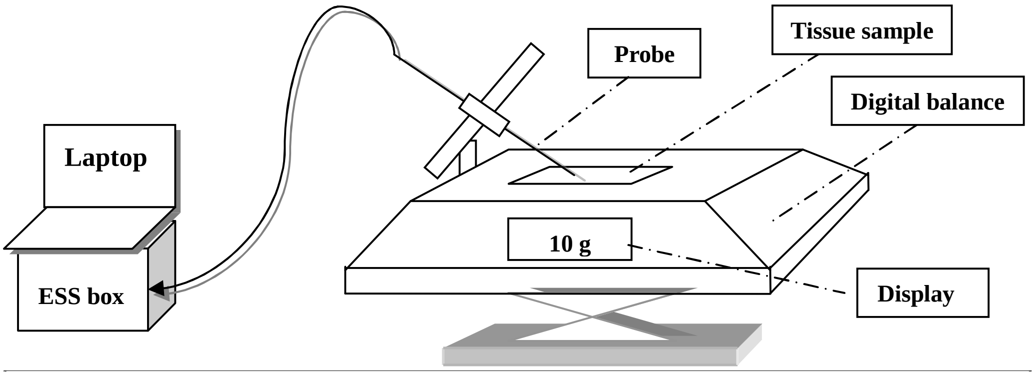


Figure 4.18: Schematic representation of the experimental set-up.

4.5.3 Experimental data

Our laboratory data show that different levels of pressure and angle have a substantial effect on the spectra. There is a clear ordering in the mean spectra in Figure 4.19. The first three principal component loadings based on the pooled within-site covariance matrix are plotted for both the experimental data and *in vivo* Barrett's data for comparison. For each principal component, there is quite a good match between the experimental and *in vivo* data as shown in Figure 4.20. For the experimental data, the pooled within-sample covariance captures the variation caused by pressure and angle under controlled laboratory conditions. For the *in vivo* Barrett's data, it describes the variability between replicate measurements. The similarity in the loadings supports the contention that the measurement variability removed by EROS comes from differences in pressure and angle when measurements were taken *in vivo*.

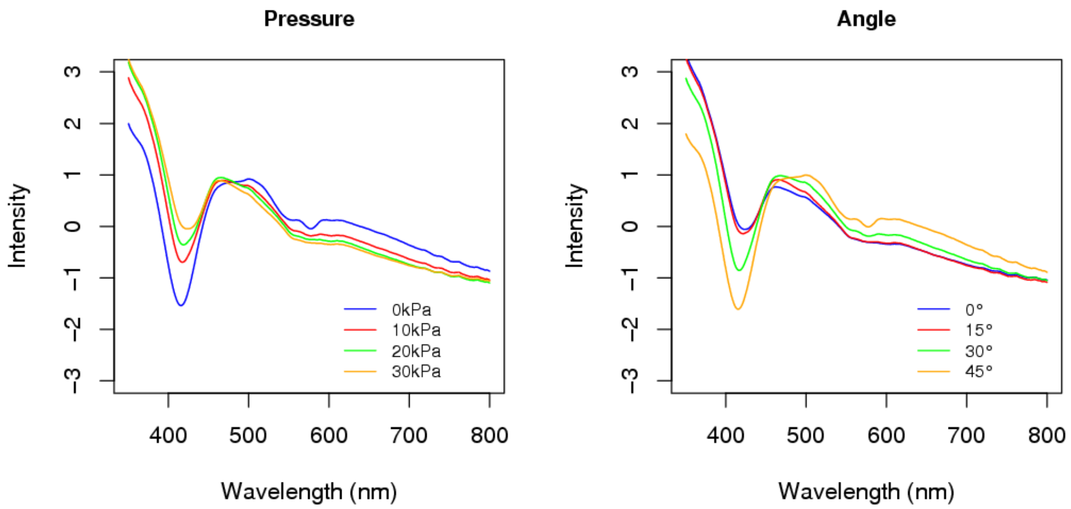


Figure 4.19: Spectral pattern of squamous tissue. Left: blue, red, green and orange lines: mean spectra at pressures 0kPa, 10kPa, 20kPa, 30kPa, respectively (Results are combined for all angles). Right: blue, red, green and orange lines: mean spectra at angles 0°, 15°, 30°, 45°, respectively (Results are combined for all pressures).

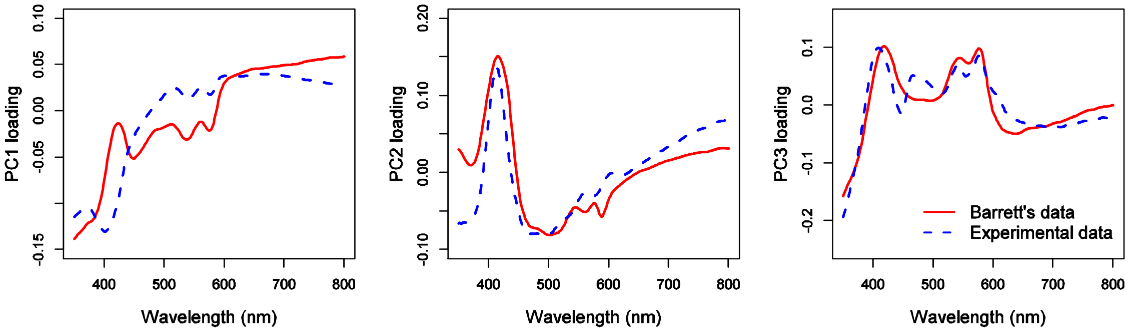


Figure 4.20: The first three principal component loadings of experimental data (dotted blue line) and *in vivo* Barrett's data (solid red line).

4.6 Prospective prediction

A new Barrett's dataset with a total of 68 matched optical and histological biopsies (corresponding to 276 spectra) was later collected from another 20 patients. It includes 50 biopsies (corresponding to 202 spectra) from low grade dysplasia or no dysplasia (low risk) and 18 biopsies (corresponding to 74 spectra) from high grade dysplasia or cancer (high risk). To test the robustness of the EROS models introduced in the Section 4.4.1.4, the classification rules trained on the Barrett's dataset described in Section 4.4.1.3 were applied to this independent test set.

The same spectral pretreatment procedures were implemented on the test data as on the training data and the classification rules were applied using the cutoffs learnt from the training data.

All the models (i.e. all the choices of *k* and *q*) were applied on the test set. The prospective prediction results are shown in Table 4.5. In general, the models with EROS give higher prediction accuracy than that without using EROS. The models with fewer number of PC scores in LDA, *q*, are the ones that work best with EROS. The comparisons of the two best validation models using $k = 0$, $q = 20$ (no EROS) and $k = 5$, $q = 5$ (EROS) (described in Section 4.4.1.4) show that the model with $k = 5$, $q = 5$ (EROS) gives much better prediction results with sensitivity of 83% and specificity of 84%, while the model with $k = 0$, $q = 20$ (no EROS) gives prediction results with sensitivity of 83% and specificity of 68%. The prediction results here show some evidence that the simple EROS model with fewer parameters gives more robust prediction accuracy for discrimination between spectra from high risk and low risk sites.

Table 4.5: Results of prospective testing prediction on an independent Barrett's dataset by applying the classification rules trained on the training set using various combinations of *k* (number of dimensions removed by EROS pre-treatment) and *q* (number of PC scores used in the LDA).

k **(EROS)**	***q*** **(PCDA)**	**Prospective Prediction Accuracy**	
		Sensitivity (%)	**Specificity (%)**
0	5	83	28
0	10	100	28
0	**20**	***83***	***68***
0	30	78	66
2	5	83	67
2	10	72	80
2	20	67	84
2	30	72	82
3	5	83	76
3	10	72	86
3	20	78	84
3	30	89	74
5	**5**	***83***	***84***
5	10	67	86
5	20	78	80
5	30	72	80
7	5	78	78
7	10	83	58
7	20	83	64
7	30	89	42

4.7 Discussion and conclusions

EROS works well as an effective pre-treatment for both Colon and Barrett's *in vivo* data, characterising the measurement variability between the spectra from repeated measurement, and ameliorating the effect of it. In comparison to using the original spectra for classification we can extract the most useful information from the first few principal components after eliminating this structured measurement variation from the total variation of the spectra. The designed experiment gave a plausible explanation for the measurement variability that much of the measurement variability comes from the differences of pressure and angle of the probe. The remaining variation left in the spectra contributed the most to the classification. The LDA loadings derived from the pre-treated spectra show a very smooth and simple structure while the LDA loading of the original spectra looks quite noisy. The simple structure is likely to be much more meaningful and instructive for the physical interpretation of the loadings. The approach, which was successful in this application, is quite general, and could be used in many other situations.

4.7.1 How EROS works

EROS worked well in the diagnostic application described above because the spectra contain a substantial amount of structured noise that can be removed by a simple projection without destroying the information of interest. This is only possible because the spectra are high-dimensional observations, but it will not always hold. The approach makes no attempt to orthogonalise the subtracted dimensions to *y*, and could thus remove some or all of the signal along with the noise. If there is a lot of interfering variability in the same dimensions as the signal of interest, then neither EROS nor any other mathematical treatment is likely to help. If it is in different dimensions, EROS will help. The situation in which EROS will have a negative effect is when it is used to remove small amounts of interfering variability that coincide in direction with the signal. In any given application, the best approach is to try it and see what happens.

An alternative approach to constructing calibrations robust to replication variability is to include replicate spectra in the training set, all with the same reference value. The fitting process then rewards prediction formulae that ignore the replication variability and tends to produce calibrations robust to this variability. The problem with this approach is that it interacts badly with the use of factor-based methods such as principal components regression or PLS to construct calibrations. Including the variability in the training set results in factors being constructed that describe it. These factors may then be downweighted in the regression part of the procedure, but the damage is already done in terms of complexity and interpretability. Removing the variability, as EROS does, removes it from the constructed factors, reducing complexity and improving interpretability.

The EROS pre-treatment is a subtraction. It removes entirely the targeted dimensions in the spectral space, but leaves the rest unaffected. It cannot therefore play the same role as a multiplicative scatter correction, and it will often be beneficial to use such a correction as well as EROS, as in the applications of Section 4.4.

An alternative approach to cope with effect of replicates by hierarchical mixed models to model structured within-sample variability in a nested or hierarchical fashion might also be interesting, but could become much more complicated, and has not been tried.

4.7.2 How EROS works with PCDA and PLSDA

Although neither the Barrett's data set nor the colon data set is large, we can still draw some conclusions about the comparison of PCDA and PLSDA. Figures 4.21-4.22 (top panels) show that before EROS pretreatment, PLSDA uses many fewer components (7) than PCDA (about 15-20) to achieve the best accuracy. The optimal results from both Barrett's and colon data seem vary with the method. This is probably because before EROS the variations extracted from the data using different methods to construct the components are different, and most probably neither method is completely successful in extracting the relevant information since results from both methods are improved after EROS. This also gives an example that PLSDA/PLSR does not always do better than PCA in discrimination, especially when there exist large within-group variations.

However after EROS pretreatment (as shown in the bottom panels of Figures 4.21-4.22), for the Barrett's data using 3~5 factors gives the best results for both PCDA and PLSDA models; for the colon data using 7 factors give the best results for both PCDA and PLSDA models. Both PCDA and PLSDA models can achieve comparable accuracy using a similar small number of components. This is especially the case for the colon data which shows highly consistent pattern of classification accuracy varying with model complexity. This gives some evidence that when the difference between two groups is subtle and there exist large variabilities within the groups, using EROS to eliminate large within-group variations before constructing classification models results in prominent distinction between groups and enables PCDA and PLSDA to achieve comparable discrimination results.

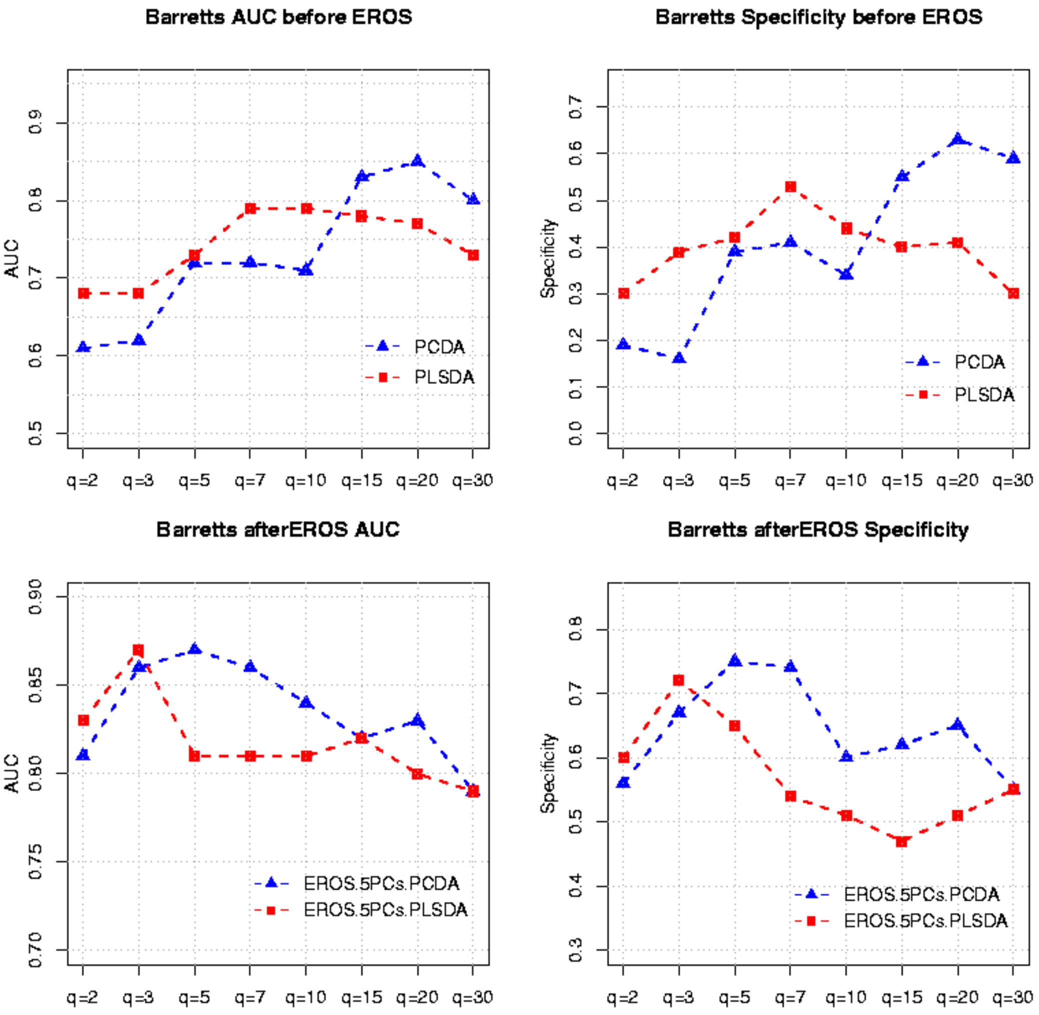


Figure 4.21: Classification accuracy of Barrett's data as measured by AUC and specificity before and after EROS pretreatment (*k* = 5 for both PCDA and PLSDA) by using PCDA (in blue) and PLSDA (in red).

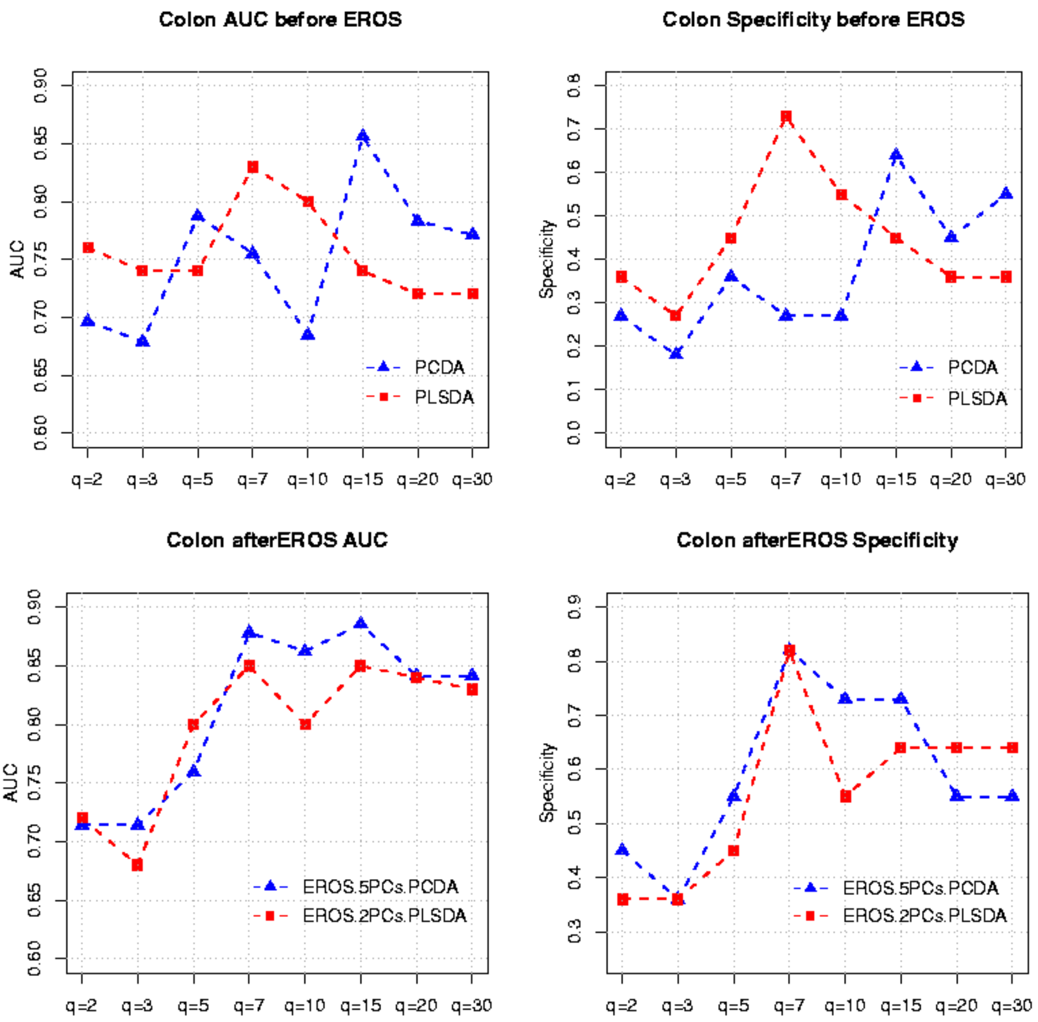


Figure 4.22: Classification accuracy of colon data as measured by AUC and specificity before and after EROS pretreatment (k = 5 for PCDA, k = 2 for PLSDA) by using PCDA (in blue) and PLSDA (in red).

In general, before EROS pretreatment, it's always worthwhile to try different classification models since their performance is variable. After EROS, the models are more parsimonious and how much variation of interest can be extracted from the EROS pre-treated data seems not to be method-dependent.

4.7.3 How to choose components used in EROS

According to our practical experience, the number of dimensions removed by EROS pretreatment should normally be small, in most cases using the first 3~5 PCs does best. This is evidence that the first few principal components stand a good chance to have contained the most undesirable variation for classification. This is to some extent warranted with reflectance spectra. For example the first PC often relates to a varying baseline of the spectra (Næs *et al.* 2002). Although in the applications above we always use the first few PCs to construct the projection matrix, it is not necessarily the only option. Two models from the colon example are compared here. One is the PCDA model with k = 5 and q = 7. The other is a PCDA model with q = 7 after EROS removed four dimensions corresponding to the first three components and the fifth component.

Figure 4.23 shows that the pretreated spectra by the second model show clearer differences and more structure (especially in the region of 400~450nm and 600~650nm) in the difference between two kinds of tissue than that for the first model. As to the PCDA loadings of both models, the second model gave the simpler structure than the first one. Removing fewer dimensions by EROS in the second model does not seem to decrease the prediction performance, as shown in Table 4.6. This is some evidence that the fourth principal component of the original spectra matrix, though capturing more measurement variability than the fifth one, might also contain information about the diagnosis. No attempt has been made to pursue this line of research any further, but there is scope for further investigation of how to select components to construct an optimal model and how to interpret them.

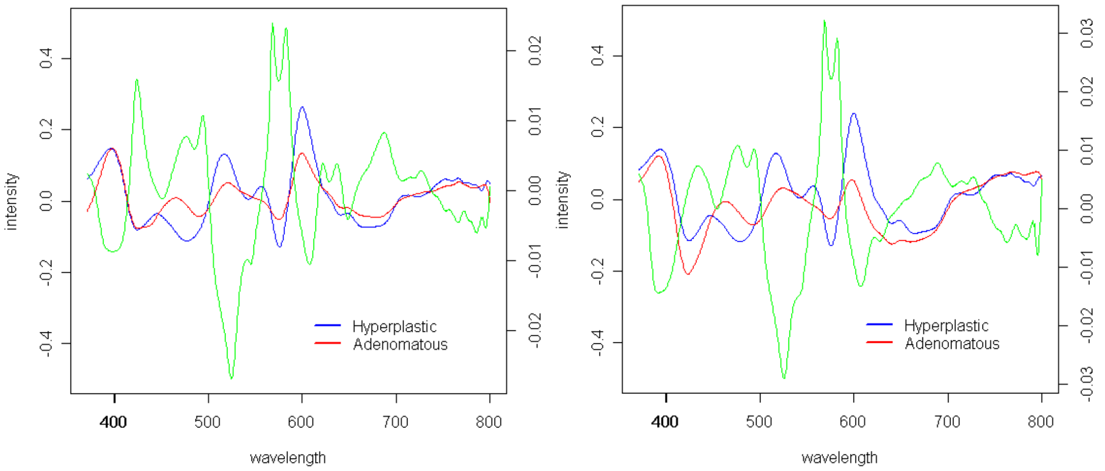


Figure 4.23: PCDA loading of the EROS pretreated spectra from colon data. Left panel is for k = 5, q = 7. Right panel is for the model with first 3 and 5th PC used for dimension removing in EROS and q = 7; Green line: PCDA loading; Red line: mean spectra from adenomatous polyps; Blue: mean spectra from hyperplastic polys.

Table 4.6: Results of leave-one-out cross-validation on colon dataset using various combination of dimension *k* (number of dimensions removed by EROS pre-treatment) and *q* (number of PCA scores used in the LDA)

k	***q***	**Prediction Accuracy**		
(EROS)	**(PCDA)**	**Sensitivity (%)**	**Specificity (%)**	**AUC**
1:5	5	63	64	0.74
-	**7**	***82***	***82***	***0.85***
-	10	81	82	0.83
-	15	82	82	0.87
	20	71	73	0.82
	30	72	73	0.76
1:3, 5	5	66	64	0.67
-	**7**	***81***	***82***	***0.86***
-	10	82	82	0.84
-	15	77	82	0.85
-	20	81	82	0.86
-	30	71	73	0.77

4.7.4 Wavelength selection

As can be seen in Figure 4.24, for the Barrett's data after pretreatment by EROS the PCDA loadings show much more importance in the spectral region where there are bigger differences between normal and cancer spectral intensity, especially in the region 500~800nm, which make it reasonable that we should interpret the LDA loading as importance of the wavelengths for interpretation.

According to the importance it appears to have for the classification, the wavelength range from 600 ~ 800 nm was selected for the PCDA model with pretreatment by EROS for $k = 5$, $q = 5$. The results of repeated holdout validation shown in Table 4.7 demonstrated that using only the information from the selected spectral interval with the range of 600 ~ 800 nm did not decrease by much the classification accuracy. This implies there exist strong spectral features in this region for classification model building. Although performance is not improved by selecting wavelength regions, the contribution of the wavelength selection to clinical diagnosis may be much more meaningful. If a small number of specific wavelengths can be found which give as much information as a full spectrum when used for classification, ESS point measurement then may possibly be developed for *in vivo* field imaging detection, which would become an invaluable tool to endoscopists.

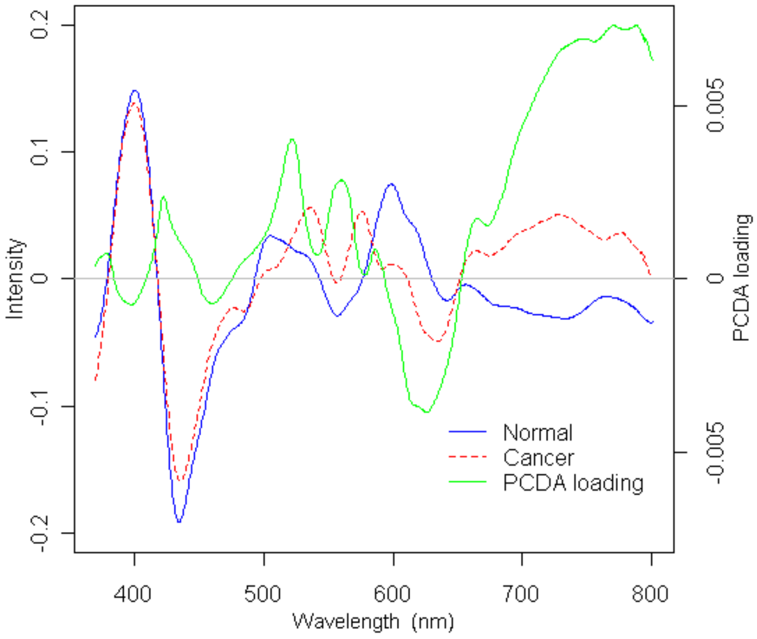


Figure 4.24: PCDA loadings for discrimination between low risk and high risk sites of Barrett's data. The PCDA loading is shown in green line, with the mean spectra for the two types superimposed (low risk solid blue line, high risk dotted red line). The PCDA model is for $k = 5$, $q = 5$.

Table 4.7: Results of repeated holdout validation on both whole range and selected range of Barrett's data by using EROS pretreatment and PCDA model with $k = 5$, $q = 5$.

		whole range: 370 ~ 800 nm			Selected range: 600 ~ 800 nm		
EROS	PCDA	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC
5PCs	5PCs	92%	75%	0.87	92%	72%	0.85

CHAPTER 5

PARTIALLY SUPERVISED BAYESIAN CLASSIFICATION OF SCANNED SENTINEL LYMPH NODES

5.1 Background and introduction

Breast cancer is the most common malignancy in women in the western world, with a reported incidence of up to 1 in 8 women. The presence or absence of metastatic cancer in the axillary lymph nodes in patients with breast cancer remains the most powerful predictor of prognosis, and plays an important role in identifying patients who are at risk of developing disease that spreads throughout the body, and thus likely to benefit from chemotherapy. Traditionally, the presence of axillary lymph node metastases has been determined by axillary lymph node dissection (ALND), which is a surgical procedure that removes all the lymph nodes under the arm. This is a substantial surgical procedure, however, which can be associated with several serious side effects, the most significant being lymphoedema (persistent swelling of the arm) and shoulder dysfunction, which adversely affect the patient’s quality of life. In current surgical practice, most patients present with early disease as a result of increased public awareness of breast cancer and mammography screening programs. Hence most patients do not have axillary lymph node metastases at presentation, and while the staging information is crucial for their future management, they get no therapeutic benefit from ALND, while still being at risk of developing the complications associated with the procedure.

The sentinel node is the first node to be invaded by cancer spreading from the breast as shown in Figure 5.1. It has been well documented that if cancer cannot be detected in sentinel nodes, the chance of there being any cancer in nodes further down the chain draining the breast is exceedingly small (Turner *et al.* 1997). Thus if the sentinel node can be easily identified, removed, and examined for cancer and no cancer is found, there is no need to remove the rest of the axillary nodes. This markedly reduces the risk of complications associated with full axillary node clearance (Veronesi 2003, Swenson 2002).

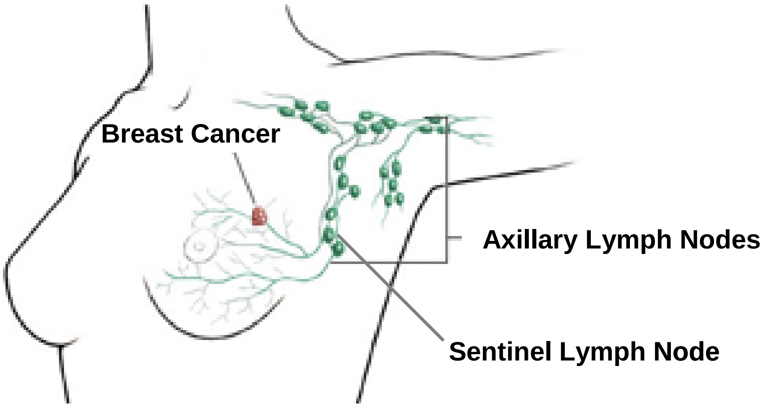


Figure 5.1: Illustration of sentinel and axillary lymph node in breast cancer.

To get the maximum benefit from sentinel node biopsy, it is important to be able to determine rapidly whether or not cancer is present. If the assessment of the node cannot be completed intraoperatively (while the patient is still on the operating table), the subsequent discovery of cancer in the node will necessitate a second operation to perform the ALND. This is technically more difficult, delays the start of adjuvant radiotherapy and chemotherapy, gives rise to additional costs and causes further anxiety to the patient.

Conventional intraoperative analysis is undertaken by touch imprint cytology or frozen section histology. Both require sample preparation and interpretation of the findings by an experienced pathologist (Johnson *et al.* 2004).

The lack of a more generally available and reliable intraoperative tool to establish the sentinel node status remains an obstacle to the routine practice of sentinel node biopsy. A real-time optical method for determining sentinel node involvement would provide significant benefits to patients undergoing surgery for breast cancer.

Elastic scattering spectroscopy (ESS), as described in Chapter 2, when performed using an appropriate optical geometry, is sensitive to the sizes, indices of refraction, and structures of the subcellular components (e.g., nucleus, nucleolus, and mitochondria) that change when cells become malignant (Mourant *et al.* 1998). Employing multivariate statistical techniques on spectroscopic measurements may enable *in vivo* examination and computer generated diagnosis without tissue processing and pathologist interpretation.

We have shown in Chapter 4 that ESS manual measurements are able to detect precancerous and early cancerous changes in Barretts oesophagus, and identify pre-cancerous polyps in the colon. ESS is applied in this chapter to diagnose sentinel lymph node metastases in breast cancer.

Conventional manually measured ESS spectra contain considerable experimental variation caused by angle or pressure of the hand-held probe as described in Chapter 4. To avoid this problem and to produce much more and reliable information for rapid intraoperative diagnosis of sentinel node metastases, an automated two-dimensional ESS scanning device was developed (as an advance on ESS manual measurements) to take measurements from the entire cut surface of the excised nodes. Instead of determining whether an individual point of tissue is abnormal or not, the ESS scanner examines a larger area of tissue by taking measurements at 400 points (pixels) in a 20×20 grid, and has the ability to produce diagnostic images (of any tissue that can be optically scanned) to assess the cancer risk. The assumption is that more extensive and reliable optical sampling of the tissue under examination should increase the accuracy of diagnosis.

However additional problems arise in the analysis and interpretation of the ESS scanned measurement data. The commonly used supervised classification methods, e.g., linear discriminant analysis, require training data to train the algorithm. Since the histology data for sentinel lymph nodes is on a per node basis, there is no reference pathology available for each pixel in the image. This causes problems when deriving a classification algorithm for pixels. We do however have manual measurements on totally normal and totally metastatic nodes from an earlier stage of this research. An obvious approach is to use the manual measurement data to train the classification algorithm for use with the scanning data. However, apart from the normal and metastatic groups, the scan also includes a third group, which is a non-nodal group from a background area, possibly contaminated by blood or lipid. There are no training data available for this non-nodal group which varies considerably from node to node, which causes further difficulty. Using the discriminant algorithm developed from manual data, the spectra from background usually gives a canonical score indicative of cancer. The non-node area may, as a result, be misclassified as metastatic and impede the recognition of normal and metastatic nodes.

As well as the problems described above, there is also one interesting opportunity arising in the ESS scanning data analysis. Since the scanner generates a spectral image, and it should therefore be possible to use a smoothness assumption about the image to improve the classification of individual pixels.

To solve the problems and make use of the advantage of having an image, a partially supervised image classification model framework has been built to recognize the non-nodal area automatically and enable clinicians to make a rapid intraoperative diagnosis of sentinel node metastases from the images. The basic idea is that the classification of points in the image will be done by clustering, using a mixture model, with the training data providing prior distributions for the groups in this model.

To simplify the presentation and to demonstrate the effects of different aspects of the model, the whole model framework is split into two parts. The first part, presented in this chapter, concerns the dimension reduction and Bayesian multivariate finite mixture image classification model. Without consideration of the spatial correlation in the image, this part of the model generates a preliminary solution. The second part is addressed in Chapter 6 with additional information on spatial correlation in the image added into the model. Taking into account the spatial correlation between contiguous pixels in the image, this part of the model aims at generating a more realistic image. All the results and discussions relevant to the joint model framework are illustrated in Chapter 7.

A list of notations used throughout the Chapters 5-7 are listed in Appendix A.

5.2 Instrumentation system

In the case of this project, ESS is being developed as a technique to discriminate between lymph nodes which contain metastatic tissue, and lymph nodes which are normal. In the first stage of the work, spectra were collected from nodes using a hand-held probe. Later on an automated two-dimensional ESS scanning device (an optical scanner) was constructed to increase the sampling effort per-node. This takes comprehensive measurements from the entire cut surface of the nodes. The ESS scanner instrumentation consists of a xenon arc lamp, a static ESS probe, a mobile stage, a spectrometer and a computer to control the various components and record the spectra (see Figure 5.2).

After excision, nodes are bivalved along their long axis and touch imprint cytology performed. ESS measurements are taken from the same cut surface. The cut surface of the node is placed under a specially designed thin fibre optic plate attached to a motorised stage that moves back and forth incrementally in two dimensions under a stationary ESS probe in a raster scanning pattern (Figures 5.3-5.4) to enable multiple ESS measurement of the cut surface of the node. The probe is optically coupled to the fibre optic plate with a drop of immersion oil. Prior to use, the system response is calibrated against a standard diffuse reflector as a reference (Spectralon) through the fibre optic plate. A 10×10 mm area of the cut surface of the each node is scanned as the fibre optic plate repeats the procedure of “move, stop and measure” pixel by pixel.

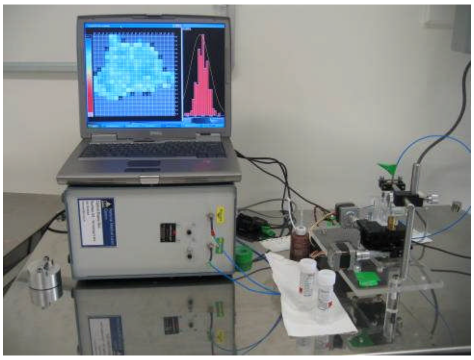


Figure 5.2: ESS scanning system

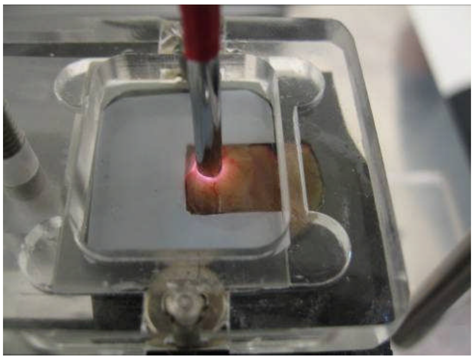


Figure 5.3: ESS scanner stage

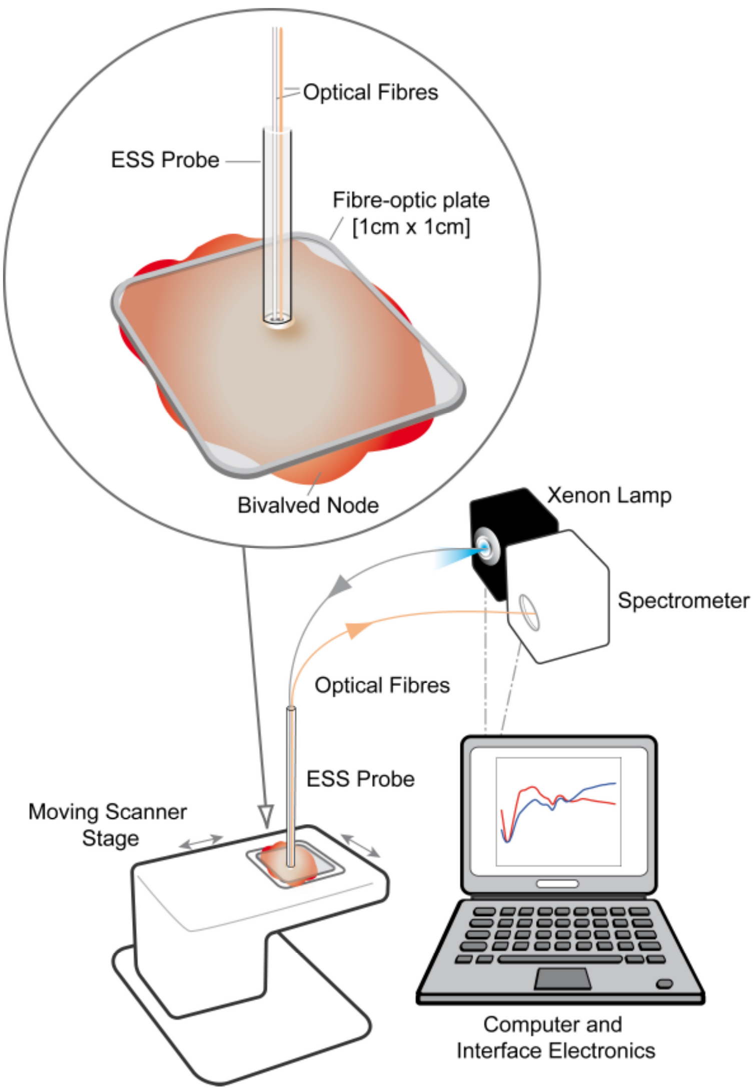


Figure 5.4: Schematic diagram of elastic scattering spectroscopy scanning device system.

A 20×20 ESS measurement pixel image is generated by the system, i.e. 400 pixels, each of size 0.5×0.5 mm. For each node a photograph with a microscopic view is taken for its area being scanned, usually including some area of fibre optic plate uncovered by the node. These photos will be later used in Chapter 7 to compare with the images generated by our model proposed in this chapter and Chapter 6 for the purpose of assessing the model's success.

5.3 Data description

The study was conducted in two phases, generating two datasets. Initially spectra taken manually from nodes known to be entirely normal or completely replaced by cancer were gathered in order to develop an algorithm to distinguish cancer from normal nodal tissue. In the second phase, an independent set of nodes was scanned, giving a 20×20 image for each node.

All spectra were recorded as a ratio to a calibration spectrum taken from the spectrally flat material Spectralon (as a reference), thus rendering the data independent of the spectral response of the system. The spectra are recorded at 1801 equidistant wavelengths in the visible and near infrared (NIR) region of 320-919 nm.

For the first part of the study, a total of 3,525 spectra from 404 nodes, 356 normal nodes and 39 totally metastatic nodes, were collected by manual point measurement. Each spectrum was measured by placing the optical probe manually at up to 16 random sites on the bivalved node. The mean spectra from these manual measurements are shown in Figure 5.5. Each node, and therefore each point is classified as normal or metastatic by histological experts. No spectra from background area are available in this dataset.

For the second part of the study, to assess nodes comprehensively, a total of 48 axillary nodes, 21 normal nodes and 27 partially metastatic nodes or totally metastatic nodes, with matched histology from 26 patients were scanned in our study, giving a 20×20 pixel image of 400 spectra for each node. The reference pathology is only provided on per node base, but not available for individual pixels in the image.

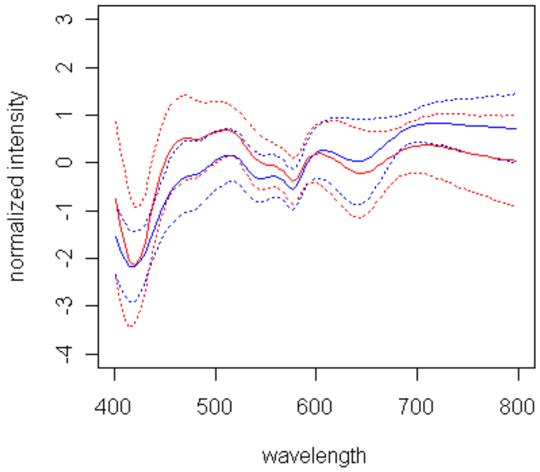


Figure 5.5: Mean spectra from normal (blue solid line) and metastatic (red solid line) node with one standard deviation on either side of the mean (dashed lines) after standard pre-processing.

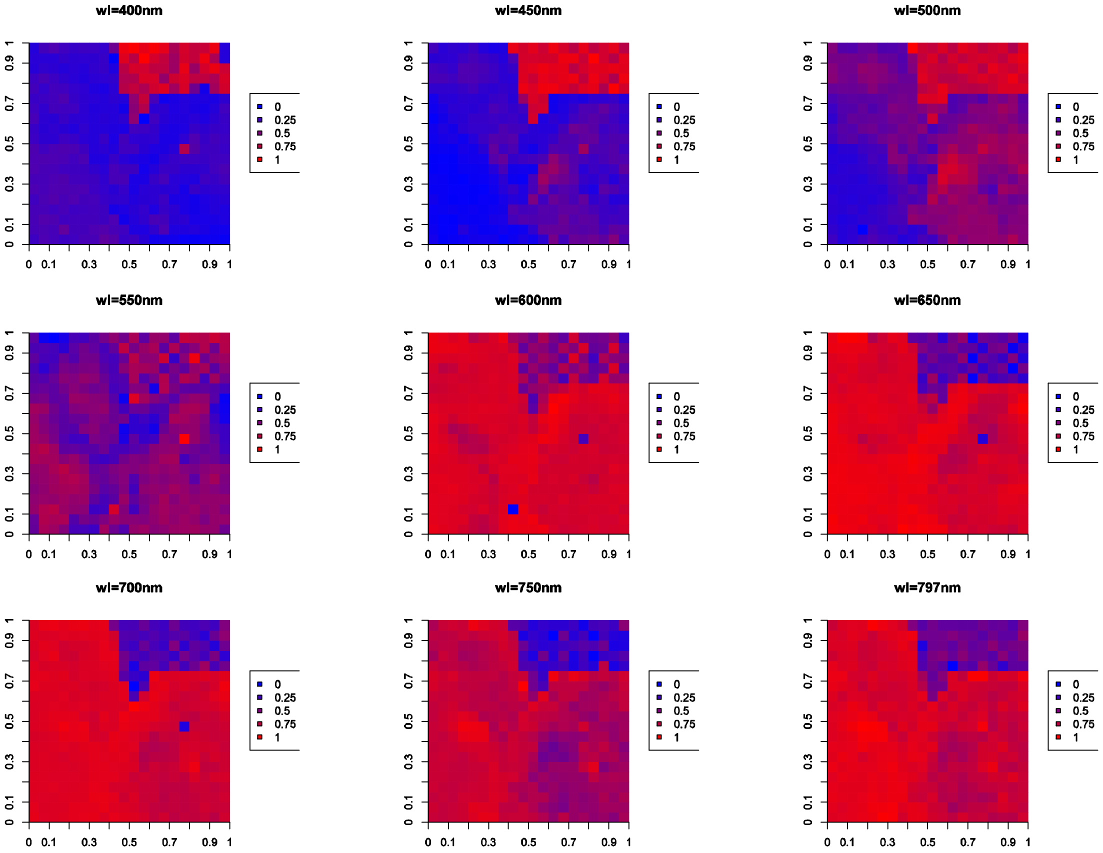


Figure 5.6: One example of ESS scanning measurement pixel image of one partially metastatic node generated by spectral intensity at nine different wavelengths.

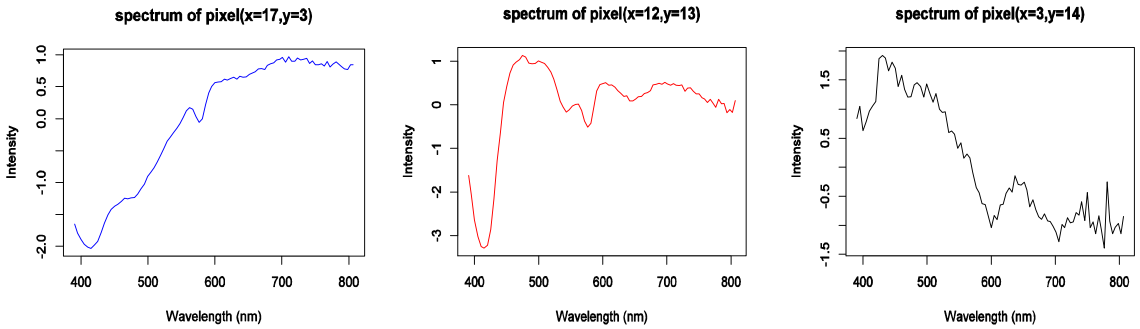


Figure 5.7: Typical spectrum from normal tissue (blue), metastatic tissue (red) and non-nodal area (black) of one partially metastatic node (the pixel position index (*x*,*y*) of each spectrum refers to the pixel location in the image of Figure 5.6).

The cut surface of the node with bivalved shape is placed under a square fibre optic plate as shown in Figure 5.4, and in most cases the scanning data contains not only the areas from the normal or metastatic groups, but also a non-nodal group from a background area, possibly contaminated by blood or lipid, which varies considerably from node to node. No training set from the manual data are available for this non-nodal group.

For illustration Figure 5.6 shows the image of one partially metastatic node at nine different wavelengths with a colour map for intensity. Typical spectral patterns of one pixel of normal tissue, metastatic tissue and non-nodal area from this node are shown in Figure 5.7.

Figure 5.6 shows this is a three-way data with two spatial dimensions and one wavelength dimension where the response is intensity. One approach to dealing with this kind of data is to begin with a second-order calibration model (Linder and Sundberg 1998). Instead of using this method, our analysis essentially ignored the spatial dimensions to start the analysis and then later on takes them into account.

5.4 Theory

In this section we give a brief review of the relevant theory on multivariate finite mixture models which will be employed in the partially supervised image classification model introduced in Section 5.5.

5.4.1 Multivariate finite mixture model

Finite mixture models, as an increasingly important tool in multivariate statistics, are widely used in semi-parametric probability density function (PDF) estimation problems or clustering tasks (e.g. McLachlan and Basford 1988, McLachlan and Peel 2000). The unknown PDF is approximated by a weighted sum of mixture distributions. Finite Gaussian mixture models (FGM) are widely used, showing good performance in many applications (McLachlan and Peel 2000, Fraley and Raftery 2002, Titterington *et al.* 1985). However, finite Student's *t* mixture models (FTM) are efficient alternatives that can deal with a limited number of outliers.

In a finite mixture model for a *k*-dimensional random vector X the PDF is a linear combination of *g* component densities $f(x \mid \theta_j)$:

$$f(x \mid \phi) = \sum_{j=1}^{g} \pi_j f(x \mid \theta_j), \tag{5.1}$$

where the mixing proportions π_j s are non-negative and must sum to one, and θ_j s are the parameters of the components in the mixture. Consider an i.i.d. realization $\{x_i\}_{i=1}^{n}$ of X. We may define the corresponding log-likelihood function:

$$L(\phi) = \log \prod_{i=1}^{n} f(x_i \mid \phi), \tag{5.2}$$

where $\phi = (\pi_1, \ldots, \pi_g, \theta_1, \ldots, \theta_g)$ summarizes the model parameters.

5.4.1.1 Multivariate Gaussian distribution

In multivariate analysis, a popular choice for the component density $f(x|\theta_j)$ is the *k*-dimensional multivariate Gaussian distribution. The *j*th component in the mixture is then characterized by its mean vector μ_j and its covariance matrix Σ_j, so that $\theta_j = \{\mu_j, \Sigma_j\}$ and

$$f(x|\theta_j) = \frac{1}{(2\pi)^{k/2} \, |\Sigma_j|^{1/2}} \exp\left\{ -\tfrac{1}{2}(x-\mu_j)^T \, \Sigma_j^{-1}(x-\mu_j) \right\}. \tag{5.3}$$

5.4.1.2 Multivariate t distribution

An alternative choice is the multivariate *t* distribution. The family of *t* distributions provides a heavy-tailed alternative to the normal family. Hence it provides a more robust approach to the fitting of normal mixture models, as observations that are atypical of a component are given reduced weight in the calculation of its parameters. Also, the use of *t* components gives less extreme estimates of the posterior probabilities of component membership of the mixture models, as demonstrated in Peel and McLachlan (2000).

A random vector X is said to have a k-variate t distribution with location parameter μ_j, positive definite scale matrix Σ_j, and the degrees of freedom v_j if its joint probability density function is given by

$$f(x|\theta_j) = \frac{\Gamma\,[(v_j + k)/2]}{(\pi\, v_j)^{k/2} \left| \Sigma_j \right|^{1/2} \Gamma(v_j/2)} \left(1 + \frac{(x-\mu_j)^T \Sigma_j^{-1} (x-\mu_j)}{v_j} \right)^{-(v_j+k)/2} . \qquad (5.4)$$

If $v_j > 1$, μ_j is the mean of *x*, and if $v_j > 2$, $v_j(v_j{-}2)^{-1}\Sigma_j$ is its covariance matrix. The degrees of freedom parameter v_j is also referred to as the shape parameter, because the Kurtosis of Equation (5.4) depends on v_j. When $v_j = 1$, it corresponds to a k-variate Cauchy distribution for which neither the mean nor the variance exists. As v_j tends to infinity, *x* becomes multivariate normally distributed with mean vector μ_j and covariance matrix Σ_j. It can be fixed in advance or it can be inferred from the data for each component thereby providing an adaptive robust procedure (Lange *et al.* 1989). In our applications we will fix, rather than estimate, v_j, so the parameters are $\theta_j = \{\mu_j, \Sigma_j\}$.

5.4.2 Maximum likelihood estimation (MLE)

The maximum likelihood approach to parameter estimation in finite mixture models obtains a point estimate $\hat{\phi}$ of the parameter ϕ by attempting to maximize the log-likelihood function in Equation (5.2), that is,

$$\hat{\phi} = \arg\max_{\phi} L(\phi) . \qquad (5.5)$$

Quantities of interest may then be estimated by “plugging in” the point estimates. For example, the density $f(x|\theta_j)$ may be estimated by $f(x|\hat{\theta}_j)$.

Although popular, the maximum likelihood approach to mixture models suffers from problems mostly caused by the fact that for many choices of parametric family *f* the likelihood in Equation (5.2) is unbounded. Point estimates of the parameters corresponding to singularities in the likelihood surface are of no real interest, and it is usual to seek a parameter estimate which corresponds to a large local maximum of the likelihood surface. Aside from the computational problems associated with finding such maxima, there may be several reasonable local maxima between which to choose, each of which may give quite different plug-in estimates for quantities of interest. In many cases it will be difficult to justify choosing one of these point estimates of the parameter above the others. Such problems have encouraged the development of a Bayesian approach to the mixture models problem.

5.4.3 The Bayesian approach

In Bayesian paradigm, parameters are treated as random quantities, and point estimates for parameters are replaced by distributions on the parameter space which represent our knowledge or belief about the value of the parameters. A comprehensive description is given by Bernardo and Smith (1994) and Gelman *et al.* (1995). Fully Bayesian inference seeks quantities of interest by integration over the parameter space, weighting by the posterior distribution of the parameters, which is implemented by Markov chain Monte Carlo (MCMC). However when a spatial prior is taken into account, the computation becomes very difficult for MCMC treatment. Here we just use posterior modes as point estimates rather than using a fully Bayesian paradigm. Since the posterior distribution is expected to contain multiple modes even in the normal case, and much more critically in the student-*t* case, here we expect to find a local maximum as the posterior mode. The choice of starting value is therefore important. This is why the cluster analysis was used to provide a good starting point.

Using a Bayesian approach here allows us to incorporate prior information (from the manual measurements) into our model. Even though we only compute modes, this will help with the problems described above. The informative prior distributions for the θ_j help to stabilize the computations. The realization of the Bayesian approach in our classification model is given in Section 5.5.2.2, where priors for the parameters θ_j in the model are chosen for tractability.

5.4.4 Expectation maximization (EM) algorithm

Our model fitting is based on the expectation maximization (EM) algorithm. The EM algorithm, first introduced by Dempster, Laird and Rubin (1977), is an iterative method for finding maximum likelihood estimates (MLE) or posterior modes, maximum a posteriori (MAP) estimates, in incomplete-data problems. In the mixture application, the 'missing data' are the unknown group memberships. The E-step of each iteration involves taking expectations of the complete-data log-likelihood function, or log-posterior density function in Bayesian analysis, given the observed data and the current parameter estimates. The M-step of each iteration re-estimates the parameters by maximizing this (expected) complete data likelihood or posterior. The EM algorithm has appealing properties relative to other iterative algorithms. Firstly, it is typically easily implemented because it relies on complete-data computations, which are often in a simple closed form. Secondly, its convergence is stable, with each iteration increasing the likelihood or posterior density.

EM is widely used in imaging analysis with the finite mixture model. An early EM algorithm applied to the Gaussian mixture model can be found in Little and Rubin (1987) and subsequently in McLachlan and Krishnan (1997). An EM algorithm for mixtures of *t*-distributions was presented by Peel and McLachlan (2000). This algorithm introduces an additional set of weights, estimated during the E step, which essentially performs a soft rejection of outliers. A significant advantage of using the *t* distribution is the ability to tune the model's robustness to a particular application or even a particular data set, by varying the degrees of freedom parameter.

5.5 Partially supervised image classification model framework

Motivated by the misclassification problems described in Section 5.1, a partially supervised image classification model employing a Bayesian multivariate finite mixture model is developed in this chapter to recognize the non-nodal area in the image and to further classify each node as normal or metastatic. Prior distributions derived from manual data for normal and metastatic components are imposed onto a multivariate mixture model derived from the scanning data, along the directions of a low-dimensional space which, in one of the approaches, is constructed from both data sets. Although the analysis is essentially a form of unsupervised clustering, it takes place in a space partially determined by the manually measured training data, and uses prior distributions derived from these data. It is therefore referred to as ‘partially supervised’.

5.5.1 Dimension reduction and variable construction

In the context of hyperdimensional image classification using a multivariate finite mixture model, some dimension reduction is essential to enable the feasibility of the multivariate distribution fitting. Reduction to a very small number of dimensions will also avoid excessive memory storage and speed up computation during the process of model fitting and parameter estimation. The advantages of dimension reduction are particularly apparent in the application of real-time prediction.

The high-dimensional spectra are projected into a low-dimensional space constructed to include most of the variations in the original directions of data space. Two different options for dimension reduction are explored here. One, which we call PCA dimension reduction, applies PCA to the spectra from the node under study. The other, which we call discriminant dimension reduction, takes as the first dimension the canonical variate derived from the training data, and adds to this a small number of dimensions derived from a PCA of the variability orthogonal to this in the spectra of the node under study.

In both cases we apply a preliminary PCA dimension reduction to the spectra from the entire set of scanned nodes thus simplifying the calculation and storage of training set means and variances for later use in prior distributions.

5.5.1.1 PCA dimension reduction by global PCA projection followed by local PCA projection

A two-step PCA dimension reduction method projects the spectra of the scanned node of interest onto a low-dimensional space through a global PCA projection followed by a local PCA projection as described below.

A global PCA reduction step extracts the general features of normal, metastatic and non-nodal spectra from all the scanned nodes by projecting the spectra via PCA onto a low-dimensional space (typically with dimensions less than 20) capturing most variations from all the scanned nodes. The global PCA loadings are then used to project the manual measurement data onto the same low-dimensional space. With known class membership for each spectrum in the manual data, the PC scores of manual data are used to derive prior distributions for the parameters of the normal and metastatic groups in the scanning data.

The only point of the global PCA reduction is to reduce the dimension before calculating the priors. If the method is programmed and implemented for real time prediction, these priors need to be stored, and the smaller the number of dimensions the better so long as the space spanned by these dimensions captures most of the important variability in the nodes. Instead of taking large memory space to store all the high-dimensional spectral data, here we only need to store the mean and variance of normal and metastatic groups of manual data in the reduced dimensions for later use as priors for scanning data.

For each node, a local PCA reduction step starting with the scores from the previous step extracts the individual features by finding a small number of axes, typically less than 5 dimensions. The priors are then projected into the space spanned by these axes.

We now describe the procedure in detail.

Step 1: Global PCA projection

A principal component analysis is carried out on the spectra from all the scanned nodes and the first k_g, typically about 15 to 20, PCs are chosen to capture most of the variability (>99%) on all the nodes. The loadings for these PCs are applied to the manually measured spectral data from normal and metastatic tissue as below, and the PC scores are used to derive prior distributions (by calculating mean and variance of the transformed spectral matrix in the space of these k_g PCs) for parameters of the normal and metastatic groups.

$$Z_{all.nodes} = X_{all.nodes}\ L_g \tag{5.6}$$

$$Z_{train} = X_{train}\ L_g \tag{5.7}$$

where $X_{all.nodes}$ is a $c\,n \times p$ spectral matrix of all the scanned nodes (*n* is the number of observations from one scanned node, *c* is the number of scanned nodes, *cn* is the number of observations from all the scanned nodes, *p* is the number of wavelength points). Each row of $X_{all.nodes}$ is a *p*-dimensional spectral vector. L_g is a $p \times k_g$ global PCA loading matrix of the first k_g PCs of $X_{all.nodes}$. $Z_{all.nodes}$ is a $cn \times k_g$ score matrix, the columns of which are the first k_g PC scores of $X_{all.nodes}$.

For the manual measurement data, X_{train}, a $m \times p$ spectral matrix (*m* is the number of spectra) is then converted into a $m \times k_g$ score matrix Z_{train} using the global PCA loading L_g. The $k_g \times 1$ mean vectors $m_n^{'}$, $m_c^{'}$ and the $k_g \times k_g$ covariance matrices $V_n^{'}$, $V_c^{'}$ of Z_{train} are calculated for normal and metastatic spectra.

Step 2: Local PCA projection

For the node of interest, PC scores are calculated in the same k_g-dimensional space as those in step 1 by using the global PCA loadings L_g as below

$$Z_{one.node} = X_{one.node}\ L_g \tag{5.8}$$

where $X_{one.node}$ is a $n \times p$ spectral matrix for this scanned node, $Z_{one.node}$ is a $n \times k_g$ score matrix whose columns contain k_g PC scores (actually the matrix $Z_{one.node}$ is a submatrix of $Z_{all.nodes}$ already calculated). $Z_{one.node}$ is then considered as raw data for this node. A PCA is carried out on $Z_{one.node}$ with the first k_l , typically less than 5, PCs extracted as below

$$S_{one.node} = Z_{one.node}\ L_l \tag{5.9}$$

where L_l is a $k_g \times k_l$ local PCA loading matrix of the first k_l PCs of the score matrix $Z_{one.node}$ of the node, $S_{one.node}$ is a $n \times k_l$ score matrix whose columns consist of k_l PC scores. Each row of $S_{one.node}$ is a k_l-dimensional spectral vector $s_{one.node}$ which will be the spectral data used in our image classification model.

The local PCA loadings L_l are then used as below to convert the $k_g \times 1$ means $m_n^{'}, m_c^{'}$ and $k_g \times k_g$ variances $V_n^{'}, V_c^{'}$ derived in the global PCA step to $k_l \times 1$ means $m_n^{''}, m_c^{''}$ and $k_l \times k_l$ variances $V_n^{''}, V_c^{''}$ for later use in priors for the parameters of the mixture distributions.

$$m_n^{''} = L_l^T m_n^{'}; \qquad m_c^{''} = L_l^T m_c^{'}; \tag{5.10}$$

$$V_n^{''} = L_l^T V_n^{'} L_l; \qquad V_c^{''} = L_l^T V_c^{'} L_l. \tag{5.11}$$

5.5.1.2 Discriminant dimension reduction using an external LDA loading and an internal PCA

The two-step discriminant dimension reduction method described here projects the spectra of each scanned node onto a low-dimensional space composed of an external variable and a small number of internal variables as described below.

The canonical variate derived from a linear discriminant analysis (LDA) on the manually measured data is used as the first axis in the reduced dimensional space for each node. This axis gives maximum separation between normal and metastatic groups in the manual data.

To allow the method to adapt to each node, a small number (e.g., 1-3) of local or internal variables will be added to this external one. However, as with the previously described method, we proceed in two steps. First we extract and store a substantial number (e.g., 15-20) of PCs derived from a PCA of the variance in all the scanned nodes orthogonal to the external variable, capturing as far as possible the variations left by the external variable. As before the main reason from this step is to be able to save in convenient form the means and variances from the manual data.

For each individual node, we then calculate its scores on these 15-20 global PCs and take these as raw data to further extract a small number (e.g., 1-3) of PCs as local variables. The loadings are then used to convert the 15-20 dimensional priors to 1-3 dimensional priors. The external variable and internal (local) variables are combined for use in the image analysis.

We now describe the procedure in detail.

Step 1: Constructing the external variable

A PCDA (a PCA followed by a LDA) is carried out on the manual measurement data, X_{train}, and the first k_{ext} principal components (carrying most of the variability of manual data) are chosen to construct a scalar canonical variable, t_{train}, as a common dimension along which there is the maximum separation between the normal and metastatic groups. The discriminant algorithm is developed by leave-one-out cross-validation on the manual data. The scores on this variable are calculated as

$$T_{train} = X_{train}\, q_{ext}\,, \tag{5.12}$$

where each element of the vector T_{train} is a 1-dimensional canonical score for one spectrum in the manual training data, and q_{ext} is a p-dimensional PCDA loading vector.

By applying the PCDA loading, q_{ext} derived from the manual data to the spectral data from the node of interest, we compute the external variable for this node

$$T_{node.ext} = X_{one.node}\, q_{ext} \tag{5.13}$$

where each point of $T_{node.ext}$ is a 1-dimensional canonical score, $X_{one.node}$ is a $n \times p$ spectral matrix for the scanned node of interest (each row of $X_{one.node}$ is a p-dimensional spectrum). We call q_{ext} the external variable loading here.

Step 2: Constructing the internal variable(s)

Before constructing the internal variables, the first $k_{int.0}$ (e.g., 15-20) PCs of all the scanned nodes are extracted from a space orthogonal to the external variable, capturing as much as possible the variations left by the external variable. Thus

$$T_{all.nodes.ext} = X_{all.nodes}\ q_{ext} \tag{5.14}$$

$$\tilde{X}_{all.nodes} = (I - T_{all.nodes.ext}({T_{all.nodes.ext}}^T T_{all.nodes.ext})^{-1}{T_{all.nodes.ext}}^T)\ X_{all.nodes} \tag{5.15}$$

$$T_{all.nodes.int.0} = \tilde{X}_{all.nodes}\ Q_{int.0} \tag{5.16}$$

where I is a $cn \times cn$ identity matrix $T_{all.nodes.ext}$ is a vector of canonical scores for all the scanned nodes, $\tilde{X}_{all.nodes}$ is a $cn \times p$ spectral matrix whose columns lie in the $p-1$ dimensional subspace orthogonal to the external variable, $Q_{int.0}$ is a matrix, the columns of which are the first $k_{int.0}$ principal component loadings of $\tilde{X}_{all.nodes}$. $T_{all.nodes.int.0}$ is a $cn \times k_{int.0}$ score matrix whose columns contain $k_{int.0}$ PC scores.

Then for the manual measurement data

$$T_{train.int.0} = X_{train}\ Q_{int.0} \tag{5.17}$$

converts X_{train}, a $m \times p$ spectral matrix (m is the number of observations in the manual data) into a $m \times k_{int.0}$ score matrix $T_{train.int.0}$. Means $m_n^{'}$, $m_c^{'}$ and variance matrices $V_n^{'}$, $V_c^{'}$ of the $(1+k_{int.0})$ variables $(T_{train}, T_{train.int.0})$ are calculated for later use in constructing priors.

For the node of interest, its scores on the $k_{int.0}$ PCs are calculated and subjected to a further PCA to extract the first k_{int} (e.g., 1-3) PCs as internal variable(s) by

$$T_{node.int} = T_{node.int.0}\ Q_{int} \tag{5.18}$$

where $T_{node.int.0}$ is a $n \times k_{int.0}$ submatrix of score matrix $T_{all.nodes.int.0}$ already calculated in Equation (5.16) whose columns contain $k_{int.0}$ PC scores, Q_{int} is a matrix the columns of which are the first k_{int} PC loadings derived from a PCA of $T_{node.int.0}$. Each row of $T_{node.int}$ is k_{int}-dimensional internal variable, $t_{node.int}$ which is composed of the first k_{int} PCs in the subspace orthogonal to the direction of external variable $t_{node.ext}$.

The dimension-reduced data vector used in the later data analysis is composed of the external variable and 1-3 internal variables, given by $x = (t_{node.ext},\ t_{node.int})$ with a dimension of $k = 1 + k_{int}$.

Using the internal variable(s) loadings Q_{int}, the $(1 + k_{int.0}) \times 1$ means $m_n^{'}, m_c^{'}$ and the $(1 + k_{int.0}) \times (1 + k_{int.0})$ covariance matrices $V_n^{'}, V_c^{'}$ of $(T_{train}, T_{train.int.0})$ are converted into $(1 + k_{int})$-dimensional means $m_n^{''}, m_c^{''}$ and $(1 + k_{int}) \times (1 + k_{int})$ covariance matrices $V_n^{''}$, $V_c^{''}$ for later use in the priors for normal and metastatic groups of scanning data as below:

$$m_n^{''} = \begin{pmatrix} 1 & 0 \\ 0 & Q_{\text{int}}^T \end{pmatrix} m_n^{'}; \quad m_c^{''} = \begin{pmatrix} 1 & 0 \\ 0 & Q_{\text{int}}^T \end{pmatrix} m_c^{'} \quad ; \tag{5.19}$$

$$V_n^{''} = \begin{pmatrix} 1 & 0 \\ 0 & Q_{\text{int}}^T \end{pmatrix} V_n^{'} \begin{pmatrix} 1 & 0 \\ 0 & Q_{\text{int}} \end{pmatrix}; \quad V_c^{''} = \begin{pmatrix} 1 & 0 \\ 0 & Q_{\text{int}}^T \end{pmatrix} V_c^{'} \begin{pmatrix} 1 & 0 \\ 0 & Q_{\text{int}} \end{pmatrix}. \tag{5.20}$$

By using the external variable we impose one dimension from the training data which we believe can separate normal from metastatic tissue. Using internal variables to add local features preserves the variability specific to this node, which is the unsupervised aspect of the model.

All the models constructed below are based on the data reduced by either the PCA dimension reduction method or the discriminant dimension reduction method.

5.5.2 Multivariate finite mixture model

5.5.2.1 The model

Suppose $x_1,...,x_n$ are k-dimensional random observations generated independently from a mixture of g underlying populations (groups) with unknown mixing proportions $\pi_1,...,\pi_g$ so that

$$f(x_i \mid \phi) = \sum_{j=1}^{g} \pi_j f(x_i \mid \theta_j) \tag{5.21}$$

where $f(x \mid \theta_j)$ denotes the conditional probability density function of x belonging to the jth group, parameterized by θ_j, and where each π_j is nonnegative and $\sum_{j=1}^{g} \pi_j = 1$.

Here $\phi = (\pi_1,...,\pi_g,\theta_1,...,\theta_g)$ denotes the set of unknown parameters. In this thesis, we assume that $g = 3$ with normal, metastatic and non-nodal groups, and x_i is the dimension-reduced spectral data measured at pixel *i* of the image for one node. We will use multivariate Gaussian and *t* distribution for *f*, as described in Section 5.4.1.1 and 5.4.1.2

We introduce the membership indicator variable, $z_i = (z_{i1},...,z_{ig})$, whole role is to encode which component has generated the *i*th observation. This will be the 'missing data' in the EM algorithm. The indicators z_i $(i = 1,...,n)$ are a set of binary variables $z_{ij} \in \{0,1\}$ $(j = 1,...,g)$ with

$$z_{ij} = \begin{cases} 1 & \text{if observation } i \text{ belongs to group } j \\ 0 & \text{otherwise} \end{cases}$$

and hence $\sum_{j=1}^{g} z_{ij} = 1$. If we knew z_i we could write

$$f(x_i \mid z_i, \phi) = \prod_{j=1}^{g} \{f(x_i \mid \theta_j)\}^{z_{ij}} \tag{5.22}$$

Given the mixing probabilities $\pi = (\pi_1,...,\pi_g)$, the indicator variables $z_1,...,z_n$ are independent, with multinomial densities,

$$f(z_i \mid \pi) = \prod_{j=1}^{g} \pi_j^{z_{ij}}, \text{ for } i = 1,...,n. \tag{5.23}$$

Then the joint density of x_i and z_i is given by

$$f(x_i, z_i \mid \phi) = \prod_{j=1}^{g} \{\pi_j f(x_i \mid \theta_j)\}^{z_{ij}} \tag{5.24}$$

5.5.2.2 Priors and posteriors for parameters π_j and θ_j

We take a uniform prior for the mixing probability π_j. For the finite mixture model the normal inverse Wishart prior (Gelman *et al.*1995) is used here as a prior for the parameters μ_j and Σ_j of both the *k*-dimensional multivariate Gaussian distribution (Fraley and Raftery 2007) and the *k*-dimensional multivariate *t* distribution (Lin *et al.* 2004). We use normal priors on the mean vectors μ_j conditional on the scale matrices Σ_j (Richardson and Green 1997, Stephens 1997):

$$\mu_j \,\big|\, \Sigma_j \;\sim\; N(\mu_{jp}, \kappa_{jp}^{-1}\Sigma_j)$$

so that $$p(\mu_j \mid \Sigma_j, \mu_{jp}, \kappa_{jp}) \propto \left|\Sigma_j\right|^{-1/2} \exp\left\{-\frac{\kappa_{jp}}{2}\left[(\mu_j - \mu_{jp})^T \Sigma_j^{-1}(\mu_j - \mu_{jp})\right]\right\}, \qquad (5.25)$$

and inverse Wishart priors on the scale matrices Σ_j, as suggested in Raftery (1996):

$$\Sigma_j \;\sim\; IW(v_{jp}, \Lambda_{jp})$$

so that $$p(\Sigma_j \big| v_{jp}, \Lambda_{jp}) \;\propto\; \left|\Sigma_j\right|^{-\frac{v_p+k+1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\Sigma_j^{-1}\Lambda_{jp}^{-1}\right]\right\}. \qquad (5.26)$$

The hyperparameters μ_{jp}, κ_{jp} and v_{jp} $(j = 1, 2, 3)$ are called the mean, shrinkage and degrees of freedom respectively, of the prior distribution. The hyperparameter Λ_{jp}^{-1}, which is a matrix of the same dimension as Σ_j, is called the scale of the inverse Wishart prior. Here the suffix *p* does not refer to the number of wavelength, but is used to indicate a hyperparameter.

The joint prior density $p(\theta_j)$ is therefore

$$p(\mu_j, \Sigma_j \,\big|\, \mu_{jp}, \kappa_{jp}, v_{jp}, \Lambda_{jp})$$
$$\propto \left|\Sigma_j\right|^{-(v_{jp}+k+2)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Sigma_j^{-1}\Lambda_{jp}^{-1})\right\} \exp\left\{-\frac{\kappa_{jp}}{2}\left[(\mu_j - \mu_{jp})^T \Sigma_j^{-1}(\mu_j - \mu_{jp})\right]\right\}$$

(5.27)

Integrating joint prior density in Equation (5.27) with respect to Σ we get the marginal prior density for the mean:

$$p(\mu_j) \propto t_{v_p - k+1}(\mu_{jp}, \Lambda_{jp}^{-1}/(\kappa_{jp}(v_{jp} - k + 1))), \qquad (5.28)$$

where t_{v_p-k+1} is a multivariate *t*-distribution with $v_p - k + 1$ degrees of freedom.

In the case of the multivariate Gaussian distribution $N(\mu_j, \Sigma_j)$ the normal inverse Wishart prior is conjugate so that the posterior can also be expressed as the product of a normal distribution and an inverse Wishart distribution with different parameters. Given data $x = \{x_1, ..., x_n\}$ from this Gaussian, i.e., data known to be from this group, we have

$$\mu_j \mid \Sigma_j, x \sim N(\tilde{\mu}_{jp}, \tilde{\kappa}_{jp}^{-1}\Sigma_j), \tag{5.29}$$

$$\Sigma_j \mid x \sim IW(\tilde{v}_{jp}, \tilde{\Lambda}_{jp}), \tag{5.30}$$

where

$$\tilde{\kappa}_{jp} = \kappa_{jp} + n, \quad \tilde{v}_{jp} = v_{jp} + n, \tag{5.31}$$

$$\tilde{\mu}_{jp} = \left(\frac{n}{\kappa_{jp} + n}\right)\bar{x} + \left(\frac{\kappa_{jp}}{\kappa_{jp} + n}\right)\mu_{jp}, \tag{5.32}$$

and

$$\tilde{\Lambda}_{jp}^{-1} = \Lambda_{jp}^{-1} + \left(\frac{\kappa_{jp} n}{\kappa_{jp} + n}\right)(\bar{x} - \mu_{jp})(\bar{x} - \mu_{jp})^T + \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T \tag{5.33}$$

which can also be expressed as

$$\tilde{\Lambda}_{jp}^{-1} = \Lambda_{jp}^{-1} + \kappa_{jp}(\hat{\mu}_j - \mu_{jp})(\hat{\mu}_j - \mu_{jp})^T + \sum_{i=1}^{n}(x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^T. \tag{5.34}$$

Here $\bar{x} = \sum_{i=1}^{n} x_i$ is sample mean of one scanned node, x_i is the dimension-reduced spectral data measured at pixel *i* of the image for one node, and *n* is the number of observations from one scanned node.

The joint posterior modes of the mean vector and the covariance matrix are

$$\hat{\mu}_j = \tilde{\mu}_{jp} \tag{5.35}$$

and

$$\hat{\Sigma}_j = \frac{\tilde{\Lambda}_{jp}^{-1}}{\tilde{v}_{jp} + k + 2}. \tag{5.36}$$

In the case of the multivariate *t* distribution $t(\mu_j, \Sigma_j, v_j)$, the normal inverse Wishart prior is not conjugate. However an EM algorithm can be used to estimate μ_j and Σ_j. An additional set of weights, viewed as missing data here,

$$\{u_i,\ i = 1,\ldots,n\} \tag{5.37}$$

are introduced, one corresponding to each of the observations x_i, so that

$$x_i \mid u_i \sim N(\mu_j, \Sigma_j / u_i), \tag{5.38}$$

independently for $i = 1, \ldots, n,$ and

$$u_i \sim G(v_j/2, v_j/2) \tag{5.39}$$

are independently distributed for $i = 1, \ldots, n$.

The complete-data log-posterior density function (distribution) is

$$\ell(\theta_j \mid \{x_i\}\{u_i\}) = \sum_{i=1}^{n} [\log p(x_i \mid u_i, \theta_j) + \log\, p(u_i \mid v_j)] + \log P(\theta_j)\,. \qquad (5.40)$$

In the E-step the conditional expectation of the complete-data log-posterior function

$$Q(\theta_j) = E[\ell(\theta_j \mid \{x_i\}\{u_i\})] = E\left\{\sum_{i=1}^{n} [\log p(x_i \mid u_i, \theta_j) + \log\, p(u_i \mid v_j)] + \log P(\theta_j)\right\} \qquad (5.41)$$

is with respect to the conditional distribution of $\{u_i\}$ given $\{x_i\}$ and current estimates $\hat{\theta}_j$.

In general it is possible to estimate v_j, but in our application to mixture models we will take v_j as fixed. In this case the part of $Q(\theta_j)$ in Equation (5.41) we have to compute in M-step (i.e., ignoring the terms not involving θ_j) expands to

$$E\left\{\sum_{i=1}^{n} [\log |\Sigma_j|^{-1/2} - \tfrac{1}{2}\, u_i\, (x_i - \mu_j)^T\, \Sigma_j^{-1} (x_i - \mu_j)] + \log P(\theta_j)\right\},$$

which has a linear dependence on u_i. Thus in the (t+1)st iteration of the EM algorithm, computing $Q(\theta_j)$ in the E-step involves replacing u_i by its expectation $E(u_i \mid x, \hat{\theta}^{(t)})$.

$$\hat{u}_i^{(t+1)} = E(u_i \mid x, \hat{\theta}_j^{(t)})\,. \qquad (5.42)$$

Since the gamma distribution is conjugate,

$$u_i \mid x_i \sim G(\tfrac{1}{2}(v_j + k), \tfrac{1}{2}(v_j + (x_i - \hat{\mu}_j^{(t)})^T\, \hat{\Sigma}_j^{(t)^{-1}} (x_i - \hat{\mu}_j^{(t)})))\,, \qquad (5.43)$$

so that

$$\hat{u}_i^{(t+1)} = \frac{v_j + k}{v_j + (x_i - \hat{\mu}_j^{(t)})^T\, \hat{\Sigma}_j^{(t)^{-1}} (x_i - \hat{\mu}_j^{(t)})}\,. \qquad (5.44)$$

In the M-step at the (t+1)st iteration of EM algorithm, we substitute $\hat{u}_i^{(t+1)}$ for u_i in the first term on the right hand side of Equation (5.40), then the posterior estimates of μ_j and Σ_j are updated by computing the modal estimates:

$$\hat{\mu}_j^{(t+1)} = \frac{\kappa_{jp}\mu_{jp} + \sum_{i=1}^{n} \hat{u}_i^{(t+1)} x_i}{\kappa_{jp} + \sum_{i=1}^{n} \hat{u}_i^{(t+1)}} \qquad (5.45)$$

$$\hat{\Sigma}_j^{(t+1)} = \frac{\Lambda_{jp}^{-1} + \kappa_{jp}(\hat{\mu}_j^{(t+1)} - \mu_{jp})(\hat{\mu}_j^{(t+1)} - \mu_{jp})^T + \sum_{i=1}^{n} \hat{u}_i^{(t+1)}(x_i - \hat{\mu}_j^{(t+1)})(x_i - \hat{\mu}_j^{(t+1)})^T}{v_{jp} + n + k + 2} \quad (5.46)$$

In the estimation of the mean and variance the observation x_i is weighted by $u_i^{(t+1)}$, the current estimate of the scale factor. When u_i is equal to 1, Equations (5.45) and (5.46) are equivalent to Equation (5.35) and (5.36), which makes the Gaussian distribution a special case of *t* distribution.

For a single *t* distribution, the computations are more complicated than for a single Gaussian, since we need to introduce an EM iteration. However in the mixture model, the iterative estimation of u_i can be incorporated in the EM we will be doing anyway for the mixture, and so costs nothing extra.

5.5.3 Model fitting algorithm and parameter estimation

Since both the class label and the parameters are unknown and they are strongly inter-dependent, the problem of parameter estimation is regarded as an “incomplete-data” problem. Many techniques have been proposed to solve this problem, among which the expectation-maximization (EM) algorithm described in Section 5.4.4 is the one most widely used. Given initial values for the parameters ϕ of the multivariate finite mixture model, in the E-step we calculate the conditional probability that pixel *i* belongs to the *j*th component in the mixture and use this to compute the conditional expectation of the complete-data log posterior density function; in M-step we then update the parameters by maximizing this log density. The process is iterated until convergence. In this chapter the class membership y_i of pixel *i* in the image is assumed independent of all the other pixels, without considering the spatial correlation of the image. The complete algorithm is described in a flowchart in Figure 5.8. To distinguish this algorithm from the one with spatial prior introduced in Chapter 6 (Figure 6.3), we refer to the algorithm in this chapter as the first stage algorithm and the one in Chapter 6 as the second stage algorithm.

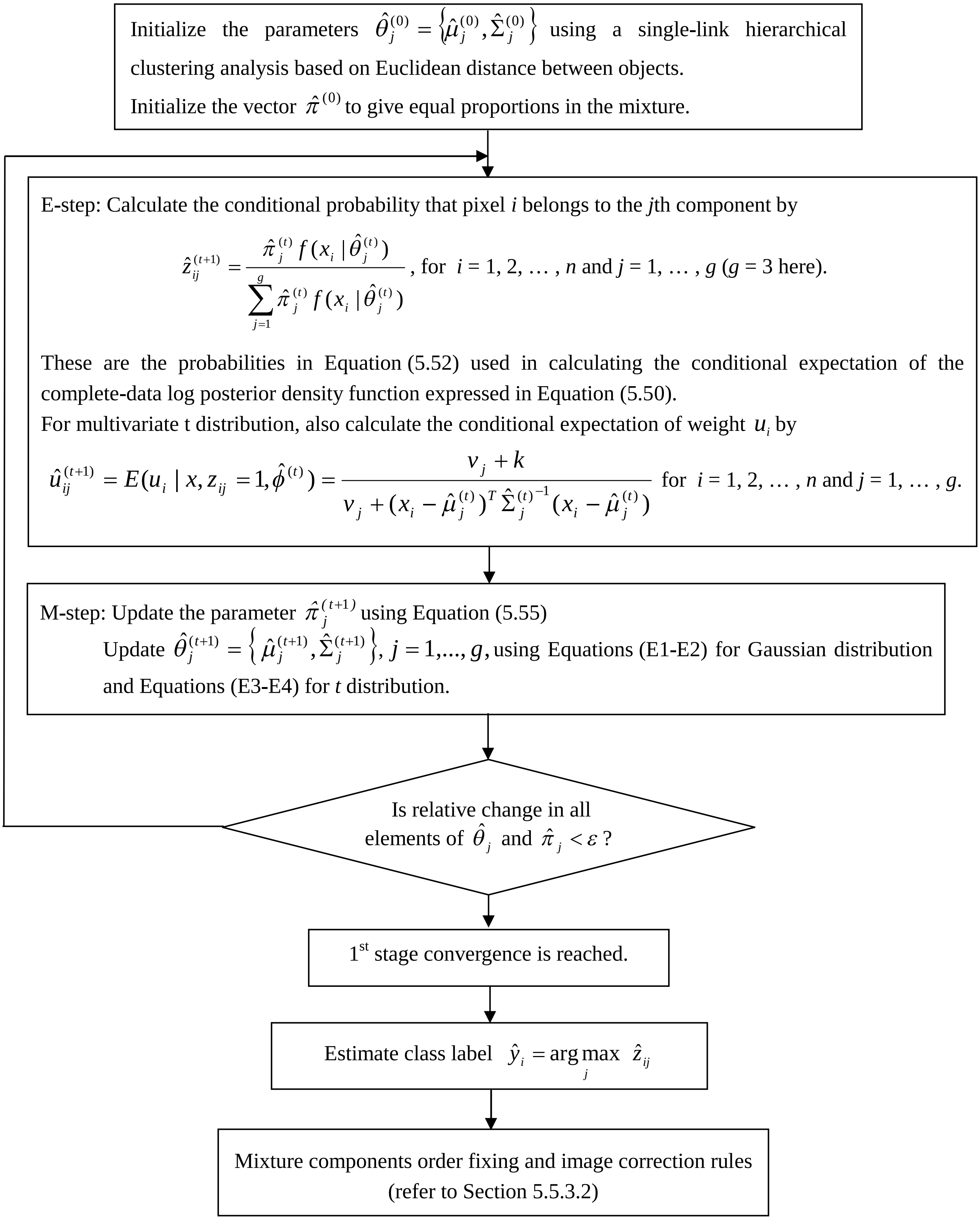

Initialize the parameters $\hat{\theta}_j^{(0)} = \left\{\hat{\mu}_j^{(0)}, \hat{\Sigma}_j^{(0)}\right\}$ using a single-link hierarchical clustering analysis based on Euclidean distance between objects.
Initialize the vector $\hat{\pi}^{(0)}$ to give equal proportions in the mixture.
E-step: Calculate the conditional probability that pixel *i* belongs to the *j*th component by
$\hat{z}_{ij}^{(t+1)} = \frac{\hat{\pi}_j^{(t)} f(x_i \mid \hat{\theta}_j^{(t)})}{\sum_{j=1}^{g} \hat{\pi}_j^{(t)} f(x_i \mid \hat{\theta}_j^{(t)})}$, for i = 1, 2, … , *n* and *j* = 1, … , *g* (*g* = 3 here).
These are the probabilities in Equation (5.52) used in calculating the conditional expectation of the complete-data log posterior density function expressed in Equation (5.50).
For multivariate t distribution, also calculate the conditional expectation of weight u_i by
$\hat{u}_{ij}^{(t+1)} = E(u_i \mid x, z_{ij} = 1, \hat{\phi}^{(t)}) = \frac{v_j + k}{v_j + (x_i - \hat{\mu}_j^{(t)})^T \hat{\Sigma}_j^{(t)^{-1}} (x_i - \hat{\mu}_j^{(t)})}$ for i = 1, 2, … , *n* and *j* = 1, … , *g*.
M-step: Update the parameter $\hat{\pi}_j^{(t+1)}$ using Equation (5.55)
Update $\hat{\theta}_j^{(t+1)} = \left\{\hat{\mu}_j^{(t+1)}, \hat{\Sigma}_j^{(t+1)}\right\}, j = 1, ..., g,$ using Equations (E1-E2) for Gaussian distribution and Equations (E3-E4) for *t* distribution.
Is relative change in all elements of $\hat{\theta}_j$ and $\hat{\pi}_j < \varepsilon$?
1st stage convergence is reached.
Estimate class label $\hat{y}_i = \arg\max_j \hat{z}_{ij}$
Mixture components order fixing and image correction rules (refer to Section 5.5.3.2)

Figure 5.8: The first stage algorithm flowchart**.**

5.5.3.1 EM for mixtures of multivariate Gaussian and t distributions via Bayesian theory

(1) The E step evaluates the conditional expectation of the complete data log-posterior density given the observed data and the current parameter estimates.

For the mixtures of multivariate Gaussian distribution, the "complete data" here are $(\{x_i\},\{z_i\})$, where $z_i = (z_{i1},...,z_{ig})$ is the unobserved portion of the data (the unknown class label), with z_{ij}, a binary indicator variable which takes the value 1 when observation *i* came from mixture component *j* and zero otherwise. From Equation (5.24) the complete-data likelihood for a mixture model with *g* components is

$$L_{mix}(\{x_i\},\{z_i\}\,|\,\phi) = \prod_{i=1}^{n}\prod_{j=1}^{g}\{\pi_j f(x_i\,|\,\theta_j)\}^{z_{ij}} . \tag{5.47}$$

Thus the Bayesian posterior density is of the form

$$L(\phi\,|\,\{x_i\},\{z_i\}) = \prod_{i=1}^{n}\prod_{j=1}^{g}\{\,\pi_j f(x_i\,|\,\theta_j)\}^{z_{ij}}\prod_{j=1}^{g}p(\theta_j) \tag{5.48}$$

and the log-posterior distribution is

$$\ell(\phi,\big|\,\{x_i\},\{z_i\}) = \sum_{i=1}^{n}\sum_{j=1}^{g}z_{ij}(\log f(x_i\,|\,\theta_j) + \log\pi_j) + \sum_{j=1}^{g}\log p(\theta_j)\ , \tag{5.49}$$

where $P(\theta_j)$ is the normal inverse Wishart prior distribution on the parameters θ_j and we have taken a uniform prior for the π_j.

The conditional expectation of the complete-data log posterior density function is

$$Q(\phi) = E[l(\phi\big|\,\{x_i\},\{z_i\})] = E\left\{\sum_{i=1}^{n}\sum_{j=1}^{g}z_{ij}(\log f(x_i\,|\,\theta_j) + \log\pi_j) + \sum_{j=1}^{g}\log p(\theta_j)\right\} \tag{5.50}$$

where the expectation is with respect to the conditional distribution of $\{z_i\}$ given $x = \{x_i\}$ and the current estimate $\hat{\phi}$.

Since $Q(\phi)$ in Equation (5.50) has a linear dependence on z_{ij}, calculating the expectation of Q simply requires replacing z_{ij} by their expectations,

$$\hat{z}_{ij} = E(z_{ij} \mid x, \hat{\phi}) = P(y_i = j \,\big|\, x, \hat{\phi}), \tag{5.51}$$

where $\hat{z}_{ij}$ is the conditional probability that pixel *i* belongs to the *j*th component given data x and the current parameter estimates $\hat{\theta}, \hat{\pi}$. At the (*t*+1)st iteration, the updating equation for $\hat{z}_{ij}$ is as follows:

$$\hat{z}_{ij}^{(t+1)} = \frac{\hat{\pi}_j^{(t)} f(x_i \mid \hat{\theta}_j^{(t)})}{\sum_{j=1}^{g} \hat{\pi}_j^{(t)} f(x_i \mid \hat{\theta}_j^{(t)})}. \tag{5.52}$$

For mixtures of multivariate *t* distributions, given additional missing data $\{u_i\}$ as defined in Equations (5.37)-(5.39), the "complete data" are $(\{x_i\}, \{z_i\}, \{u_i\})$. The conditional expectation of the complete-data log posterior density function is

$$Q(\phi) = E[l(\phi|\ \{x_i\}, \{z_i\}, \{u_i\})]$$

$$= E\left\{ \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij} (\log f(x_i \mid u_i, z_{ij} = 1, \theta_j) + \log f(u_i \mid v_j) + \log \pi_j) + \sum_{j=1}^{g} \log p(\theta_j) \right\}. \tag{5.53}$$

Calculation of Equation (5.53) in the E-step can be effected by first taking the expectation of u_i conditional on $\{z_i\}$ and $\{x_i\}$, and then finally over the z_i given $\{x_i\}$. For these calculations we need $E(z_{ij} \mid x, \hat{\phi})$ as given in Equation (5.52) above and $E(u_i \mid x, z_{ij} = 1, \hat{\phi}^{(t)})$ for $i = 1,...,n$ and $j = 1,...,g$,

$$\hat{u}_{ij}^{(t+1)} = E(u_i \mid x, z_{ij} = 1, \hat{\phi}^{(t)}) = \frac{v_j + k}{v_j + (x_i - \hat{\mu}_j^{(t)})^T \hat{\Sigma}_j^{(t)^{-1}} (x_i - \hat{\mu}_j^{(t)})}. \tag{5.54}$$

(2) The M step for the multivariate Gaussian distribution involves maximizing Equation (5.50) over π_j and θ_j with z_{ij} fixed at the values $\hat{z}_{ij}$ computed in the E step. For π_j, let $\hat{n}_j^{(t+1)} = \sum_{i=1}^{n} \hat{z}_{ij}^{(t+1)}$, then we have

$$\hat{\pi}_j^{(t+1)} = \frac{\hat{n}_j^{(t+1)}}{n}. \tag{5.55}$$

The results for θ_j are given in Equations (E1-E2) of Table 5.1.

For the multivariate t distribution, the conditional expectation function to be maximized in Equation (5.53) involves the additional weight $\hat{u}_{ij}$. The result for π_j is the same as in Equation (5.55). The results for θ_j are given in Equations (E3-E4) of Table 5.1.

The whole EM algorithm is judged to have converged when the relative changes in all elements of θ_j and π_j are less than ε. Here we choose the value 0.1 for ε in order to achieve a weak convergence in this stage, which merely provides a stationary configuration for the second stage.

An alternative procedure for fitting mixture models by using the classification Expectation Maximization (CEM) (Celeux and Govaert 1992, Dean *et al.* 2006) algorithm is almost identical to the EM algorithm, except that it incorporates a classification step between the E-step and M-step replacing the $\hat{z}_{ij}$ by:

$$\text{new } \hat{z}_{ij}^{(k+1)} = \begin{cases} 1, & \text{if } \hat{z}_{ij}^{(k+1)} > \hat{z}_{ij'}^{(k+1)} \text{ for all } (j' \neq j), \\ 0, & \text{otherwise.} \end{cases} \tag{5.56}$$

That is, a discrete classification is made where each object is assigned to a unique group. For example, if EM algorithm produces $\hat{z}_{ij}^{(k+1)} = (0.8, 0.05, 0.15)$ then the equivalent CEM step would transform this to a new $\hat{z}_{ij}^{(k+1)} = (1, 0, 0)$. These values of $\hat{z}_{ij}^{(k+1)}$ are used in the M-step of the CEM algorithm.

The CEM algorithm converges faster, but has a tendency to eliminate genuine small groups.

Table 5.1: M-step estimators for the mean and variance of multivariate mixture models under the normal inverse Wishart conjugate priors. Here $\hat{z}_{ij}^{(t+1)}$ is the conditional probability that observation *i* belongs to the *j*th component at (*t*+1)st iteration.

$$\hat{n}_j^{(t+1)} \equiv \sum_{i=1}^{n} \hat{z}_{ij}^{(t+1)} \ , \ \hat{u}_{ij}^{(t+1)} = \frac{k + v_j}{(x_i - \hat{\mu}_j^{(t)})^T \hat{\Sigma}_j^{(t)^{-1}} (x_i - \hat{\mu}_j^{(t)}) + v_j}$$

Distribution	Parameter	M-step estimates of mean and variance with Prior
Multivariate Gaussian distribution	μ_j	$\hat{\mu}_j^{(t+1)} = \frac{\kappa_{jp}\mu_{jp} + \sum_{i=1}^{n} \hat{z}_{ij}^{(t+1)} x_i}{\kappa_{jp} + \hat{n}_j^{(t+1)}}$ (E1)
	Σ_j	$\hat{\Sigma}_j^{(t+1)} = \frac{\Lambda_{jp}^{-1} + \kappa_{jp}(\hat{\mu}_j^{(t+1)} - \mu_{jp})(\hat{\mu}_j^{(t+1)} - \mu_{jp})^T + \sum_{i=1}^{n} \hat{z}_{ij}^{(t+1)} (x_i - \hat{\mu}_j^{(t+1)})(x_i - \hat{\mu}_j^{(t+1)})^T}{v_{jp} + \hat{n}_j^{(t+1)} + k + 2}$ (E2)
Multivariate *t* distribution	μ_j	$\hat{\mu}_j^{(t+1)} = \frac{\kappa_{jp}\mu_{jp} + \sum_{i=1}^{n} \hat{z}_{ij}^{(t+1)} \hat{u}_{ij}^{(t+1)} x_i}{\kappa_{jp} + \sum_{i=1}^{n} \hat{z}_{ij}^{(t+1)} \hat{u}_{ij}^{(t+1)}}$ (E3)
	Σ_j	$\hat{\Sigma}_j^{(t+1)} = \frac{\Lambda_{jp}^{-1} + \kappa_{jp}(\hat{\mu}_j^{(t+1)} - \mu_{jp})(\hat{\mu}_j^{(t+1)} - \mu_{jp})^T + \sum_{i=1}^{n} \hat{z}_{ij}^{(t+1)} \hat{u}_{ij}^{(t+1)} (x_i - \hat{\mu}_j^{(t+1)})(x_i - \hat{\mu}_j^{(t+1)})^T}{v_{jp} + \hat{n}_j^{(t+1)} + k + 2}$ (E4)

5.5.3.2 Mixture components order fixing and image correction rules

The mixture components order fixing and image correction rules at the end of algorithm as shown in Figure 5.8 came into existence here because of two problems in the Bayesian mixture model, one is a label-switching problem, the other is a number of components problem.

The existence of multiple MLE solution to the mixture problem, corresponding to permutations of the labels, is well known (Titterington *et al.* 1985). A Bayesian analysis with informative prior distributions is less prone to such problems, but our use of a diffuse prior for the background component does cause difficulties here.

The number of components problem arises because we assume for each node there are three components in the mixture. The model is thus trying to fit the data with three components, which is usually but not always correct. This may result in the misclassification of some small cluster of data which is actually not a separate group.

When a node is analysed, we first run the model described in this chapter, then use the result as a starting point for the more sophisticated model of Chapter 6. In between, we apply the rules described below to fix the problems of incorrect labelling and spurious groups.

(a) Mixture component order fixing

Since non-nodal component is more likely to appear on the fringe area of the node, in order to easily recognize the non-nodal component first, here we introduce a distance d_i and a background score ω_i . The d_i is used to measure the distance between each pixel and the central point of the image at (10.5, 10.5), and varies from 0.7 for the most central points to 13.4 in the corners. As a function of d_i , ω_i gives a background score to each pixel as below:

$$\omega_i = f(d_i) = \begin{cases} (d_i/13.4)^{1/\rho}, & \text{if } d_i > \sqrt{7.5^2 + 0.5^2} \\ d_i/13.4, & \text{otherwise} \end{cases}$$

$$d_i = \sqrt{(r_i - 10.5)^2 + (s_i - 10.5)^2}\ ,\ r_i \in \{1,2,...,20\},\ s_i \in \{1,2,...,20\} \qquad (5.57)$$

where r_i, s_i denote the row position (index) and the column position (index) of the pixel *i* in the image, the power ρ (for which we have tried values in the range 1-5) is used to emphasize the scores for the pixels on the edge. Here we choose $\sqrt{7.5^2 + 0.5^2}$ as a threshold for the piecewise function of d_i. This represents the radius of a circle that reaches to two pixels from the edge of the image. By these definitions, pixels on the corner or the edge of the image are given higher background score than those in the centre. These scores will be used to identify the non-nodal component in the mixture, as the one receiving the highest average background score.

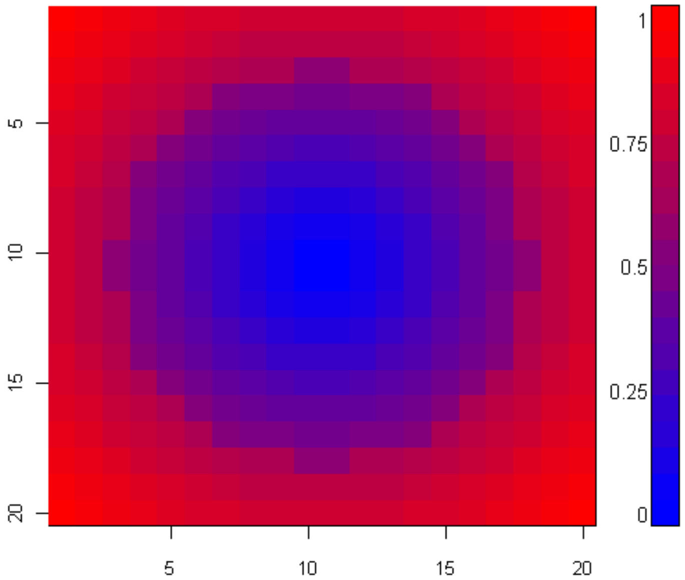


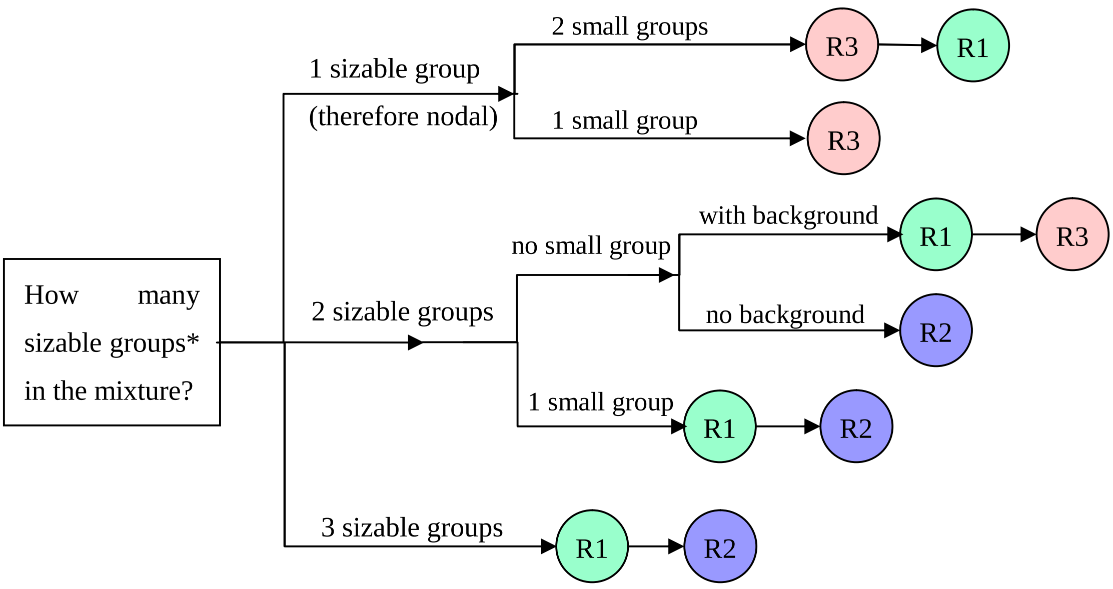
Figure 5.9: The mapping of the position weight of each pixel with the scores of distance parameter d_i plotted with false coded colour (red shows high probability being background pixel, blue shows low probability being background pixel).

The effect of the choices of ω_i on the probability of each pixel being background is represented in Figure 5.9 with false coded color (red corresponding to pixel indicative of background with high probability, blue corresponding to pixel indicative of background with low probability). Here we use $\rho = 1.5$

The two nodal components can then be distinguished by comparing their mean scores with prior means for the components: for the discriminant reduction, the prior mean of the first dimensional variable (external variable) of the metastatic component is higher than that of normal component; for PCA reduction, the reverse is true. These rules work well for the situation where there exist three components in the mixture and they are all sizable groups.

However there are some situations where a very small nodal group concentrated on the fringe of the image will have higher average background scores than a more spread out non-nodal group, and the nodal component might then be misclassified as a non-nodal component. To avoid this situation, we make some adjustments to the background score for small groups. For the situation where only one sizable nodal component exists, we find its class label by comparing its mean score with the prior means of both nodal components. The complete rules for mixture component order fixing are given in detail in Figure 5.10.

Step 1: Mixture components order fixing



* Sizable group refers to the group with size larger than or equal to 30.

R1 Calculate $g_j = \sum_i \omega_i / n_j$, with background score ω_i at pixel i, summed over the n_j points assigned to the jth component of the mixture. If exactly 2 sizable groups exist and 1 small group with current label *k*, then adjust $g_k = 0.7 g_k$. Background index = $\arg\max_j \{g_j\}$.

R2 Let k_1, k_2 be the labels of the two nodal groups. For discriminant reduction, meta index $= \arg\max_{j=k_1,k_2} \{\hat{\mu}_j[1]\}$; for PCA reduction, meta index $= \arg\min_{j=k_1,k_2} \{\hat{\mu}_j[1]\}$.

R3 For one sizable nodal group with current label k, if $\arg\min_{j=1,2}\{|\hat{\mu}_k[1] - \hat{\mu}_{jp}[1]|\} = 1$, then normal index $= k$; if $\arg\min_{j=1,2}\{|\hat{\mu}_k[1] - \hat{\mu}_{jp}[1]|\} = 2$, then meta index $= k$.

If 1 small group also exists, and if its current label is background, leave it; if not background, it belongs to the other nodal group.

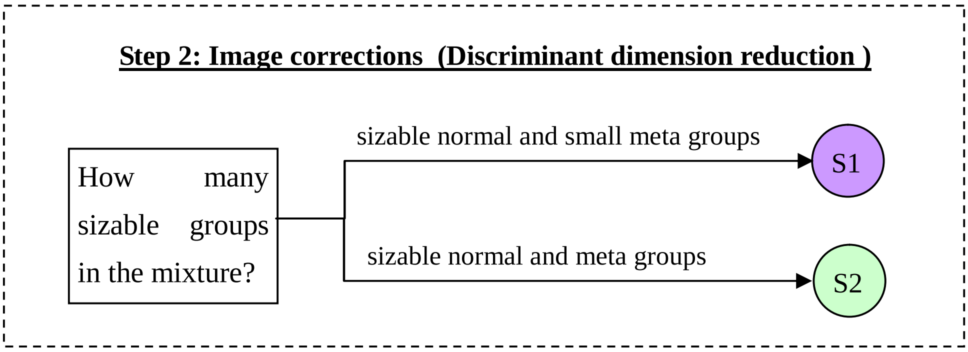
Figure 5.10: Mixture components order fixing algorithm flowchart. The set of 3 labels {1, 2, 3} corresponds to normal, metastatic and non-nodal groups, respectively.

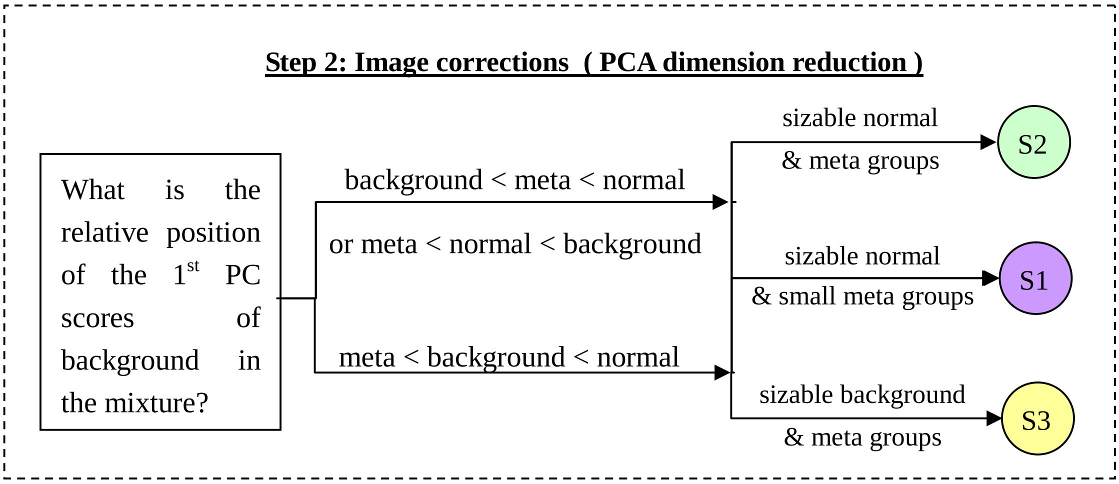
(b) Image correction rules

Given three components in the mixture, the model will try to find three groups in the image. However, when there are really only one or two components in the node, some misclassification might happen, especially to some small nodal groups.

The image correction rules attempt to detect and rectify this situation. For the case where there is one sizable normal group and one small metastatic group, if amongst the pixels from fitted metastatic group the maximum conditional probability of a pixel belonging to metastatic group is very low, then we consider the metastatic group actually belongs to normal group or non-nodal group. For the case where both normal and metastatic groups are sizable, if the mean score on the first dimension of the metastatic component is much closer to the mean score of its neighbour component than to the prior mean of metastatic component, we will consider the fitted metastatic group may belong to normal group.

For the PCA dimension reduction method, since for each node the spectra and the prior means are all projected onto a different low-dimensional space, especially for the variable non-nodal component, the image correction rules given here also depend on the different relative positions of the first dimensional scores of the non-nodal component in the mixture. The image correction rules are given in detail in Figure 5.11.


Step 2: Image corrections (Discriminant dimension reduction)
How many sizable groups in the mixture?
sizable normal and small meta groups
S1
sizable normal and meta groups
S2


Step 2: Image corrections (PCA dimension reduction)
What is the relative position of the 1st PC scores of background in the mixture?
background < meta < normal
or meta < normal < background
meta < background < normal
sizable normal
& meta groups
S2
sizable normal
& small meta groups
S1
sizable background
& meta groups
S3

S1: If 1 sizable normal group and 1 small meta group with current label k,

if $\max_i (z_{ik}) < 4.5 / \hat{n}_k$, then merge meta group to the closest group, either normal or non-nodal group.

S2: Both normal and meta sizable groups with labels of k_1 and k_2. If $|\hat{\mu}_{k_1}[1]-\hat{\mu}_{k_2}[1]|<|\hat{\mu}_{2p}[1]-\hat{\mu}_{k_2}[1]|$, then merge meta group to normal group.

S3: Both meta and background sizable groups with labels of k_2 and k_3. If $|\hat{\mu}_{k_3}[1]-\hat{\mu}_{k_2}[1]|<|\hat{\mu}_{2p}[1]-\hat{\mu}_{k_2}[1]|$, then merge meta group to background group.

Figure 5.11: Image correction algorithm flowchart for discriminant dimension reduction and PCA dimension reduction.

CHAPTER 6

IMAGING CLASSIFICATION WITH SPATIAL PRIOR ON SCANNED SENTINEL LYMPH NODES

6.1 Introduction

The imaging model presented in Chapter 5 takes the class memberships of the pixels in the image to be independent, thus ignoring the spatial contextual information. In the true image adjacent pixels tend to belong to the same group. The model in this chapter takes into account the spatial correlation in the image by adding a Markov random field spatial prior to the previous model. Since the property of neighbourhood contiguity of the image is now considered, the classification model framework aims at generating an image with smooth pattern, comparable to the real tissue structure of the node.

6.2 Theory

6.2.1 Markov random fields

MRF theory

Markov random fields (MRF) provide a powerful and robust framework for modeling spatial interactions between neighboring or nearby pixels. The local correlations provide a mechanism for modeling a variety of image properties (Chellappa and Jain 1993, Li 2001). In medical imaging, they are typically used because most pixels belong to the same class as their neighboring pixels. In physical terms, this implies that any anatomical structure that consists of only one pixel has a very low probability of occurring.

In an MRF, the sites in S are related to one another via a neighbourhood system, $N = \{N_i, i \in S\}$, where N_i denotes the set of neighbours of i. The system is symmetric, so that $i \in N_j \Leftrightarrow j \in N_i$ and no site is its own neighbour, so that $i \notin N_i$. A random field Y is said to be an MRF on S with respect to a neighborhood system N if and only if

$$p(y_i \mid y_{S\setminus\{i\}}) = p(y_i \mid y_{N_i}) \tag{6.1}$$

$$p(y) > 0\,,\ \forall y \in Y \tag{6.2}$$

where $y_{S\setminus\{i\}}$ denotes a realization of the field restricted to $S \setminus \{i\} = \{j \in S, j \neq i\}$.
The neighbour set of $s = (i, j)$ for a regular lattice S is commonly defined as

$$N_s = \{r = (k,l) \in S : 0 < (k-i)^2 + (l-j)^2 \le o\,, \tag{6.3}$$

where o is the order of the neighbourhood system. Figure 6.1 shows the typical examples used in image analysis, the two dimensional lattice with a first order and second order neighbourhood systems.

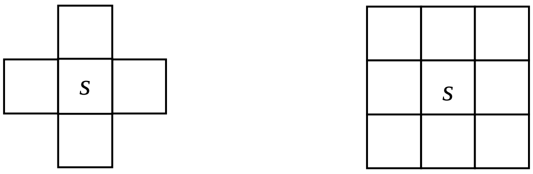


Figure 6.1: MRF neighbourhood systems for site $s \in S$. Left: first-order; Right: second-order.

Property (6.1) means that in an MRF, the interactions between site i and the other sites actually reduce to interactions with its neighbours, and that according to the Hammersley-Clifford theorem, established by Hammersley and Clifford (1971) and further developed by Besag (1974), the joint probability distribution of a MRF (an image labeling configuration) can be characterized by a Gibbs distribution given by

$$P(y) = Z^{-1} \exp(-U(y)), \qquad (6.4)$$

where Z is a normalizing constant called the partition function, and $U(y)$ is an energy function which is a sum of clique potentials over all possible cliques c. A clique c for (S, N) is a subset of pixels in S chosen to represent the spatial interaction. The computation of Z involves all possible realizations y of the MRF, and therefore, it is in general not computationally feasible. More detail on MRFs and the role of the Gibbs distribution can be found in the paper by Geman and Geman (1984)

Hidden Markov Random Fields (HMRF)

The HMRF model is popular in the statistical analysis of pixellated images (Geman and Geman 1984). In image analysis, problems involving incomplete data are common. As described in Chapter 5, the complete data here include the observations representing measurements, e.g. multivariate variables recorded for each pixel of an image, and the missing data (hidden data) consisting of unknown class assignment to be estimated from the observations for each pixel. The complete likelihood is given by

$$f(x, y \mid \theta, \beta) = f(x \mid y, \theta) P(y \mid \beta) \qquad (6.5)$$

where θ are parameters of the distribution generating X, and β are parameters of the MRF.

In hidden Markov models, the observations X are assumed to be conditionally independent given the image configuration $Y = y$, we then have the joint conditional probability of x given y

$$f(x \mid y, \theta) = \prod_{i \in S} f(x_i \mid y_i, \theta). \qquad (6.6)$$

Equation (6.5) is then expressed as

$$f(x, y \mid \theta, \beta) = Z^{-1} \exp\{-U(y \mid \beta) + \sum_{i \in S} \log f(x_i \mid y_i, \theta)\}, \qquad (6.7)$$

Thus the conditional field Y given $X = x$ is a Markov field as is Y, with posterior energy function:

$$U(y \mid x,\theta,\beta) = U(y \mid \beta) - \sum_{i \in S} \log f(x_i \mid y_i,\theta)\,. \tag{6.8}$$

For known parameters θ, β, the maximum a posteriori (MAP) estimation of the image is equivalent to minimizing the posterior energy function

$$\hat{y} = \arg\min_{y \in Y} \{U(y \mid x,\theta,\beta)\} \tag{6.9}$$

The difference between the concept of an HMRF and that of an MRF is that the former is defined with respect to a pair of random variables (Y, X) while the latter is only defined with respect to Y.

6.2.2 Parameter estimation in HMRF-based models

In a typical application of an HMRF-based model we face the problem that both the class labels and the parameters are unknown and are strongly interdependent. A major difficulty of parameter estimation then arises due to the introduction of the dependence in the MRF. In our application we will take the parameters β of the MRF to be fixed, thus avoiding some complications. However, even with this restriction, it is not simple to estimate θ. The classical EM algorithm applied to the situation in Chapter 5 runs into problems because the conditional expectations required in the E-step cannot be simply and directly calculated.

Various approximations have been introduced in order to make the problem tractable. Techniques such as simulated annealing with Gibbs sampling (Geman and Geman 1984), and a Monte Carlo method based on the Gibbs sampler (Chalmond 1989) have been proposed to attempt to compute the true mode of $p(y)$, thereby obtaining the MAP restoration.

The iterated conditional modes (ICM) algorithm proposed by Besag (1986) does not claim to find a global maximum. It uses an iterative local minimization for which convergence is rapid, but will be to a local mode only. Given the data x and the other labels $y_{S\setminus\{i\}}^{(k)}$, the algorithm sequentially updates each $y_i^{(k)}$ into $y_i^{(k+1)}$ by minimizing $U(y_i \mid x, y_{S\setminus\{i\}})$, i.e. maximizing the conditional posterior probability, with respect to

y_i.

Other approximation techniques have been suggested. The mean field approximation (Celeux *et al*. 2003, Alfo *et al*. 2008) implements a standard EM algorithm after approximating the hidden field of class labels. A similar strategy underlies the EM algorithm given by Zhang *et al*. (2001) which estimates the missing data as $\hat{y}$ given the current θ estimate by ICM and then uses it to form the complete data $\{\hat{y}, x\}$ followed by a new θ estimation via E-M steps.

A number of approaches are suggested by Qian and Titterington (1991) for the image-analysis context. These are based on iterative restoration of the true scene, perhaps using the ICM algorithm, alternating with parameter estimation, with the use of Besag's pseudolikelihood (1975) when the parameter β of $p(y)$ is to be estimated.

There are many suggested approaches in the literature, and in some cases distinguishing between them is not easy. However in the case where β is not estimated they generally alternate between estimating the image, using approaches that vary in their level of sophistication, and estimating the parameters θ using an EM-type algorithm. Here we will adopt the simplest version of such a scheme, as described below.

6.3 Partially supervised Bayesian imaging classification with Markov random field prior

6.3.1 Markov random field model

We assume that the true configuration y is a realization of a locally dependent Markov random field (MRF). Following the suggestions of Besag (1986) we model the conditional prior probability of pixel i having class label j, given the class labels of all other pixels, in the following way:

$$\pi_{ij} = p(y_i = j \mid y_{\partial i}) \propto \alpha_{ij} \exp\{-\beta\, \gamma_{ij}(y)\} \qquad (6.10)$$

where ∂i is the set of neighbours of i, and $\gamma_{ij}(y)$ is the proportion of neighbours having class memberships different to j. In our model we always consider a second order neighbourhood, that is, the eight pixels surrounding each single pixel of the image.

In Equation (6.10), β is a fixed regularization parameter (smoothness parameter), which, when positive, discourages neighbours having different labels. In this model, the size of the parameter β will decide the local smoothness imposed by the prior distribution. Figure 6.2 shows the effect of β on the prior probability π_{ij} by plotting the relationship between the prior probability and the number of neighbours having different labels when different β are given, fixing $\alpha_{ij} = 1$.

Using the background score, ω_i, already defined in Equation (5.57) of Chapter 5, the position parameter α_{ij} is defined using the form given by Equation (6.11). The label coding is 1, 2 and 3 for normal, metastatic and background (non-nodal) components, respectively. The α_{ij} allows the probability that the pixel is background ($j = 3$) to depend on its position in the image (measured by its distance d_i from the centre of the image, using the function $f(d_i)$ in Equation (5.57) with fixed $\rho = 1.5$). The remaining probability is split equally between the two nodal groups.

$$\alpha_{ij} = \begin{cases} \omega_i, & if\ j = 3 \\ (1-\omega_i)/2, & if\ j = 1 \\ (1-\omega_i)/2, & if\ j = 2 \end{cases} \tag{6.11}$$

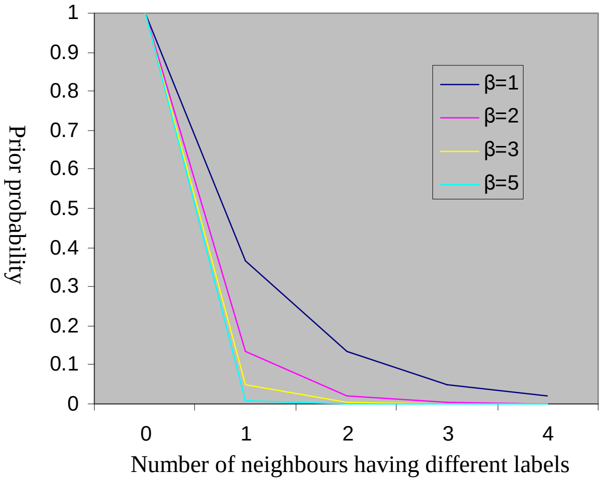
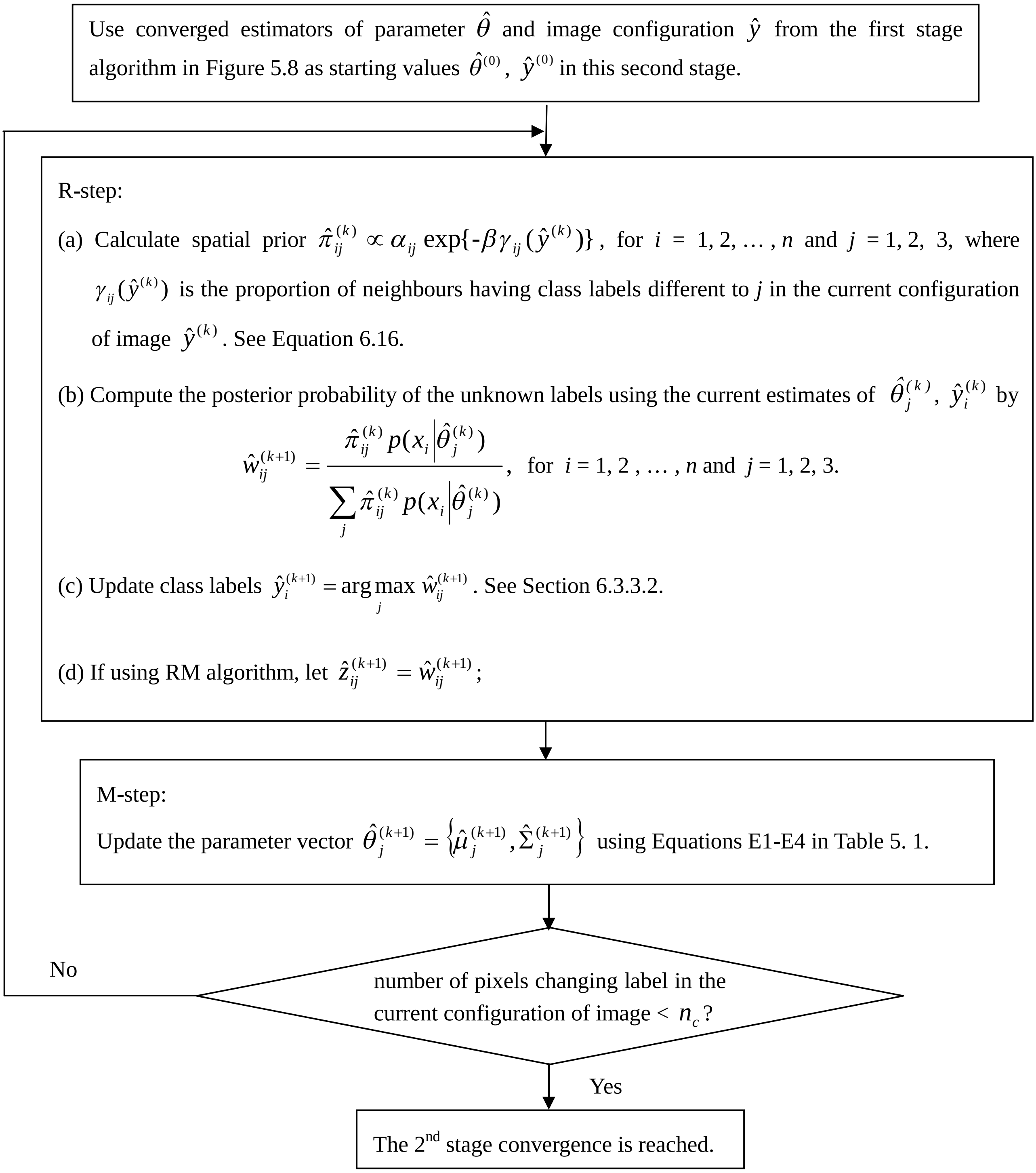


Figure 6.2: The relationship between the prior probability and the number of neighbours having different labels as β varies.

6.3.2 Model fitting algorithm and parameter estimation

We extend the fitting algorithm of Chapter 5 to incorporate the locally dependent MRF prior as shown in Figure 6.3.



Use converged estimators of parameter $\hat{\theta}$ and image configuration $\hat{y}$ from the first stage algorithm in Figure 5.8 as starting values $\hat{\theta}^{(0)}$, $\hat{y}^{(0)}$ in this second stage.

R-step:

(a) Calculate spatial prior $\hat{\pi}_{ij}^{(k)} \propto \alpha_{ij} \exp\{-\beta\gamma_{ij}(\hat{y}^{(k)})\}$, for i = 1, 2, … , n and j = 1, 2, 3, where $\gamma_{ij}(\hat{y}^{(k)})$ is the proportion of neighbours having class labels different to j in the current configuration of image $\hat{y}^{(k)}$. See Equation 6.16.

(b) Compute the posterior probability of the unknown labels using the current estimates of $\hat{\theta}_j^{(k)}$, $\hat{y}_i^{(k)}$ by

$$\hat{w}_{ij}^{(k+1)} = \frac{\hat{\pi}_{ij}^{(k)} p(x_i|\hat{\theta}_j^{(k)})}{\sum_j \hat{\pi}_{ij}^{(k)} p(x_i|\hat{\theta}_j^{(k)})}, \quad \text{for } i = 1, 2, \ldots, n \text{ and } j = 1, 2, 3.$$

(c) Update class labels $\hat{y}_i^{(k+1)} = \arg\max_j \hat{w}_{ij}^{(k+1)}$. See Section 6.3.3.2.

(d) If using RM algorithm, let $\hat{z}_{ij}^{(k+1)} = \hat{w}_{ij}^{(k+1)}$;

M-step:

Update the parameter vector $\hat{\theta}_j^{(k+1)} = \left\{\hat{\mu}_j^{(k+1)}, \hat{\Sigma}_j^{(k+1)}\right\}$ using Equations E1-E4 in Table 5. 1.

number of pixels changing label in the current configuration of image < n_c?

The 2nd stage convergence is reached.

Figure 6.3: The second stage algorithm flowchart.

6.3.3.1 Details of the algorithm

A restoration maximization (RM) algorithm and classification RM algorithm are used here for fitting the multivariate finite mixture model with MRF prior.

The joint probability of the observations *x* and pixel labels *y* is

$$f(x, y \mid \theta) = f(x \mid y, \theta)\, p(y \mid \alpha_{ij}, \beta) \tag{6.12}$$

where $p(y \mid \alpha_{ij}, \beta)$ is joint density of MRF model from a Gibbs distribution, defined in Equation (6.4) and (6.10), and $\theta = \{\theta_1, ..., \theta_g\}$ are the parameters of the component distributions in the mixture.

As described in Chapter 5, the binary membership indicator variable, $z_i = (z_{i1}, ..., z_{ig})$, was introduced as 'missing data' in the EM algorithm, indicating which component the *i*th observation belongs to. If y_i is known (that is, z_i is known), the x_i are conditionally independent, using Equation (5.24) we have

$$f(x \mid y, \phi) = \prod_{i=1}^{n} \prod_{j=1}^{g} \{ f(x_i \mid \theta_j) \}^{z_{ij}} \ . \tag{6.13}$$

Incorporating the joint prior density $P(\theta_j)$ defined in Equation (5.27), the complete-data log-posterior density function is

$$\ell(\theta \big| \{x_i\}, \{z_i\}) = \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij} \log f(x_i \mid \theta_j) + \sum_{j=1}^{g} \log P(\theta_j) + \log p(y \mid \alpha_{ij}, \beta), \tag{6.14}$$

which reduces to the expression in the form of Equation (5.49) when the y_i 's are independent of each other.

The EM algorithm applied in Chapter 5 will not work here. The M-step is straightforward, since we are only estimating θ and not α_{ij} or β , but the E-step requires the z_{ij} in Equation (6.14) to be replaced by their conditional expectation given the current parameter estimate $\hat{\theta}$ and the observed data x ,

$$\hat{z}_{ij}^{(k+1)} = E(z_{ij} \big| x, \hat{\theta}^{(k+1)}) \ . \tag{6.15}$$

This is non-trivial, because of the dependence structure introduced by the MRF.

In order to deal with the above situation, a practicable approach called the restoration-maximization (RM) algorithm by Qian and Titterington (1991) is used here. This generates a sequence of pairs $\{\theta^{(k+1)}, y^{(k+1)}\}$ of iterates for θ and y such that $y^{(k+1)}$ is updated on the basis of x and $\theta^{(k)}$, and $\theta^{(k+1)}$ is updated from x and $y^{(k+1)}$. Given the current $\{\theta^{(k)}, y^{(k)}\}$, the general procedure of the RM algorithm is as follows,

(1) ***The R-step (the E-like step)*** updates $y^{(k+1)}$ from $p(y\,|\,x,\theta^{(k)})$.

In order to update each pixel y_i, we adopt the approach suggested by Besag (1986) in his iterated conditional modes (ICM) algorithm. Given the data x and the current realization of the neighbourhood $y_{\partial i}$, the algorithm updates each pixel y_i by the class label which maximizes the conditional posterior probability:

$$\hat{w}_{ij}^{(k+1)} = \frac{\hat{\pi}_{ij}^{(k)} p_j(x_i \big| \hat{\theta}_j^{(k)})}{\sum_{j=1}^{G} \hat{\pi}_{ij}^{(k)} p_j(x_i \big| \hat{\theta}_j^{(k)})} \quad \text{for } j=1,2,\ldots,g \text{ and } i=1,2,\ldots,n. \tag{6.16}$$

where
$$\pi_{ij}^{(k)} = p(y_i = j \,|\, y_{\partial i}) \propto \alpha_{ij} \exp\{-\beta\gamma_{ij}(\hat{y}^{(k)})\}. \tag{6.17}$$

Thus
$$\hat{y}_i^{(k+1)} = \arg\max_j \hat{w}_{ij}^{(k+1)} \tag{6.18}$$

Then z_{ij} in Equation (6.14) is replaced by either $\hat{w}_{ij}^{(k+1)}$, for an EM-type algorithm or by

$$\hat{z}_{ij}^{(k+1)} = \begin{cases} 1, & \text{if } \hat{y}_i^{(k+1)} = j \\ 0, & \text{otherwise} \end{cases} \tag{6.19}$$

for a CEM-type algorithm.

(2) ***The M-step*** updates $\theta^{(k+1)}$ by maximizing Equation (6.14) with z_{ij} replaced by $\hat{z}_{ij}$ with respect to θ.

For mixtures of multivariate Gaussian or *t* distributions we have the same update equations as given in Chapter 5 (Equations E1-E4 in Table 5.1) for $\hat{\mu}_j, \hat{\Sigma}_j$.

6.3.3.2 Some issues in the implementation of the algorithm

1) Image updating by ICM

The R-step corresponds to a single cycle of Besag's ICM, where each pixel is updated by the class label which has maximum conditional probability. Here a synchronous updating is used in which all the pixels are updated simultaneously rather than cycling through the image. The main advantage of this approach is computational simplicity. Also the updating results will not be affected by the updating sequence of pixels. However convergence can no longer be guaranteed and small oscillations may occur.

One possible modification to this pixel updating would be to update them sequentially. Another would be to use a number of cycles of updating the image before updating the estimators of the parameters. Neither of these has been tried since the option chosen appears to work satisfactorily.

2) Parameter updating

Two approaches are used to update the estimators of the parameters. One is to use conditional posterior probability $\hat{w}_{ij}$ given the spectral records x and the current realization of the neighbourhood $y_{\partial i}$ to replace the z_{ij} in Equation (6.14). The other, CEM-type approach is to use a z_{ij} that corresponds to $\hat{y}_i$, via Equation (6.19). The former method keeps the algorithm converging smoothly. The latter method can be very much affected by the initial estimate $\hat{y}$. If the initial crude estimate of the image from the first stage algorithm is plausible and close to the true image convergence will be reached quickly, otherwise it may take a long time or be very difficult to work up to the true scene.

Because it assigns each pixel definitely to one group for parameter estimation, the CEM algorithm tends to produce more compact and non-overlapping distributions. In some applications this may be an advantage. Here it is not cleat that it will be.

CHAPTER 7

APPLICATION OF IMAGE CLASSIFICATION METHODS TO SCANNED SENTINEL LYMPH NODES

In Chapters 5 and 6 two customized dimension reduction methods have been described to project the spectral data from the scanned lymph nodes into a low-dimensional space and a partially supervised image classification algorithm employing a Bayesian multivariate finite mixture model with Markov random field spatial prior has been developed on the low-dimensional data to model the three unknown groups in the image. In this chapter we explore the application of this image classification model to the diagnosis of sentinel lymph node metastases in breast cancer from the ESS images. The models are studied by exploring the two dimension reduction methods, choosing model parameters and priors, assessing classification performance and analyzing model sensitivity. A hierarchical structure of the model framework is shown in Figure 7.1.

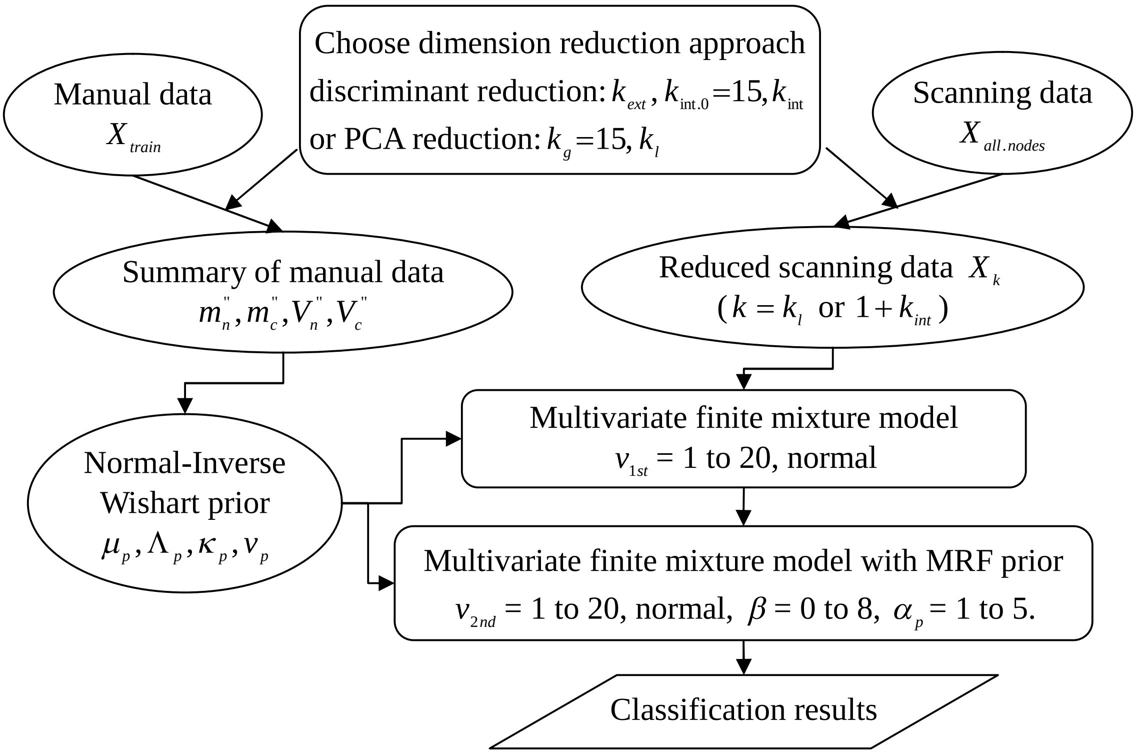

Choose dimension reduction approach
discriminant reduction: k_{ext}, $k_{int.0}$=15, k_{int}
or PCA reduction: k_g=15, k_l
Manual data
X_{train}
Scanning data
$X_{all.nodes}$
Summary of manual data
$m_n'', m_c'', V_n'', V_c''$
Reduced scanning data X_k
$(k = k_l$ or $1 + k_{int})$
Normal-Inverse
Wishart prior
$\mu_p, \Lambda_p, \kappa_p, v_p$
Multivariate finite mixture model
v_{1st} = 1 to 20, normal
Multivariate finite mixture model with MRF prior
v_{2nd} = 1 to 20, normal, β = 0 to 8, α_p = 1 to 5.
Classification results

Figure 7.1: Hierarchical structure of observations, parameters, priors, and values of constants used in our analysis of the sentinel lymph node data. In the square boxes are tuning parameters or choices, in the ellipses are data or fixed constants.

7.1 Data description and pre-processing

As described in Chapter 5, two datasets collected in different way were used in this study and each performs its own function in training the classification algorithm. One is the ESS manual measurement data, including 3,523 spectra measured from 365 totally normal nodes and 39 totally metastatic nodes from 241 patients, all with known reference pathology. For each node spectra were manually measured at up to 16 random sites on the bivalved node with one spectrum taken at one site. Reference pathology is available on a per-site (or per-spectrum) basis. The mean spectral pattern of the manual data is shown in Figure 5.5. The other is the ESS scanning data, including another 48 sentinel lymph nodes from 26 patients of which 21 are normal and 27 partially metastatic or totally metastatic, with a 20×20 pixel image of 400 spectra for each node. Reference pathology is available on a per-node basis, but not on a per-spectrum basis for individual pixels in the image. For each node, the scanning data contains not only normal or metastatic nodal group, but usually a third group which is a very variable non-nodal group from a background area, possibly contaminated by blood or lipid. No training data are available for this group.

Standard data pre-processing was carried out on spectra from both manual measurements and scanning measurements to improve signal quality. This involved spectral smoothing using the Savitzky–Golay filter (Savitzky and Golay 1964), cropping the noisy ends of the spectra before 400nm and after 800nm and normalizing using the standard normal variate (SNV) method (Barnes *et al.* 1989).

A preliminary analysis on the manual measurement data was first carried out by a PCA followed by a LDA to find the canonical variate, the direction maximizing the discrimination between normal and metastatic nodes. Leave-out-one-site cross-validation was undertaken to access the accuracy of the LDA analysis on a per-site basis. Figure 7.2 shows the distribution of canonical scores from normal (blue) and metastatic spectra (red) in the manual measurement data derived from a LDA using 20 principal components. There are two dominant peaks, the first at a score of 0, corresponding to normal nodes and some part of metastatic nodes, and the second at 4 corresponding to metastatic nodes. Scores from metastatic nodes have a multimodal distribution. It may be that despite the selection of totally metastatic nodes, there remains a mix of normal and metastatic spectra, presumably related to remaining normal structures within the metastatic nodes, or it may be that the metastatic areas are genuinely very variable.

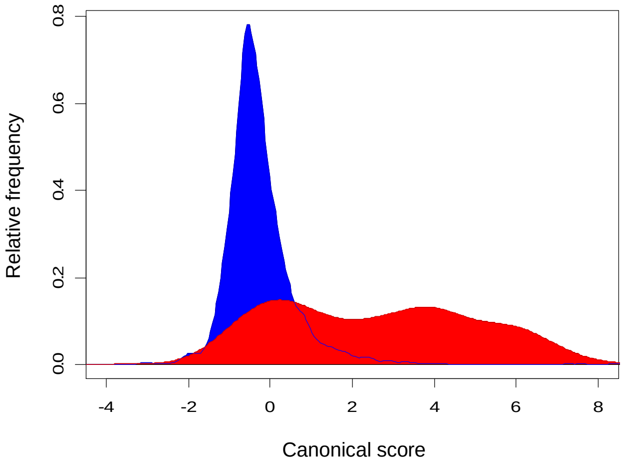


Figure 7.2: Distribution plot of LDA canonical scores of spectra from manual measurement data. The frequency is plotted as a proportion of class. Normal nodes shown in blue, metastatic nodes in red.

7.2 Dimension reduction

The two different options for dimension reduction described in Section 5.5.1 were explored on both manual and scanning data.

7.2.1 Discriminant dimension reduction

For the discriminant dimension reduction method, a variable external to the scanning data was constructed by projecting each scanned node data onto the direction of the canonical variable derived from the LDA on the first k_{ext} PC scores of the manual data (using Equations 5.10-5.11). Internal variable(s) were constructed in two steps. First do a preliminary PCA to extract the first $k_{int.0}$ PCs of all the scanned nodes in a subspace orthogonal to the direction of the external variable (using Equations 5.12-5.14). Then for the node of interest its scores on the $k_{int.0}$ PCs are calculated for further extracting the first k_{int} PCs as internal variable(s) (using Equations 5.16-5.17). The dimension-reduced scanning data of each node thus contain one external variable and one or more internal variable(s), with dimensions reduced from $n \times p$ (n = 400, p = 1801 here) to $n \times (k_{int} + 1)$.

The manual data were then projected into the space spanned by the directions of the external and internal variables for this node. Along these directions means and variances derived from normal and metastatic spectra were calculated to be used in the priors for normal and metastatic components (using Equations 5.12, 5.19-5.20). In the first dimension the prior mean is -0.21 for the normal component, and 2.44 for the metastatic component. More details about these priors can be found in Section 7.3.

External and internal variables were constructed on each scanned node for a grid of values of k_{ext}, the number of principal components constructing the external variable by using LDA , and k_{int}, the number of internal variable(s), with k_{ext} ranging from 10 to 20 and k_{int} from 1 to 5. Here we fix $k_{int.0}$ at 15 since this preliminary PCA is really just used to reduce the dimension for convenience and 15 PCs capture the variability of the data in the subspace orthogonal to the external variable.

7.2.2 PCA dimension reduction

The PCA dimension reduction method was carried out in two stages as described in Section 5.5.1.1. A global PCA first projected each scanned node data into a space spanned by the first k_g PCs derived from a preliminary PCA reduction on the pooled scanning data (using Equations 5.6-5.7); then a local PCA projection used a PCA on the k_g PC scores of the node of interest to extract the local variables, the first k_l PCs for each node (using Equation 5.9). After this, the dimension-reduced scanning data of each node contain k_l PC scores, with dimensions reduced from $n \times p$ (n = 400, p = 1801 here) to $n \times k_l$.

The manual data were first converted into the same k_g-dimensional space as above for calculating means and variances for normal and metastatic components. These means and variances were then converted into the same k_l-dimensional space (using Equations 5.10-5.11) as that from the node of interest for later use in the priors for normal and metastatic components.

A global PCA and a local PCA were carried out on each scanned node with k_g fixed at 15 and k_l varying from 2 to 5.

7.3 Bayesian multivariate finite mixture model application (stage 1)

The partially supervised image classification algorithm employing a Bayesian multivariate finite mixture model described in Chapter 5 was then applied to the low-dimensional data (reduced by either PCA dimension reduction method or discriminant dimension reduction method) to model the three unknown groups (normal, metastatic and non-nodal) in the images from the scanned measurements.

Both the multivariate Gaussian distribution and multivariate *t* distribution with degrees of freedom *v* ranging from 3 to 20 were tried for the component density of the mixture model. For both the Gaussian and *t* distributions, the normal inverse Wishart prior with parameters derived from the manual measurements was used as a prior on the parameters of the components of the mixture, with Gaussian prior for component mean vector μ conditional on scale matrix Σ and inverse Wishart prior for Σ (Equations 5.26-5.27).

The following choices were made following the suggestions from Fraley and Raftery (2007) for the prior hyperparameters for multivariate mixtures. Here the prior hyperparameters, mean μ_p, scale Λ_p and shrinkage κ_p are different for each component of the mixture with $\mu_p = (\mu_{1p}, \mu_{2p}, \mu_{3p})$, $\Lambda_p = (\Lambda_{1p}, \Lambda_{2p}, \Lambda_{3p})$ and $\kappa_p = (\kappa_{1p}, \kappa_{2p}, \kappa_{3p})$ for normal, metastatic and non-nodal components respectively.

- $\mu_{1p}, \mu_{2p}, \mu_{3p}$: For the normal and metastatic components, we take $m_n^{''}, m_c^{''}$, the mean of normal and metastatic groups from the manual data (Equations 5.10 and 5.19) as prior means μ_{1p}, μ_{2p}. For discriminant reduction, the first dimension mean is $\mu_{1p}[1] = -0.21$ for the normal component, and $\mu_{2p}[1] = 2.44$ for the metastatic component; for PCA reduction, the prior distributions for the parameters of normal and metastatic groups are projected into a different space for each individual scanned node. To aid interpretation and simplify the image correction rules described in Section 5.5.3.2 the sign of the first PC was chosen so that $\mu_{2p}[1] < \mu_{1p}[1]$. For the non-nodal component, a few spectra selected by an experienced physicist from a non-nodal area of a scanned node were used to generate a prior mean μ_{3p}.
- $\Lambda_{1p}^{-1}, \Lambda_{2p}^{-1}, \Lambda_{3p}^{-1}$: For the normal and metastatic components, we take as prior scales $\Lambda_{1p}^{-1} = V_n^{''}\ g^{-2/k}, \Lambda_{2p}^{-1} = V_c^{''}\ g^{-2/k}$, i.e., the empirical covariance matrix of normal and metastatic groups from the manual data (Equations 5.11 and 5.20) divided by the square of the number of components, g ($g = 3$), to the power 1/*k*, *k* being the reduced dimension of the spectra. For the non-nodal component, the same selected spectra from a non-nodal area as described above for prior mean were used to generate a prior scale Λ_{3p}^{-1} (more details are given in Section 7.7.2.3).
- v_p: The marginal prior distribution of μ is a multivariate *t* distribution centred at μ_p with $v_p - k + 1$ degrees of freedom. The mean of this distribution is μ_p provided that $v_p > k$, and it has a finite covariance matrix provided that $v_p > k + 1$ (Schafer 1997). Here we choose $v_p = k + 2$, the smallest integer

value for the degrees of freedom that gives a finite covariance matrix, using the same degrees of freedom for all components.

- κ_p : The shrinkage vector of the prior distribution, $(\kappa_{1p}, \kappa_{2p}, \kappa_{3p})$, gives weights on the contributions of the means of the prior distribution to the posterior means (as shown in Equation 5.45).
  - For discriminant dimension reduction, there are considerable variations in the metastatic group, with two obvious peaks for the distribution of the canonical scores in Figure 7.2 for example. Therefore more certainty is made about the prior information on the normal component than that on the metastatic one by choosing the prior weight vector $\kappa_p = (2,0.1,0.1)$ with a larger value, 2, for the normal component and a smaller value, 0.1, for both metastatic and non-nodal components.
  - For PCA dimension reduction (after a change of sign where necessary, as described above) the first dimension PC scores of normal spectra show positive values while the scores of metastatic spectra vary considerably in the negative region. Thus a stronger prior for the normal component and weaker priors for both metastatic and non-nodal components are chosen for the mixture model by giving κ_p the values of (5, 2, 1) for normal, metastatic and non-nodal components, respectively. These specific values were arrived at by experiments.

The model fitting was implemented by the iterated EM algorithm of Figure 5.8. Rather than using the discrete classification result (via CEM), here the M-step uses the conditional expectation as the value of $\hat{z}_{ij}^{(k+1)}$. This was suggested by some preliminary results running both algorithms. Although CEM may converge faster, the EM algorithm seems always to give an image closer to the real picture.

7.4 Markov random field spatial prior (stage 2)

Spatial interactions between neighbouring or nearby pixels modelled by a Markov random field were then imposed onto the Bayesian multivariate finite mixture model using the prior in Equation (6.10) with smoothing parameter β ranging from 0 to 30, and with position parameter α_{ij} as specified in Section 6.3.1. The other prior specifications were unchanged from the model without spatial interaction.

The model fitting was implemented by the RM (the EM-like) algorithm of Figure 6.3 starting from the configuration reached after the stage 1 fitting. In this stage also the conditional expectations were used in the M-step for the values of $\hat{z}_{ij}^{(k+1)}$ instead of the classification values (via CEM). Apart from the fact that EM generates a better image, this is also because the CEM result can depend strongly on the initial estimate from the stage 1.

7.5 Image classification performance assessment

Since reference pathology is not available for individual pixel of the images, but only available for each node, the classification was carried out on a per node basis. To define conditions for labelling each node as metastatic or non-metastatic, we simply counted the number of positive (metastatic) pixels in the node. The likelihood of scattered false positive pixels occurring over a node is very slim since the spatial correlation between adjacent pixels of image has been taken into consideration in the model fitting, hence we classify a node as metastatic if it has even one pixel thus classified.

Leave-out-one-node cross-validation was used to assess performance, the classification accuracy being measured by sensitivity, specificity and AUC, the area under the ROC curve. The global dimension reduction was carried out inside the loop, that is, on the pooled scanning data with the omitted node not included in the procedure. An image was generated by plotting a 20×20 matrix of probabilities with the following colour codings. Black indicates the pixels classified as non-nodal component; for normal or metastatic component we used the posterior probability of the pixel belonging to the metastatic component to generate a colour between red (represents positive) and blue (represents negative) for each pixel. This image was compared by eye with the photograph of the node to assess the method's success in reconstructing its shape.

7.6 Classification results

7.6.1 Using discriminant dimension reduction

For discriminant dimension reduction, in order to search for optimal combinations of $k_{ext}, k_{\text{int}}, v_{1st}, v_{2nd}$ and β (with ρ fixed at 1.5 as mentioned in Section 5.5.3.2), instead of an exhaustive searching of all the possible combinations of these parameters, a more restricted approach was adopted. First we varied combinations of k_{ext} and k_{int} for a restricted number of values of β, v_{1st} and v_{2nd}, and then we fixed k_{ext} = 20 and k_{int} = 1, the values which gave the best results in this limited search. The leave-out-one-node cross-validation results for the various combinations of v_{1st} and v_{2nd} (the number of degrees of freedom of the multivariate *t* distribution or multivariate Gaussian distribution in the first and second stage algorithms) and β with these fixed values of k_{ext} and k_{int} are shown in Table 7.1. The whole computation is very time-consuming, even just for generating the results in Table 7.1 with one combination of k_{ext} and k_{int} it took several weeks and several computers to run the algorithm with many times. During the experiments, we found when v_{2nd} is smaller than v_{1st}, the algorithm does not always converge, so results are only given for the combinations with $v_{1st} \leq v_{2nd}$.

For discriminant dimension reduction, the combination k_{ext} = 20, k_{int} = 1, $v_{1st} = 4$, $v_{2nd} = 4$ with a wide range of β starting from 8 (or other choices with v_{2nd} = 10, 20 and a correspondingly higher β) gave the best results, with sensitivity, specificity and AUC of 81%, 90.4% and 0.86, respectively. The scanned nodal spectra data are here reduced to a space with two dimensions, one external variable, and one internal variable. In the rest of the chapter we will refer to the combination of k_{ext} = 20, k_{int} = 1, $v_{1st} = 4$, $v_{2nd} = 4$ and $\beta = 8$ as the optimal model for the discriminant reduction method, with a ROC curve shown in the left panel of Figure 7.5. Some examples of the mapping images of metastatic nodes and normal nodes generated by this optimal model are shown in Figure 7.3.

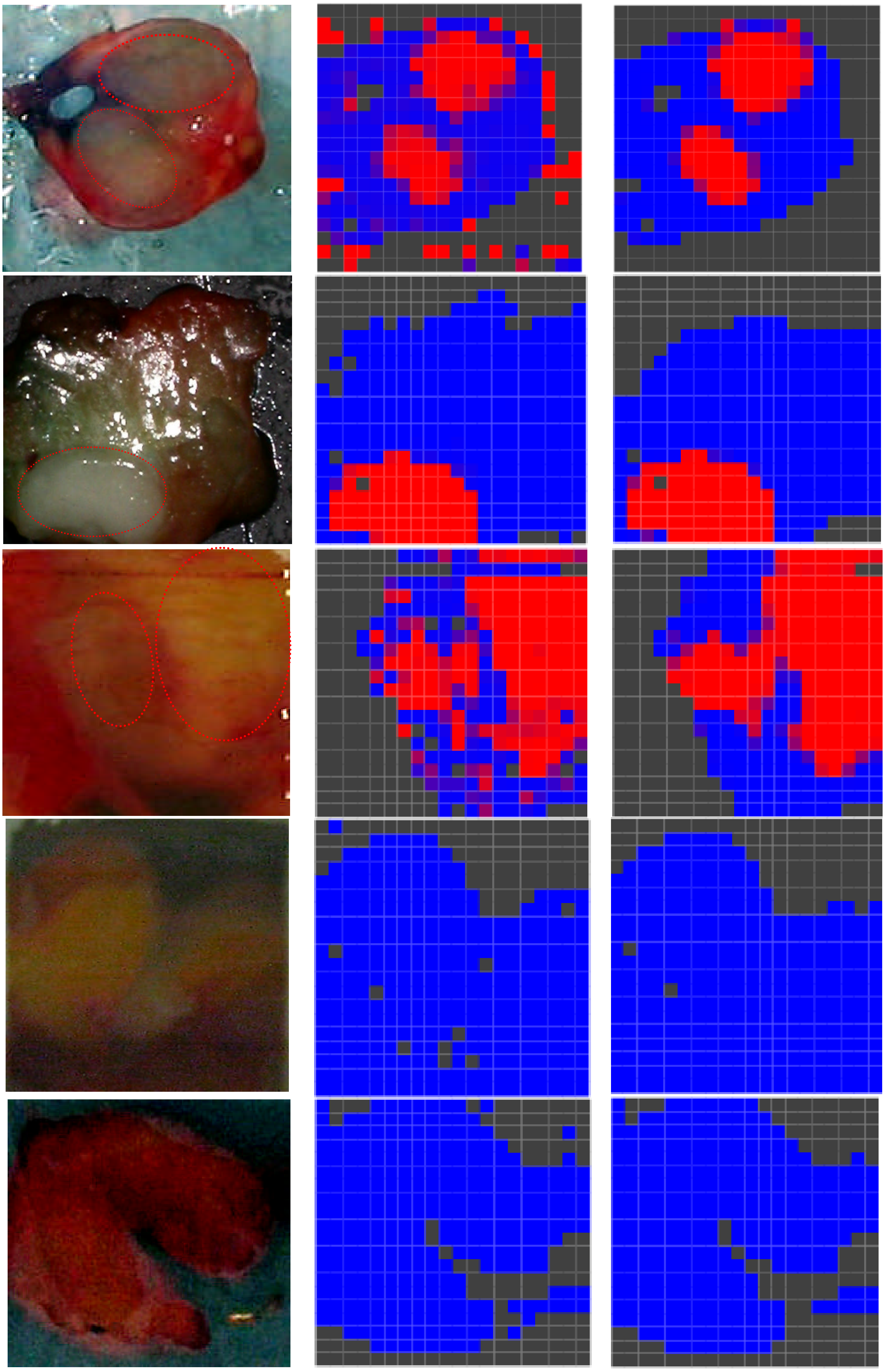


Figure 7.3: Examples of the mapping images from three partially or totally metastatic nodes (first three rows) and two totally normal nodes (last two rows) using the partially supervised image classification model via discriminant reduction. The left column shows the photographs

of the nodes (red dotted line circles the supposed metastatic area of the partially metastatic node). The middle and right columns are the 1st stage and 2nd stage resulting maps of the scanned nodes, with posterior probability scores plotted as colours. Red corresponds to spectra indicative of metastases, blue corresponds to spectra indicative of normal lymph tissue, and black corresponds to non-nodal spectra.

Table 7.1: Accuracy results for discriminant reduction method using various combinations of v_{1st} and v_{2nd} (degrees of freedom of t or Gaussian distribution in the 1st and 2nd stage algorithms respectively) and β (the smoother parameter) with $k_{ext} = 20$ and $k_{int} = 1$.

v_{1st}	v_{2nd}	β	Accuracy		
			Sensitivity (%)	**Specificity (%)**	**AUC**
4	4	4	78	90	0.84
4	4	5	81	90	0.85
4	4	6	81	90	0.85
4	***4***	***8***	81	90	0.86
4	***4***	***10***	81	90	0.86
4	***4***	***15***	81	90	0.86
4	***4***	***20***	81	90	0.86
4	***4***	***30***	81	90	0.86
4	10	4	78	90	0.84
4	10	5	78	90	0.84
4	10	6	78	90	0.84
4	10	8	78	90	0.84
4	10	10	78	90	0.84
4	***10***	***15***	81	90	0.86
4	***10***	***20***	81	90	0.86
4	***10***	***30***	81	90	0.86
4	20	4	78	76	0.80
4	20	5	78	76	0.81
4	20	6	78	81	0.83
4	20	8	78	86	0.83
4	20	10	78	86	0.83
4	20	15	78	90	0.84
4	20	20	78	90	0.84
4	***20***	***30***	81	90	0.86
4	norm	4	78	76	0.82
4	norm	5	78	76	0.82
4	norm	6	78	76	0.82
4	norm	8	78	76	0.82
4	norm	10	78	76	0.82
4	norm	15	78	90	0.84

4	norm	20	78	90	0.84
4	norm	30	78	90	0.84
10	10	4	78	71	0.77
10	10	5	78	71	0.77
10	10	6	78	71	0.77
10	10	8	78	71	0.78
10	10	10	78	71	0.79
10	10	15	81	71	0.80
10	10	20	81	71	0.81
10	10	30	81	71	0.81
10	20	4	78	71	0.78
10	20	5	78	71	0.77
10	20	6	78	71	0.77
10	20	8	78	71	0.78
10	20	10	78	71	0.79
10	20	15	78	71	0.79
10	20	20	78	71	0.79
10	20	30	81	71	0.81
10	norm	4	78	62	0.75
10	norm	5	78	62	0.77
10	norm	6	78	62	0.77
10	norm	8	78	62	0.77
10	norm	10	78	62	0.78
10	norm	15	78	71	0.78
10	norm	20	78	71	0.79
10	norm	30	78	71	0.79
20	20	4	78	67	0.78
20	20	5	78	67	0.78
20	20	6	78	67	0.79
20	20	8	78	67	0.80
20	20	10	78	67	0.80
20	20	15	78	67	0.79
20	20	20	78	71	0.80
20	20	30	81	71	0.81
20	norm	4	78	62	0.77
20	norm	5	78	62	0.78
20	norm	6	78	62	0.78
20	norm	8	78	62	0.79
20	norm	10	78	62	0.79
20	norm	15	78	67	0.79
20	norm	20	78	67	0.79
20	norm	30	78	67	0.79

norm	norm	4	70	43	0.76
norm	norm	5	70	43	0.76
norm	norm	6	70	43	0.76
norm	norm	8	70	43	0.76
norm	norm	10	70	43	0.77
norm	norm	15	70	43	0.77
norm	norm	20	70	43	0.77
norm	norm	30	70	43	0.76

7.6.2 Using PCA dimension reduction

For the PCA dimension reduction method, the same procedure as used in discriminant reduction was used to search the optimal combinations of $k_g, k_l, v_{1st}, v_{2nd}$ and β, that is, k_g and k_l were first decided by experiment using restricted choices of v_{1st}, v_{2nd} and β, and then fixed to explore the other parameters in more detail. Among all the choices, the model worked best in a two-dimensional space with $k_g = 15$ and $k_l = 2$.

Table 7.2 Shows the leave-out-one-node cross-validation results for various combinations of v_{1st}, v_{2nd} and β for these choices of dimension. As in discriminant reduction, the results here are only given for combinations with $v_{1st} \le v_{2nd}$ since otherwise the model may fail to converge. For PCA dimension reduction, the combination k_g = 15, k_l = 2, v_{1st} =10, $v_{2nd} = 10$ and $\beta = 20$ gave the best results, with sensitivity, specificity and AUC of 88.9%, 76% and 0.81, respectively, and a ROC curve shown in the right panel of Figure 7.5. Some examples of the mapping images from metastatic nodes and normal nodes generated by this optimal image classification model via PCA reduction method are shown in Figure 7.4. The same nodes as shown in Figure 7.3 are used here for convenience of comparison.

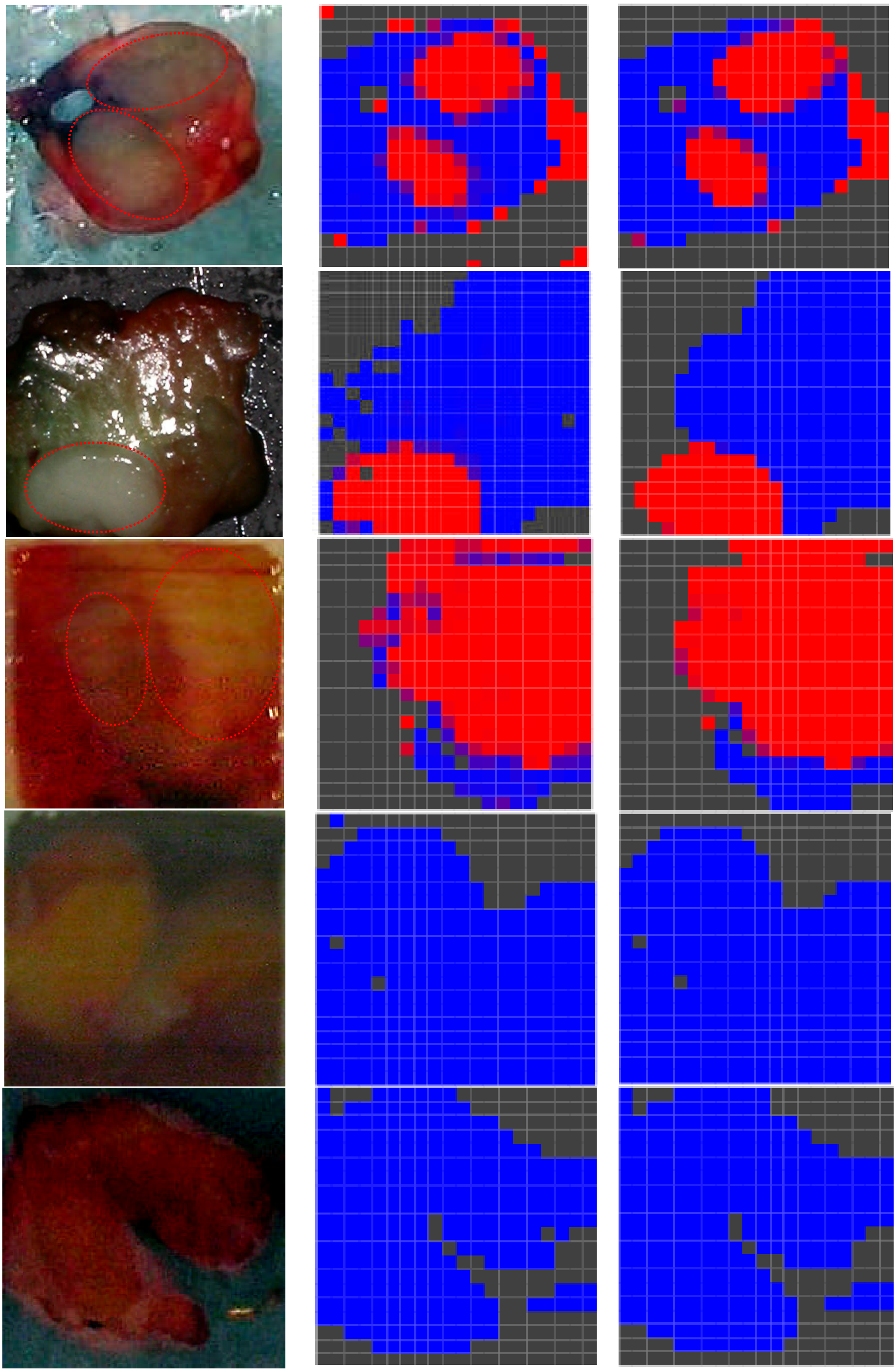


Figure 7.4: Examples of mapping images from three partially or totally metastatic nodes (first three rows) and two totally normal nodes (last two rows) using the partially supervised image classification model via PCA reduction. The left column shows the photographs of the node (red

dotted line circles the supposed metastatic area of the partially metastatic node). The middle and right columns are the resulting maps of spectra using the 1st and 2nd stage algorithms of the model, with posterior probability scores plotted as colours. Red corresponds to spectra indicative of metastases, blue corresponds to spectra indicative of normal lymph tissue, and black corresponds to non-nodal spectra.

Table 7.2: Accuracy results for PCA reduction method using various combinations of v_{1st} and v_{2nd} (degrees of freedom of t or Gaussian distribution in the 1st and 2nd stage algorithms respectively) and β (the smoother parameter) with $k_g = 15$ and $k_l = 2$.

v_{1st}	v_{2nd}	β	**Accuracy**		
			Sensitivity (%)	**Specificity (%)**	**AUC**
3	3	4	85	62	0.76
3	3	5	85	62	0.77
3	3	6	85	62	0.78
3	3	8	85	62	0.79
3	3	10	85	62	0.79
3	3	15	85	62	0.79
3	3	20	85	62	0.79
3	3	30	81	62	0.77
4	4	4	78	67	0.68
4	4	5	78	67	0.68
4	4	6	78	67	0.68
4	4	8	78	67	0.68
4	4	10	78	67	0.69
4	4	15	78	67	0.68
4	4	20	78	67	0.68
4	4	30	74	67	0.66
4	10	4	78	62	0.66
4	10	5	78	67	0.68
4	10	6	78	67	0.68
4	10	8	78	67	0.67
4	10	10	78	67	0.68
4	10	15	78	67	0.69
4	10	20	78	67	0.68
4	10	30	78	67	0.67
4	20	4	78	48	0.63
4	20	5	78	48	0.63
4	20	6	78	48	0.63
4	20	8	78	48	0.65
4	20	10	78	52	0.66
4	20	15	78	52	0.67

4	20	20	78	67	0.69
4	20	30	78	67	0.68
4	norm	4	78	33	0.64
4	norm	5	78	33	0.65
4	norm	6	78	33	0.65
4	norm	8	78	33	0.65
4	norm	10	78	33	0.67
4	norm	15	78	38	0.67
4	norm	20	78	38	0.67
4	norm	30	78	38	0.67
10	***10***	***4***	89	71	0.80
10	***10***	***5***	89	71	0.79
10	***10***	***6***	89	71	0.79
10	***10***	***8***	89	71	0.80
10	***10***	***10***	89	71	0.80
10	***10***	***15***	89	71	0.81
10	***10***	***20***	89	76	0.81
10	***10***	***30***	85	71	0.80
10	20	4	89	62	0.79
10	20	5	89	62	0.78
10	20	6	89	67	0.79
10	20	8	89	67	0.79
10	20	10	89	62	0.79
10	***20***	***15***	89	71	0.81
10	***20***	***20***	89	71	0.81
10	***20***	***30***	89	71	0.81
10	norm	4	89	38	0.76
10	norm	5	89	38	0.79
10	norm	6	89	43	0.78
10	norm	8	89	43	0.80
10	norm	10	89	43	0.80
10	norm	15	89	43	0.81
10	norm	20	89	43	0.81
10	norm	30	89	48	0.81
20	20	4	85	67	0.78
20	20	5	85	67	0.78
20	20	6	85	67	0.78
20	20	8	85	67	0.78
20	20	10	85	67	0.78
20	20	15	85	67	0.79
20	20	20	85	67	0.80
20	20	30	85	71	0.80

20	norm	4	85	43	0.77
20	norm	5	85	43	0.78
20	norm	6	85	48	0.78
20	norm	8	85	48	0.78
20	norm	10	85	48	0.78
20	norm	15	85	48	0.80
20	norm	20	85	48	0.80
20	norm	30	85	57	0.80

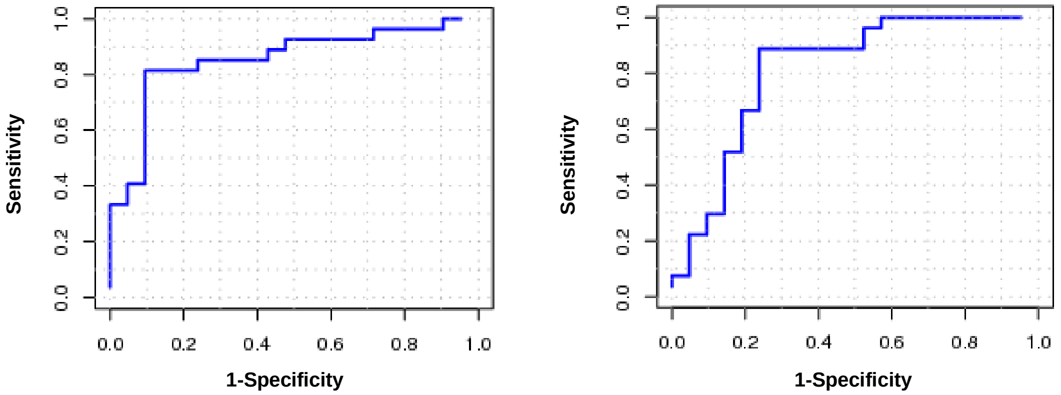


Figure 7.5: ROC curves from leave-out-one-node cross-validation using the optimal image classification model after discriminant dimension reduction (left, k_{ext} = 20, k_{int} = 1, $v_{1st} = 4$, $v_{2nd} = 4$ and $\beta = 8$) and PCA dimension reduction (right, k_g = 15, k_l = 2, $v_{1st} = 10$, $v_{2nd} = 10$ and $\beta = 20$).

7.6.3 Discussion of the results

The method gives acceptable accuracy using both of the dimension reduction approaches. Some conclusions that apply in either case are:

- Although the spaces themselves are different, two dimensions are sufficient to separate the three groups using either of the dimension reduction approaches.
- In both stages of the model, the multivariate *t* distribution gives better results, with higher AUC and specificity than the multivariate Gaussian distribution (This is discussed further in Section 7.7.4).
- Higher degrees of freedom for the *t* distribution in the first stage of the model, v_{1st}, require at least the same level of degrees of freedom in the second stage of the model, v_{2nd}. The algorithm may run into problems when $v_{2nd} < v_{1st}$, not always converging.

- Higher degrees of freedom for the t distribution in the second stage of the model need to be combined with a stronger spatial prior with higher smoother parameter β to work well.
- Large v_{1st} with subsequently large v_{2nd} gives high sensitivity; small v_{1st} with small v_{2nd} gives high specificity.
- The image generated by the first stage of the model contains scattered spots that are probably not genuine. Introducing spatial priors in the second stage of the model normally does not improve the classification accuracy in terms of sensitivity, specificity and AUC on per-node basis, but does improve the image quality in terms of image smoothness and probable closeness to the real picture.

Some specific conclusions and comparisons for the two dimension reduction options are drawn as follows:

- For the PCA dimension reduction method, classification models with good accuracy result from using moderate or high degrees of freedom for the multivariate t distributions and a spatial prior with a high value of the smoother parameter β . The combinations of $v_{1st} = 10$ with $v_{2nd} = 10$ and $\beta \geq 4$ (or with $v_{2nd} = 20$ and $\beta \geq 15$) give the better accuracy than other choices with sensitivity, specificity and AUC of 89%, 76% (or 71%) and 0.81.
- For the discriminant dimension reduction method, classification models with good accuracy result from lower degrees of freedom of the multivariate t distribution for a wide range of values of β . The combinations of v_{1st}= 4 with v_{2nd}= 4, 10, 20 all give good accuracy with sensitivity, specificity and AUC of 81%, 90% and 0.86, respectively. Some interactions between v_{2nd} and β can be seen from Table 7.1. For the model with v_{1st} = 4, as v_{2nd} increases from 4 to 20, the minimum value for a strong enough smoother β to give good accuracy increases from 8 to 30.
- In general the discriminant dimension reduction method works better than the PCA dimension reduction method.
    - Using discriminant dimension reduction, the two extracted dimensions are feature-oriented and thus more interpretable. The external variable

works in a direction capturing most of the variation between the normal and metastatic groups, and the internal variable works in a direction capturing most of variation between the nodal and non-non-nodal groups. Using PCA dimension reduction, the separation of the three groups by the first two dimensions of the local PCA projection varies with different nodes, the first dimension can be either the direction capturing most of the variation between the nodal and non-nodal groups or the direction capturing most of the variation between the normal and metastatic groups. More details can be found in Section 7.7.1. One consequence is the need to use higher degrees of freedom in the *t*-distribution to impose the prior grouping more tightly.

- Models using PCA dimension reduction and higher degrees of freedom tend always to find three components in the mixture although in some cases there are only one or two components. It may give a better representation of a node with all three of normal, metastatic and background components, and thus gives higher sensitivity. However, there is also the possibility to misclassify some of the normal pixels as metastic when in fact there are only normal and background groups, thus it gives lower specificity.
- An option that works well is to use lower degrees of freedom in the first stage of the model with the discriminant dimension reduction method. This results in rapid convergence of the first stage to a rough solution. With an appropriate choice of degrees of freedom and smoother parameter in the second stage of the model this stage then converges to a good image.
- The image results generated by the model following the discriminant reduction method are closer to the real picture than that generated by the model following PCA reduction method.

7.7 Parameter tuning, model sensitivity, and comparison of model choices

7.7.1 Comparison of the dimension reduction methods

The classification models described in Section 7.6 derived from two-dimensional data extracted by either discriminant or PCA dimension reduction method, though built in the same way, work in a completely different space. Two examples are given here to show how the classification works in the two different two-dimensional spaces constructed from the same scanned node data.

Figure 7.6(a) demonstrates how the three groups of a partially metastatic node are classified in a two-dimensional space constructed by discriminant dimension reduction method. The classification of the pixels in these figures is that resulting from the application of the model: there is no reference pathology at a pixel level. Along the first dimension, the external variable (in the left panel), the metastatic component can be discriminated from the normal and non-nodal components; along the second dimension, the only internal variable here (in the right panel), the non-nodal component can be discriminated from the normal and metastatic components. The peak of the distribution of the external variable scores from the metastatic component is close to the higher mode of the distribution of canonical scores from metastatic spectra in Figure 7.2.

For the same partially metastatic node Figure 7.6(b) illustrates how three groups are classified in the different two-dimensional space constructed by the PCA dimension reduction method. Along the first dimension, the metastatic component can be discriminated from normal and non-nodal components; along the second dimension only some of the spectra from the non-nodal component can be discriminated from the nodal component, and some misclassification of the nodal component as non-nodal component occurs in this case. The mapping image derived from the data reduced by discriminant dimension reduction method is much closer to the real picture of the node than that derived from the data reduced by the PCA dimension reduction method.

Another example given in Figure 7.7 (a) shows three groups from a metastatic node being fairly well separated in the two-dimensional space constructed by the discriminant dimension reduction method. In Figure 7.7 (b) the spectra from the same node are projected to a two-dimensional space via PCA dimension reduction method. Along the direction of the first dimension in Figure 7.7 (b), the three groups are already well separated.

Using either reduction method, the best performing models all use a two-dimensional space. Using two dimensions in directions orthogonal to each other seems to be a reasonable and sufficient choice to do discrimination between three groups. Following the discriminant dimension reduction method, the first and second dimensions (i.e. the external and internal variables) typically work as two classifiers, the first discriminates between the two nodal (metastatic and normal) components and the second between non-nodal and nodal components, Following the PCA dimension reduction method, the first and second dimensions (i.e. the two local variables) work in a different two-dimensional space for each node, and how three components in the mixture are separated in this two-dimensional space really depends on the features of each individual node, especially the behaviour of the variable non-nodal component. The first dimension itself may or may not be enough to discriminate one of three components from the others, or may even be good enough for the discrimination between three components.

In these two examples, using discriminant reduction the peak of the distribution for the metastatic component from the scanning data is either around a score of 4.5 (Figure 7.6(a)) or around 0 (Figure 7.7(a)). This is consistent with the appearance of two dominant peaks in the LDA analysis of the manually measured metastatic nodes (as shown in Figure 7.2).

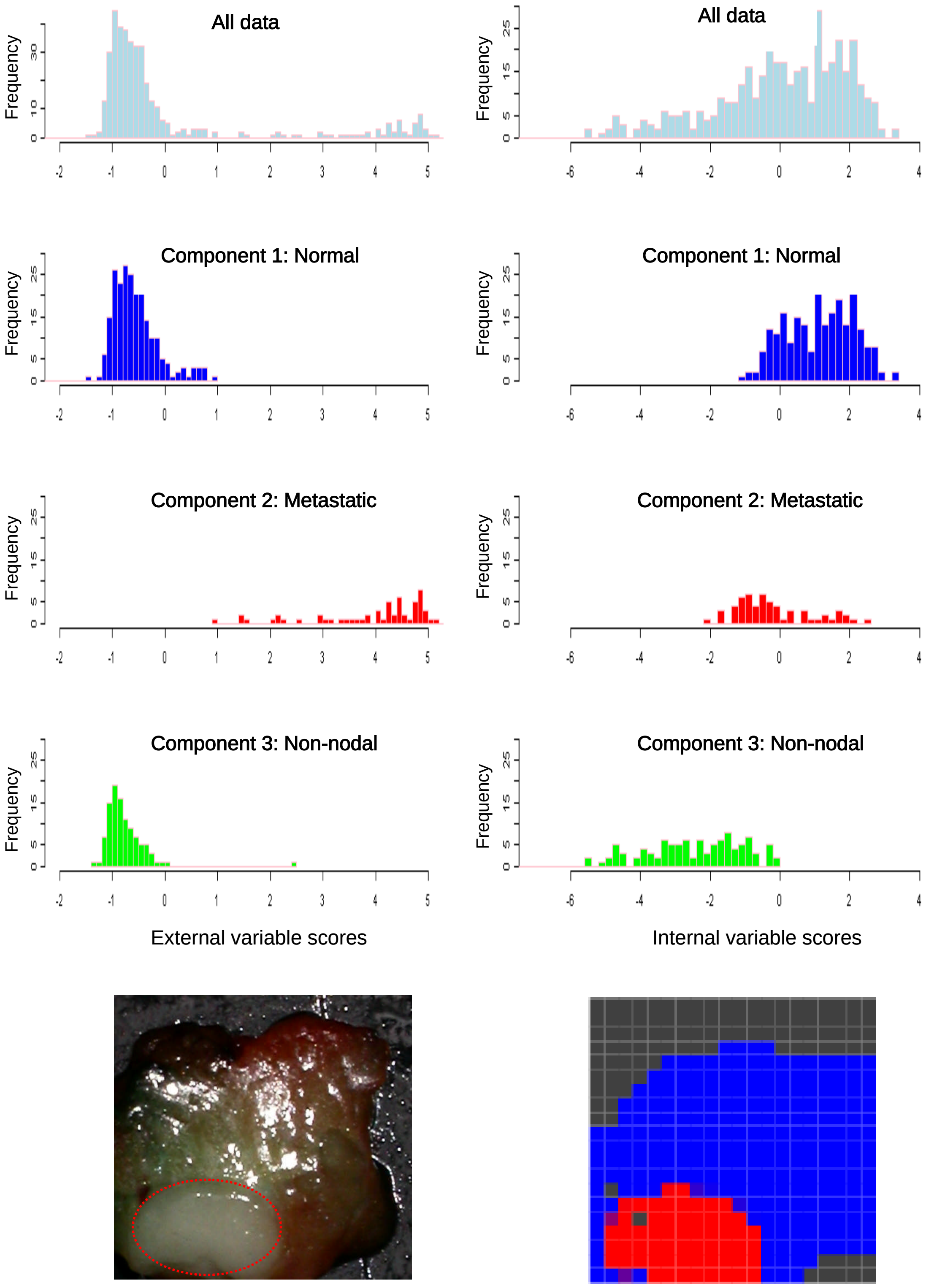


Figure 7.6 (a): Plots of external variable scores (left panel) and internal variable scores (right panel) of spectra from a partially metastatic node (sc8471) after discriminant dimension reduction. In rows 2, 3, 4 histogram of normal spectra (i.e. spectra classified as normal) is shown in blue, metastatic in red and non-nodal in green. Last row shows the photograph (left) and the mapping image (right, red indicative of metastatic spectra, blue of normal and black of non-nodal) of this node.

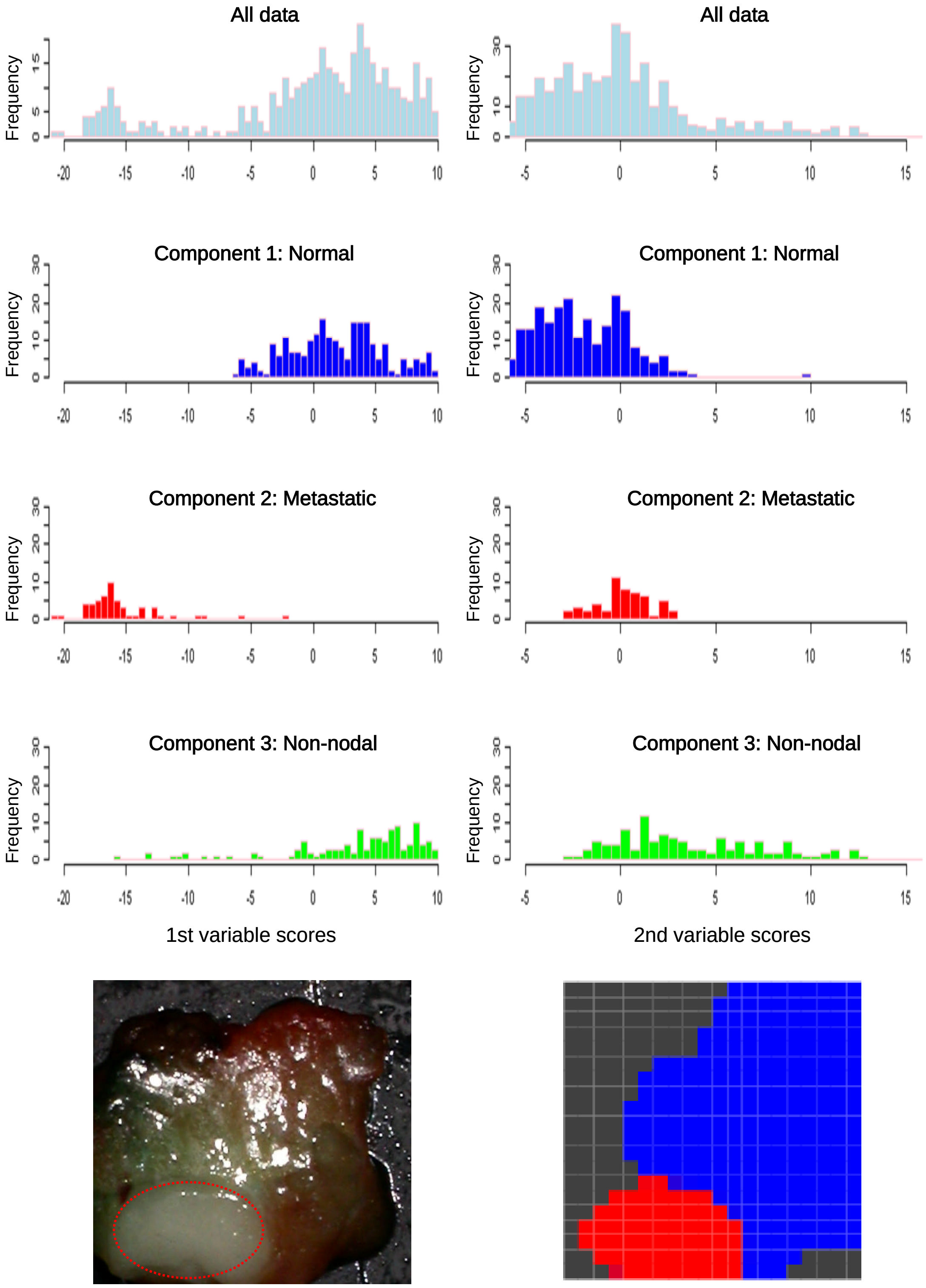


Figure 7.6 (b): Plots of the 1st dimension PC scores (left panel) and the 2nd dimension PC scores (right panel) of spectra from a partially metastatic node (sc8471) after PCA dimension reduction. In rows 2, 3, 4 histogram of normal spectra (i.e. spectra classified as normal) is shown in blue, metastatic in red and non-nodal in green. Last row shows the photograph (left) and the mapping image (right, red indicative of metastatic spectra, blue of normal and black of non-nodal) of this node.

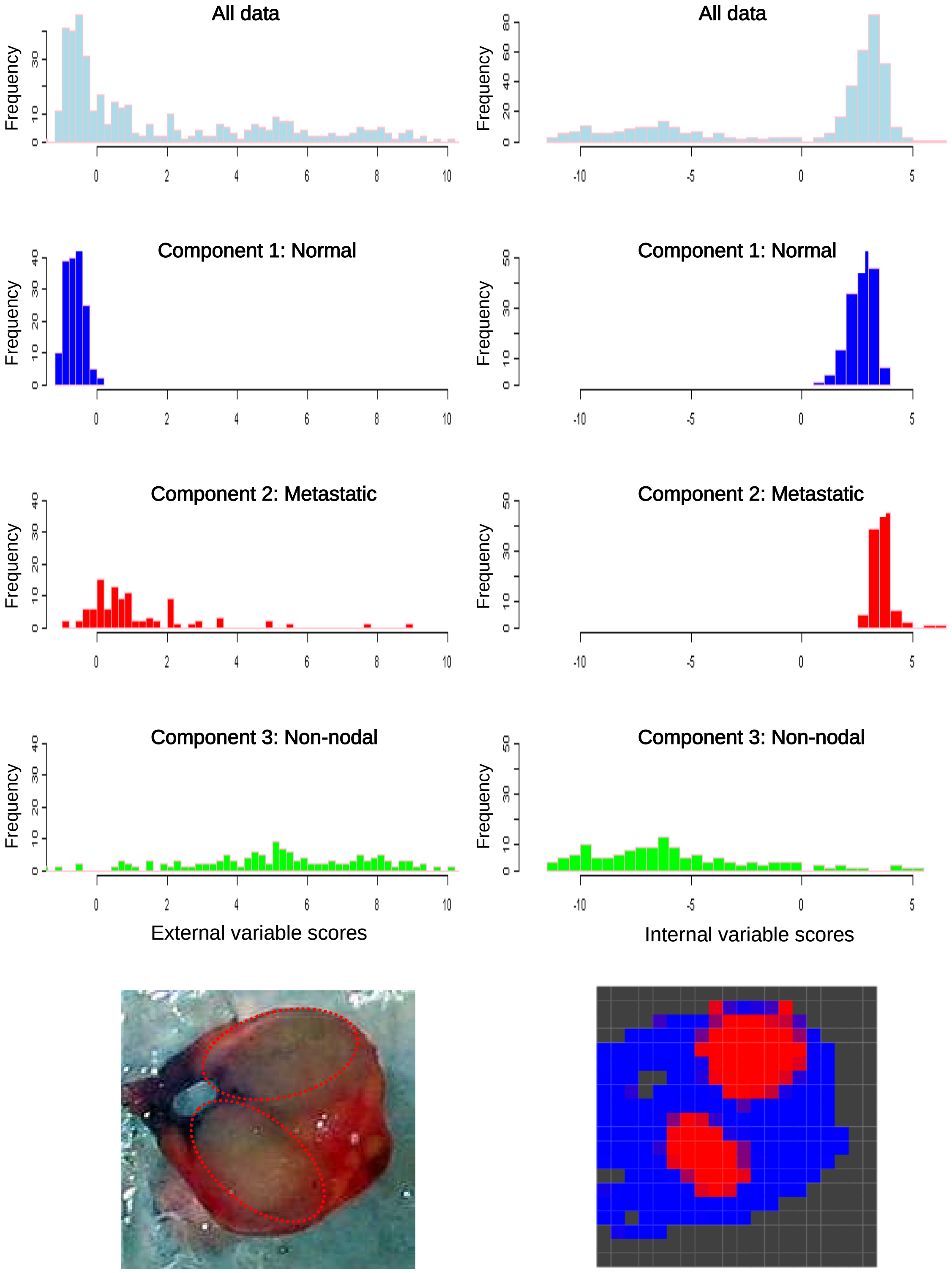


Figure 7.7 (a): Plots of external variable scores (left panel) and internal variable scores (right panel) of spectra from a metastatic node (sc8841) after discriminant dimension reduction. In rows 2, 3, 4 histogram of normal spectra (i.e. spectra classified as normal) is shown in blue, metastatic in red and non-nodal in green. Last row shows the photograph (left) and the mapping image (right, red indicative of metastatic spectra, blue of normal and black of non-nodal) of this node.

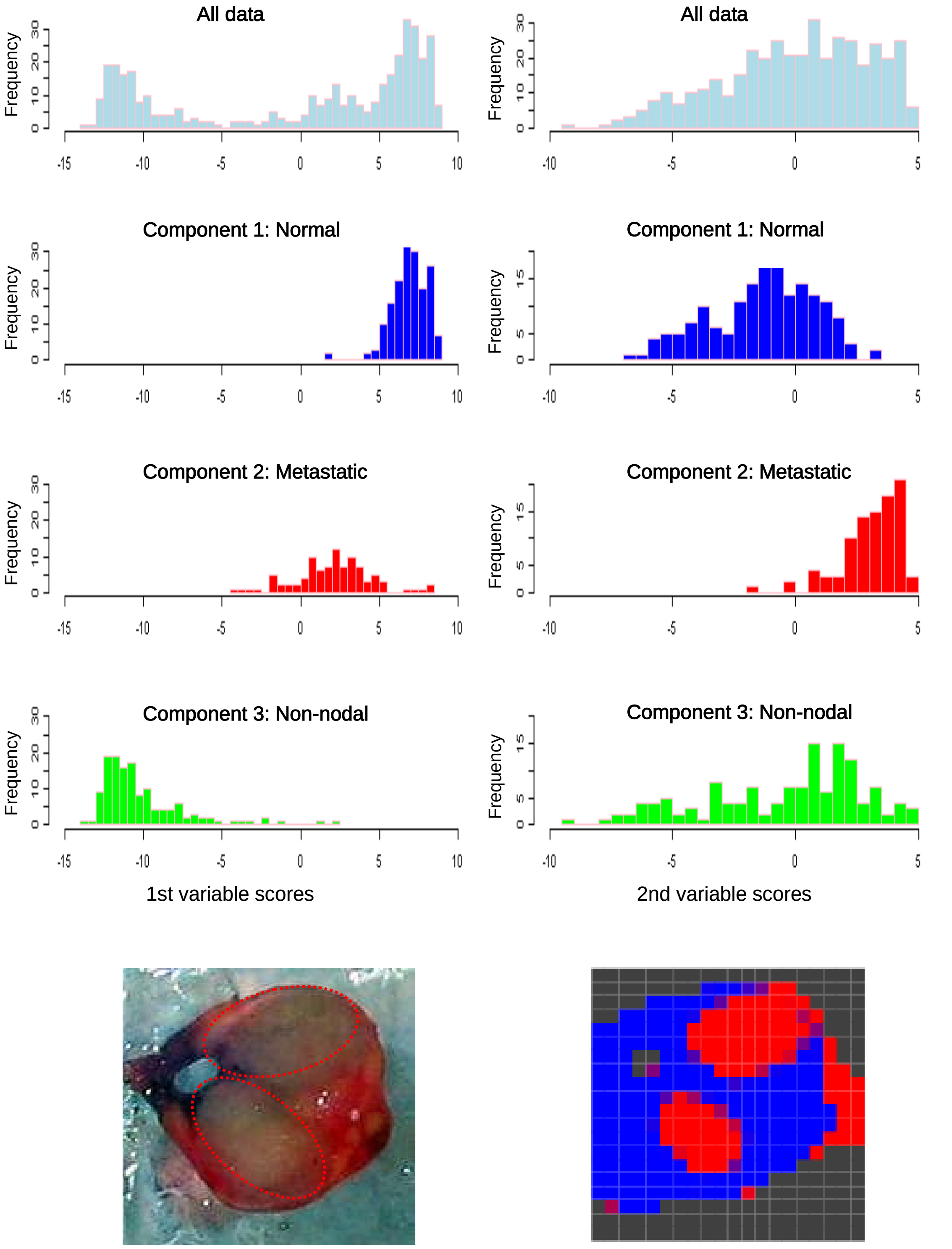


Figure 7.7 (b): Plots of the 1st dimension PC scores (left panel) and the 2nd dimension PC scores (right panel) of spectra from a metastatic node (sc8841) after PCA dimension reduction. In rows 2, 3, 4 histogram of normal spectra (i.e. spectra classified as normal) is shown in blue, metastatic in red and non-nodal in green. Last row shows the photograph (left) and the mapping image (right, red indicative of metastatic spectra, blue of normal and black of non-nodal) of this node.

7.7.2 Choice of prior and sensitivity of results to prior assumptions in mixture model

Choice of prior distributions can be a contentious issue in any situation where Bayesian methods are applied, and the mixture model and our application provide particular problems in this aspect. As described in Section 7.3, in our application, informative priors based on the manual data are given for the normal and metastatic components, although the prior for the metastatic component makes no attempt to model the apparent multimodal distribution seen in Figure 7.2. For the non-nodal component a diffuse prior distribution was used with a prior mean derived from some non-nodal spectra selected from the scanned nodes by an experienced physicist and a large prior variance. A shrinkage parameter (prior weight) is also included for each component in the mixture with a relatively strong value for the normal component and weak values for both metastatic and non-nodal components.

However, we found that even when the priors are relatively diffuse, inference can still be influenced by the priors. The sensitivity of the Bayesian inference to the priors is investigated here by comparing the effect of different choices.

7.7.2.1 Sensitivity to the shrinkage parameter, κ_p , of the prior distribution

As described in Equations (5.32)-(5.33), the conditional posterior mean $\tilde{\mu}_{jp} = \left(\frac{n}{\kappa_{jp}+n}\right)\bar{x} + \left(\frac{\kappa_{jp}}{\kappa_{jp}+n}\right)\mu_{jp}$ is a weighted average of the prior mean μ_{jp} and the sample mean $\bar{x}$ of one scanned node (*n* is the number of pixels from one scanned node, *n* = 400), which can be viewed as adding κ_{jp} observations with value μ_{jp} to group *j* in the data. The shrinkage parameter of the prior distribution, κ_{jp}, also called the prior weight here, thus gives the weight on the contributions of the prior mean μ_{jp} to the posterior mean $\tilde{\mu}_{jp}$. It is standard to refer to κ_{jp} as the shrinkage parameter, as it governs the extent to which the sample mean $\bar{x}$ is shrunk toward its prior mean, μ_{jp} .

The choices of κ_{jp} will crucially influence the posterior distribution of means and variances as in Equations (5.32)-(5.33). Strong weight (larger κ_{jp}) on the priors will aid convergence and help to avoid the label switching problem. However, too strong a weight will not allow enough flexibility for components that vary from node to node.

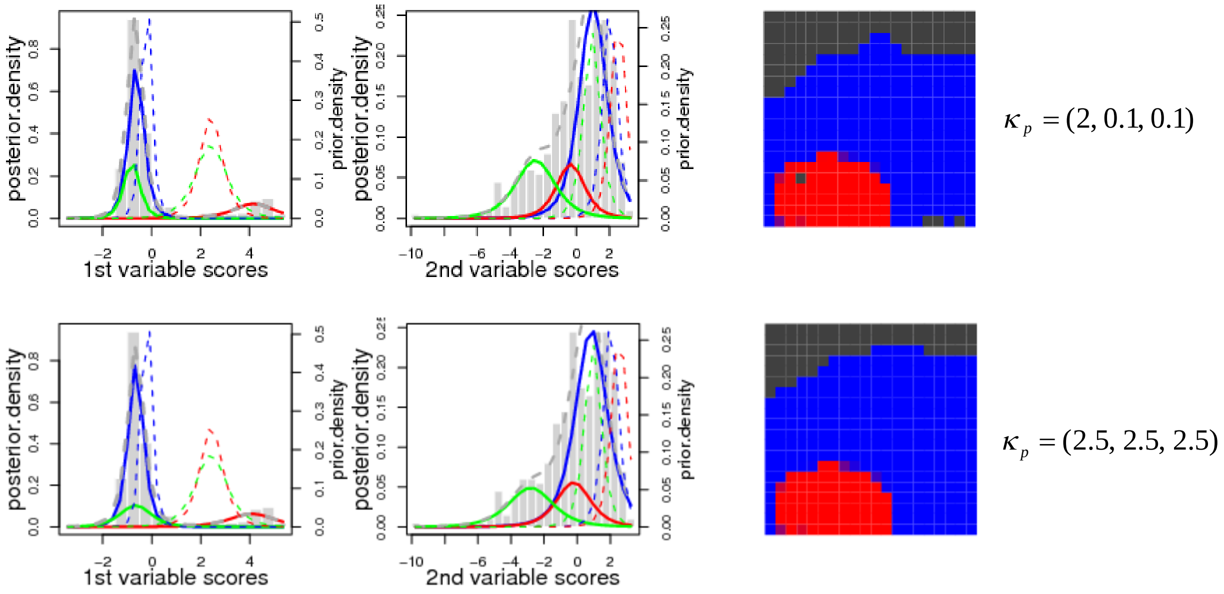
To reflect the variability of the metastatic and non-nodal components, we chose a larger prior weight for normal component and a smaller one for metastatic and non-nodal components. As described in Section 7.3 for the discriminant dimension reduction method we give $\kappa_p = (\kappa_{1p}, \kappa_{2p}, \kappa_{3p})$ the values of (2, 0.1, 0.1) for normal, metastatic and non-nodal components, and for the PCA dimension reduction method we chose the values of (5, 2, 1).

One example is given in Figure 7.8 to demonstrate how different choices of prior weight affect the results of the imaging classification. Since the analysis following the dimension reduction via the discriminant method is the same as that PCA reduction method, but works in different spaces, here the example is given using the analysis via discriminant reduction method using the optimal model described in Section 7.6.1 with a combination of parameters k_{ext} = 20, k_{int} = 1, $v_{1st} = 4$, $v_{2nd} = 4$ and $\beta = 8$. In the first dimension (the external variable), the prior mean was given a value of -0.21 for the normal component, and a value of 2.44 for the metastatic component. For the non-nodal component, the first dimension prior mean, with a value of 2.42, is very close to that for the metastatic component.

Figure 7.8 shows an example of a partially metastatic node with three components in the mixture and the non-nodal component being very similar to the normal component, on the first dimension at least. From the top row in Figure 7.8 we can see that given prior weights of (2, 0.1, 0.1) for the three components, the normal component (solid blue line) converged to a value very close to its prior distribution (dotted blue line). For the metastatic and non-nodal components, although their prior means are very close to each other, their posterior distributions (solid red line and solid green line) depart substantially from their prior distributions (dotted red line and dotted green line) converging to what seem to be correct values because the resulting image is plausible.

When the prior weights κ_p rise (row 2 in Figure 7.8) the posterior density of the non-nodal component on the first dimension starts shifting to the right, and the area from the non-nodal component in the image reduces. When κ_p increases to (3, 3, 3), rather than the correct posterior shown in row 1, here a posterior distribution between the normal and metastatic components is fitted to the non-nodal component as shown in row 3 of Figure 7.8. As a result, some areas from both normal and metastatic groups are taken up by the non-nodal component while the real non-nodal group which should appear at the top of the image has been completely misclassified as normal. When κ_p increases to (5, 5, 5) and (20, 20, 20), the non-nodal component further shifts to the right, closer to its prior distribution, and more areas from the metastatic component rather than from the normal component are then graduately encroached by the non-nodal component. Hence, with κ_{jp} increasing to a higher value, it tends to cause the severe problem of misclassifying a metastatic node or partially metastatic node as a normal node with non-nodal area.

Our experimental experiences show that weak prior weights on prior mean of metastatic and non-nodal components combined with a strong prior weight for the normal component gives the most satisfactory image classification results.



$\kappa_p = (2, 0.1, 0.1)$

$\kappa_p = (2.5, 2.5, 2.5)$

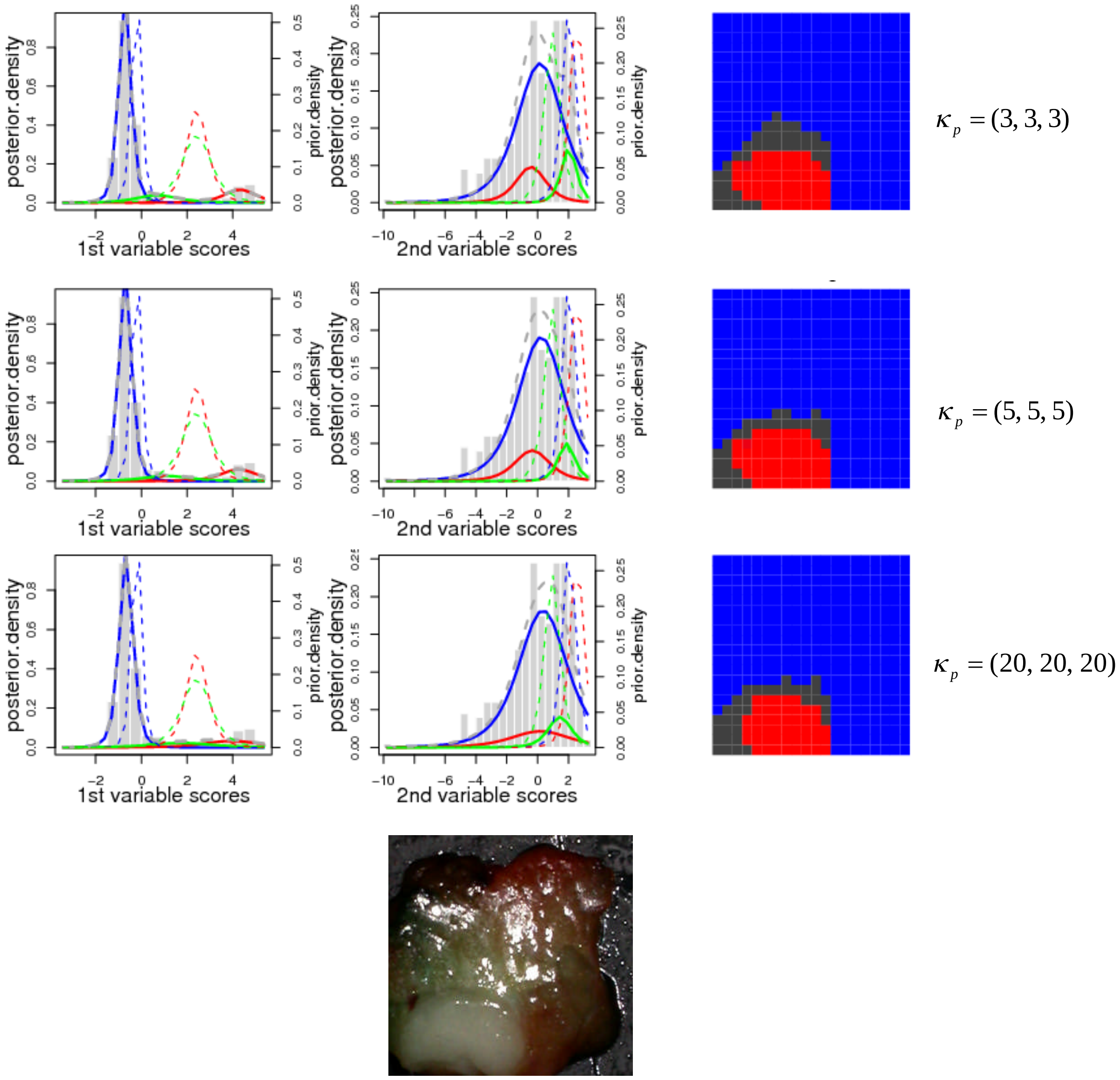


Figure 7.8: Posterior density plots and the mapping image of multivariate mixture model on the spectra from one partially metastatic node showing the model sensitivity to the prior weight. The left and middle columns show the histograms for the 1st and the 2nd variable scores of discriminant-dimension-reduced spectral data. The prior (dotted line) and posterior (solid line) densities are superimposed on the histogram for the three components with red for metastatic, blue for normal, green for non-nodal component and grey for the converged curve of the whole mixture model. The right column shows the mapping image from the 2nd stage algorithm with colour coding at each pixel (red for metastatic, blue for normal and black for non-nodal spectra). The rows 1-5 correspond to the results when the shrinkage parameter κ_p takes values of (2, 0.1, 0.1), (2.5, 2.5, 2.5), (3, 3, 3), (5, 5, 5) and (20, 20, 20), respectively for prior mean of normal, metastatic and non-nodal components. The last row shows a photograph of this node.

7.7.2.2 Sensitivity to prior distribution of mean for the non-nodal component

An important component of the mixture model is the prior model for the means μ_j, which is taken to be a multivariate normal distribution $N(\mu_{jp}, \kappa_{jp}^{-1}\Sigma_j)$ conditional on the covariance matrix Σ_j.

Different choices of prior means μ_{jp} may affect the result of the image convergence. Since the prior means for normal and metastatic components are given by manual measurement data we are not going to explore the choices of these prior means. However, the prior mean of non-nodal component, which is constructed from some spectra selected from the non-nodal areas of scanned nodes by an experienced physicist as described in Section 7.3, is much more speculative so that different choices of prior mean for the non-nodal component are worth exploring.

Two examples of partially metastatic nodes in Figures 7.9-7.10 show how different choices of prior mean for the non-nodal component affect the image classification results, taking some plausible values of -10, -5, 0, 5 and 10 for both dimensions of μ_{3p} (a two-dimensional prior mean vector for the non-nodal component). The analysis here is based on the optimal classification model via discriminant dimension reduction method with the prior weight $\kappa_p = (2, 0.1, 0.1)$ described in Section 7.6.1.

From Figure 7.9 we can see that along the axis of the external (1st) variable the group of non-nodal component is almost covered by the normal group but the two nodal components can separate from each other. The non-nodal component can be separated from nodal components along the axis of of the 2nd, internal, variable. Different choices of prior mean μ_{3p} do not seem to have much impact on the image results.

In contrast, in Figure 7.10, when neither external nor internal variable can easily discriminate the non-nodal component from the nodal components, we see that taking different values of the prior mean really affect the image results of the non-nodal component, and also the metastatic component which seems easier to confuse with non-nodal than does the normal component.

Although the real mean for the non-nodal component varies tremendously from node to node, when the prior mean takes values around 0 or between 0 and the real mean for both external and internal variable, the parameters for the non-nodal component converge well to a sensible looking picture, as do those for the nodal components (which can be seen from a good match between the mapping image and the real picture shown in the bottom row of Figure 7.9). The two-dimensional prior mean for the non-nodal component constructed from some selected non-nodal spectra, with a value around 2.42 for the external variable, is a reasonable compromise choice

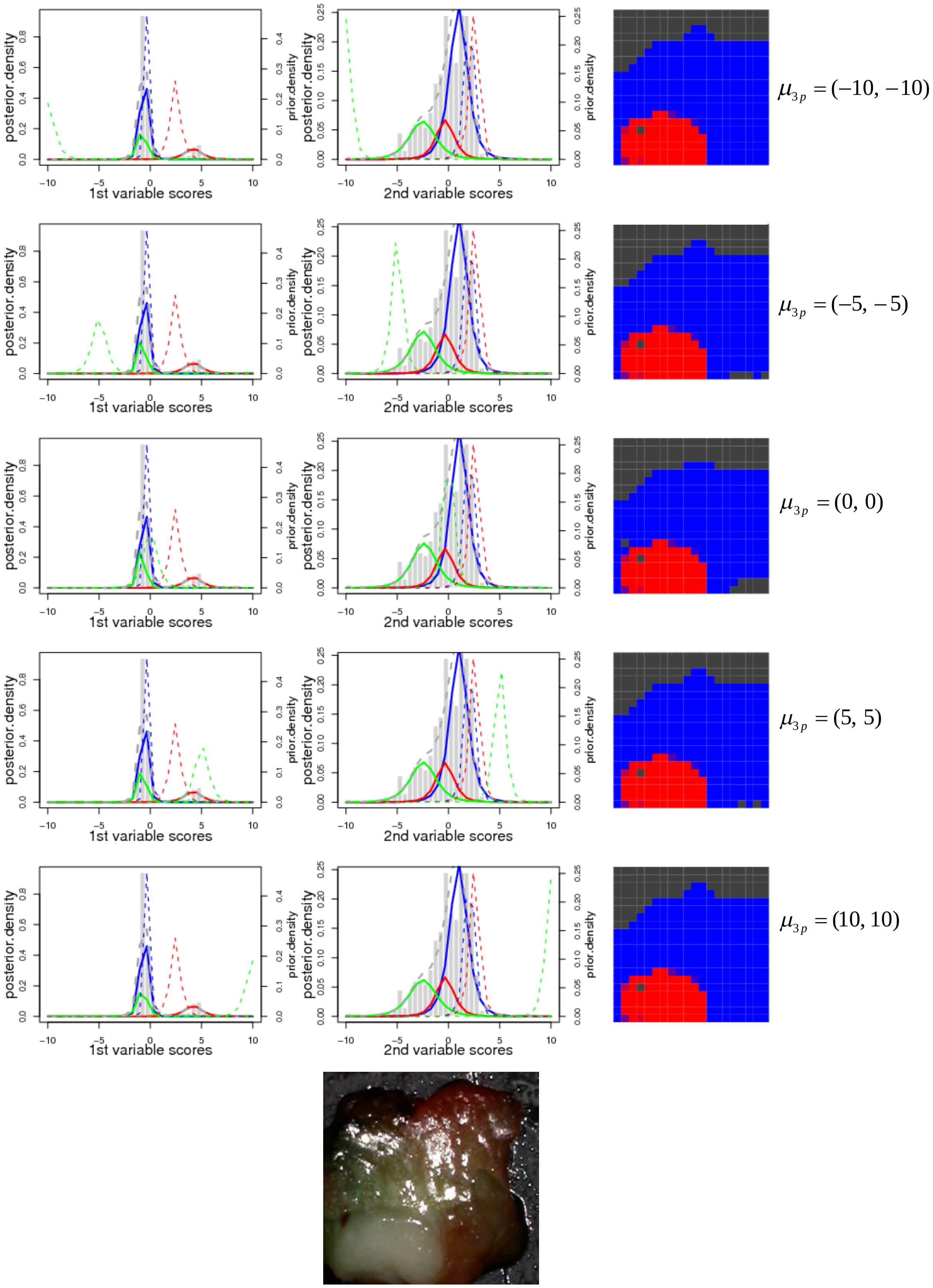


Figure 7.9: Posterior density plots and the mapping image of multivariate mixture model on the spectra from one partially metastatic node for investigating the model sensitivity to the prior mean. The left and middle columns show the histograms for the 1st and the 2nd variable scores of

discriminant-dimension-reduced spectral data with prior (dotted line) and posterior (solid line) densities superimposed for three components in the mixture with red for metastatic, blue for normal, green for non-nodal area and grey for the converged curve of the whole mixture model. The right column shows the mapping image from the 2nd stage algorithm with colour coding at each pixel (red for metastatic, blue for normal and black for non-nodal). Rows 1-5 correspond to the results when the prior mean for the non-nodal component takes values of $\mu_{3p}[1]=\mu_{3p}[2]=-10,-5,0,\ 5,10,$ with the shrinkage parameter $\kappa_p=(2,0.1,0.1)$. The last row shows a photograph of this node.

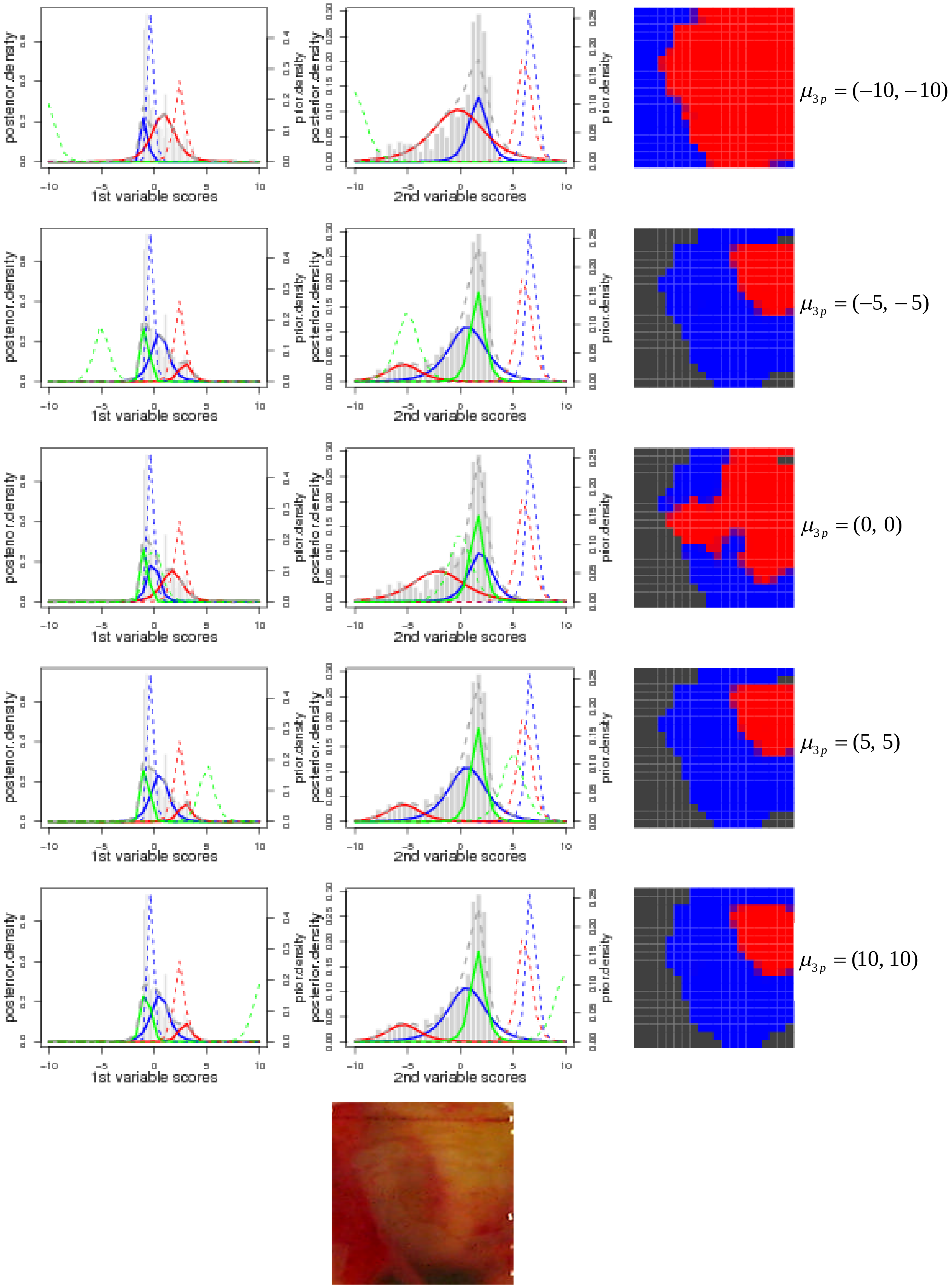


Figure 7.10: Posterior density plots and the mapping image of multivariate mixture model on the spectra from one partially metastatic node for investigating the model sensitivity to the

prior mean. The left and middle columns show the histograms for the 1st and the 2nd variable scores of discriminant-dimension-reduced spectral data with prior (dotted line) and posterior (solid line) densities superimposed for three components in the mixture with red for metastatic, blue for normal, green for non-nodal area and grey for the converged curve of the whole mixture model. The right column shows the mapping image from the 2nd stage algorithm with colour coding at each pixel (red for metastatic, blue for normal and black for non-nodal). Rows 1-5 correspond to the results when prior mean for non-nodal component takes values of $\mu_{3p}[1] = \mu_{3p}[2] = \text{-}10, \text{-}5, 0,\ 5, 10,$ with the shrinkage parameter $\kappa_p = (2, 0.1, 0.1)$. The last row shows a photograph of this node.

7.7.2.3 Sensitivity to prior distribution of variance for the non-nodal component

Three possible choices of prior scale Λ^{-1} were investigated for the non-nodal component : (a) using the total variation of all the dimension-reduced scanning data with three unlabelled components; (b) using the variation of the selected background spectra determined by the physicist expert; or (c) using $\Lambda_{3p}^{-1} = c\,(\Lambda_{1p}^{-1} + \Lambda_{2p}^{-1})$, where c is a coefficient (taking the value of 1 or 0.5) and Λ_{1p}^{-1} and Λ_{2p}^{-1} are the scale matrices of the prior distributions for normal and metastatic components, respectively, derived from the manual data as described in Section 7.3. Both (a) and (c) result in a widely spread prior distribution.

Investigations, not reported in detail here, suggest that the eventual classification and image are not particularly sensitive to its choice, and we have settled for using method (b) in our analysis.

7.7.2.4 Interactions between the shrinkage parameter κ_p and the prior mean μ_p for the non-nodal component

The interactions between the shrinkage parameter (prior weight) κ_p and the prior mean μ_p are investigated here on an example of a partially metastatic node.

When weak prior weight is given to the prior mean, especially for metastatic and non-nodal components, different choices of the prior mean don't have much impact on the posterior mean, at least for this node, as shown in Figure 7.11 (a). When strong prior weight is imposed on the prior mean, different choices of prior mean do have an effect on the posterior distributions and, as a result, in the mapping image the component far away from its prior mean might disappear or be misclassified as an other component with closer mean, as shown in Figures 7.11 (b)-(d).

With a relatively strong prior weight of 2 for all three components in the mixture as shown in Figure 7.11 (b) , when the 1st dimensional value of prior mean for non-nodal component rises from -10 to 0 (first 3 rows of Figure 7.11 (b)) a non-nodal component appears at the top of the image. When the prior mean increases from 0 to 5 (rows 3 and 4 of Figure 7.11 (b)), the non-nodal component begins to interfere with the metastatic component and at the same time the metastatic component encroaches on some of the area of the normal component. When this prior mean increases from 5 to 10 (rows 4 and 5 in Figure 7.11 (b)), the non-nodal component disappears altogether.

When the prior weight is getting stronger with k_{3p} increasing to 5 and 10 as shown in Figure 7.11 (c) (d), the non-nodal component only appears when its prior mean takes values close to the data mean.

A strong prior weight allows less flexibility and adaptation of the data to the choices of prior mean (and variance). As a result, there are problems when the prior and data do not agree. When modest prior weight is given, observations give greater contributions to the updating of the posterior mean and variance, and some bias in the choices of priors won’t spoil the image results. Although this may result in a relatively slow convergence, it enables the parameters converge in a proper and accurate way, especially with the preferred choices of strong prior weight for normal component and weak prior weight for metastatic and non-nodal components as shown in Figure 7.11(a).

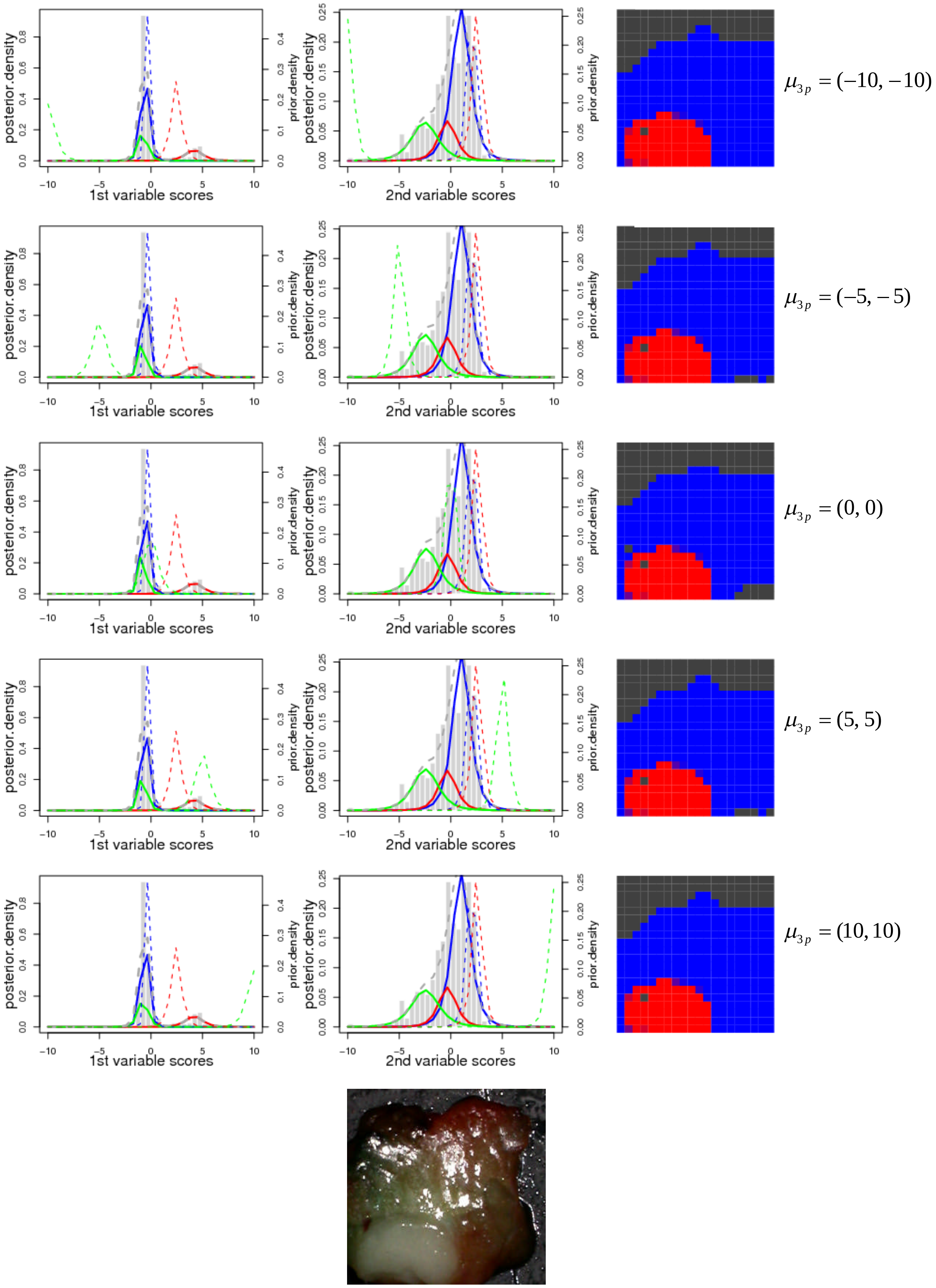


Figure 7.11 (a): Posterior density plots and the mapping image of multivariate mixture model on the spectra from one partially metastatic node for investigating the interactions between

shrinkage parameter and prior mean. Here the shrinkage parameter $\kappa_p = (2,0.1,0.1)$. The left and middle columns show the histograms for the 1st and the 2nd variable scores of discriminant-dimension-reduced spectral data with prior (dotted line) and posterior (solid line) densities superimposed for the three components in the mixture with red for metastatic, blue for normal, green for non-nodal area and grey for the converged curve of the whole mixture model. The right column shows the mapping image from the 2nd stage algorithm with colour coding at each pixel (red for metastatic, blue for normal and black for non-nodal). Rows 1-5 correspond to the results with prior mean for non-nodal component taking values of $\mu_{3p}[1] = \mu_{3p}[2] = -10, -5, 0, \ 5, 10$. The last row shows a photograph of this node.

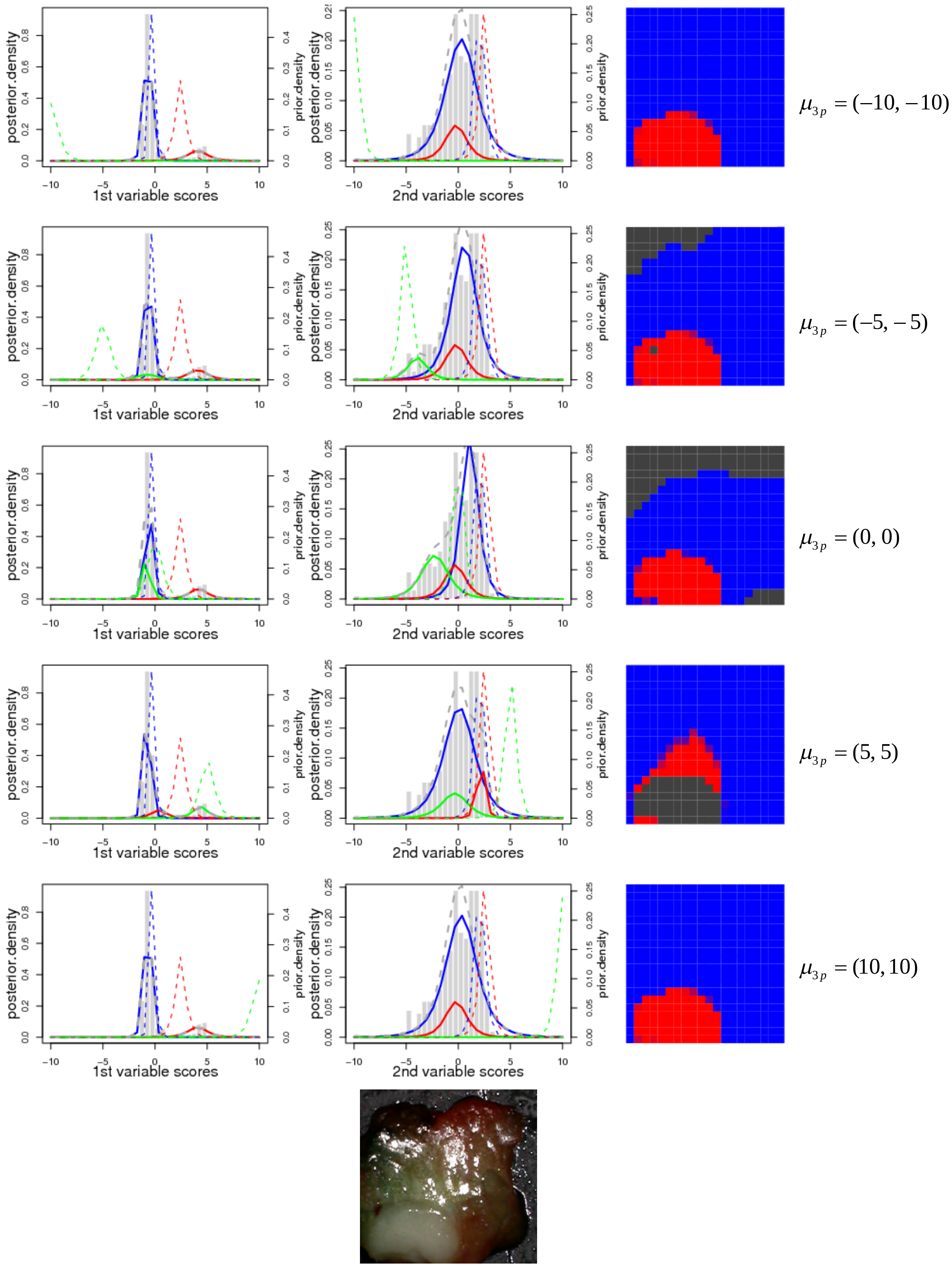


Figure 7.11 (b): Posterior density plots and the mapping image of multivariate mixture model on the spectra from one partially metastatic node for investigating the interactions between shrinkage parameter and prior mean. Here the shrinkage parameter $\kappa_p = (2, 2, 2)$. Rows 1-5 correspond to the results with prior mean for non-nodal component taking values of $\mu_{3p}[1] = \mu_{3p}[2] = $ -10, -5, 0, 5, 10 . The last row shows a photograph of this node.

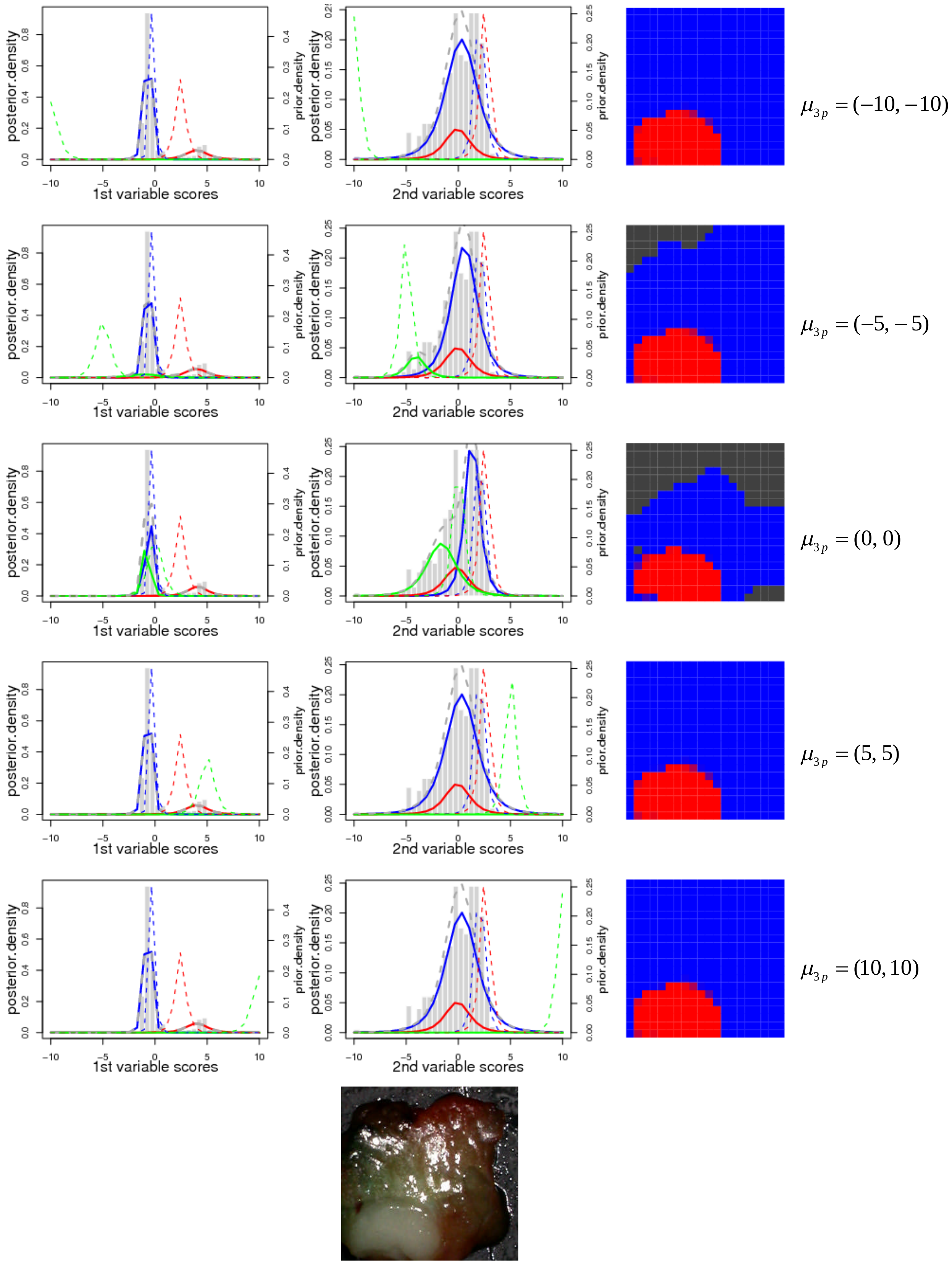


Figure 7.11 (c): Posterior density plots and the mapping image of multivariate mixture model on the spectra from one partially metastatic node for investigating the interactions between shrinkage parameter and prior mean. Here the shrinkage parameter $\kappa_p = (5, 5, 5)$. Rows 1-5 correspond to the results with prior mean for non-nodal component taking values of $\mu_{3p}[1] = \mu_{3p}[2] = \text{-}10, \text{-}5, 0, 5, 10$. The last row shows a photograph of this node.

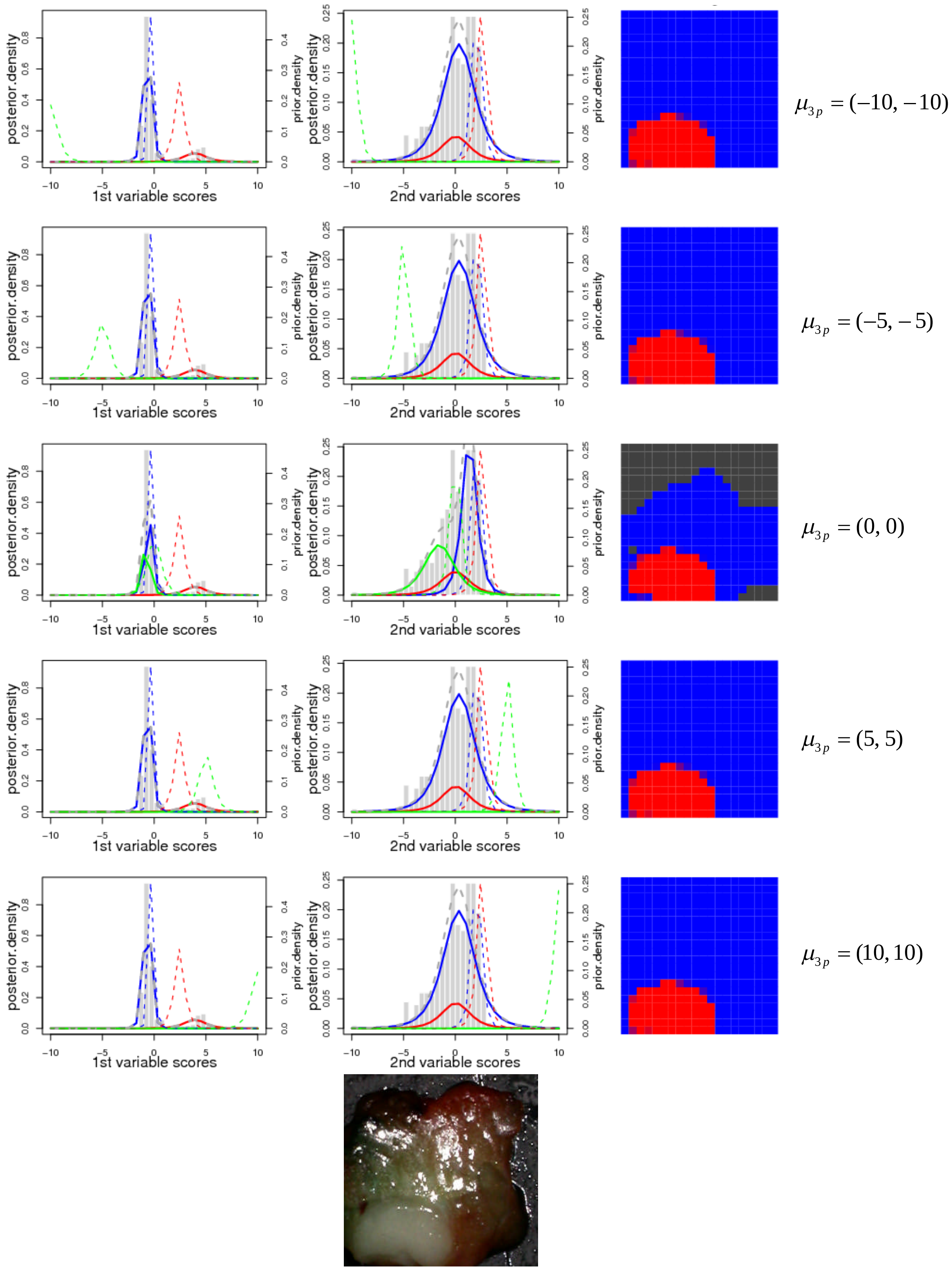


Figure 7.11 (d): Posterior density plots and the mapping image of multivariate mixture model on the spectra from one partially metastatic node for investigating the interactions between shrinkage parameter and prior mean. Here the shrinkage parameter $\kappa_p = (10, 10, 10)$. Rows 1-5 correspond to the results with prior mean for non-nodal component taking values of $\mu_{3p}[1] = \mu_{3p}[2] = \text{-}10, \text{-}5, 0,\ 5, 10$. The last row shows a photograph of this node.

7.7.2.5 How the two-stage model fitting algorithm works for image classification

Two examples are given here to show how the two-stage algorithm works. The images are generated from two scanned nodes data after discriminant dimension reduction.

Figure 7.12 gives an example of a partially metastatic node with three components. Points in blue, red and green refer to the pixels classified as normal, metastatic and non-nodal component, respectively, at the current stage.

The model uses prior distributions derived from manual data for normal and metastatic components as described in Section 7.3, and a widely spread prior distribution for the non-nodal component as shown in the top panel of Figure 7.12. Guided by the prior information, the centres of the three components develop along the directions of the external (1^{st}) and internal (2^{nd}) variables, being updated by the observed data from the individual node. After the first stage fitting, shown in the middle panel of Figure 7.12, all three groups, but especially the metastatic one, have moved substantially from the prior positions. The groups seem to be fairly well separated from each other, although the variances are still large for each group.

In this stage the features of the three components are seen from their distributions to have fairly well separated means (the peak of the posterior distribution of the 1^{st} variable scores from the metastatic component matches one of the two modes of the distribution of canonical scores from metastatic spectra in Figure 7.2) and a relatively concentrated dispersion for the non-nodal component. The image configuration at this stage shows a rough match to the photograph, with some background spots on the upper left corner being misclassified as normal or metastatic.

In stage 2, on incorporating the spatial information into the model the groups do not move very much, but isolated pixels are tidied up, and those misclassified pixels from the background area all recover as shown in the bottom panel of Figure 7.12. Although the fitted posterior distributions of the three components show some overlap with each other, the resulting image from this stage shows a much better match to the photograph. This example shows clearly the effect of incorporating spatial information into the model. The estimated groups become less concentrated and overlap more, but the resulting image is a better one.

Figure 7.13 gives an example of a node including only normal and non-nodal components. Given relatively informative prior distributions for two nodal components and a diffuse prior distribution for non-nodal component, three groups are fitted in stage 1 with generally more concentrated posterior distributions. However, when compared with the photograph, the image result (on the mid-right panel of Figure 7.13) shows that normal and non-nodal components are separated well, but the observations in between classified as metastatic are all false positives since metastatic group does not exist in this node.

After stage 2 converges, all the misclassifications from stage 1 are corrected with the effect that the whole fitted metastatic group disappears and three scattered normal spots misclassified as non-nodal in stage 1 all flip back to normal. With spatial information incorporated in this stage, the improvement here is significant, but not surprising, since the smoothing parameter can eliminate a scattered small group (misclassified metastatic group). With no class membership in the metastatic group in stage 2, the posterior distribution for the metastatic component is the same as its prior distribution. The between-stages image correction rules described in Section 5.5.3.2 which might also have removed the metastatic group were not applied here in order to show clearly that the stage 2 algorithm can also fix the problem of some small group misclassifications in the stage 1.

From both examples we can see that the two-stage algorithm works well here in a flexible way. Stage 1 focuses on the distribution convergence with the result of tight fitted groups, and a rough convergence is enough to generate plausible starting points for the stage 2 fitting, avoiding wasting time in fitting a partial model at this stage. With the MRF spatial prior incorporated in stage 2, although the distribution density fitting might be not as good as the previous stage, the image becomes much smoother and some misclassification caused by the small group fitting and distribution density fitting might be able to recover. When small groups survive this stage, they are probably real. In both cases, priors calculated from manual measurement data for the normal component seem fairly informative and plausible. The prior for the metastatic component seems to be less useful, or at least rather wide of the mark in Figure 7.12.

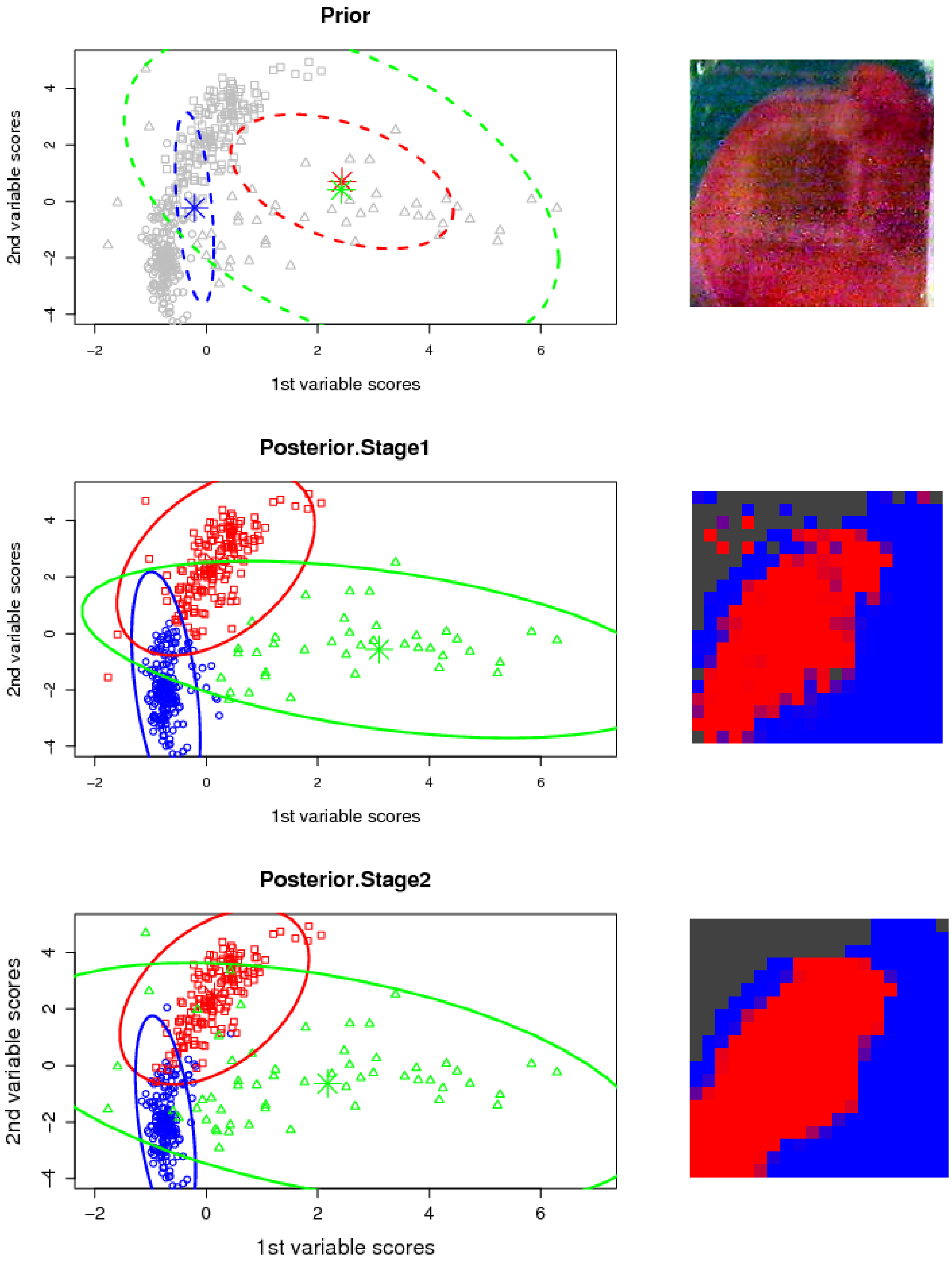


Figure 7.12: Plot of the two-stage imaging result from a partially metastatic node after discriminant dimension reduction. Two-dimensional prior (top left panel) and posterior (middle and bottom left panels) probability density contour plots showing the effect of stage 1 (middle panel) and stage 2 (bottom panel) model fitting for a mixture of three components (normal in blue, metastatic in red and non-nodal in green). The points show the fitted class membership of each pixel at current stage, the stars show the prior mean and posterior mean, and the ellipses represent (95%) probability contours of the estimated probability distribution for each component. The right panel shows the photograph of the node and the fitted images from stage 1 and stage 2 with red indicative of metastases, blue of normal, and black of non-nodal spectra.

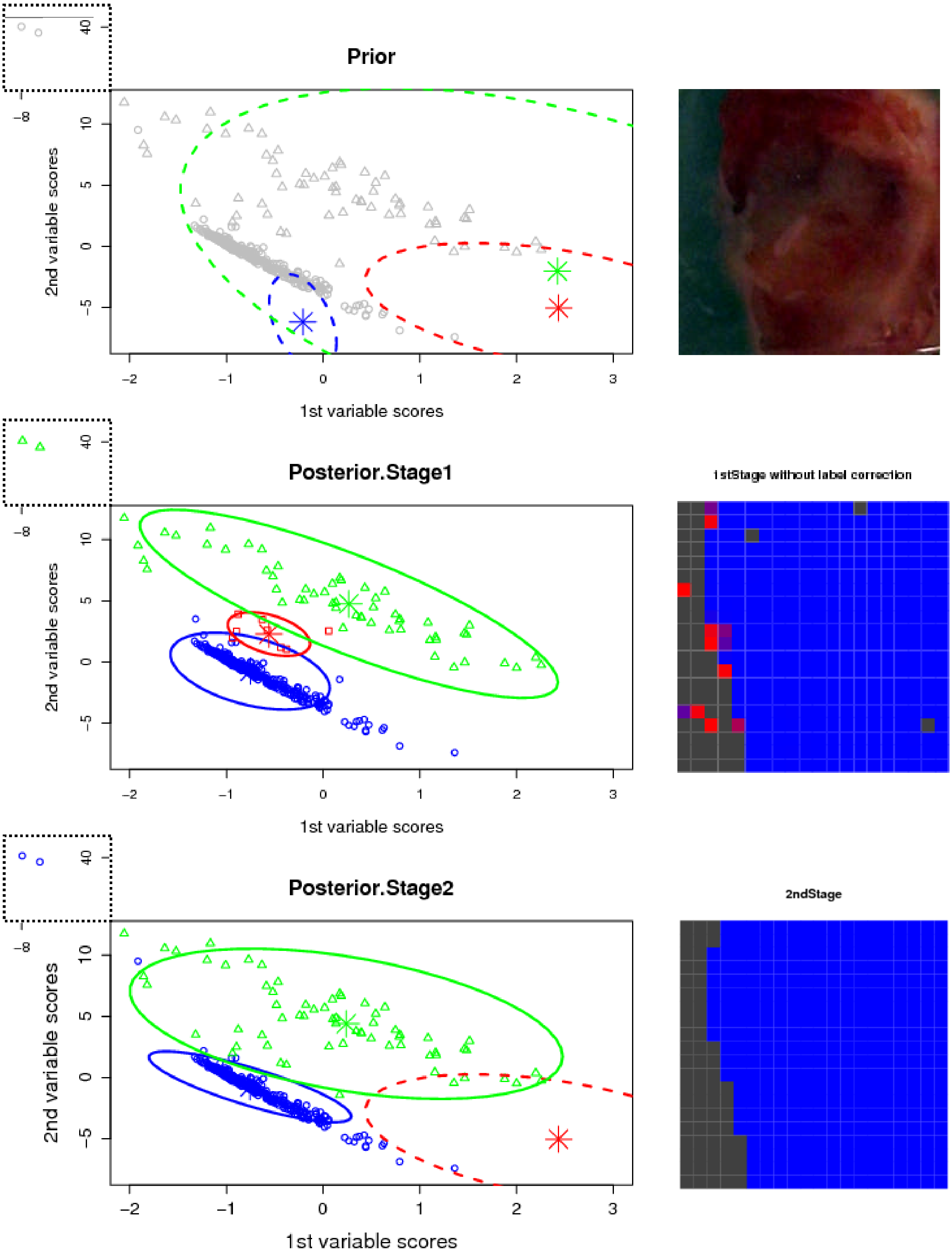


Figure 7.13: Plot of the two-stage imaging result from a totally normal node after discriminant dimension reduction. Two-dimensional prior (top left panel) and posterior (middle and bottom left panels) probability density contour plots showing the effect of stage 1 (middle panel) and stage 2 (bottom panel) model fitting for a totally normal node with two components (normal in blue and non-nodal in green) in the mixture. The points show the estimated class membership of each pixel at current stage, the stars show prior and posterior means, and the ellipses represent (95%) probability contours of the estimated probability distribution for each component. The right panel shows the photograph of the node and the two mapping images from 2 stages with red indicative of metastases, blue of normal, and black of non-nodal.

7.7.3 Sensitivity of second-stage results to MRF prior distribution

In the MRF spatial prior distribution, $\pi_{ij} = p(y_i = j | y_{\partial i}) \propto \alpha_{ij} \exp\{-\beta\, u_{ij}(y)\}$, as described in Equation (6.10), β rewards smoothness, and α_{ij} reflects the position of each pixel in the image, giving more probability to the non-nodal group for the pixels on the corner or edge than that in the centre, according to the definition of α_{ij} in Equations (5.56) and (6.11).

Figure 7.14 gives an example of a partially metastatic node to show how the two-stage algorithm works when β varies. When $\beta = 0$ and $\alpha_{ij} \equiv 1$, it is equivalent to generating the result from the 1st stage algorithm, which we use to initialize the 2nd stage algorithm. When β increases, the area of metastatic component extends, as does the normal component. Some isolated pixels, generally associated with the metastatic component, vanish. When single pixels have different labels from their contiguous pixels, only the ones with large posterior probability of occurrence can keep the current label. Once β reaches 8 or 10 the image is as smooth as the pixel size will allow. Here we take the value of 8 for discriminant dimension reduction and the value 4 for PCA dimension reduction as compromise choices of β to keep the mapping images closer to the real pictures in general**.**

Figure 7.15 shows the effect of the position parameter α_{ij} (defined by Equations (6.11) and (5.57)) on the image as the power parameter ρ in Equation (5.57) is varied. For a given value of β, when the value of ρ increases, more non-nodal component appears, especially in the lower right corner of the image. Here we chose $\rho = 1.5$.

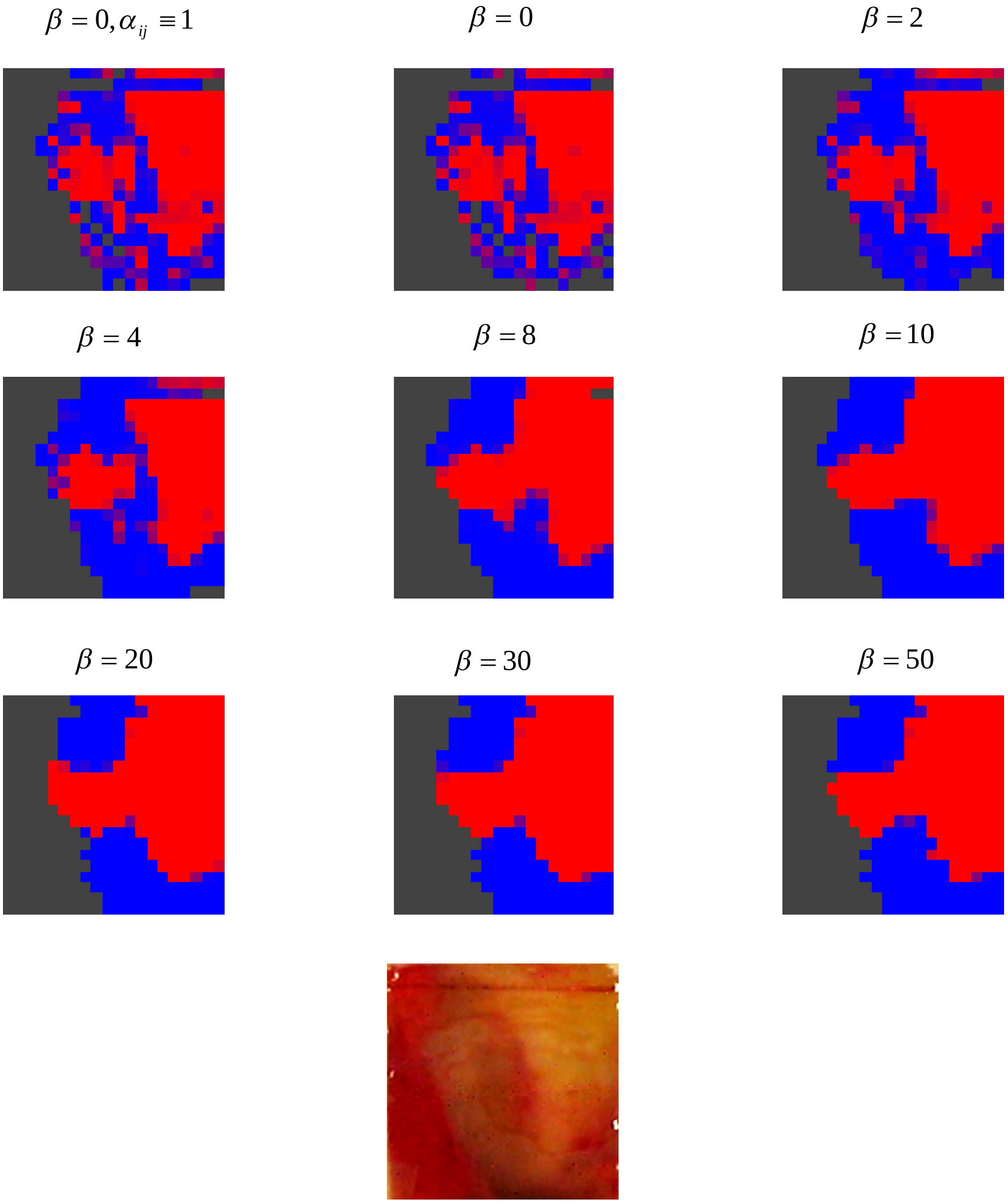
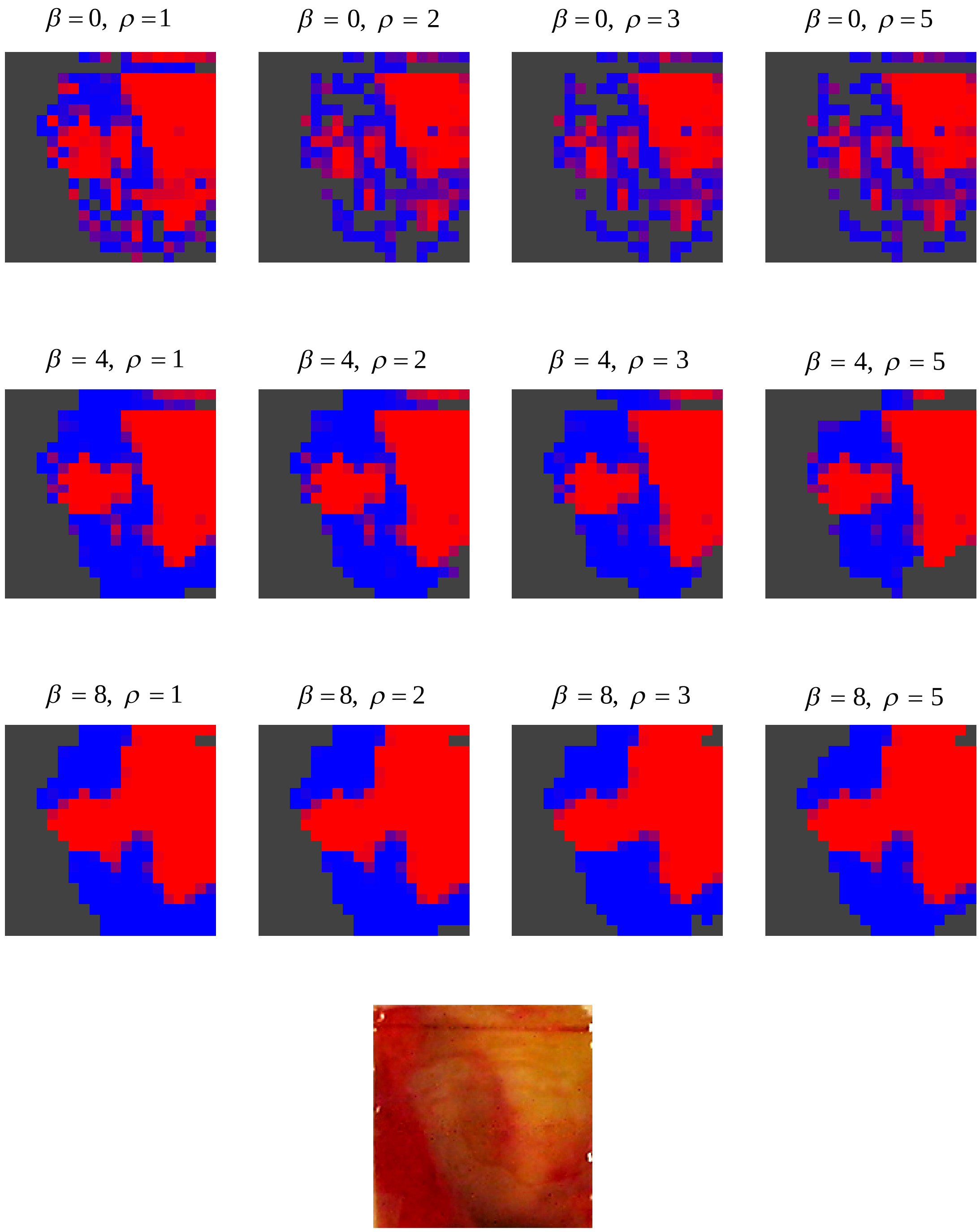


Figure 7.14: Second-stage results for one partially metastatic node after discriminant reduction with $\beta = 0$, 2, 4, 6, 8, 10, 20, 30, 50. ($\alpha_{ij} \equiv 1$ for the 1st image and $\rho = 1.5$ for the others.) The first three rows are the resulting maps of the spectra with red indicative of metastatic, blue normal and black non-nodal. The last row is a photograph of the node.



$\beta = 0,\ \rho = 1$ $\beta = 0,\ \rho = 2$ $\beta = 0,\ \rho = 3$ $\beta = 0,\ \rho = 5$

$\beta = 4,\ \rho = 1$ $\beta = 4,\ \rho = 2$ $\beta = 4,\ \rho = 3$ $\beta = 4,\ \rho = 5$

$\beta = 8,\ \rho = 1$ $\beta = 8,\ \rho = 2$ $\beta = 8,\ \rho = 3$ $\beta = 8,\ \rho = 5$

Figure 7.15: Second-stage results for one partially metastatic node with all the combinations of $\beta = 0$, 4, 8 and ρ = 1, 2, 3, 5. The first three rows are the resulting maps of the spectra with red indicative of metastatic, blue normal and black non-nodal. The last row is a photograph of the node.

7.7.4 Sensitivity to the choices of multivariate *t* distributions or multivariate normal distributions

In this image analysis, our experiments show that the multivariate *t*-distribution with heavier tails, as an alternative to multivariate Gaussian distribution, shows robustness towards outliers by minimizing their impact on the estimation of the model parameters, which increases flexibility and robustness of the model. In general the multivariate *t*-distribution provides an overall more accurate image classification result than multivariate Gaussian distribution for the scanning data.

An example of a totally normal node after discriminant dimension reduction is given in Figure 7.16. The rows correspond to different choices of degrees of freedom of *t* distribution (varying from 4 to 20) and normal distribution. It shows that using multivariate normal distributions or multivariate *t* distributions with large degrees of freedom leads to the appearance of three components in the mixture. The second stage eliminates some of the isolated areas, but at least one clump always persists.

Although the multivariate normal distribution always gives better fitting to the histogram of the data than the multivariate *t* distribution, it gives worse classification accuracy. Good histogram convergence doesn't mean goodness of model fitting since the histogram only reflects the marginal distribution and does not include spatial information.

When using large degrees of freedom v_{1st} of *t* or a normal distribution in the first stage, we should be very careful with the model fitting and the choice of prior. When there are three components in the mixture, the metastatic group is unlikely to be missed by the classification and this gives high sensitivity. However when there are fewer than three components in the mixture, the extra component misclassified in the first stage using a large v_{1st} is usually difficult to remove in the second stage algorithm since a degrees of freedom, v_{2nd}, larger than that in the first stage will be used in the second stage for the reasons of model convergence as described in Section 7.6.3. In this case small v_{1st} will be preferred and thus lead to a high specificity. However in the real prediction we won't be able to know how many components there are in the mixture, using a modest v_{1st} in the 1st stage to achieve a vague model fitting may give more possibility for further accurate fitting in the 2nd stage.

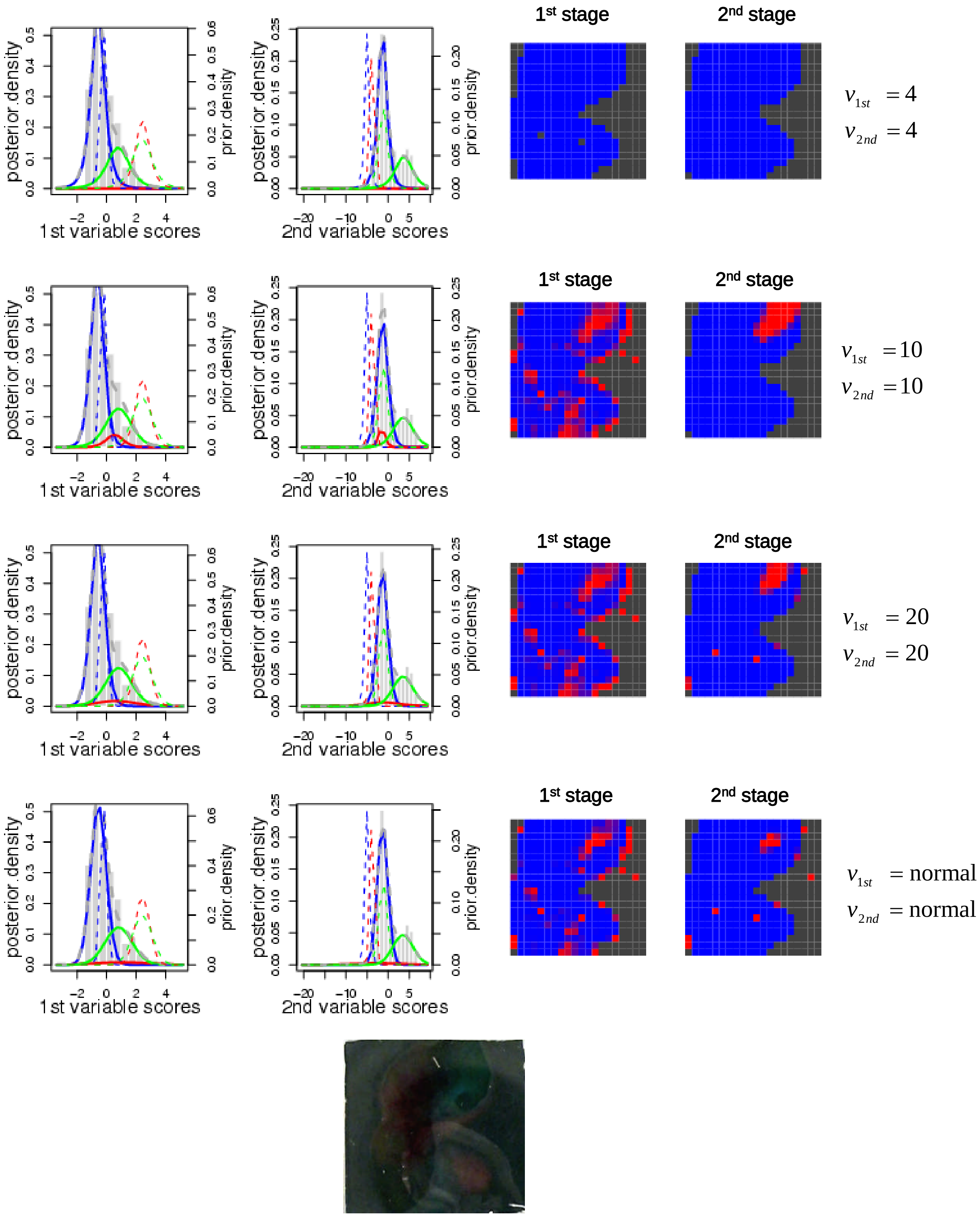


Figure 7.16: Histogram for the 1st (1st column) and the 2nd (2nd column) variable scores of spectral data from one totally normal node after discriminant reduction, with prior (dotted line) and posterior (solid line) densities superimposed for three components of multivariate mixture model

with red for metastatic, blue for normal and green for non-nodal component. The corresponding mapping images from the 1st stage and 2nd stage algorithms are shown in the last two columns with colour coding at each pixel (red for metastatic, blue for normal and black for non-nodal spectra). The rows correspond to different choices of *t* distribution with degrees of freedom varying from 4 to 20, and normal distribution. The last row is a photograph of this node.

7.7.5 Label-switching problem

The label switching problem, mainly due to the use of a diffuse prior for the background has been described in Section 5.5.3.2.

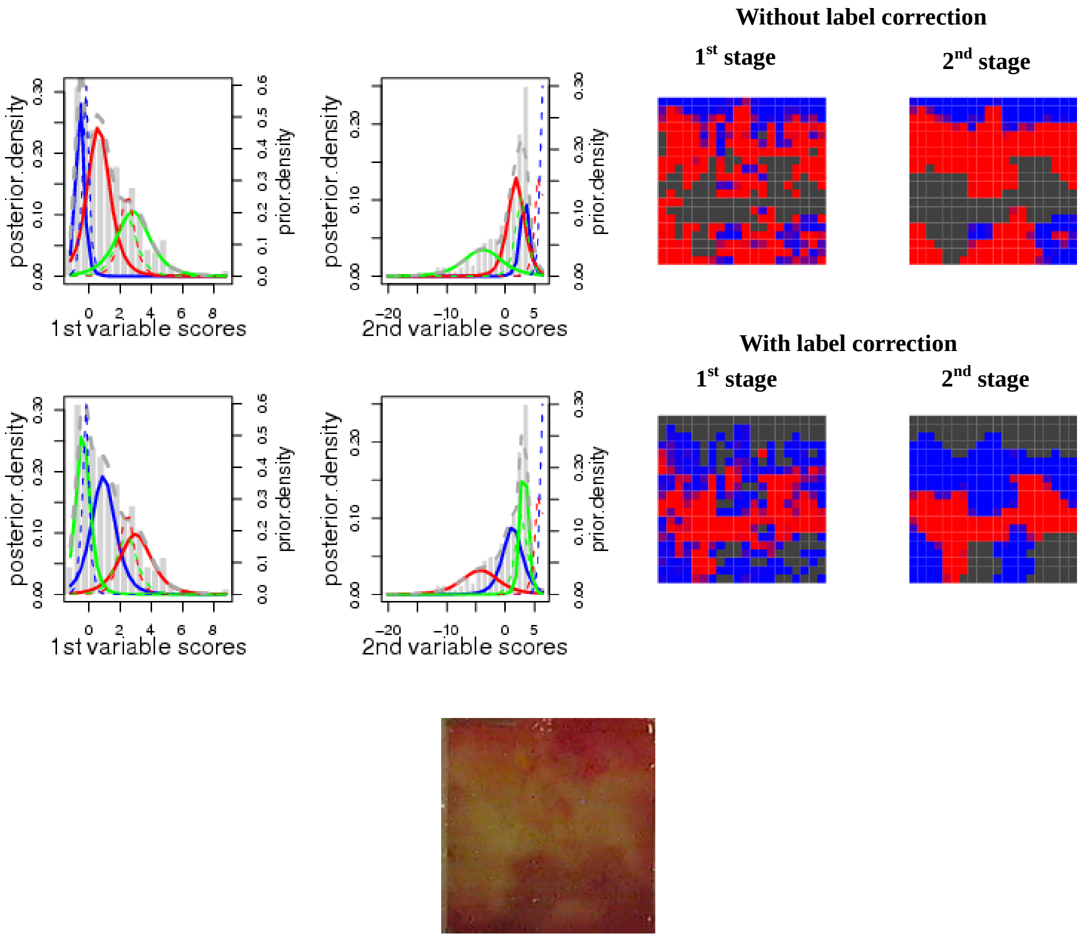


Figure 7.17: Histogram for the 1st (1st column) and the 2nd (2nd column) variable scores of spectral data from one partially metastatic node after discriminant dimension reduction, with prior (dotted line) and posterior (solid line) densities superimposed for three components of multivariate mixture model with red for metastatic, blue for normal and green for non-nodal component. The corresponding mapping images from the 1st stage and the 2nd stage are shown in the last two columns with colour coding at each pixel (red for metastatic, blue for normal and black for non-nodal). The 1st and 2nd rows show the results without and with class label correction rule applied between two stages. The last row shows a photograph of this node.

Figure 7.17 shows an example of a partially metastatic node with all three components mislabeled after discriminant dimension reduction. This is essentially because on the external variable the non-nodal and metastatic components are given similar prior means with a large prior variance for the background and low prior weights. The non-nodal component is incorrectly identified, and the other two groups then remain in the same relative positions as suggested by the prior means, but both wrong, as shown in the top panel of Figure 7.17. Even though the second stage algorithm introduces the position parameter α_{ij} which should help it to identify the background, it is still unable to recover.

The class label correction rules described in Section 5.5.3.2 fix the mislabeling problems in between stages. By giving higher scores to the pixels on the corner or edge, the non-nodal component can usually be recognized, followed by the recognition of the other two groups. This gives correct starting points to the smooth fitting in the second stage as shown in the bottom panel of Figure 7.17.

Although they are easy to implement, having to use these rules is not ideal. One possible alternative would be to improve the first stage fitting by incorporating a prior using α_{ij}.

7.7.6 Number of components problem

Although our model can deal successfully with the nodes with less than three components, it often attempts to find an extra group, normally with only a small number of pixels.

Figure 7.18 is an example of a totally normal node (with two components of normal and non-nodal in the mixture). The results and image are generated from the optimal model via PCA reduction described in Section 7.6.2, with k_g = 15, k_l = 2, $v_{1st} = 10$, $v_{2nd} = 10$ and $\beta = 20$. Without using the image correction rules of Section 5.5.3.2 between the stages three components were fitted by the two-stage model, some pixels from the normal component being misclassified as metastatic ones in the first stage algorithm, and persisting in the second stage as shown in the top panel of Figure

7.18. For both dimensions, the PC score distribution of the false positive spectra looks like a tail of the score distribution of the normal spectra. Although the scores of both nodal components are very close, the large prior variance given to the metastatic component, especially in the second dimensional scores, allows its convergence to a separate component in the mixture. Different parameter choices were tried on this node: using a heavy-tailed multivariate *t* distribution with lower degrees of freedom did not seem to help; building the model in a space with more dimensions gave more chance to fit the unnecessary multimodality of the spectra data.

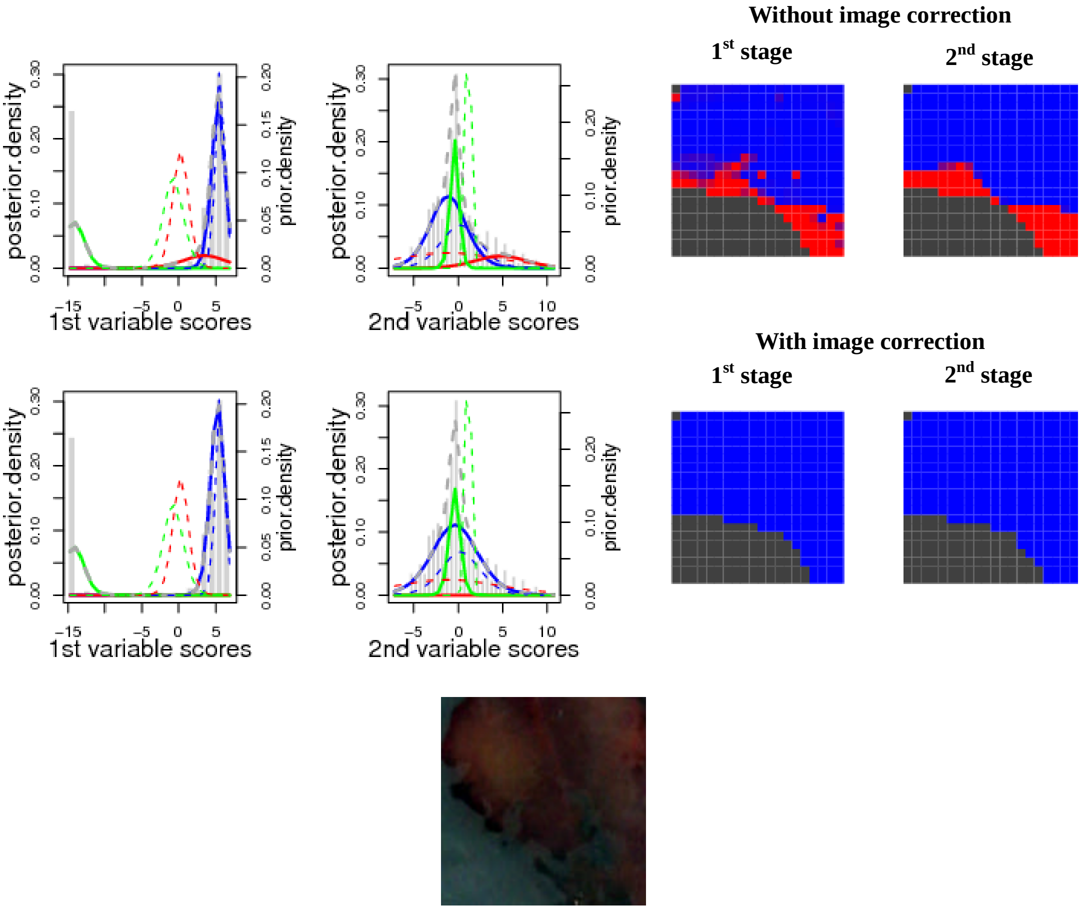


Figure 7.18: Histogram for the 1st (1st column) and the 2nd (2nd column) variable scores of spectral data from one normal node after PCA dimension reduction, with prior (dotted line) and posterior (solid line) densities superimposed for three components of multivariate mixture model with red for metastatic, blue for normal and green for non-nodal component. The corresponding images from the 1st stage and the 2nd stage are shown in the last two columns with colour coding at each pixel (red for metastatic, blue for normal and black for non-nodal). The 1st and 2nd rows show the results without and with image correction rules applied between two stages. The last row shows a photograph of this node.

To avoid encouraging the small group fitting in the mixtures with a very close mean to that of another sizable group, the image correction rules described in Section 5.5.3.2 are applied. The false positive metastatic component is thus merged to the normal component at the end of the first stage algorithm and the model fitting continues working in a right way in the second stage algorithm.

7.8 Discussions and conclusions

In Chapters 5 and 6, a partially supervised image classification algorithm based on a composite Bayesian multivariate finite mixture model with MRF spatial prior was developed to represent a scanned node image. This chapter has applied the model to the scanned nodes and explored its behaviour.

Two datasets with different structure and knowledge have been used for model construction: one is the manual measurement data with known reference for spectra from totally normal and totally metastatic nodes; the other is scanning data with three unknown groups (normal, metastatic and non-nodal) in the images from scanned measurements.

A traditional supervised classification method applied directly to the scanned data is not suitable here to derive an algorithm to classify pixels, since there is no reference pathology available for individual pixels in the image and, furthermore, the required training set for the non-nodal group from background areas is not available, and these background areas are very variable from node to node.

The key issues addressed in Chapters 5-7 are the representation of knowledge and inference methods for using the available knowledge to infer the correct image. The main idea is to enable an integration of a priori knowledge from manual measurement data, with accumulated evidence from scanning data, encoded in terms of a joint posterior probability distribution with Markov random field, through a Bayesian formalism.

The spectral data are high-dimensional and are reduced to a low dimensional space, two dimensions work best, before constructing an image classification model.

Two customized dimension reduction methods have been explored. The model on two-dimensional data reduced by discriminant dimension reduction works better than that via PCA dimension reduction. For discriminant reduction, the two axes (of external and internal variables) are function-specific and interpretable. Typically the first axis separates the normal and metastatic groups; the second axis allows the model to capture the remaining individual nodal features, particularly from the non-nodal component. For PCA reduction, the directions of both axes vary for different scanned nodes. How the model works in this two-dimensional space depends on the individual features of the node. Since the model via PCA reduction has more flexibility we need put more constraint on it by using high degrees of freedom for the multivariate *t* distributions representing the components.

Based on the low-dimensional data, the image classification model is fitted in two stages. In the first stage, a Bayesian multivariate finite mixture model is employed to model three unknown groups (normal, metastatic and non-nodal) in the images from scanned measurements. The normal inverse Wishart prior distributions for the parameters of normal and metastatic groups derived from the manual data are projected into the low-dimensional model space. Since the class memberships in the mixture here are not interchangeable, the prior knowledge given here works as identifiability constraint for normal and metastatic groups. In the second stage, considering the spatial correlation between adjacent pixels of image, a spatial prior based on a Markov random field (MRF) is then incorporated into the model from the first stage to represent the continuity of the image, and the first stage result used as a starting point for fitting this model.

Some recommendations may be made on the choices of distributions, prior distributions and parameters of these distributions in an attempt to balance flexibility and the incorporation of prior information:

(a) The multivariate *t* distribution works better than the normal distribution. For discriminant reduction, using low degrees of freedom in stage 1, $v_{1st} = 4$, combined with a modest or at least the same degrees of freedom in stage 2,

$v_{2nd} = 4, 10, 20$, may achieve optimal accuracy. For PCA reduction, using modest or high degrees of freedom in both two stages, $v_{1st} = 10$ and $v_{2nd} = 10, 20$, may achieve good accuracy. The *t* distribution gives higher classification accuracy and increases the flexibility and robustness of the model by minimizing the impact of outliers on the model parameters estimation. Especially the rough convergence achieved in the first stage algorithm using the *t* distribution gives more flexibility for model fitting in the second stage.

(b) The choice of prior distributions on the parameters of non-nodal component is more difficult than that for the nodal groups because of variability in this component and lack of training data. Here we chose a fairly flat prior to make the model more flexible.

(c) The prior weight for three components is a combination of a strong value for normal component and weak values for metastatic and non-nodal components, which makes the model more flexible to variability in these components between nodes.

(d) Weak convergence in the first stage leaves more flexibility for model fitting in the second stage and does not waste time.

(e) Some interactions exist between different parameters of the model. Optimising the parameter choices requires careful experimentation.

Label switching and an incorrect number of components are two common problems from mixture models for image classification. Misclassifications do happen, especially to some small groups of pixels but occasionally to whole groups. Some proposed correction rules have been applied between the two stages of the model fitting algorithms to deal with these problems. These work, but it has to be admitted that this is not an elegant solution. It would be nicer to construct a first-stage model that avoided the problems in the first place. Better modeling of the distribution of the metastatic spectra, perhaps even as a mixture itself, would be one possible approach. Incorporating the positional prior, via α_{ij}, into the first stage model would also be worth investigating.

Our experiments show that good classification results can be achieved by different options. Working in a PCA reduced two-dimensional space with multivariate *t* distributions on moderate degrees of freedom gives higher sensitivity. Working in a discriminant reduced two-dimensional space with lower degrees of freedom multivariate *t* distribution gives higher specificity. The model with higher degrees of freedom for the multivariate *t* distributions requires a higher value of the smoother parameter β to modify the discontinuity of the image pattern. For the models with lower degrees of freedom, a lower value of β is more suitable. Although the high degrees of freedom and smoother normally give better convergence to any arbitrary distribution with multimodal structure, we attempt to keep the balance between preserving the nodal structure of each node (with the feature of multimodality) and not misclassifying normal as metastatic spectra.

Our results so far suggest that this model can be successful in recognising the variable non-nodal areas automatically and distinguishing between normal and metastatic nodes with sort of accuracy required to enable clinicians to make a rapid intraoperative diagnosis of sentinel node metastases in breast cancer. As a general method, this model may also be applied to many other situations for both noise/background recognition and multi-group tissue classification.

CHAPTER 8

CONCLUSION

In this chapter, a summary of the achieved objectives in this work is given, focussing on the specific statistical issues in the context of elastic scattering spectroscopy (ESS) applications, as well as some brief comments on the possible directions for future methodological research.

8.1 Summary of thesis

This thesis has focused on the statistical issue of how to eliminate irrelevant variations in high-dimensional ESS data and extract the most useful information to enable the classification of tissue as normal or abnormal. The classification models, either built on an individual spectrum basis or further developed for an image analysis, have been employed for detection of pre-cancerous and early cancerous changes in human tissue, with applications in Barrett's oesophagus, colon lesions and sentinel lymph nodes.

ESS, as a non-invasive and real-time *in vivo* optical diagnosis technique, is sensitive to changes in the physical properties of human tissue resulting from dysplasia. The evidence given in Chapter 2 that normal and abnormal tissues have different spectral patterns of scattering and absorbance makes it possible to use measured spectra to classify areas of tissue as normal or abnormal for purpose of cancer diagnosis.

Multivariate statistical methods reviewed in Chapter 3 are used in ESS spectral analysis, amongst which principal component discriminant analysis (PCDA) and partial least squares discriminant analysis (PLSDA) are the most explored methods throughout the thesis as general tools for dimension reduction and supervised classification. Based on these classical methods, customized multivariate methods are proposed in this thesis driven by two ESS clinical diagnosis applications: One application is gastrointestinal cancer diagnosis in Barretts esophagus and colon lesions, to which the major statistical contribution is a spectral pretreatment method called error removal by orthogonal subtraction (EROS) as described in Chapter 4. This is proposed for the situation where there exists large measurement variability while the differences between normal and abnormal tissues are subtle. The other application is breast cancer diagnosis in scanned sentinel lymph nodes, to which the main statistical contributions are two customized dimension reduction methods and a partially supervised image classification model as presented in Chapters 5-7. This is proposed for the situation where there exist large variations for the components of the mixture model and the reference pathology for algorithm training are not completely available.

Spectral pretreatment (EROS)

Spectral pretreatment works as an important step to eliminate the irrelevant variations from spectra before constructing an effective regression or classification model. When ESS spectra are measured *in vivo* by a hand-held optical probe, small changes in angle and pressure may cause considerable measurement variability which should be removed or at least reduced from the spectra.

A customized spectral pre-treatment called error removal by orthogonal subtraction (EROS) was designed to model the measurement variability and to ameliorate the effects of it from the spectra in the following way. Data on replicate measurements were analysed by PCA to identify the dominant structure in the variability, and the corresponding factors were subtracted from the spectra by orthogonal projection. Pre-treatment with EROS improved the classification accuracy and substantially reduced the model complexity, resulting in the use of a small total number of factors to attain good levels of accuracy in the classification. The success of the model has been proved in two applications with clinical diagnosis in Barrett's oesophagus and colon lesions. The robustness of the classification model after EROS pretreatment is also proved by a maintained improvement in an independent prospective prediction. The nature and possible causes of measurement variability were better understood by a laboratory experiment for helping with the model interpretation. This approach, as a general pretreatment method, can be used in many other situations.

However the success of any spectral pretreatment method is always subject to the sources of the interfering variability and the spectral structure. In our situation EROS worked well because the spectra contain a substantial amount of structured noise (interfering variability) in the different dimensions as the signal of interest that can be removed with targeted dimensions without destroying the useful information. However, if the situation is that interfering variability coincides in the same direction of the signal, neither EROS nor any other mathematical pretreatment will help.

Dimension reduction

Dimension reduction has been a fundamental topic of this thesis. Several approaches to looking for a representation of high-dimensional data in a low-dimensional space have been considered. A classical linear projection approach by principal components analysis (PCA) has been used throughout the thesis to reduce the number of spectroscopic variables. PCA plays a key role in constructing a multivariate classification/discrimination model by solving the collinearity and overfitting problems. Based on PCA, two customized dimension reduction options have been developed to solve the problems in ESS image analysis. An automated ESS scanner was developed to take measurements from a larger area of tissue to produce ESS images for cancer diagnosis. Problems arise due to the existence of background area in the scanned image with considerable between-node variation and no training data available. Our customized methods for dimension reduction try to capture as much as possible the structured nodal (normal and metastatic) variations by new components. One called discriminant dimension reduction uses an external variable, the canonical variable from a linear discriminant analysis on a separate set of manual data for which we do have reference pathology, to capture variations between normal and metastatic groups, and uses a small number of internal variables, the principal components on the individual scanning data orthogonal to the external variable, capture the remaining individual nodal features, particularly from the non-nodal component. The other, called PCA dimension reduction, uses a global PCA projection to capture most of the common information from all the scanned nodes followed by a local PCA projection to extract individual nodal features. In our application, discriminant reduction works better than PCA reduction in image classification model. Although both methods allow adaptation to individual nodal features, discriminant reduction has one common direction, adding extra model robustness, while the directions extracted by PCA reduction all depend on the individual features of each node.

The PCA dimension reduction methods provide a general option for a preliminary dimension reduction. The discriminant dimension reduction methods can be used in a general situation where some specific direction can be used as a guide for constructing a low-dimensional space.

Image classification

As an extension of the point measurement, ESS imaging generated by ESS scanning device examines a larger area of tissue and here smoothness assumptions make it possible to improve the classification of individual pixels. A partially supervised image classification model based on a Bayesian multivariate finite mixture model with a Markov random field (MRF) spatial prior was developed to represent the scanning data of each node by a statistical image containing information on both between-group features and spatial correlation features. The main idea of the composite model is to enable an integration of a priori knowledge, via normal inverse Wishart priors, from manual measurement data on the two main components, normal and metastatic, with observational evidence from the scanning data of the individual node. The smoothness of the image is modeled by a spatial prior using a Markov random field. The whole model fitting is implemented in two stages by EM and ICM algorithms, with the feature priors and the spatial prior imposed onto the model space in turn. In the first stage, a multivariate finite mixture model with three components uses informative prior distributions derived from the manually measured data for the parameters of the normal and metastatic components in the mixture, and a diffuse prior distribution for the variable non-nodal component. This allows the flexibility and adaptation to the background component varying from node to node. In the second stage, with a MRF spatial prior incorporated into the model to represent the spatial correlation between adjacent pixels of image, a rough image generated from the first stage then develops to an image showing the quality of continuity and smoothness with a good match to the nodal shape and good accuracy in diagnosis.

Our results show that from the mapping image the variable non-nodal areas can be recognized automatically and normal and metastatic nodes can then be distinguished easily with sort of accuracy required to help clinician with a rapid intraoperative diagnosis of sentinel nodal metastases in breast cancer. As a flexible and robust model, this can also be applied to many other multi-group image classification situations with incomplete information, such as group features and class memberships.

8.2 Future work

The subject of this thesis being extensive, we do not pretend to have covered it exhaustively. Several interesting issues arising from classification by ESS analysis could not be dealt with, due to lack of time. We mention some of them.

Spectral wavelength selection for classification is of essential interest. The LDA loading, which measures how intensity at each wavelength contributes to the classification, shows more important weight on certain regions of the wavelength range for discrimination between normal and abnormal tissue. The challenge is how to select the most informative spectral features with minimum redundancy and noise to enhance the classification performance without losing useful information. The wavelength regions with high-ranked loadings selected by the model for the classification task could be highly correlated among themselves. Simply combing one effective variable or region with another doesn’t necessarily form a better subset, because it might contain certain redundant features. Also calculations of all possible wavelength combinations require enormous computations. This concern is of particular interest for dimension reduction, and for the purpose of building cheaper instruments.

The topic of *ordered multi-group classification* needs to be explored with the aim of grading the level of abnormality. A sequential multi-group classification based on the two-way classification model described in Chapter 4 has already been put into practice for the real-time clinical prediction. However it would be of great interest to classify the different groups simultaneously enabling a better understanding of misclassification errors and helping a rapid real-time diagnosis. One possible approach is using the Soft Independent Modeling of Class Analogy (SIMCA) (Wold and Sjöström 1977) to capture the local structure of each group to train the classification model. However, this supervised model is very sensitive to the training data, especially when the between-group differences are subtle. Further study will focus on how to balance the flexibility and robustness of the model, avoiding misclassifications.

Concerning *imaging classification* methods, several aspects of the model can be further investigated. In the multivariate Gaussian or *t* mixture model, we have, so far, used separate and unstructured covariance matrices for each group. Choosing a common covariance matrix or a diagonal but different covariance matrix may give the opportunity to use more dimensions in the model. Although some problems of label-switching and incorrect number of components can be tackled by our model, some rules in between stages are still applied to correct some problems. Ideally the first stage model should be improved to make these rules unnecessary. Since the model fitting is implemented by a composite two-stage algorithm, it is possible to use different fitting algorithms in the two stages and alternatives could be explored.

Another issue for further study relates to *classification assessment* stimulated by the problems occurring when AUC and ROC curves are used in conjunction with cross-validation. The assumption made in the pooling strategy that the classifier outputs are comparable across all the left-out segments is generally not valid and thus large biases can be caused. How to estimate AUC and ROC curves with less bias for the purpose of model selection and comparison in practice is of potential interest. There is also a need for defining general measures for assessing multi-group classification accuracy, perhaps via an appropriate 3D-ROC curve. Also there is scope for exploring the assessment of the image classification, for the cases with both discrete and continuous values for each pixel.

APPENDIX A

NOTATION

Throughout the Chapters 5-7, the notations commonly used are listed as below:

n : the number of observations from one scanned node.

m: the number of observations from point measurement data.

c: the number of scanned nodes.

p : the number of wavelength point.

$X_{all.nodes}$: a $cn\times p$ spectral matrix of all the scanned nodes.

X_{train} : a $m\times p$ spectral matrix of point measurement spectral data.

$X_{one.node}$: a $n\times p$ spectral matrix of one scanned node.

k_{ext} : the number of principal components constructing the external variable by discriminant reduction method (the dimension of external variable).

$k_{int.0}$: the number of principal components capturing most of the variations left by the external variable by discriminant reduction method (before constructing the internal variables).

k_{int} : the number of principal components constructing the internal variables by discriminant reduction method (the dimension of internal variable).

k_g : the number of principal components constructing the global variables by PCA reduction method (the dimension of global variable).

k_l : the number of principal components constructing the local variables by PCA reduction method (the dimension of local variable).

k : the dimension of spectral matrix of one scanned node when projected to a low dimensional space ($k = k_l$ for PCA reduction, $k = 1 + k_{int}$ for discriminant reduction).

X_k : a $n \times k$ spectral matrix of one scanned node with reduced dimensions of k.

$Z_{all.nodes}$: a $cn \times k_g$ score matrix whose columns are the first k_g PC scores of $X_{all.nodes}$.

Z_{train} : a $m \times k_g$ score matrix whose columns are the first k_g PC scores of X_{train} , for the use of k_g -dimensional priors.

$Z_{one.node}$: a $n \times k_g$ score matrix whose columns contains the first k_g PC scores of $X_{one.node}$, a $n \times p$ spectral matrix for one scanned node, each row of which is a p-dimensional spectrum.

$S_{one.node}$: a $n \times k_l$ matrix whose columns consist of k_l PC scores.

L_g : the global PCA loadings of the first k_g PCs of all the scanned nodes $X_{all.nodes}$.

L_l : the local PCA loadings of the first k_l PCs of the scores matrix $Z_{one.node}$.

T_{train} : a $n \times 1$ dimensional canonical score vector of the spectra from manual data.

$T_{node.ext}$: a $n \times 1$ dimensional external variable score vector of the spectra from one scanned node.

$T_{all.nodes.ext}$: a $cn \times 1$ dimensional external variable score vector of the spectra from all the scanned nodes.

$\tilde{X}_{all.nodes}$: a $cn \times p$ spectral matrix whose columns lie in the $p-1$ dimensional subspace orthogonal to the external variable $T_{all.nodes.ext}$.

$T_{all.nodes.int.0}$: a $cn \times k_{int.0}$ score matrix whose columns contain $k_{int.0}$ PC scores.

$T_{train.int.0}$: a $m \times k_{int.0}$ score matrix whose columns contain $k_{int.0}$ PC scores.

$T_{node.int.0}$: a $n \times k_{int.0}$ score matrix whose columns contain $k_{int.0}$ PC scores.

$T_{node.int}$: a $n \times k_{int}$ -dimensional internal variable score matrix.

q_{ext} : the $p \times 1$-dimensional PCDA loading vector of manual data X_{train} .

Q_{ort} : a $cn \times cn$ projection matrix constructing $\tilde{X}_{all.nodes}$.

$Q_{int.0}$: a $p \times k_{int.0}$ loading matrix, the columns of which are the first $k_{int.0}$ principal component loadings of $\tilde{X}_{all.nodes}$.

Q_{int} : a $k_{int.0} \times k_{int}$ loading matrix, the columns of which are the first k_{int} PC loadings derived from a PCA of $T_{node.int.0}$.

$m_n^{''}, m_c^{''}$: k -dimensional means derived from manual data for use with priors on normal and metastatic groups of scanning data.

$V_n^{''}, V_c^{''}$: $k \times k$ variances derived from manual data for later use with priors on normal and metastatic groups of scanning data.

v_{1st} : the degree of freedom of multivariate t distribution in the 1st step EM algorithm.

v_{2nd} : the degree of freedom of multivariate t distribution in the 2nd step EM algorithm.

y_i : a multi-dimensional variable with reduced dimensions at pixel *i*.

π_j : the proportion of the *j*th component in the mixture.

μ_j : the mean vector of *j*th component in the mixture of multivariate *t* distribution or multivariate Gaussian distribution.

$\sum_j$: the covariance matrix of *j*th component in the mixture of multivariate *t* distribution or multivariate Gaussian distribution.

μ_p : the mean of the normal priors (of normal inverse Wishart priors) on the mean vectors μ_j conditional on covariance matrices $\sum_j$ of multivariate *t* distribution or multivariate Gaussian distribution.

Λ_p : the scale of the inverse Wishart priors (of normal inverse Wishart priors) on the covariance matrices Σ_j of multivariate *t* distribution or multivariate Gaussian distribution.

κ_p : the shrinkage parameter of the normal inverse Wishart prior for multivariate *t* distribution or multivariate Gaussian distribution.

v_p : the degrees of freedom of the normal inverse Wishart prior for multivariate *t* distribution or multivariate Gaussian distribution.

β : a regularization parameter (smoothness parameter) used in spatial label prior expression.

α_{ij} : a position parameter used in spatial label prior expression.

d_i : a distance score measuring the distance between each pixel and the central point of the image.

ω_i : a background score of each pixel i emphasizing the scores for the pixels on the edge.

ρ : a power parameter used in background scores and spatial label prior expression.

BIBLIOGRAPHY

Alfò, M., L. Nieddu, and D. Vicari (2008). A finite mixture model for image segmentation. *Statistics and Computing 18* (2), 137-150.

Andrew, A. and T. Fearn (2004). Transfer by orthogonal projection: making near-infrared calibrations robust to between-instrument variation. *Chemom. Intell. Lab. Syst. 72,* 51-56.

Backman, V., M. B. Wallace, L. T. Perelman, J. T. Arendt, R. Gurjar, M. G. Muller, Q. Zhang, G. Zonios, E. Kline, J. A. McGilligan, S. Shapshay, T. Valdez, K. Badizadegan, J. M. Crawford, M. Fitzmaurice, S. Kabani, H. S. Levin, M. Seiler, R. R. Dasari, I. Itzkan, J. Van Dam, and M. S. Feld (2000). Detection of preinvasive cancer cells. *Nature 406* (6791), 35-36.

Barker, M. and W. Rayens (2003). Partial least squares for discrimination. *J. Chemometrics 17* (3), 166–173.

Barnes, R. J., M. S. Dhanoa, and S. J. Lister (1989). Standard normal variate tansformation and detrending of near infrared diffuse reflectance. *Appl. Spectrosc. 43*, 772-777.

Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian Theory.* Wiley, Chichester.

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *J. Roy. Stat. Soc. B 36* (2), 192-236.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician 24* (3), 179-195.

Besag, J. (1986). On the statistical analysis of dirty pictures, *J. Roy. Stat. Soc. B 48* (3), 259-302.

Bigio, I. J. and J. R. Mourant (1997). Ultraviolet and visible spectroscopies for tissue diagnostics: Fluorescence spectroscopy and elastic scattering spectroscopy. *Phys. Med. Biol. 42* (5), 803-81.

Bradley, A. P. (1997). The use of area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition 30* (7), 1145-1159.

Braga-Neto, U. M. and E. R. Dougherty (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics 20*, 374-380.

Celeux, G. and G. Govaert (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis 14*, 315-332.

Celeux, G., F. Forbes, and N. Peyrard (2003). EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern Recognition 36* (1), 131-144.

Chalmond, B. (1989). An iterative Gibbsian technique for reconstruction of m-ary images. *Pattern Recognition 22* (6), 747-761.

Chellappa, R. and A. Jain (Eds.) (1993). *Markov Random Fields: Theory and Applications*. Academic Press.

Dean, N., T. B. Murphy, and G. Downey (2006). Using unlabelled data to update classification rules with applications in food authenticity studies. *J. Roy. Stat. Soc. C 55* (1), 1-14.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood form incomplete data via EM algorithm. *J. Roy. Stat. Soc. B 39*, 1-38.

Dhar, A., K. S. Johnson, M. R. Novelli, S. G. Bown, I. J. Bigio, L. B. Lovat, and S. L. Bloom (2006). Elastic scattering spectroscopy for the diagnosis of colon lesions: initial results of a novel optical biopsy technique. *Gastrointestinal Endoscopy 63* (2), 257-261.

Drezek, R., A. Dunn, and R. Richards-Kortum (1999). Light scattering from cells: finite-difference time-domain simulations and goniometric measurements. *Applied Optics 38* (16), 3651-3661.

Dudoit, S., J. Fridlyand, and T. Speed (2002). Comparison of discrimination methods for the classification of tumors using expression data. *J. Amer. Stat. Assoc. 97* (457), 77-87.

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for data mining researchers. *Tech report HPL-2003-4.* HP Laboratories, Palo Alto, CA, USA.

Fearn, T. (2000). On orthogonal signal correction. *Chemom. Intell. Lab. Syst. 50*, 47-52.

Flach, P. (2004). The many faces of ROC analysis in machine learning. *Tutorial at ICML'04.*

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Stat. Assoc. 97* (458), 611-631.

Fraley, C. and A. E. Raftery (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification 24* (2), 155-181.

Geladi, P., D. McDougall, and H. Martens (1985). Linearisation and scatter correction for near infrared reflectance spectra of meat. *Appl. Spectrosc. 39*, 491-500.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian Data Analysis*. Chapman and Hall.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence 6* (6), 721-741.

Georgakoudi, I., B. C. Jacobson, J. Van Dam, V. Backman, M. B. Wallace, M. G. Muller, Q. Zhang, K. Badizadegan, D. Sun, G. A. Thomas, L. T. Perelman, and M. S. Feld (2001). Fluorescence, reflectance, and light-scattering spectroscopy for evaluating dysplasia in patients with Barrett's esophagus. *Gastroenterology 120* (7), 1620-1629.

Gonzalez-Correa, C. A., B. H. Brown, R. H. Smallwood, N. Kalia, C. J. Stoddard, T. J. Stephenson, S. J. Haggie, D. N. Slater, and K. D. Bardhan (2000). Assessing the conditions for in vivo electrical virtual biopsies in Barrett’s oesophagus. *Med. Biol. Eng. Comput. 38*, 373–376.

Gurjar, R. S., V. Backman, L. T. Perelman, I. Georgakoudi, K. Badizadegan, I. Itzkan, R. R. Dasari, and M. S. Feld (2001). Imaging human epithelial properties with polarized light-scattering spectroscopy. *Nat. Med. 7* (11), 1245-1248.

Hale, G. M. and M. R. Querry (1973). Optical constants of water in 200-nm to 200-μm wavelength region. *Applied Optics 12* (3), 555-563.

Hameeteman, W., G. N. Tytgat, H. J. Houthoff, and J. G. van den Tweel (1989). Barrett's esophagus: development of dysplasia and adenocarcinoma. *Gastroenterology 96* (5), 1249-1256.

Hammersley, J. M. and P. Clifford (1971). Markov field on finite graphs and lattices. unpublished.

Hanley, J. A. and B. J. McNeil (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology 143,* 29-36.

Hansen, P. W. (2001). Pre-processing method minimizing the need for reference analyses. *J. Chemometrics 15* (2),123-131.

Jain, R. K. (1988). Determinants of tumor blood flow: a review. *Cancer Res. 48,* 2641-2658.

Jöbsis, F. F. (1977). Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science 198* (4323),1264-1267.

Johnson, K. S., D. W. Chicken, D. C. Pickard, A. C. Lee, G. Briggs, M. Falzon, I. J. Bigio, M. R. Keshtgar, and S. G. Bown (2004). Elastic scattering spectroscopy for intra-operative determination of sentinel lymph node status in the breast. *Journal of Biomedical Optics 9* (6), 1122-1128.

Kaufman, L. and P. J. Rousseeuw (1990). *Finding groups in data*. John Wiley & Sons, New York, USA.

Lange, K. L., R. J. A. Little, and J. M. G. Taylor (1989). Robust statistical modeling using the *t* distribution. *J. Amer. Stat. Assoc. 84* (408), 881-896.

Levine, D. S., R. C. Haggitt, P. L. Blount, P. S. Rabinovitch, V. W. Rusch, and B. J. Reid (1993). An endoscopic biopsy protocol can differentiate high-grade dysplasia from early adenocarcinoma in Barrett's esophagus. *Gastroenterology 105* (1), 40-50.

Li, S.Z. (2001). *Markov Random Field Modeling in Image Analysis*. Tokyo: Springer.

Lin, T. I., J. C. Lee and H.F. Ni (2004). A Bayesian analysis of mixture modeling using the multivariate t distribution. *Statistics and Computing 14,* 119-130.

Linder, M. and R. Sundberg (1998). Second order calibration: Bilinear least squares regression and a simple alternative. *Chemom. Intell. Lab. Syst. 42,* 159-178.

Little, R. J. A. and D. B. Rubin (1987). *Statistical Analysis with Missing Data*. John Wiley and Sons, New York.

Lovat, L. B. and S. G. Bown (2004). Elastic scattering spectroscopy for detection of dysplasia in Barrett's esophagus. *Gastrointest. Endosc. Clin. N. Am. 14* (3), 507-517.

Lovat, L. B., K. Johnson, G. D. Mackenzie, B. R. Clark, M. R. Novelli, S. Davies, M. O'Donovan, C. Selvasekar, S. M. Thorpe, D. Pickard, R. Fitzgerald, T. Fearn, I. Bigio, and S. G. Bown (2006). Elastic scattering spectroscopy accurately detects high-grade dysplasia and cancer in Barrett's oesophagus. *Gut 55,* 1078-1083.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979). *Multivariate Analysis*. Academic Press, London, UK.

Martens, H. and T. Næs (1998). *Multivariate Calibration*. Wiley, Chichester.

McLachlan, G. L. and K. E. Basford (1988). *Mixture Models: Inference and Applications to Clustering.* New York: Marcel Dekker.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, Chichester, UK.

McLachlan, G. J. and T. Krishnan (1997). *The EM Algorithm and Extensions*. Wiley, New York, NY.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. Wiley, New York, NY.

Messmann, H., R. Knuchel, W. Baumler, A. Holstege, and J. Scholmerich (1999). Endoscopic fluorescence detection of dysplasia in patients with Barrett's esophagus, ulcerative colitis, or adenomatous polyps after 5-aminolevulinic acid-induced protoporphyrin IX sensitization. *Gastrointestinal Endoscopy 49* (1), 97-101.

Molinaro, A. M., R. Simon, and R. M. Pfeiffer (2005). Prediction error estimation: A comparison of resampling methods. *Bioinformatics 21* (15), 3301-3307.

Mourant, J. R., T. Fuselier, J. Boyer, T. M. Johnson, and I. J. Bigio (1997). Predictions and measurements of scattering and absorption over broad wavelength ranges in tissue phantoms. *Applied Optics 36* (4), 949-957.

Mourant, J. R., A. H. Hielscher, A. A. Eick, T. M. Johnson, and J. P. Freyer (1998). Evidence of intrinsic differences in the light scattering properties of tumorigenic and nontumorigenic cells. *Cancer Cytopathol 84* (6), 366-374.

Mourant, J. R., M. Canpolat, C. Brocker, O. Esponda-Ramos, T. M. Johnson, A. Matanock, K. Stetter, and J.P. Freyer (2000). Light scattering from cells: the contribution of the nucleus and the effects of proliferative status. *J. Biomed. Opt. 5* (2), 131-137.

Næs, T., T. Isaksson, T. Fearn, and T. Davies (2002). *A User-friendly Guide to Multivariate Calibration and Classification*. NIR publications, Chichester, UK.

Newnham, A., M. J. Quinn, P. Babb, J. Y. Kang, and A. Majeed (2003). Trends in the subsite and morphology of oesophageal and gastric cancer in England and Wales 1971-1998. *Aliment Pharmacol. Ther. 17* (5), 665-676.

Peel, D. and G. J. McLachlan (2000). Robust mixture modelling using the *t* distribution, *Statistics and Computing 10* (4), 339-348.

Phillips, R. W. and R. K. Wong (1991). Barrett's esophagus. Natural history, incidence, etiology, and complications. *Gastroenterol. Clin. N. Am. 20* (4), 791-816.

Polkowski, W., J. P. Baak, J. J. van Lanschot, G. A. Meijer, L. T. Schuurmans, F. J. ten Kate, H. Obertop, and G. J. Offerhaus (1998). Clinical decision making in Barrett's oesophagus can be supported by computerized immunoquantitation and morphometry of features associated with proliferation and differentiation. *J. Pathol. 184* (2), 161-168.

Qian, W. and D.M. Titterington, Estimation of parameters in hidden Markov models (1991). *Philos. Trans. Roy. Soc. Lond. Ser. A 337*, 407-428.

Rabinovitch, P. S., G. Longton, P. L. Blount, D. S. Levine, and B. J. Reid (2001) Predictors of progression in Barrett's esophagus III: baseline flow cytometric variables. *Am. J. Gastroenterol. 96* (11), 3071-3083.

Raftery, A. E. (Eds.) (1996). Hypothesis testing and model selection via posterior simulation. In: W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in practice*,163-188. Chapman & Hall, London.

Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Stat. Soc. B 59*, 731-792.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, UK.

Roger, J. M., F. Chauchard, and V. Belon-Maurel (2003). EPO-PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits. *Chemom. Intell. Lab. Syst. 66*, 191-204.

Sandick, J. W. van, J. J. van Lanschot, B. W. Kuiken, G. N. Tytgat, G. J. Offerhaus, and H. Obertop (1998). Impact of endoscopic biopsy surveillance of Barrett's oesophagus on pathological stage and clinical outcome of Barrett's carcinoma. *Gut 43* (2), 216-222.

Savitzky, A. and M. J. E. Golay (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry 36*, 1627-1639.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.

Sharma, V. K., K. K. Wang, B. F. Overholt, C. J. Lightdale, M. B. Fennerty, P. J. Dean, D. K. Pleskow, R. Chuttani, A. Reymunde, N. Santiago, K. J. Chang, M. B. Kimmey, and D. E. Fleischer (2007). Balloon-based, circumferential, endoscopic radiofrequency ablation of Barrett's esophagus: 1-year follow-up of 100 patients. *Gastrointest. Endosc.65* (2), 185-195.

Sjöblom, J., O. Svensson, M. Josefson, H. Kullberg, and S. Wold (1998). An evaluation of orthogonal signal correction applied to the calibration transfer of near-infrared spectra. *Chemom. Intell. Lab. Syst. 44* (1), 229-244.

Stephens, M. A. (1997). *Bayesian Method for Mixtures of Mormal Distributions*. Ph.D. thesis, University of Oxford.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *J. Roy. Stat. Soc. B 36* (2), 111-147.

Sun, J. (1997). Statistical analysis of NIR data: data pre-treatment. *J. Chemometrics 11*, 525-532.

Swenson, K. K., M. J. Nissen, C. Ceronsky, L. Swenson, M. W. Lee, and T. M. Tuttle (2002). Comparison of side effects between sentinel lymph node and axillary lymph node dissection for breast cancer. *Ann. Surg. Oncol. 9* (8), 745–753.

Swets, J. and R. Pickets (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory.* Academic Press.

Swets, J. (1988). Measuring the accuracy of diagnostics systems. *Science 240* (4857), 1285-1293.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, Chichester, UK.

Toms, J. R. (Eds.) (2004). *CancerStats Monograph 2004*. Cancer Research UK, London.

Trygg, J. and S. Wold (2002). Orthogonal projections to latent structures (O-PLS). *J. Chemometrics 16* (3), 119-128.

Turner, R. R., D. W. Ollila, D. L. Krasne, and A. E. Giuliano (1997). Histopathologic validation of the sentinel lymph node hypothesis for breast carcinoma. *Ann. Surg. 226* (3), 271–276.

van den Berg, T. J. T. P. and H. Spekreijse (1997). Near infrared light absorption in the human eye media. *Vision Research 37* (2), 249-253.

Veronesi, U., G. Paganelli, G. Viale, A. Luini, S. Zurrida, V. Galimberti, M. Intra, P. Veronesi, C. Robertson, P. Maisonneuve, G. Renne, C. De Cicco, F. De Lucia, and R. Gennari (2003). A randomized comparison of sentinel-node biopsy with routine axillary dissection in breast cancer. *N. Engl. J. Med. 349* (6), 546–553.

Wallace, M. B., L. T. Perelman, V. Backman, J. M. Crawford, M. Fitzmaurice, M. Seiler, K. Badizadegan, S. J. Shields, I. Itzkan, R. R. Dasari, J. Van Dam, and M. S. Feld (2000). Endoscopic detection of dysplasia in patients with Barrett's esophagus using light-scattering spectroscopy. *Gastroenterology 119* (3), 677-682.

Webb, C. E. and J. D. C. Jones (Eds.) (2004). *Handbook of Laser Technology and Applications Vol.III: Applications.* Bristol : Institute of Physics Publishing Ltd.

Westerhuis, J. A., S. de Jong, and A. K. Smilde (2001). Direct orthogonal signal correction. *Chemom. Intell. Lab. Syst. 56*, 13-25.

Williamson, W. A., F. H. Jr. Ellis, S. P. Gibb, D. M. Shahian, H. T. Aretz, G. J. Heatley, and E. Jr. Watkins (1991). Barrett's esophagus. Prevalence and incidence of adenocarcinoma. *Arch. Intern. Med. 151* (11), 2212-2216.

Witten, I. H. and E. Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, second edition.

Wold, S. and M. Sjöström (1977). SIMCA: A method for analyzing chemical data in terms of similarity and analogy, in: Kowalski B. R. (Ed.), *Chemometrics: Theory and Applications. Am. Chem. Soc. Symp. Ser. 52*, 243-282.

Wold, S., H. Antii, F. Lindgren, and J. Öhman (1998). Orthogonal signal correction of near-infrared spectra. *Chemom. Intell. Lab.Syst. 44* (1), 175-185.

Zhang, Y., M. Brady, and S. Smith (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Medical Imaging 20* (1), 45-57.

Zhu, Y., T. Fearn, D. Samuel, A. Dhar, O. Hameed, S. G. Bown, and L. B. Lovat (2008). Error removal by orthogonal subtraction (EROS): a customised pre-treatment for spectroscopic data. *J. Chemometrics 22* (2), 130-134.

Zijlstra, W. G., A. Buursma, and O. W. van Assendelft (2000). *Visible and Near Infrared Absorption Spectra of Human and Animal Haemoglobin: Determination and Application*. VSP International Science Publishers, Netherlands.

Zonios, G., L. T. Perelman, V. M. Backman, R. Manoharan, M. Fitzmaurice, J. Van Dam, and M. S. Feld (1999). Diffuse reflectance spectroscopy of human adenomatous colon polyps in vivo. *Applied Optics 38* (31), 6628-6637.