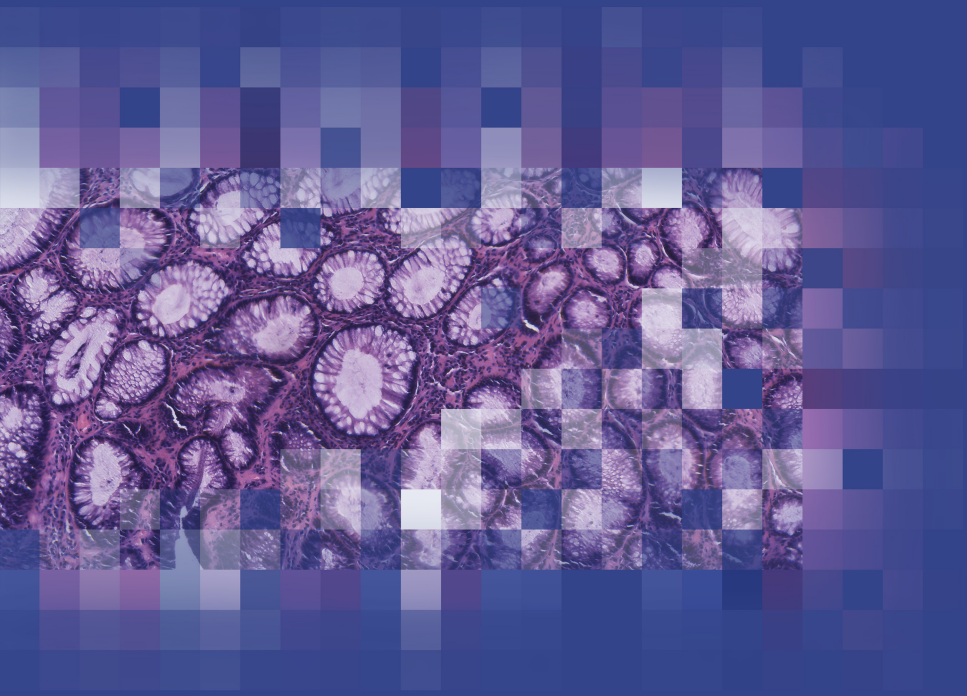
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Application of artificial vision algorithms to images of microscopy and spectroscopy for the improvement of cancer diagnosis



Francisco José Peñaranda Gómez

Supervisor:
Valery Naranjo Ornedo

PhD Thesis
February, 2018

Departamento de Ingeniería Electrónica

Contents

Acknowledgements

To all the partners and friends of the UPV who have given me the opportunity to learn together during all these years.

To all the partners of the MINERVA project who have being patient and have taught me to be practical.

To Valery, for believing in me.

To all my family, especially to my mother Josefa, for their unconditional support.

To Clara and Tambor, for their love.

Abstract

The final diagnosis of most types of cancers is performed by an expert clinician in anatomical pathology who examines suspicious tissue or cell samples extracted from the patient. Currently, this assessment largely relies on the experience of the clinician and is accomplished in a qualitative manner by means of traditional imaging techniques, such as optical microscopy. This tedious task is subject to high degrees of subjectivity and gives rise to suboptimal levels of discordance between different pathologists, especially in early stages of cancer development.

Fourier Transform infrared (FTIR) spectroscopy is a technology widely used in industry that has recently shown an increasing capability to improve the diagnosis of different types of cancer. This technique takes advantage of the ability of mid-infrared light to excite the vibrational modes of the chemical bonds that form the biological samples. The main generated signal consists of an absorption spectrum that informs of the chemical composition of the illuminated specimen. Modern FTIR microspectrometers, composed of complex optical components and high-sensitive array detectors, allow the acquisition of high-quality hyperspectral images with spatially-resolved chemical information in a common research laboratory. FTIR images are information-rich data structures that can be analysed alone or together with other imaging modalities to provide objective pathological diagnoses. Hence, this emerging imaging technique presents a high potential to improve the detection and risk stratification in cancer screening and surveillance.

This thesis studies and implements different methodologies and algorithms from the related fields of *image processing*, *computer vision*, *machine learning*, *pattern recognition*, *multivariate analysis* and *chemometrics* for the processing and analysis of FTIR hyperspectral images. Those images were acquired with a modern benchtop FTIR microspectrometer from tissue and cell samples affected by colorectal and skin cancer, which were prepared by following protocols close to the current clinical practise. The most relevant concepts of FTIR spectroscopy are thoroughly investigated, which ought to be understood and considered to perform a correct interpretation and treatment of its special signals. In particular, different physicochemical factors are reviewed and analysed, which influence the spectroscopic measurements for the particular case of biological samples and can critically affect their later analysis.

All these knowledge and preliminary studies come into play in two main applications. The first application tackles the problem of registration or alignment of FTIR hyperspectral images with colour images acquired with traditional microscopes. The aim is to fuse the spatial information of distinct tissue samples measured by those two imaging modalities and focus the discrimination on regions selected by the pathologists, which are meant to be the most relevant areas for the diagnosis of colorectal cancer. In the second application, FTIR spectroscopy is pushed to their limits of detection for the study of the smallest biomedical entities. The aim is to assess the capabilities of FTIR signals to reliably discriminate different types of skin cells containing malignant phenotypes. The developed studies contribute to the improvement of objective decision methods to support the pathologist in the final diagnosis of cancer. In addition, they reveal the limitations of current protocols and intrinsic problems of modern FTIR technology, which should be tackled in order to enable its transference to anatomical pathology laboratories in the future.

Resumen

El diagnóstico final de la mayoría de tipos de cáncer lo realiza un médico experto en anatomía patológica que examina muestras tisulares o celulares sospechosas extraídas del paciente. Actualmente, esta evaluación depende en gran medida de la experiencia del médico y se lleva a cabo de forma cualitativa mediante técnicas de imagen tradicionales como la microscopía óptica. Esta tarea tediosa está sujeta a altos grados de subjetividad y da lugar a niveles de discordancia inadecuados entre diferentes patólogos, especialmente en las primeras etapas de desarrollo del cáncer.

La espectroscopía infrarroja por Transformada de Fourier (siglas FTIR en inglés) es una tecnología ampliamente utilizada en la industria que recientemente ha demostrado una capacidad creciente para mejorar el diagnóstico de diferentes tipos de cáncer. Esta técnica aprovecha las propiedades del infrarrojo medio para excitar los modos vibratorios de los enlaces químicos que forman las muestras biológicas. La principal señal generada consiste en un espectro de absorción que informa sobre la composición química de la muestra iluminada. Los microespectrómetros FTIR modernos, compuestos por complejos componentes ópticos y detectores matriciales de alta sensibilidad, permiten capturar en un laboratorio de investigación común imágenes hiperespectrales de alta calidad que aúnan información química y espacial. Las imágenes FTIR son estructuras de datos ricas en información que se pueden analizar individualmente o junto con otras modalidades de imagen para realizar diagnósticos patológicos objetivos. Por lo tanto, esta técnica de imagen emergente alberga un alto potencial para mejorar la detección y la evaluación del riesgo del paciente en el cribado y vigilancia de cáncer.

Esta tesis estudia e implementa diferentes metodologías y algoritmos de los campos interrelacionados de *procesamiento de imagen*, *visión por ordenador*, *aprendizaje automático*, *reconocimiento de patrones*, *análisis multivariante* y *quimiometría* para el procesamiento y análisis de imágenes hiperespectrales FTIR. Estas imágenes se capturaron con un moderno microscopio FTIR de laboratorio a partir de muestras de tejidos y células afectadas por cáncer colorrectal y de piel, las cuales se prepararon siguiendo protocolos alineados con la práctica clínica actual. Los conceptos más relevantes de la espectroscopía FTIR se investigan profundamente, ya que deben ser comprendidos y tenidos en cuenta para llevar a cabo una correcta interpretación y tratamiento de sus señales especiales. En particular, se revisan y analizan diferentes factores fisicoquímicos que influyen en las mediciones espectroscópicas en el caso particular de muestras biológicas y pueden afectar críticamente su análisis posterior.

Todos estos conceptos y estudios preliminares entran en juego en dos aplicaciones principales. La primera aplicación aborda el problema del registro o alineación de imágenes hiperespectrales FTIR con imágenes en color adquiridas con microscopios tradicionales. El objetivo es fusionar la información espacial de distintas muestras de tejido medidas con esas dos modalidades de imagen y centrar la discriminación en las regiones seleccionadas por los patólogos, las cuales se consideran más relevantes para el diagnóstico de cáncer colorrectal. En la segunda aplicación, la espectroscopía FTIR se lleva a sus límites de detección para el estudio de las entidades biomédicas más pequeñas. El objetivo es evaluar las capacidades de las señales FTIR para discriminar de manera fiable diferentes tipos de células de piel que contienen fenotipos malignos. Los estudios desarrollados contribuyen a la mejora de métodos de decisión objetivos que ayuden al patólogo en el diagnóstico final del cáncer. Además, revelan las limitaciones de los protocolos actuales y los problemas intrínsecos de la tecnología FTIR moderna, que deberían abordarse para permitir su transferencia a los laboratorios de anatomía patológica en el futuro.

Resum

El diagnòstic final de la majoria de tipus de càncer ho realitza un metge expert en anatomia patològica que examina mostres tissulars o cel·lulars sospitoses extretes del pacient. Actualment, aquesta avaluació depèn en gran part de l'experiència del metge i es porta a terme de forma qualitativa mitjançant tècniques d'imatge tradicionals com la microscòpia òptica. Aquesta tasca tediosa està subjecta a alts graus de subjectivitat i dóna lloc a nivells de discordança inadequats entre diferents patòlegs, especialment en les primeres etapes de desenvolupament del càncer.

L'espectroscòpia infraroja per Transformada de Fourier (sigles FTIR en anglès) és una tecnologia àmpliament utilitzada en la indústria que recentment ha demostrat una capacitat creixent per millorar el diagnòstic de diferents tipus de càncer. Aquesta tècnica aprofita les propietats de l'infraroig mitjà per excitar els modes vibratoris dels enllaços químics que formen les mostres biològiques. El principal senyal generat consisteix en un espectre d'absorció que informa sobre la composició química de la mostra il·luminada. Els microespectrómetres FTIR moderns, compostos per complexos components òptics i detectors matricials d'alta sensibilitat, permeten capturar en un laboratori d'investigació comú imatges hiperespectrals d'alta qualitat que uneixen informació química i espacial. Les imatges FTIR són estructures de dades riques en informació que es poden analitzar individualment o juntament amb altres modalitats d'imatge per a realitzar diagnòstics patològics objectius. Per tant, aquesta tècnica d'imatge emergent té un alt potencial per a millorar la detecció i la avaluació del risc del pacient en el cribratge i vigilància de càncer.

Aquesta tesi estudia i implementa diferents metodologies i algoritmes dels camps interrelacionats de *processament d'imatge*, *visió per ordinador*, *aprenentatge automàtic*, *reconeixement de patrons*, *anàlisi multivariant* i *quimiometria* per al processament i anàlisi d'imatges hiperespectrals FTIR. Aquestes imatges es van capturar amb un modern microscopi FTIR de laboratori a partir de mostres de teixits i cèl·lules afectades per càncer colorectal i de pell, les quals es van preparar seguint protocols alineats amb la pràctica clínica actual. Els conceptes més rellevants de l'espectroscòpia FTIR s'investiguen profundament, ja que han de ser compresos i tinguts en compte per dur a terme una correcta interpretació i tractament dels seus senyals especials. En particular, es revisen i analitzen diferents factors fisicoquímics que influeixen en els mesuraments espectroscòpiques en el cas particular de mostres biològiques i poden afectar críticament la seua anàlisi posterior.

Tots aquests conceptes i estudis preliminars entren en joc en dues aplicacions principals. La primera aplicació aborda el problema del registre o alineació d'imatges hiperespectrals FTIR amb imatges en color adquirides amb microscopis tradicionals. L'objectiu és fusionar la informació espacial de diferents mostres de teixit mesurades amb aquestes dues modalitats d'imatge i centrar la discriminació en les regions seleccionades pels patòlegs, les quals es consideren més rellevants per al diagnòstic de càncer colorectal. En la segona aplicació, l'espectroscòpia FTIR es porta als seus límits de detecció per a l'estudi de les entitats biomèdiques més xicotetes. L'objectiu és avaluar les capacitats dels senyals FTIR per discriminar de manera fiable diferents tipus de cèl·lules de pell que contenen fenotips malignes. Els estudis desenvolupats contribueixen a la millora de mètodes de decisió objectius que ajuden al patòleg en el diagnòstic final del càncer. A més, revelen les limitacions dels protocols actuals i els problemes intrínsecs de la tecnologia FTIR moderna, que haurien d'abordar per permetre la seva transferència als laboratoris d'anatomia patològica en el futur.

Derived publications

Journals

- **Francisco Peñaranda**, Valery Naranjo, Rafael Verdú-Monedero, Gavin R. Lloyd, Jayakrupakar Nallala, Nicholas Stone. (2017) *Multimodal registration of optical microscopic and infrared spectroscopic images from different tissue sections: an application to colon cancer.* Digital Signal Processing, Volume 68, Pages 1-15, ISSN 1051-2004, DOI:10.1016/j.dsp.2017.04.014.
- **Francisco Peñaranda**, Valery Naranjo, Gavin R. Lloyd, Lena Kastl, Björn Kemper, Jürgen Schnekenburger, Jayakrupakar Nallala, Nicholas Stone. (2018) *Discrimination of skin cancer cells using Fourier transform infrared spectroscopy.* (In preparation).

Book chapters

- Valery Naranjo, **Francisco Peñaranda**, Mariano Alcañiz, Bruce Napier, Mark Farries, Gary Stevens, John Ward, Cestmir Barta, Radek Hasal, Angela Seddon, Slawomir Sujecki, Samir Lamrini, Uffe Møller, Ole Bang, Peter M. Moselund, Munir Abdalla, Danny De Gaspari, Rosa M. Vinella, Hedda Malm, Gavin R. Lloyd, Nick Stone, Jayakrupakar Nallala, Juergen Schnekenburger, Lena Kastl, and Björn Kemper. (2015). *MINERVA Project, mid- To near Infrared Spectroscopy for Improved Medical Diagnostics.* In European Project Space on Intelligent Systems, Pattern Recognition and Biomedical Systems - EPS Lisbon, ISBN 978-989-758-095-6, pages 53-69. DOI: 10.5220/0006162400530069

Conferences

- **Francisco Peñaranda**, Fernando López-Mir, Valery Naranjo, Jesús Angulo, Lena Kastl and Juergen Schnekenburger. (2015) *New Spectral Representation and Dissimilarity Measures Assessment for FTIR-spectra using Unsupervised Classification.* In Proceedings of the International Conference on Bio-inspired Systems and Signal Processing - Volume 1: BIOSIGNALS, (BIOSTEC 2015) ISBN 978-989-758-069-7, pages 172-177. DOI: 10.5220/0005188001720177.

- **Francisco Peñaranda**, Valery Naranjo, Rafael Verdú, Gavin R. Lloyd, Jayakrupakar Nallala, Nick Stone. (2016) *A two-step framework for the registration of HE stained and FTIR images.* Proc. SPIE 9703, Optical Biopsy XIV: Toward Real-Time Spectroscopic Imaging and Diagnosis, 970305. DOI:10.1117/12.2208869.

- Lena Kastl, Björn Kemper, Gavin R Lloyd, Jayakrupakar Nallala, Nick Stone, Valery Naranjo, **Francisco Peñaranda**, Jürgen Schnekenburger. (2016) *Potential of mid IR spectroscopy in the rapid label free identification of skin malignancies.* Proc. SPIE 9703, Optical Biopsy XIV: Toward Real-Time Spectroscopic Imaging and Diagnosis, 970307. DOI: 10.1117/12.2208897.

- **Francisco Peñaranda**, Valery Naranjo, Lena Kastl, Björn Kemper, Gavin R. Lloyd, Jayakrupakar Nallala, Nicholas Stone, Jürgen Schnekenburger. (2016) *Multivariate Classification of Fourier Transform Infrared Hyperspectral Images of Skin Cancer Cells.* In Proceedings of the 24th European Signal Processing Conference (EUSIPCO 2016). ISBN 978-0-9928-6266-4, p. 1328-1332.

- **Francisco Peñaranda**, Valery Naranjo. *Diagnóstico citopatológico de cáncer de piel mediante análisis de imágenes hiperespectrales infrarrojas.* XXXIV Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2016), 2016. p. 130-133.

- Lena Kastl, Björn Kemper, Gavin R Lloyd, Jayakrupakar Nallala, Nick Stone, Valery Naranjo, **Francisco Peñaranda**, Jürgen Schnekenburger. (2017) *Performance of mid infrared spectroscopy in skin cancer cell type identification.* Proc. SPIE 10060, Optical Biopsy XV: Toward Real-Time Spectroscopic Imaging and Diagnosis, 10060. DOI: 10.1117/12.2253314.

List of figures

List of tables

Acronyms

AJCC	American Joint Committee on Cancer
AO	Acousto-Optic
ATR	Attenuated Total Reflection
CCD	Charge-Coupled Device
CT	Computed Tomography
CV	Cross-Validation
DA	Discriminant Analysis
DNA	Deoxyribonucleic Acid
DoG	Difference-of-Gaussian
DTGS	Deuterated Triglycine Sulfate
EMSC	Extended Multiplicative Signal Correction
FOV	Field of View
FPA	Focal Plane Array
FT	Fourier Transform
FTIR	Fourier Transform Infrared
GHFT	Gloucestershire Hospitals NHS Foundation Trust
H&E	Hematoxylin and Eosin
IARC	International Agency for Research on Cancer
INT	Intermediate
IQR	Interquartile Range
IR	Infrared

ISO	International Organization for Standardization
LDA	Linear Discriminant Analysis
MCT	Mercury Cadmium Telluride
MRI	Magnetic Resonance Imaging
NOR	Normal
PBS	Phosphate Buffered Saline
PC	Principal Component
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PLS	Partial Least Squares
PMMA	Poly(methyl methacrylate)
QDA	Quadratic Discriminant Analysis
RANSAC	Random Sample Consensus
RGB	Red-Green-Blue
RMieS	Resonant Mie Scattering
RMSE	Root-Mean-Square Error
ROI	Region of Interest
SCP	Spectral Cytopathology
SEER	Surveillance, Epidemiology, and End Results
SG	Savitzky-Golay
SI	International System of Units
SIFT	Scale Invariant Feature Transform
SNR	Signal-to-Noise Ratio
SNV	Standard Normal Variate
TMA	Tissue Microarray
TUM	Tumoral
UICC	Union for International Cancer Control
UK	United Kingdom
UoE	University of Exeter
UPV	Universitat Politècnica de València

US	United States
UV	Ultraviolet
WHO	World Health Organization
WWU	Westfälische Wilhelms-Universität

Symbols

Δ	Difference
Ω	Set of pixels
α	Weight of regulariser energy term
$\varepsilon(\tilde{\nu})$	Molar attenuation coefficient or absorptivity
η	Iteration index
η_{max}	Maximum number of iterations
λ	Wavelength
$\tilde{\nu}$	Wavenumber
$\boldsymbol{\omega}$	Two-dimensional variable in the frequency domain
$\rho(\tilde{\nu})$	Size parameter
θ	Angle of rotation
ξ	Displacement of moving mirror in interferometer
$A(\tilde{\nu})$	Absorbance or absorption spectrum
a	Offset parameter (EMSC)
$A_{app}(\tilde{\nu})$	Apparent absorbance
$A_{chem}(\tilde{\nu})$	Idealised chemical contributions to the absorbance spectrum
$A_{corr}(\tilde{\nu})$	Corrected absorbance spectrum
A_i	Absorbance value
a_m	Scattering coefficient in full Mie theory
$\overline{A}$	Mean absorbance value
a_n	Approximation of the average real refractive index (RMieS-EMSC)

$A_{phys}(\tilde{\nu})$	Physical contribution to the absorbance spectrum
$A_{ref}(\tilde{\nu})$	Reference absorbance spectrum (EMSC)
$\mathbf{A}$	Absorbance spectrum (vector)
b	Multiplicative scaling parameter (EMSC)
$\mathcal{B}$	Blue component of RGB color space
BA	Balanced Accuracy
$\mathbf{b}$	Regression coefficient vector (PLS-DA)
B_λ	Blackbody spectral radiance
b_m	Scattering coefficient in full Mie theory
b_n	Amplification factor for n_{KK} (RMieS-EMSC)
C	Concentration
c	Speed of light
CR	Correlation ratio
d	Spatial resolution
$\mathcal{D}$	Energy term to measure distance between images
d_{Mah}	Mahalanobis distance
d_n	Coefficients of the polynomial baseline (EMSC)
E	Energy
$e(\tilde{\nu})$	Unmodelled residual absorbance (EMSC)
$\mathbf{E}$	Matrix of residuals (PCA)
f	Frequency; Coefficient of Mie scattering contribution (EMSC)
$\mathbf{f}$	Residual of response vector (PLS)
FN	False Negatives
$\tilde{\mathbf{f}}(\boldsymbol{\omega})$	Fourier transform of the external forces field
FP	False Positives
FRE	Fiducial registration error
$\mathrm{FRE_{FB}}$	Fiducial registration error of feature-based registration
$\mathrm{FRE_{PA}}$	Fiducial registration error of Procrustes analysis
G_0	Quiescent stage in cell cycle
G_1	Gap 1 phase in cell cycle

G_2	Gap 2 phase in cell cycle
$\mathcal{G}$	Green component of RGB colour space
g_i	Coefficients of metamodel of Mie scattering (EMSC)
h	Planck's constant
$H(\boldsymbol{\omega})$	Low-pass filter in the frequency domain
H_R	Entropy of image R
H_{RT}	Joint entropy of images R and T
H_T	Entropy of image T
$I(\tilde{\nu})$	Transmitted intensity spectrum
$I_0(\tilde{\nu})$	Reference background spectrum
IDcomb	Identifier of combination of SIFT parameters
IDsample	Identifier of sample
$I_s(\tilde{\nu})$	Intensity of stray light
$\mathcal{J}$	Energy functional
$J(\xi)$	Interferogram
K	Number of retained components
$k(\tilde{\nu})$	Absorption index or imaginary refractive index
k_B	Boltzmann's constant
ℓ	Effective optical path length
M	Mitosis phase in cell cycle
MI	Mutual information
n	Real refractive index; Width of descriptor (SIFT)
$\tilde{n}(\tilde{\nu})$	Complex refractive index
n_∞	Average real refractive index
NA	Numerical aperture
$\mathbb{N}$	Set of natural numbers
n_{KK}	Output of the first Kramers-Kronig transform
N_p	Number of pixels
N_w	Number of wavenumbers
$\mathcal{P}$	Cauchy principal value of improper integrals

$\mathbf{P}$	Matrix of loadings or autovectors (PCA)
P_i	Probability of intensity level i in R
p_i	Loadings or autovectors of principal components
P_{ij}	Joint probability of intensity levels i in R and j in T
P_j	Probability of intensity level j in T
$\mathbf{Q}$	Matrix of $Q_{ext}(\tilde{\nu})$ curves
$\mathbf{q}$	Loadings vector of the response vector (PLS)
$Q_{ext}(\tilde{\nu})$	Extinction cross section (Mie scattering)
R	Restriction point in cell cycle; Reference image in registration
r	Radius of the scattering sphere (RMieS-EMSC); Number of orientations of descriptor (SIFT)
R^2	Coefficient of determination
$\mathcal{R}$	Red component of RGB color space
$\mathbb{R}$	Set of real numbers
rms	Root-mean-square level
rod	Ratio of distances closest/next closest (SIFT)
S	Syntesis phase in cell cycle
s	Number of scale samples per octave (SIFT)
$\mathbf{S}$	Covariance matrix
$\mathcal{S}$	Regulariser energy term
Sn	Sensitivity
Sp	Specificity
std	Standard deviation
$T(\tilde{\nu})$	Transmittance spectrum
T_1	Original target image
T_2	Intermediate target image transformed by rigid transformation
T_3	Final target image transformed by non-rigid transformation
$\mathbf{T}$	Matrix of scores (PCA)
T_i	Deformed intermediate states of target image
T	Generic target image

T_k	Absolute temperature
TN	True Negatives
TP	True Positives
$\mathcal{T}$	Rigid spatial transformation
$\{\mathcal{T}(\mathbf{x}_i)\}$	Generic set of transformed points
$\{\mathcal{T}_{\mathrm{FB}}(\mathbf{x}_i)\}$	Set of points transformed by the feature-based registration step
$\{\mathcal{T}_{\mathrm{PA}}(\mathbf{x}_i)\}$	Set of points transformed by Procrustes analysis
t_x	Translation in horizontal direction
t_y	Translation in vertical direction
$\mathbf{u}$	Non-rigid displacement field
$\mathbf{w}$	Weight vector (PLS)
$\mathbf{X}$	Data or measurements matrix, independent variables
x	Horizontal spatial dimension or horizontal pixel position
$\hat{\mathbf{X}}$	Data or measurements matrix of new samples to predict
$\mathbf{x}$	Pixel or spatial point position (vector)
$\{\mathbf{x}_i\}$	Set of selected points in T_1
y	Vertical spatial dimension or vertical pixel position
$\mathcal{Y}$	Luma component of YIQ color space
$\mathbf{y}$	Response or labels vector, dependent variable
$\hat{\mathbf{y}}$	Predicted response or labels vector
$\{\mathbf{y}_i\}$	Set of selected points in R

Chapter 1

Introduction

Contents

1.1 Cancer

According to the World Health Organization (WHO), *cancer* is a generic term which includes a large group of pathologies that may affect any part of the body [1]. Cancer is a genetic disease caused by mistakes or *mutations* during Deoxyribonucleic Acid (DNA) replication. The likelihood of suffering these remaining mutations is mainly determined by environmental factors and lifestyle (around 90-95% of all cancer cases) and in a lower degree by inherited genetic defects (only 5-10% of all cancer cases) [2]. The key characteristic of cancer is the rapid growth of abnormal cells capable of dividing out of control and spreading beyond their natural boundaries. This invasion of other tissues is called *metastasis* and it is the first cause of death in cancer patients.

Tumour is the common word to refer to *neoplasms*, that is, abnormal growths of tissue with an increase in volume [3]. Tumours can be generally classified in two major types: *benign* and *malignant*. Benign tumours only grow in a local and restricted region without invading nearby tissues, meanwhile malignant tumours can spread to adjacent and distant tissues (*metastasis*). Despite these differences, the word tumour is commonly used with implicit malignant connotations. Most primary tumours in humans are benign and do not cause any harm, unless their expansion presses a vital organ and affects its correct operation. The majority of deaths connected with cancer are caused by malignant tumours in the more advanced metastatic state.

Tumours are created by malfunctioning cells, which are incapable of building tissues with normal morphology and functioning, disobeying the rules that control the correct construction and maintenance. Cancer cells only have one target: making more copies of themselves [4]. They can achieve this purpose by acquiring new capabilities (Sec. 1.1.2) as a result of accumulated genetic mutations, which are transferred to their descendants [5].

Some of these unsuitable alterations are the deactivation of *apoptosis* and the *invisibility* to immune system. Apoptosis is the programmed mechanism of cellular suicide by which all normal cells destroy themselves when they perform an incorrect biological functioning. When this mechanism fails the immune system is responsible for killing the defective cells. However, cancerous cells are able to *hide* from immune system through different processes or remain unaffected by their defence mechanisms [6].

1.1.1 Development of cancer

The development of a malignant tumour is a slow process comprising different steps where *proper* mutations get accumulated. In most cases the progression of a tumour requires several years, in the order of decades. That is why cancer is considered as a chronic disease more associated with old people.

The typical steps of cancer can be illustrated with the evolution of *carcinoma*. Carcinoma is a cancerous lesion that begins in the epithelial tissue and is the most common malignant tumour. Epithelial tissue is a sheet of cells that forms the inner and outer surfaces of the body, such as the external layers of the bowel or the skin. As illustrated in Fig. 1.1, the main stages in the development of carcinoma are [4, 7]:

1. *Genetically altered cell*: the tumour development starts when a normal cell undergoes a genetic change that increases its predisposition to proliferate.

2. *Hyperplasia*: is the result of an unconstrained division of one altered cell into a high number of cells in a limited region of tissue. The structure in these cells is not altered and they look normal although their size may be larger. Hyperplasia is a reversible process and is normally produced as a response to an irritant stimulus. After years in this stage, one cell in a million suffers an additional mutation that makes it grow more uncontrollably.

3. *Dysplasia*: literally means *disordered growth* and is the result of an excessive proliferation of cells in hyperplasia, which lose their normal structure and get disorganised due to additional genetic changes. In this stage, cells keep growing in an abnormal way. Just as hyperplasia, cells in dysplasia can recover their normal state but sometimes, after a time, a rare mutation occurs that changes the cell behaviour. Therefore, tissues in dysplasia must be specially controlled and, in some cases, receive treatment.

4. *In situ cancer*: this stage is reached when cells do not recover their normal state and accumulate more and more alterations. The tumour presents an irregular structure. In this state, the cells are not able to invade other tissues and remain confined in the primary location, maybe indefinitely. However, *in situ* cancers must be promptly resected or treated before they acquire invasive properties because of further mutations.

5. *Invasive cancer*: is the final stage of tumour formation. Cells become malignant and invade adjacent tissue and even spread to distant sites through the blood vessels or the lymphatic system. The spread cells can get attached to different tissues and create new tumours, producing *metastasis*. This stage normally becomes lethal because spread cells, which typically cannot be totally eradicated, finally invade and disrupt a vital organ.

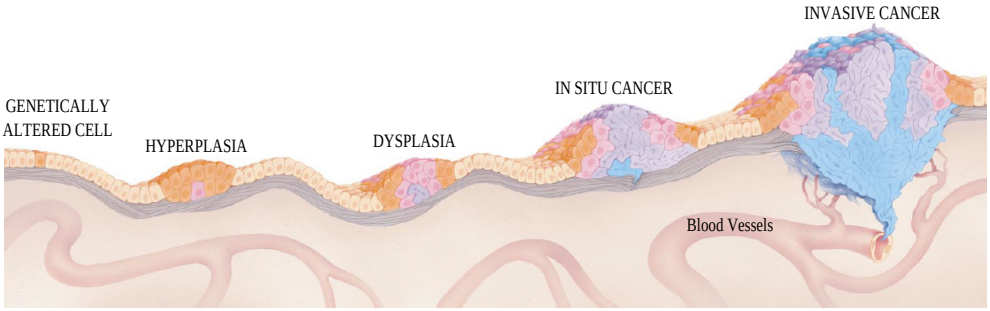


Fig. 1.1: Illustration of the development of a cancerous lesion in the epithelial tissue. Reproduced from [7] with permission by courtesy of Scientific American, Inc.

1.1.2 The characteristics of cancer

The cell cycle

Before explaining the main properties that distinguish cancer at the cellular level, it is interesting to describe the cell cycle. All living cells reproduce by dividing in two daughter cells following a coordinated cell cycle that tries to guarantee the conservation of the genome. As depicted in Fig. 1.2, in most human cells this cycle is composed of four main stages [8]:

1. *G_1 phase (Gap 1)*: the cell grows in volume by synthesising proteins and making copies of its subcellular components called *organelles*, but excluding the nucleus, which contains DNA.

2. *S phase (Synthesis)*: the *chromosomes*, the macromolecules of DNA that encodes the genetic information, are replicated into two identical entities called

sister chromatids that remain attached. This is the longest stage of the cycle. Once entered this phase of replication, cells must finish the complete cycle.

3. G_2 *phase (Gap 2)*: the cell growth continues by making more proteins and organelles. The cellular content is also reorganised in preparation for the next step. In particular, the chromatids of chromosomes become more compact.

4. *M phase (Mitosis)*: the replicated chromosomes and the *cytoplasm* (the non-nuclear part of the cell) finally split to produce two identical new cells.

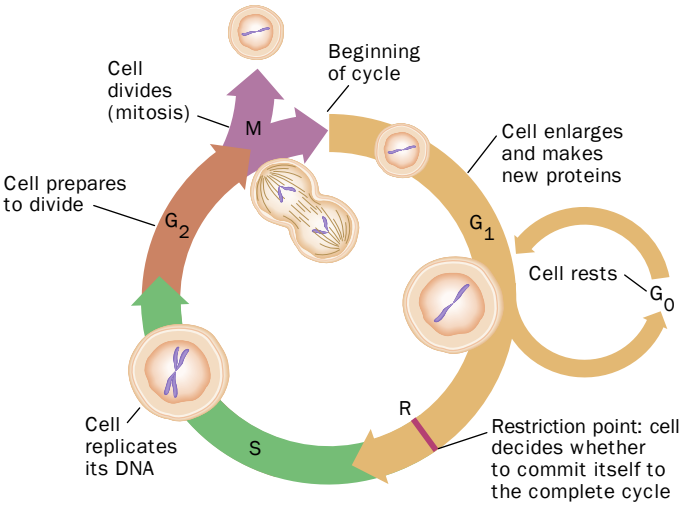


Fig. 1.2: Stages of the cell cycle. Reproduced from [7] with permission by courtesy of Scientific American, Inc.

The whole cycle is strictly regulated by specific *signals* both coming from outside and inside the cell, which ensure a correct cell division. For instance, at the end of G_1 phase there is a specific period called *restriction point* (R) [9] where cells must receive from the outside the so-called *growth factors* to trigger the DNA replication. In normal tissues, these growth factors are not supplied if, for example, a nutritional deficiency exists in the region. From this point, cells do not depend on extracellular stimuli to finish the cell cycle. If growth factors are not supplied, cells enter in a quiescent or still stage called G_0, in which they have an active metabolism but they do not keep on growing and synthesising more proteins. In addition, each phase of the cell cycle has specific *checkpoints* [10], which prevent entering into the next phase

if the preceding phase is not correctly completed. In particular, multiple checkpoints try to guarantee that incomplete or damaged chromosomes are not replicated. As detailed below, cancer cells can break the strict rules governing the cell cycle by complex and interrelated pathways that benefit their sustained proliferation [11–15].

The hallmarks of cancer

In the past, tumours were considered as accumulations of relatively homogeneous cancer cells whose biology only depended on autonomous properties. However, now it is widely recognised that tumours are not just isolated masses of proliferating cancer cells but heterogeneous mixtures of different cell types interacting with one another and constructing a *tumour microenvironment* of high complexity, as depicted in Fig. 1.3a.

As commented, cancer is a very complex and diverse disease but research discoveries in the last decades have revealed that this diversity may be reduced to a small number of principles. Hanahan and Weinberg [16, 17] suggested a series of hallmarks or distinctive characteristics (illustrated in Fig. 1.3b) that allow cells to survive, proliferate and disseminate and which differentiate cancer cells and their microenvironment from normal cells and tissues. In 2000 [17], they initially proposed six complementary hallmarks or functional capabilities:

- *Sustaining proliferative signaling*: cancer cells can develop different mechanisms to maintain a supply of *growth factors*, which allow them to circumvent the restriction point (Fig. 1.2) and favour their chronic proliferation. In some cases, they can produce these substances themselves or they can stimulate other normal cells to provide the growth factors to them.

- *Evading growth suppressors*: cancer cells can elude regulatory circuits of cellular proliferation from extracellular and intracellular sources. For instance, they abolish the signals generated by cell-to-cell contacts that prevent excessive propagation in dense populations of normal cells.

- *Resisting cell death*: cancer cells do not follow *apoptosis*, the programmed cell death induced by physiologic stresses and DNA damage associated with hyperproliferation. The most common way is inactivating the production of the

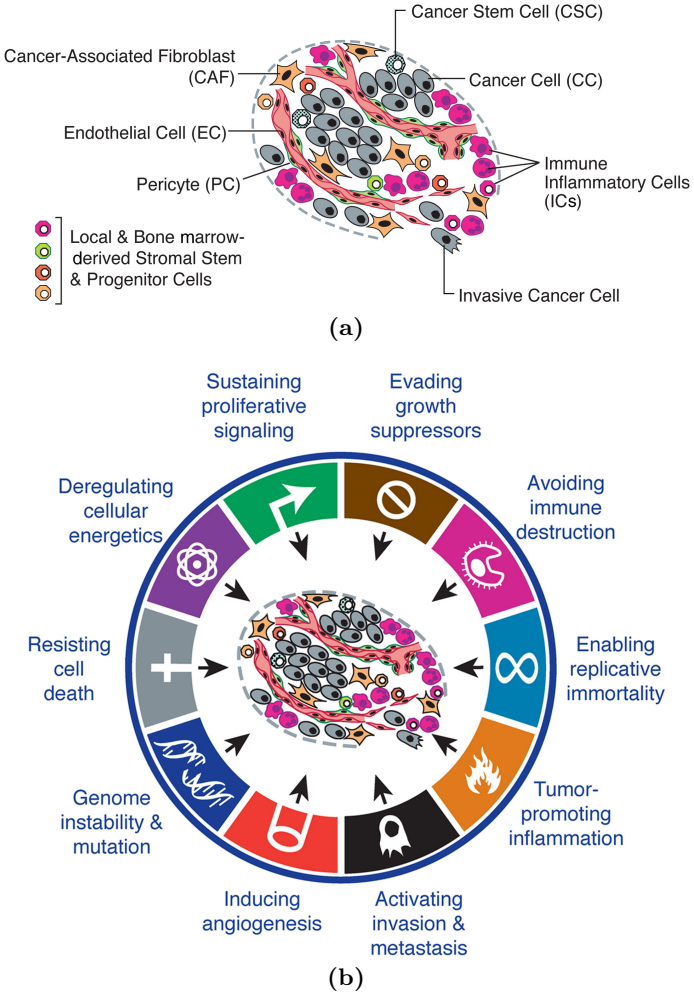


Fig. 1.3: The characteristics of cancer. (a) Illustration of a tumour microenvironment, which is composed by a complex mixture of different cell types. (b) Illustration of the hallmarks and enabling characteristics that distinguish cancer cells and their tumour microenvironment. Reproduced from [16] with permission by courtesy of Elsevier.

protein p53, nicknamed the *guardian of genome* [18], which triggers apoptosis when irreparable damage to the genome or other crucial subcellular subsystems are detected in the different stages of the cell cycle [19]. *Autophagy* is another mechanism that cancer cells develop to survive in the highly-stressed and

nutrient-limited environments of tumours by which they can break down their cellular components and recycle the obtained metabolites. Finally, a surprising characteristic is that cancer cells seem to be able to conveniently induce cell death by *necrosis*. Whereas apoptosis is an organised process which breaks cells into compact pieces that are consumed by neighbours, necrotic cells explode and release all their content into the local microenvironment. This release of cellular debris attracts immune inflammatory cells (Fig. 1.3a), which generate stimulating signals that promote the growth and invasiveness of the remaining cancer cells (see below). Therefore, some tumours may tolerate some degree of necrotic cell death in order to favour the surviving cells.

- *Enabling replicative immortality*: chromosomes have protective ends called *telomeres*, which give them structural stability. In most adult normal cells, telomeres shorten during successive replications as a programmed ageing until reaching a non-proliferative state called *senescence* or suffering apoptosis. However, cancer cells can produce enzymes that regenerate telomeres and get an unlimited number of replications, becoming immortal.

- *Inducing angiogenesis*: cancer cells can stimulate endothelial cells (Fig. 1.3a) to construct new blood vessels around the tumour region even in the early stages of development. With this ability they guarantee the supply of nutrients and oxygen and the evacuation of metabolic waste and carbon dioxide, which is essential for the tumour expansion.

- *Activating invasion and metastasis*: this is the most dangerous and lethal characteristic of cancer cells. They develop alterations that modify their attachment to other cells and to the extracellular matrix. However, the mechanisms that empower cancer cells to colonise other tissues and travel to other distant parts of the body remain quite unknown. The development of a secondary malignant tumour requires a permissive microenvironment in the new colonised zone.

Experimental advances during the first decade of this century confirmed and further clarified the described traits and allowed to add two emerging hallmarks in 2011, whose workings are not well understood yet [16]:

- *Reprogramming energy metabolism*: cancer cells can reprogram their glucose metabolism, and thus their energy production, in order to fuel cell growth and replication. The uptake of glucose is highly increased in many tumours and the modifications in its utilization seem to be key in the hypoxic conditions existing in the tumour microenvironment.

- *Evading immune destruction*: immune system constantly monitors all the body components and quickly recognise and eliminate the majority of altered cells that may become cancer cells. By still-unresolved mechanisms, cancer cells seem to manage to hide from immune system or limit its killing response.

In addition, two *enabling characteristics* that make possible the acquisition of the described hallmarks were identified:

- *Genome instability and mutation*: cancer cells often accelerate their rates of random mutations by showing an increased sensitivity to mutagenic agents. This instability is translated into a high heterogeneity between different tumour types and an extreme adaptability to new conditions.

- *Tumor-promoting inflammation*: tumour microenvironment is rich in inflammatory cells belonging to the immune system, which try to eradicate tumours. Unfortunately, inflammatory cells not only attack cancer cells without success but also produce growth factors that favour proliferation and promote angiogenesis, invasion and metastasis.

Understanding these hallmark principles and their underlying mechanisms is playing and will play a crucial role in identifying therapeutic targets that effectively stop and eradicate malignant tumours without hurting normal tissues excessively [20].

1.1.3 Colorectal cancer

Colorectal cancer is a very common human malignancy that starts in the walls of the large intestine. Before summarising some facts of this type of cancer, it is worth describing some concepts of the anatomy and physiology of the large intestine.

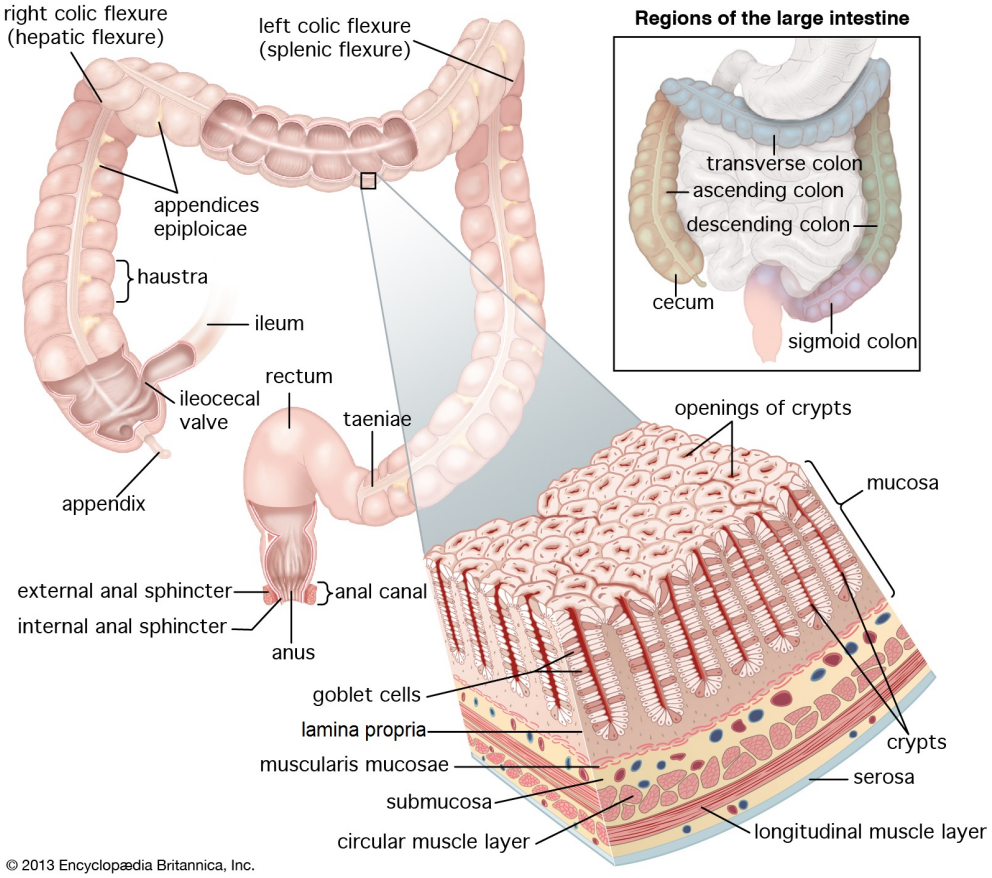


Fig. 1.4: Structures of the human large intestine, rectum, and anus. Reproduced from [22] with permission by courtesy of Encyclopaedia Britannica, Inc., copyright 2003.

Anatomy and physiology of the large intestine

The large intestine is the last part of the alimentary tract. It is about 1.5 meters long and, as shown in Fig. 1.4, can be divided into several regions namely *cecum*, *colon* (ascending, transverse, descending and sigmoid), *rectum*, *anal canal* and *appendix*. As a part of the digestive system, the main functions of the large intestine are the absorption of electrolytes, fluids and gases, and the conversion of undigested materials and dead bacteria into faeces [21].

As can be appreciated in the detail view of Fig. 1.4, the walls of most regions of

the large intestine (excluding the anal canal and the appendix) are composed of four main layers of tissue [23]:

1. *Mucosa*: is the inner part of the wall and constitutes the lining of the intestine. Mucosa consists of three sub-layers of tissue:

   (a) *Epithelium*: is the external layer, which is composed of different types of cells in a *simple columnar* configuration. The epithelium together with the *lamina propria* build invaginations in the form of simple tubular glands or *crypts*. The rectum contains fewer and deeper crypts than the colon. Between the cells composing the epithelium, *goblet cells* have an important role because they secret *mucus* into the central hole of the crypts, called *lumen*. Mucus lubricates the lining of the intestine and facilitates the movement and evacuation of faeces.

   (b) *Lamina propria*: is composed of loose connective tissue whose main function is to give structural and nutritional support to the epithelium.

   (c) *Muscularis mucosae*: is a thin outer layer of smooth (involuntary) muscle cells, which gently contract to promote the evacuation of the crypts' content.

2. *Submucosa*: is composed of fibroelastic connective tissue and contains blood, lymphatic vessels and nerves. It gives support to the rest of tissue layers.

3. *Muscle layer*: consists of two layers of smooth muscle, the inner one is circular and the outer one is longitudinal. The involuntary contraction and relaxation of these layers generate a movement called *peristalsis* that pushes the content of the intestine onwards.

4. *External layer*: is formed by connective tissue and is called *adventitia* in the ascending and descending portions of the colon, where it has a structural function. In the cecum and remainder colon, it is called *serosa* and is composed of a layer of connective tissue and another one of epithelium, which secretes lubricating fluid to reduce the friction with the muscle layer.

Main facts of colorectal cancer

Colorectal cancer [24], also referred to as *bowel cancer* [25] or sometimes differenced as cancer of the colon [26] and the rectum [27], originates in the epithelial cells of the mucosa that form the glands or crypts. This type of malignant tumour is called *adenocarcinoma* (*adeno-* is the Greek term for *gland*) and accounts for more than 90% of cancer cases whose primary site is the wall of the large intestine [28]. Another precursors of cancer in other layers of tissue are muscle cells (produce *sarcomas*), cells of the lymphatic system (give rise to *lymphomas*) or specialized hormone-making cells (produce *carcinoid tumours*), but these subtypes are normally included in categories different from colorectal cancer.

As was described in Fig. 1.1, several mutations are accumulated in the epithelial cells that develop a tumour. When these tumours appear in the lining of the colon or the rectum, they are called *polyps*. Polyps can be divided in two main types :

- *Hyperplastic polyps*: present, in general, little potential to become an invasive cancer. They are relatively common and are normally considered as benign.
- *Adenomatous polyps or adenomas*: are known as the main precursors of cancer lesions and constitute a premalignant condition. Adenomas have characteristics associated with cancer development (Sec. 1.1.1), such as epithelial overgrowth, dysplasia and abnormal differentiation. Nevertheless, few adenomas progress to invasive cancer and the process can take over one to three decades.

Several genetic pathways are already known to be responsible for the change from premalignant polyps to invasive lesions. In fact, the family history of colorectal cancer and other hereditary syndromes have relevant influence in this change. Apart from the inherited predisposition, the main risk factors associated with colorectal cancer are: obesity and sedentary life, smoking, alcohol, high-caloric diet, high red meat consumption, overcooked red meat and high-saturated fats. On the contrary, high level of physical activity and diets high in fibre, low in red and processed meat, and rich in fresh fruit and vegetables, seem to decrease the risk of colorectal cancer.

Different tests are applied in screening programs in order to detect colorectal cancer in undiagnosed individuals. The three most common tests are fecal occult blood tests and two endoscopy alternatives: *sigmoidoscopy* (endoscopy of the left side

of the colon) and *colonoscopy* (endoscopy of the entire colon). Endoscopy is also used to extract small biopsies from small portions of tissue or even remove whole polyps. It also gives valuable complementary information to perform a final diagnosis, which is accomplished by histological examination (Sec. 1.1.6).

The clinical management of colorectal cancer differs depending on the location of the tumour and its spread [25]. If tumours are localized in the colon, the involved bowel segments including the local lymphatic drainage area are normally removed by laparoscopic surgery, with high curative potential (Sec. 1.1.5). Cancers in the rectum, especially in carcinomas below the mid-sigmoid region, may also be treated with radiation therapy apart from surgical resection. In advanced stages of spread, complementary adjuvant treatments, such as chemotherapy or other systemic therapies, can also be applied to prevent the formation of new tumours in the primary or secondary sites.

1.1.4 Skin cancer

Melanoma is the most aggressive subtype of skin cancer. As in colorectal cancer, some concepts of the anatomy and physiology of the skin and its relevant constituents must be explained prior to describe some general facts of skin cancer and melanoma.

Anatomy and physiology of the skin

The skin is the outer covering and the largest organ of human body. Its functions are: to protect the body against injury, desiccation and infection; to regulate body temperature; to absorb Ultraviolet (UV) radiation; and to contain receptors for touch, temperature and pain stimuli from the external environment [23]. The skin is divided into two main layers: the *epidermis* (upper or outer layer) and the *dermis* (lower or inner layer) (Fig. 1.5). These two tissue layers are separated by the basement membrane.

The epidermis is composed of *stratified keratinised squamous epithelium*, whose predominant cell type is the *keratinocyte* [21]. Keratinocytes are arranged in five layers or *strata*, where they acquire different characteristics. In the deepest layer of epidermis, keratinocytes are also called *basal cells* and are mitotically active, that

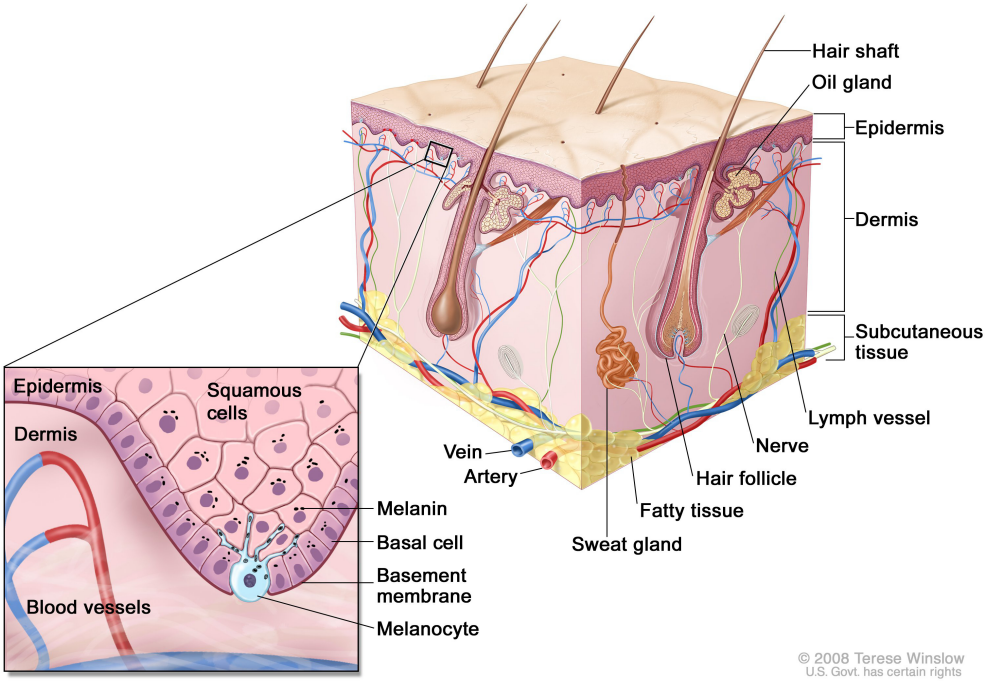


Fig. 1.5: Illustration of the anatomy of healthy skin. Reproduced from [29] with permission by courtesy of Terese Winslow (Illustrator).

is, they are in continuous cell division to regenerate the upper layers. In the rest of epidermis layers, keratinocytes are called *squamous cells* and they evolve towards the surface until becoming flattened dead *cells* filled with impermeable *keratin* in the most superficial layer.

Melanocytes are cells also present in the deepest layer of the epidermis. The mission of melanocytes is to synthesise a dark brown pigment called *melanin* and to transfer it into nearby keratinocytes (see detail view of Fig. 1.5). Melanin protects against tissue damage caused by UV radiation and its synthesis is promoted by continuous exposure to sunlight.

The dermis is formed by connective tissue that supports different structures, such as blood and lymph vessels, nerves, sweat and oil glands, and hair follicles. Connective tissue plays an important role in immune and inflammatory responses, tissue repair after injury, as well as delivering nutrients to the avascular epidermis through the

basement membrane [30]. *Fibroblasts* are the resident cells of connective tissue, whose mission is to elaborate fibres, such as collagen and elastin, and other components of the extracellular matrix.

Main facts of skin cancer

Skin cancer can be divided into two main groups: *non-melanoma* skin cancer and *melanoma* skin cancer. Among non-melanoma skin cancer, two subtypes are more common [31]:

- *Basal cell carcinoma*: is a slow-growing neoplasm that originates from keratinocytes in the basal cell layer of the epidermis. This carcinoma rarely produces metastasis but its local recurrence and invasiveness result in significant tissue destruction. As depicted in Fig. 1.6a, basal cell carcinoma normally remains confined in the epidermis and does not expand to other tissues. It typically develops on sun-exposed regions of people with lighter skin.
- *Squamous cell carcinoma*: is a neoplasm that has malignant characteristics, such as local invasion and metastatic potential. It originates from kerotynocytes of epidermis in *squamous cell* configuration and can invade nearby tissues like the dermis, as illustrated in Fig. 1.6a. Most lesions of this cancer type do not cause pain and remain undetected, growing slowly until they become invasive carcinomas.

Melanoma is a neoplastic disorder that starts with the transformation of normal melanocytes [35–37]. Although melanocytes are also present in other body locations (e.g., meninges, upper oesophagus or eyes), their malignant transformations are more common in the skin. Melanoma is the most aggressive type of skin cancer, that is why it is normally considered as an individual category. In fact, it is one of the most violent and therapy-resistant human cancers, mainly due to its high mutation capability and heterogeneity. Compared with other skin cancers, melanoma is characterised by its superior ability to grow and spread both radially and vertically, finally invading inner tissues in a very aggressive way (Fig. 1.6a).

Melanoma occurs mainly in white people with fair skin. Although not necessary occurring in all cases, some melanomas start from common accumulations of

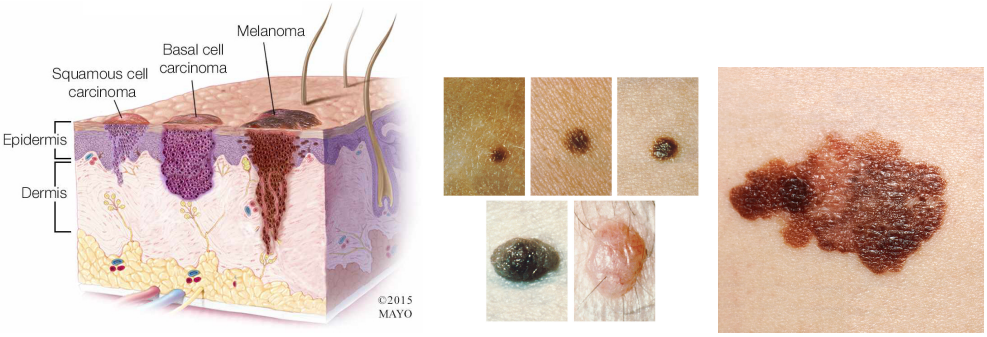


Fig. 1.6: Skin cancer. (a) Illustration of the most common types of skin cancer. Reproduced from [32] with permission by courtesy of Mayo Foundation for Medical Education and Research. All rights reserved. (b) Photographs of five common benign moles. Reproduced from [33]. (c) Photograph of a malignant melanoma. Reproduced from [34].

melanocytes called *nevi* or *moles* (Fig. 1.6b). Individuals with a high number of melanocytic nevi (more than 50) have higher odds of developing melanoma [38]. A famous rule to distinguish the features of a malignant melanoma (Fig. 1.6c) is the mnemonic ABCD method [39, 40]: Asymmetry, Border irregularity, Colour variation, Diameter >6mm.

The common major risk factor in skin cancers is sun overexposure with inadequate protection even in the absence of sunburn, especially in childhood and adolescence, or other cumulative exposures to UV radiation as artificial tanning lamps and beds. Other risk factors include inherited genetic susceptibility, immunosuppression and viral infections. The clinical management of skin cancers depends on the tumour aggressiveness and stage, and also on the current status and medical record of the patient. The common treatments expand from traditional surgical excision to destructive modalities that include cautery or electrodesiccation, cryosurgery, laser surgery, radiation therapy and chemotherapy. In the excisional surgery, the borders of the extracted mass of tissue can be studied to assess the completeness of the tumour ablation. However, the destructive techniques do not provide any information to know if the cancer has been completely removed.

1.1.5 Cancer statistics

Cancer burden

GLOBOCAN project, elaborated by the International Agency for Research on Cancer (IARC), the specialized cancer agency of WHO, provides the most updated and reliable estimates of the global cancer burden, which can be retrieved from [41, 42]. These statistics correspond to 2012 and are given in terms of estimated *incidence*, *mortality* and *prevalence* [43, 44]. In this case, these terms are defined as follows [45]:

- *Incidence*: is the number of new cases arising in a specified population within a given period of time. Here it is expressed as the absolute number of new cases per year (2012).
- *Mortality*: is the number of deaths occurring in a specified population within a given period of time. Here it is also expressed as the absolute number of deaths per year (2012).
- *Prevalence*: is a function of both incidence and survival. It is defined as the number of individuals within a specified population who have been diagnosed with a specific cancer and who are still alive at a given point in time (i.e. the survivors). In this case, it is presented as *partial prevalence*, which limits the number of patients to those diagnosed during a fixed time in the past, 5 years in particular. In most types of cancers, patients still alive at 5 years after diagnosis are usually considered *cured* because the death rates of such patients are similar to those *expected* in the general population [46]. Therefore, the *5-year prevalence* is an estimate of the *cure rate* of a specific cancer. Here it is presented the number of estimated cases diagnosed since 2004 who were still alive at the end of 2008 [44].

Fig. 1.7 graphically summarises the described measures for the corresponding top 10 cancer sites worldwide and in Europe, for both sexes. Roughly more than 14 million new cases of all cancers were diagnosed in the world in 2012 and 3.4 million of them were registered in Europe. WHO foresees a rise of 70% in this value over the next 2 decades due to population increase and ageing [47], which will result in almost 24 million new cases diagnosed worldwide per year. Regarding the relevant

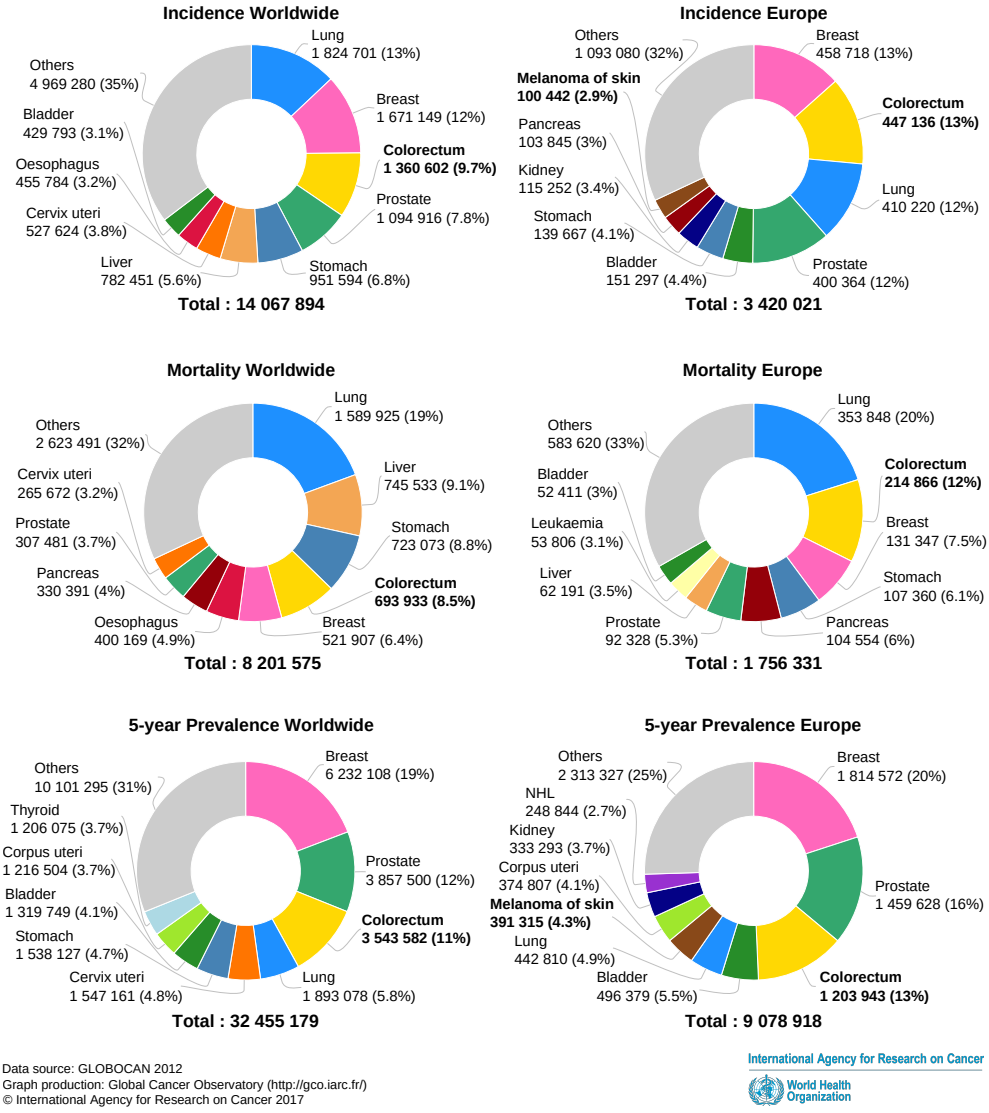


Fig. 1.7: Estimated incidence (first row), mortality (second row) and 5-year prevalence (third row) of the corresponding top 10 cancer sites worldwide (first column) and in Europe (second column), for both sexes in 2012. The relative values for each cancer site with respect to the corresponding total number of cases are shown in parentheses. The relevant cancer sites for this PhD thesis (colorectum and melanoma of skin) are highlighted with bold letters. Adapted with the graphs produced from [42].

cancer sites for this PhD thesis, colorectal cancer was the 3rd most diagnosed cancer worldwide in 2012 (9.7% of all new cases) and melanoma of skin occupied the 19th position (1.7%). The relative incidence in Europe for these types of cancer is higher, with colorectal cancer in the 2nd place (13%) and specially melanoma in the 9th position (2.9%). Some factors that may explain these local differences could be the characteristics of population (e.g., people with white skin for melanoma), their habits or the screening programs which are mainly applied in more developed countries.

Cancer is the second leading cause of mortality, just after cardiovascular diseases, both worlwide with around 8.2 million deaths and in Europe with more than 1.7 million deaths in 2012. Lung cancer deserves special mention since it is clearly the most lethal type of cancer, causing around 20% of all cancer deaths. These data are even more shocking when they are compared with the corresponding incidence. Colorectal cancer is also a very lethal cancer, ranking the 4th worlwide (6.4%) and the 2nd in Europe (12%). For its part, melanoma is in the 22nd place worlwide (0.7%) and in the 19th position in Europe (1.3%).

The estimated 5-year prevalence of all cancers rises to almost 32.5 million patients worlwide and more than 9 millions in Europe. By comparing the values for each cancer type with the respective incidence, an idea of the survival rate can be extrapolated. Again the lethality of the lung cancer can be confirmed by this way, especially in Europe. Colorectal cancer is the third most prevalent cancer both worlwide (11%) and in Europe (13%), meanwhile melanoma ranks the 11th worlwide (2.7%) and the 6th in Europe (4.3%).

According to the described figures, cancer constitutes a major public health problem. In particular, colorectal cancer and melanoma can be considered as very serious diseases with a high general impact on the population. Further details and analyses of the cancer statistics from GLOBOCAN can be found in [48].

Cancer survival by stage of diagnosis

As commented, the 5-year prevalence compared with the incidence can be used as a rough estimate of the survival rate for a specific cancer. However, knowing the exact survival rate, especially by the stage at which cancer was diagnosed, is very valuable to develop appropriate screening programs and surveillance guidelines.

These values are not easy to get because not many countries have enough resources to record high-quality national data with such richness. Despite that, some national agencies have aggregated data where diagnosed cancers are broadly classified in few categories depending on the spread of the lesion. Two main staging systems are commonly used for all types of cancer:

- *TNM Staging System*: was developed and is revised jointly by the American Joint Committee on Cancer (AJCC) [49] and the Union for International Cancer Control (UICC) [50]. It specifies five stages in Roman numerals (*0-IV*).

- *SEER Summary Staging*: is promoted by the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute of the United States (US) [51]. It specifies four categories: *in situ*, *localized*, *regional* and *distant*.

An additional stage or category called *unknown* is also considered in both staging systems to catalogue those cases without enough evidence to be assigned to the other stages. Although each cancer type may have its own variants, the overall stages are common and can be illustrated for the particular case of colorectal cancer (Fig. 1.8). In the case of melanoma, the transition from the first abnormal melanocytes to metastatic melanoma can be divided into the same stages or categories, which are mainly characterised by the vertical growth of the tumour. The stages of the TNM Staging System, related to the categories from the SEER Summary Staging, are:

- *Stage 0*: abnormal cells appear in the first tissue site, but they have not spread to other tissues. In the case of colorectal cancer, these cells appear in the epithelium of the mucosa layer without penetration of the basement membrane (Fig. 1.8). This stage is equivalent to the *in situ* category of the SEER Summary Staging.

- *Stage I*: cancer acquires invasive properties and spreads to nearby tissue layers. In colorectal cancer, the malignant cells spread from the mucosa to the submucosa and muscle layer (Fig. 1.8). In this stage, the cancer is still in the *localized* category of the SEER Summary Staging because it is limited to the organ where it started.

- *Stage II*: cancer manages to spread through all the layers of organ (all the intestine wall, including the external layers, in the case of colorectal cancer, Fig. 1.8), and even invades nearby organs. According to the SEER Summary Staging, this is an intermediate stage where the lesion may be defined as *localized* or *regional* depending on whether the lesion does or does not remain confined to the organ of origin, respectively.

- *Stage III*: this stage is more complicated and have a variety of different sub-stages. In them, the cancer may or may not have locally spread from the original organ, but it manages to access the lymphatic vessels and spread to nearby lymph nodes. In the SEER Summary Staging, the cancer clearly reaches the *regional* level but is still limited to a specific area of the body.

- *Stage IV*: cancer finally produces metastasis and spreads to other parts of the body through the lymphatic or circulatory systems, giving rise to secondary tumours. This phase corresponds to the *distant* category in the SEER Summary Staging.

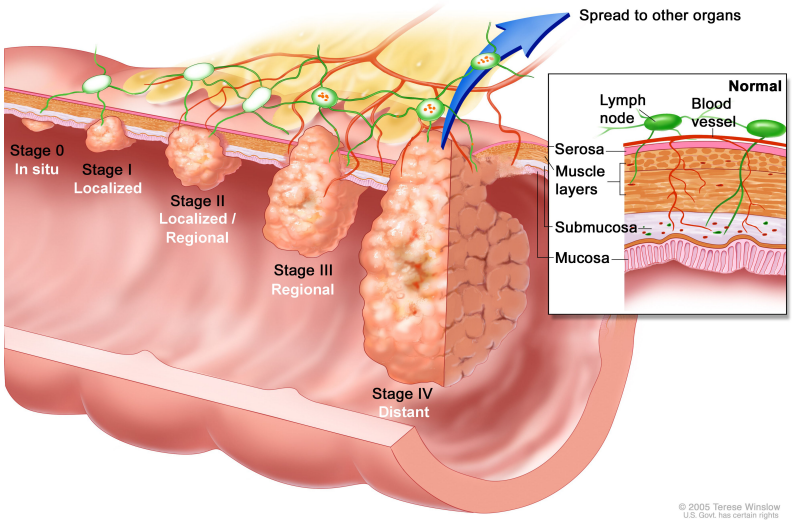


Fig. 1.8: Stages of colorectal cancer. Stages from the *TNM Staging System* are specified in black fonts and the corresponding categories from the *SEER Summary Staging* in white fonts. Adapted from [52] with permission from Terese Winslow (Illustrator).

To the author's knowledge, only two sources provide survival statistics specified by stage of diagnosis for the two cancer types of interest here (colorectal or bowel cancer and melanoma). The first source is Cancer Research UK [53], which summarises the records of the adult cancer patients (aged 15-99 years), separated by sex, from the Former Anglia Cancer Network in the period 2002-2006. The staging of this source uses the TNM Staging System. The second source is the SEER Program [54], which uses the SEER Summary Staging to distinguish the data obtained from the US population in the period 2007-2013. The most relevant data from these sources are graphically presented in Fig. 1.9.

The survival statistics in both sources are given in terms of *5-year relative survival*. It compares the survival during five years or more after diagnosis of a cohort of cancer patients with the survival of another cohort of people from the general population with similar characteristics (same age, race, and sex) but not diagnosed with cancer. If this relationship is greater than 100%, it indicates that cancer patients in a specific group have a better chance of surviving five years after diagnosis compared with the studied general population.

As can be seen in Fig. 1.9.a and Fig. 1.9.c (top graph), colorectal or bowel cancer has good expectancies of 5-year survival when it is detected at early Stage I or is still localized (90-100% relative survival). The survival gradually decreases in intermediate stages of spread ($\sim$85% in Stage II, $\sim$63% in Stage III and $\sim$71% in regional stage). But there is an abrupt drop in survival when the cancer produces metastasis (roughly 10% or less in Stage IV and distant stage). The data from US population (Fig. 1.9.c, bottom graph) also inform about the percent of patients diagnosed at each stage. According to them, colorectal cancer does not have good records of early diagnosis with 39% of cases detected in the localized stage and 35% with regional spread. Sadly, more than 20% of cases were diagnosed in the distant stage, with the worst prognosis. Only 4% of cases could not be classified in a specific stage.

Similar survival trends can be observed in melanoma (Fig. 1.9.b and Fig. 1.9.d, top graph). Cases diagnosed during first stages of localized primary tumours have a high cure rate with almost and even more than a 100% 5-year relative survival. The survival also gradually decreases in stages with intermediate spread ($\sim$80% in Stage II, $\sim$50% in Stage III and $\sim$63% in regional stage). However, patients whose cancers have advanced up to metastatic stages only reach a 5-year relative survival

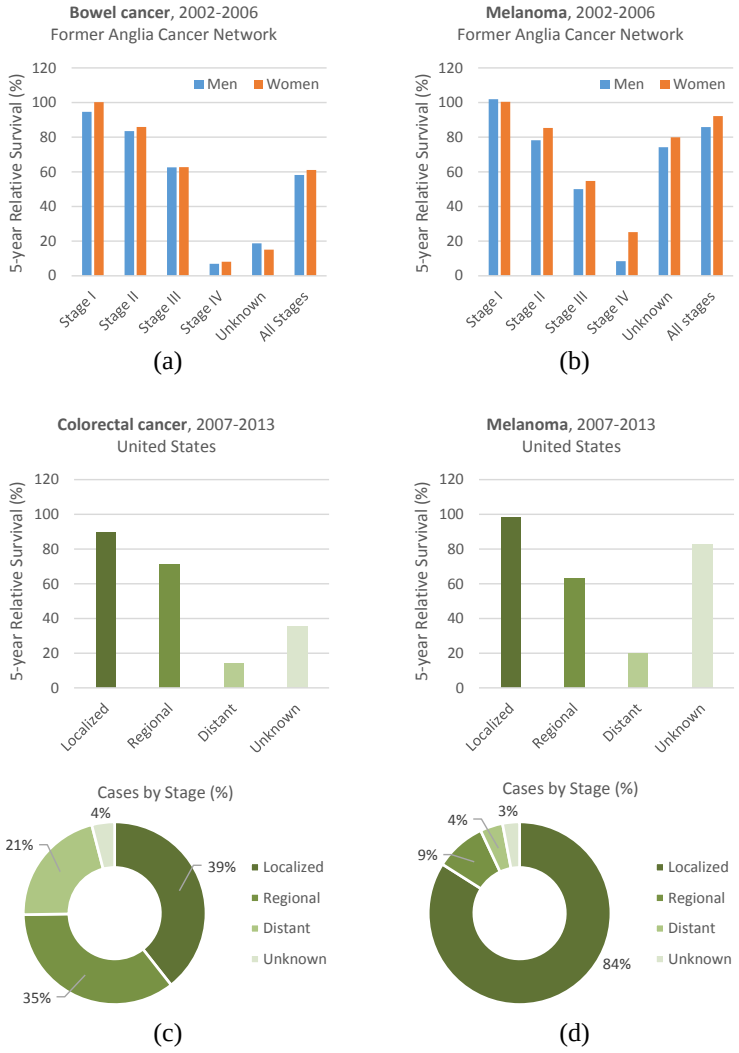


Fig. 1.9: Survival by stage of diagnosis. First row: 5-year survival rate by stage in the period 2002-2006 of the adult cancer patients (aged 15-99 years), separated by sex, from the Former Anglia Cancer Network who were diagnosed with (a) bowel cancer and (b) melanoma (Data retrieved from [53] using the TNM Staging System; no significant differences were found between men and women at any stage or type of cancer). Second and third row: 5-year survival rate by stage (second row) and percent of cases per stage (third row) in the period 2007-2013 of the US population diagnosed with (c) colorectal cancer and (d) melanoma (Data retrieved from [54] using the SEER Summary Staging).

from around 10 to 20%. Again, the percent of patients diagnosed at each stage in the US population is shown for melanoma (Fig. 1.9.d, bottom graph). The early diagnosis in melanoma seems to be much more effective than in colorectal cancer in this population: 84% of cases detected in the localized stage, 9% with regional spread and just 4% with distant metastasis. In this case, just 3% of cases were not properly catalogued.

As main conclusion of these figures, early diagnosis is the key factor to reduce mortality. Current diagnosis protocols seem to have a low rate of early diagnosis for colorectal cancer but a high rate for melanoma. Finally, it must be noticed that these statistics correspond to two leading countries in screening programs and cancer treatments.

1.1.6 Cancer diagnosis and research

Although some cancer lesions show different observable signs that allow its detection *in vivo* (directly in the patient), ultimate diagnosis and staging is currently established *ex vivo* by analysing biological samples extracted from the patient. *Histopathology* and *cytopathology* are the main medical disciplines that analyse the *ex-vivo* specimens with diagnostic purposes. On the other hand, *in-vitro* techniques, such as *cell culture*, play a crucial role in cancer research and in the development of new diagnostic technologies.

Histopathology

Histopathology is the branch of anatomical pathology that uses *histology* (the science that studies tissues) to diagnose diseases. Histopathology is considered the *gold standard* or reference test to perform a final diagnosis of cancer. That is, although different techniques such as endoscopy, radiology or blood test are used to detect the lesion or even obtain complementary diagnostic information, the final evaluation and staging depend on the histological findings.

In histopathology, small portions of tissue extracted from suspicious regions of the patient are examined under a microscope to assess different properties, especially their morphology. The pathologist must follow a series of steps in order to prepare the sample for examination under a microscope. The main steps of a typical tissue preparation are detailed below [55] and some of them are illustrated in Fig. 1.10:

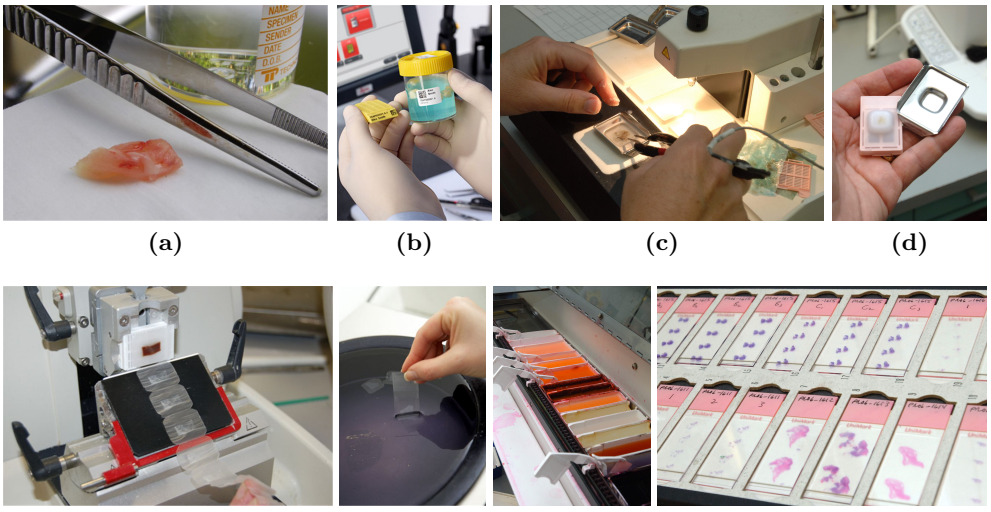


Fig. 1.10: Tissue processing for histopathology. (a) Fresh biopsy extracted from a patient. (b) Fixation of the tissue. (c) Chemically processed tissue is placed into a metal mould and embedded in molten paraffin. (d) Solidified block of paraffin with the embedded tissue. (e) Cut of thin serial sections with a microtome. (f) Paraffinated sections are placed in a warm bath to flatten them and then they are mounted on a microscope slide. (g) The mounted tissue sections are stained, typically with Hematoxylin and Eosin (H&E). (h) The final stained and mounted slides are ready to be examined with a microscope by an expert pathologist. Images (a), (b), (e), (f) reproduced from [56]. Images (c), (d), (g), (h) reproduced from [57].

- *Biopsy extraction*: a piece of tissue called *biopsy* is extracted from a patient; for example, during surgery or endoscopy (Fig. 1.10a). The fresh biopsy must be processed quickly to avoid its breakdown.

- *Fixation*: the fresh tissue is treated in order to harden it and preserve its morphological structure with the time (Fig. 1.10b). Although fixation can be accomplished by physical methods (e.g., heating, microwaving, or freeze-drying), liquid chemical fixatives are commonly used. In particular, the most typical fixative used in diagnostic pathology is formaldehyde in aqueous solution, which is normally called formalin.

- *Dehydration*: water and aqueous fixatives are removed from tissue components

by hydrophilic substances, such as ethanol. Immersion of specimens into dehydrating reagents is performed through a series of baths of increasing concentration in order to prevent cell distortion.

- *Clearing*: this is a necessary intermediate step because dehydration reagents are not miscible with infiltrating solutions. The clearing substance must be miscible with both the dehydration and infiltrating solutions. They normally have a refractive index similar to protein and they produce a translucent appearance in the tissue: hence the term *clearing agent*. One of the most used clearing substance is xylene.

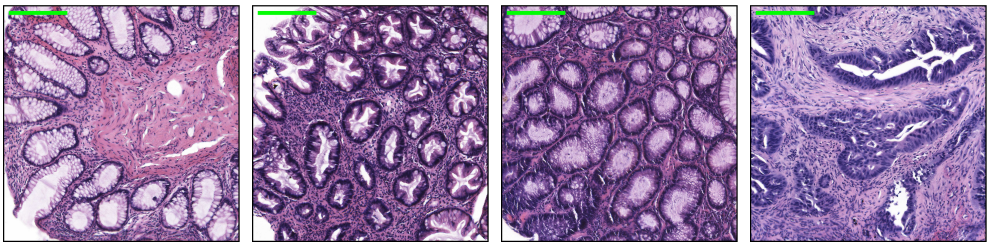
- *Infiltrating and embedding*: the clearing agent is replaced by an infiltrating substance, typically molten paraffin wax. The specimen is placed in a metal mould with a proper orientation and embedded in molten paraffin (Fig. 1.10c). When the paraffin solidifies, the block with the embedded tissue gains rigidity, which prevents distortion of its structure during microtomy (Fig. 1.10d).

- *Cutting*: the block of paraffin is cut with a *microtome* into very thin sections of around 5 µm thickness (Fig. 1.10e). The cut sections are put into a warm water bath in order to flatten them and then recovered with a microscope slide where they get attached (Fig. 1.10f).

- *Staining*: the mounted slices of tissue are treated with chemicals to remove the paraffin and then immersed in different stains (Fig. 1.10g). The excessive stains are removed with water and the final slides get ready for observation with a microscope (Fig. 1.10h).

Tissue processing is a complex procedure which employs chemical and physical processes that can potentially alter and distort the characteristics of the tissue. For diagnostic pathology, it is important to guarantee that induced undesired artifacts are at least consistent. Therefore, a strict and repeatable protocol should be followed during the whole tissue processing.

As described, the last step of the tissue processing is the staining. Processed tissues are transparent in the visible light and they must be *coloured* with stains. Stains or dyes are substances that have affinities with particular components of the tissue,



(a) **(b)** **(c)** **(d)**

Fig. 1.11: Histopathological images of colorectal biopsies with different pathological conditions and stained with H&E. Green scale bars represent 200 µm. (a) Normal sample. (b) Hyperplastic polyp (benign). (c) Adenoma (precancerous). (d) Adenocarcinoma (cancer).

promoted by specific chemical interactions. They have also absorptive properties to electromagnetic radiation, typically in the visible light range, which give them colour.

Hematoxylin and Eosin (H&E) stain is by far the most used dye in anatomical pathology and is actually considered the *gold standard* technique to diagnose tissue diseases. Its popularity is based on its relative simplicity and its ability to reveal a high number of different tissue structures. H&E is a combination of two chemical substances with complementary properties. Hematoxylin is a basic dye that binds to acidic (or basophilic) structures, such as cell nuclei, and colours them in dark blue or violet. By contrast, eosin is a dye with varying shades and intensities of pink, orange and red, which has an acidic nature and, therefore, it binds to basic (or acidophilic) substances, such as proteins in the cytoplasm of cells or fibres of the extracellular matrix. With these two dyes, the main tissue structures are revealed and, by interpreting their morphological arrangement, the pathologist can assess the underlying condition of the tissue and establish a diagnosis.

As an example, Fig. 1.11 shows four H&E histopathological images of colorectal biopsies from lesions with different pathological conditions. These kinds of histopathological samples will be employed in Ch. 4. Some typical characteristics of each condition [28, 58] can be identified in these images:

(a) *Normal sample*: nuclei of epithelial cells in the mucosa layer, stained with hematoxylin in dark shades of violet, are aligned in the outer border of glands

or crypts. As can be seen in Fig. 1.11a, glands have approximately round or elliptical shapes, depending on their orientation with respect to the microtome's cutting plane. Cytoplasms of epithelial cells, coloured in a light purple, are placed inward the glands enclosing the lumen, which mainly has white tones. This spatial orientation is called *polarity* and facilitates the secretion of mucus and other fluids into the lumen. Outside the glands, nuclei from different types of cells of the rest of layers are surrounded by their cytoplasm and extracellular matrix stained with pink eosin. Some breaks in the tissue can also be observed in white.

(b) *Hyperplastic polyp (benign)*: as in the normal state, epithelial nuclei are small, regular and round. However, hyperplastic tissue presents increased cellularity and mitotic activity, which can be mixed up with adenoma. In fact, they have some mutations in common that make them adopt similar features. Hyperplastic polyps have also simple tubular architecture with elongated and straight crypts, which cannot be properly visualised in Fig. 1.11b due to the cutting orientation. Nonetheless, this view allows to identify the *serrated* (saw-tooth or star-shaped) morphology of the glands, which characterises this abnormal and benign condition.

(c) *Adenoma (precancerous)*: this precursor lesion is defined by the presence of intraepithelial neoplasia, which is characterized by hypercellularity with enlarged, hyperchromatic nuclei, varying degrees of nuclear stratification, and loss of polarity [28]. In some types of adenoma, glands also present a serrated configuration if they are observed with a low magnification microscope, but the epithelial cells are dysplastic unlike in hyperplastic polyps. Dysplastic cells lack morphological uniformity, lose their architectural orientation and often contain large hyperchromatic nuclei with a high nuclear-to-cytoplasmic ratio [3]. In adenoma, dysplastic glandular structures normally occupy a great portion of the luminal surface, as can be observed in the bottom-left region of Fig. 1.11c.

(d) *Adenocarcinoma (cancer)*: its defining feature is invasion through the mucosa into the submucosa, where it has higher risk of metastasis. The grade of adenocarcinoma is mainly determined by the extent and appearance of the glands, which become more heterogeneous and distorted. Depending on whether the glands can be identified, the lesions can be divided into well, moderately and

poorly differentiated adenocarcinoma. Fig. 1.11d is an example of a moderately differentiated adenocarcinoma, where the epithelial cells are large and tall, and the gland lumina have almost disappeared or contain cellular debris.

Tumours present a great histological diversity, showing different regions with various degrees of differentiation, proliferation, vascularity, inflammation and/or invasiveness [16]. Histopathological comparison of normal tissue and tumours, although reveals morphological differences, also confirms common features in their microarchitecture. The pathologist must be able to identify these (sometimes subtle) differences in order to define the grade of cancer and then choose the best clinical management. This assessment is even harder in intermediate phases of development, when lesions are localised in small regions of the sample and may not show clearly distinctive characteristics.

Cytopathology

In a similar way to histopathology, *cytopathology* studies and diagnoses diseases at the cellular level [59–61]. In certain types of cancer, individual cells are examined as a complementary test to histopathology or even as the main diagnostic tool. The most typical examples are leukaemia, cervical cancer and bladder cancer, where cells can be collected in a relatively simple way. However, a significant amount of cells can be extracted from less accessible regions, for example, by fine needle aspiration as in lymph nodes. Cytopathology gives valuable information about molecular and morphological changes that helps in the diagnosis of precancerous and malignant lesions.

Cytopathology, as well as a science, is considered an art even more challenging than histopathology because changes in cells can be subtler or more hidden than in tissue samples [59]. The quality of the preparation of cytological samples is crucial to perform an accurate interpretation. As done with tissues, cell specimens collected from suspicious sites by exfoliative procedures, fine needle aspiration or even endoscopy, must be processed in different steps that typically include [62]: *cytocentrifugation*, to separate cells; *fixation*, to maintain morphology; and *staining*, to facilitate cell visibility, detection and interpretation with classical light microscopes. A famous dye in cytopathology is the Papanicolau stain, which is a mixture of five dyes including hematoxylin and eosin.

Cell culture

In parallel with traditional cytopathology, where *ex-vivo* cells collected from humans are directly studied for diagnostic purposes, *cell culture* consists of creating cytological samples under controlled laboratory conditions (*in vitro*). Cell culture is becoming more important in both academic research and industry to develop new diagnostic technologies, to generate and use test or model systems, or to produce biopharmaceuticals or other oncological products [63]. The main advantage of cell cultures is the homogeneity of replicated samples, which normally belong to a specific *cell line*. This is specially useful in the development of screening systems for diagnosis and drug testing because the pathological conditions of the cell lines are known and the obtained results can be potentially consistent and reproducible [64]. As main limitations, culture techniques require high levels of expertise and relatively expensive equipments in order to guarantee the quality of the cell culture. In addition, not only for quality but also for safety reasons, people working with cultured cells must follow standard protocols and specific practical guidelines [65–67].

As commented, cell cultures normally employ catalogued cell lines with known properties that were originally extracted from human or other animal tissues. Fig. 1.12 illustrates the evolution with the time, in terms of cumulative number of cells, of a hypothetical cell line. After tissue removal or *explantation*, cells are isolated and placed into a suitable culture environment to create the *primary culture*. These original cells grow and proliferate until reaching *confluence*, that is, cells occupy all the available substrate and make close contact with one another. At this point, cells are detached from the substrate with enzymes and moved to a new substrate to produce the first *subculture* or *passage* and give rise to the *cell line*. The process is repeated in a series of passages where the most proliferative cell lineages gradually predominate and the cell line becomes more stable.

Cell lines created from normal tissue, unlike malignant cells (Sec. 1.1.2), have a limited number of replications due to the shortening of telomeres and finally undergo *senescence* (non-proliferative state). The population of these finite cell lines reaches a maximum and then decreases when the non-proliferative existing cells gradually die (Fig. 1.12). However, some cultured cells experiment genetic transformations and originate a *continuous cell line* capable of replicating infinitely. This transformation or *immortalisation* may occur spontaneously or be chemically or virally induced. Not

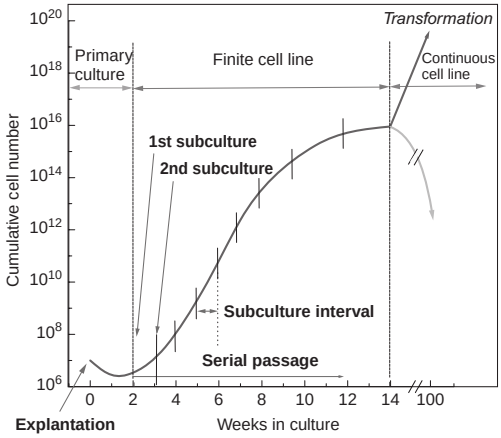


Fig. 1.12: Evolution of a cell line. The vertical axis represents total cell growth (assuming no reduction at passage) for a hypothetical cell culture. Total cell number (cell yield) is represented on this axis in a logarithmic scale, and the time in culture is shown on the horizontal axis on a linear scale. Although a continuous cell line is depicted as arising at 14 weeks, with different cells it could arise at any time. Likewise senescence may occur at any time. Reproduced from [64] with permission by courtesy of John Wiley & Sons Inc.

all types of cells have been successfully immortalised. The most famous case is normal human fibroblasts, which never give rise to continuous cell lines [64].

Established cell lines can be cryopreserved and stored with liquid nitrogen, perpetuating their characteristics indefinitely. A great diversity of cell lines are available from recognised cell banks, which guarantee their authenticity and freedom from microbial contamination [67]. Specialised laboratories, which must have trained personnel and appropriate equipment (e.g., microbial safety cabinets, incubators, autoclaves, water filtration units, etc.), can acquire batches of catalogued cell lines from recognised cell banks or distributors and create their own cell cultures for research purposes.

When culturing animal cells, it is essential to consider that they only survive under relatively constant physiological conditions, which must simulate blood plasma or interstitial fluids and avoid contamination with external microorganisms. They also require a strict control of the physicochemical microenvironments in order to develop their specific *phenotype* (distinguishable characteristics or traits). In particular, some factors that determine the correct development of cultured cells are: the incubation

temperature; the concentration of nutrients and growth factors in the supplied media; the nature of the substrate, especially for cell adhesion; and the interaction and contact between cells.

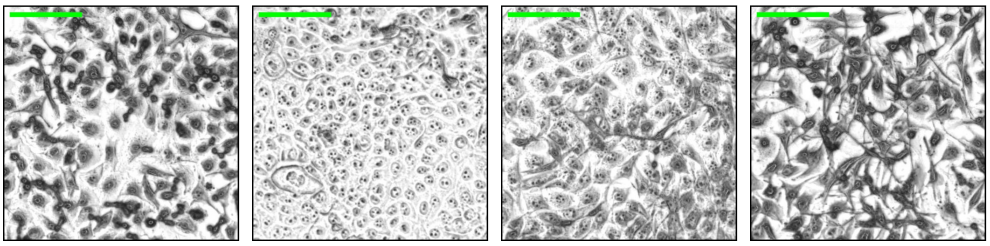
Cell proliferation is governed by the cell cycle (Fig. 1.2), which is regulated by growth factors that must be artificially supplied and other signals from the microenvironment. For instance, normal cells stop proliferating in highly dense cultures due to cell contact inhibitions, whilst cancerous and other transformed cell lines may keep dividing (Sec. 1.1.2). Each cell line, mainly specialised cells such as epidermal keratinocytes, needs specific culture conditions and handling in order to proliferate and keep differentiated.

Four different skin cell lines will be analysed in this thesis (Ch. 5): two of them (HaCaT and NIH-3T3) derived from normal or healthy skin (Sec. 1.1.4) and the other two (A-375 and SK-MEL-28) from different melanoma lesions. Fig. 1.13 shows images from specific regions of cultures of each cell line, which were acquired with a phase contrast microscope. Some characteristics of these skin cell lines can be described in alphabetical order:

(a) *A-375*: is a cell line derived from a 54-year-old female of unknown ethnicity with malignant melanoma, which was established in 1973 [68]. When inoculated subcutaneously in immunosuppressed mice *in vivo*, these cells rapidly develop tumours resembling amelanotic melanomas.

(b) *HaCaT*: is an immortal cell line that was spontaneously transformed *in vitro* from normal keratinocytes extracted form a 62-year-old Caucasian male in 1988 [69]. It presents unlimited growth potential but remains nontumorigenic and is capable of reproducing the orderly structure of normal epithelial tissue. Therefore, HaCaT cells tend to join together and form densely packed monolayers in cultivations *in vitro*.

(c) *NIH-3T3*: is an immortal non-pathological cell line created from NIH Swiss mouse embryo fibroblasts in 1969 [70]. It is considered a standard fibroblast cell line and has been used in numerous cytological studies as a substitute of human fibroblasts, which have not been successfully immortalised and present critical problems of instability *in vitro* [64].



(a) (b) (c) (d)

Fig. 1.13: Phase contrast microscopic images (with corrected background) from cell cultures of four different skin cell lines, including melanoma. Green scale bars represent 100 µm. (a) A-375 (Malignant melanoma). (b) HaCaT (Non-tumoral keratinocytes). (c) NIH-3T3 (Non-tumoral mouse fibroblasts). (d) SK-MEL-28 (Malignant melanoma).

(d) *SK-MEL-28*: is a malignant melanoma cell line that was established from an axillary lymph node of a 51-year-old male of unknown ethnicity in 1976 [71]. SK-MEL-28 cells cultivated *in vitro* tend to grow as a monolayer and normally present a polygonal morphology.

As can be observed in Fig. 1.13, each cell line has different morphological and growing properties. HaCaT cells form the most organised and connected monolayer, whereas NIH-3T3 fibroblasts and specially melanoma cells grow in a more disorganised way with frequent areas of overlapping cells. In addition, individual melanoma cells present more compact and rounded shapes, as a sign of their enhanced mitotic activity (Sec. 1.1.2).

Finally, as in traditional cytopathology, cell cultures can be processed after growing to be visualised with specific imaging methods and allow their storage for posterior studies. In particular, they can be chemically treated with fixatives to keep their morphology with the time. However, these treatments may alter the chemical characteristics of the cells.

1.2 Motivation

As was described and supported with official statistics in Sec. 1.1.5, cancer in general, and colorectal cancer and melanoma in particular, are major health problems with

high incidence and mortality. Besides, the survival of cancer patients dramatically decreases when the lesion is not diagnosed at early stages of development. As also commented before, histopathological (and in some cases cytopathological) analysis remains the *gold standard* for the diagnosis of most types of cancer.

Nowadays, the final assessment of suspicious biopsies is still performed by an expert pathologist, who visually inspects the prepared samples with a traditional optical microscope and gives a qualitative judgement. This lengthy and tedious methodology possesses a high degree of subjectivity and heavily relies on the experience and skills of the clinician. Several studies, involving pathologists with many years of experience, have revealed a suboptimal inter-observer variability in the differentiation and reporting of both colorectal polyps [72–75] and melanocytic neoplasms [76–79]. This diagnostic discordance is normally higher (even surpassing 30% of studied cases [78]) in early stages when abnormal tumours do not show clearly discriminant signs yet. This circumstance leads to changes in clinical management and gives rise to situations where a region of benign tissue is removed (overdiagnosis) or, even worse, a malignant lesion is not properly treated (underdiagnosis). The controversy in the diagnosis mostly disappears in advanced stages, but the lesion is already lethal in most of those cases. Most reported studies conclude that more objective criteria should be applied for risk stratification in screening and surveillance guidelines.

1.2.1 Infrared spectroscopy and cancer diagnosis

One of the main targets in cancer research is the discovery and validation of new biological markers (*biomarkers*), which are representative of the pathology. A cancer biomarker is any functional entity that can be objectively measured, such as genes, proteins, metabolites, morphological, cytogenetic or cytokinetic parameters, as well as any other physical feature or detectable change in body fluids or tissues that can be significantly related to a specific cancerous disease [80]. Infrared (IR) spectroscopy (Ch. 2) takes advantage of the ability of IR light to excite the vibrational modes of the chemical bonds that form the biological matter. Its main generated signal provides rich and concise information about the biochemical composition and the *omics* of the illuminated specimen, fitting all the characteristics to become a useful biomarker of cancer in clinical practice [81].

Unlike other imaging techniques well-established in medical practice, such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) [82, 83], IR spectroscopy is still an emerging technology in the biomedical field. IR spectroscopic techniques have been successfully used since the last century in chemical and pharmaceutical industry, e.g., in the characterisation of polymers and drugs [84, 85]. Some attempts were made since the 1950's to investigate biological specimens, although they are far more complex than industrial material. But it has been mainly in the current century when new advances in instrumentation, data acquisition and data analysis have shown the potential of IR spectroscopy for biomedical applications [86].

Many proof-of-concept studies have demonstrated the diagnostic capabilities of IR spectroscopy in different types of cancer over the past decades [81, 87–89]. This incipient technology potentially offers objective and automated analyses of biological specimens, which could even be adapted to routine screening. Therefore, IR spectroscopy can play a significant role at different levels of cancer diagnosis and research, from *in-vitro* applications such as analysis of cell cultures, *ex-vivo* diagnosis such as histopathology and cytopathology, and even *in-vivo* screening such as non-invasive early detection of skin cancer [90]. Nevertheless, translation to clinical practice still must be justified with statistically relevant studies of increasing complexity that gradually incorporate the current existing variability in the instrumentation, sample preparation, measurement protocol and data processing [91].

1.2.2 MINERVA project

MINERVA is a project funded by the European Commission through its Seventh Framework Programme (FP7) that brings together thirteen partners from across Europe with the common objective of developing mid-IR technology to improve the diagnosis and research of cancer [92,93]. MINERVA pursues several targets in parallel, from developing new IR instrumentation, such as fibre lasers, Acousto-Optic (AO) modulators, supercontinuum sources and detectors in the mid-IR range, to explore the performance and limitations of current IR spectroscopic technology in cancer identification.

The spectral region studied in MINERVA covers the so-called *fingerprint region* (Sec. 2.4.1), which includes the most distinctive absorption peaks of biomolecules.

By studying the pattern of IR spectroscopic signals it is possible to deduce valuable information for disease diagnosis. But this process is not straightforward because the useful molecular information is buried in the interrelated distribution of biological species, which only differ in subtle biochemical changes. In addition, IR spectroscopic signals are affected by different kinds of spectral artifacts, which may arise from the preparation of the sample, the acquisition procedure or other undesirable physicochemical effects. Therefore, advanced mathematical techniques, such as *multivariate analysis* and *machine learning*, are required to extract the relevant information from the large amount of spectral data in order to identify and differentiate cancer.

This thesis has been developed within the framework of MINERVA project. Universitat Politècnica de València (UPV), as a partner of MINERVA, has collaborated in the acquisition, processing and analysis of IR spectra from biomedical samples. In particular, two applications related to cancer diagnosis have been explored in MINERVA:

- *Histopathology*: the main aim is to evaluate the capabilities of IR spectra to discriminate colon cancer in tissue biopsy samples. In this task, UPV has collaborated with University of Exeter (UoE) and Gloucestershire Hospitals NHS Foundation Trust (GHFT), both from the United Kingdom (UK). A significant number of tissue samples, extracted from patients with colon cancer in different pathological conditions, was provided by GHFT. Hyperspectral Fourier Transform Infrared (FTIR) images from all these samples were acquired at UoE by using a state-of-the-art benchtop FTIR microspectroscope. Expert spectroscopists from UoE and GHFT have been in charge of the acquisition and analysis of this relevant dataset of hyperspectral FTIR images. UPV has mainly played a supporting role in this application, by adapting and developing algorithms from the fields of image processing and computer vision in order to fuse the information of hyperspectral FTIR images and conventional microscopic images (Ch. 4).

- *Cytopathology*: the aim is to assess the potential and limitations of IR spectroscopy to distinguish different types of catalogued skin cell lines including melanoma. To accomplish this task, UPV has collaborated with Westfälische Wilhelms-Universität (WWU), from Münster (Germany). WWU provided the

biomedical support by preparing the cell cultures of skin cell lines. UPV has played a more significant role in this application by acquiring most of the hyperspectral FTIR images from the cell cultures at UoE, as well as processing and analysing them (Ch. 5). During this process, UPV counted on the valuable help and supervision of experts from UoE and GHFT.

1.3 Objectives

This thesis aims at contributing to the improvement of the final diagnosis of cancer, a real and relevant problem in the biomedical field, by means of new optical technologies and analytical methodologies. In particular, the main aim is to contribute to the development and assessment of objective decision support systems based on images captured from tissue and cell samples with modern benchtop FTIR microspectrometers. In order to fulfil this global and generic task, several specific objectives are derived:

- Reviewing and synthesising the most relevant concepts from the biomedical field which are essential to correctly deal with pathological samples.
- Reviewing and synthesising the most updated knowledge concerning the technology of IR spectroscopy, including its instrumentation and the physicochemical phenomena involved in the creation of FTIR spectral signals for the special case of biological samples. These concepts are needed to collaborate in the correct acquisition of FTIR hyperspectral images and to perform a sounder processing of the spectral data.
- Reviewing and identifying the most relevant methods and algorithms currently employed for the processing and analysis of the recorded hyperspectral FTIR images. These methods belong to different interconnected areas, such as *image processing*, *computer vision*, *machine learning*, *pattern recognition*, *multivariate analysis*, *chemometrics*, as well as specific techniques from the spectroscopic field.
- Applying all the concepts and methodologies together with preliminary studies to real diagnostic problems based on histopathological and cytopathological

samples. Special emphasis will be given to the correct assessment of the obtained results and the possible limitations of the employed technologies and methodologies.

1.4 Outline

In this chapter the most relevant notions from the biomedical field related to cancer and its final diagnosis have been reviewed. This background helps to understand the necessity for the studied technology. In addition, its knowledge is essential to perform a correct treatment, measurement and analysis of the involved biological samples. The rest of chapters focus on the technological concepts and their applications.

Ch. 2 will present the most relevant theoretical an practical aspects concerning Infrared spectroscopy, which must be considered and understood to proceed correctly in the later applications.

Ch. 3 will describe several techniques connected with the preprocessing of FTIR spectra acquired from biological samples as well as the methods commonly used to extract the most relevant information in these high-dimensional datasets.

Ch. 4 will thoroughly detail the first of the two main applications of this thesis, which aims at helping to improve the diagnosis of histopathological samples. This application consists of the development of a methodology that automatically aligns different histological sections measured by two imaging modalities in order to combine the spatial information extracted from those two techniques. To that end, the theoretical background of FTIR spectroscopy described in previous chapters will be fused with modern methodologies from the fields of *image processing* and *computer vision*.

Ch. 5 will present the second main application of this thesis, this time with the goal of improving the diagnosis of cytopathological samples. Its main objective is to assess the generalisation capabilities of FTIR signals to reliably discriminate different types of skin cells containing malignant phenotypes. The followed methodology, which fuses techniques described in previous chapters with new methods from the areas of *machine learning* and *pattern recognition*, will be detailed together with the main reasoning behind each applied step. In addition, the final results will be quantitatively assessed and thoroughly discussed.

Finally, Ch. 6 will summarise the most relevant conclusions derived from the previous chapters with special attention to the most important findings and limitations identified in the two main applications.

Chapter 2

Fourier Transform Infrared Spectroscopy

Contents

2.1 Introduction

In this chapter, the fundamental concepts of Infrared spectroscopy, its practical instrumentation, and the key ideas related to the signal and image formation will be described. Special emphasis will be made in the practical aspects that will affect the later applications. In particular, the main sources of artifacts and errors that may hamper the posterior analysis of the measured signals will be detailed.

2.2 Infrared spectroscopy

Infrared spectroscopy together with Raman spectroscopy constitute the field of analytical chemistry called *vibrational spectroscopy*. This field has many diverse applications in biomedicine, biochemistry, forensic sciences or food and pharmaceutical industry [94].

IR band covers the *wavelengths* (λ) of the electromagnetic spectrum which are immediately superior to the visible spectrum (λ ∼380-780 nm). The division of the IR band is not precise and can vary depending on the field of application or the related publication. Considering the definitions of spectral bands specified by the International Organization for Standardization (ISO) [95], IR band can be divided into three main regions of increasing wavelength: near-IR (λ ∼780 nm-3 µm), mid-IR (λ ∼3-50 µm) and far-IR (λ ∼50 µm-1 mm). Fig. 2.1 illustrates the most useful bands of the electromagnetic spectrum with an enlarged view of the visible and IR spectrum (wavelength axis is not scaled). In this figure, schematic drawings of objects with different sizes have been added at the position of the approximate scale of wavelength. It must be pointed out that the typical size of a human cell is around tens of microns, which lies in the mid-IR region.

The most common spectral metric in spectroscopy is the *wavenumber* ($\tilde{\nu}$), which is the spatial frequency of a wave and is defined as the inverse of the wavelength:

$$\tilde{\nu} = \frac{1}{\lambda} \tag{2.1}$$

$\tilde{\nu}$ represents the number of wave cycles per unit length and its usual unit is cm^{-1}. Therefore, the mid-IR region defined by ISO [95] extends from 200 cm^{-1} to around 3300 cm^{-1}. It must be remarked that bibliographic references in the field of spectroscopy (e.g., [84, 98–101]) usually delimit the mid-IR region between 400-4000 cm^{-1} (λ ∼2.5-25 µm).

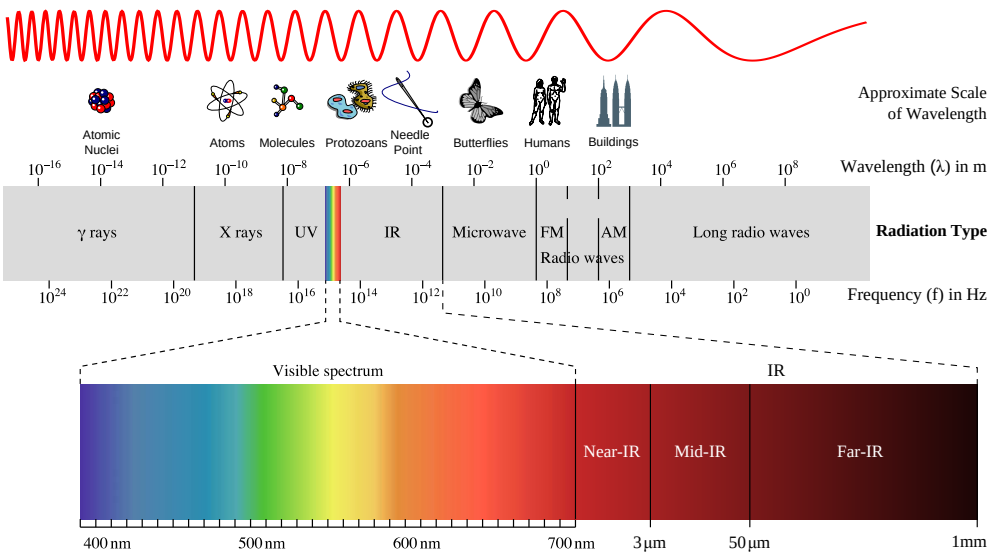


Fig. 2.1: Illustration of the electromagnetic spectrum. Adapted from [96] and [97].

The energy E of a photon is defined by the Planck-Einstein relation:

$$E = hf = h\frac{c}{\lambda} = hc\tilde{\nu} \tag{2.2}$$

where h is the Planck's constant, f is the frequency of the radiation and c is the speed of light in the vacuum. The energy of mid-IR photons extends from 0.025 to 0.4 eV (according to ISO [95]) and covers the quantum values which are able to excite most vibrational modes of matter [98].

Each kind of molecular bond has characteristic modes of vibration whose states of energy are quantised. In a traditional absorption process, a resonance interaction is produced when a photon of the incident light possesses the exact energy of a specific molecular energy level. In that case, that molecule is promoted into a state of higher excitation and the photon is absorbed [101]. Hence, if a sample of matter is placed into the path of a mid-IR beam, its molecules will absorb those photons whose frequencies match their resonant frequencies of vibration. This is the main principle of IR spectroscopy. The fundamental vibrational modes which can be detected in mid-IR spectroscopy are *stretching* (changes of bond lengths), both *symmetric* and

asymmetric, and *bending* (changes of bond angles) modes [81]. These fundamental vibrational modes are sketched in the case of a non-linear triatomic molecule, e.g., water, in Fig. 2.2.

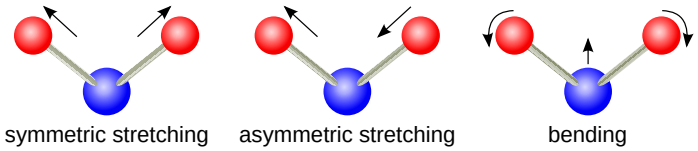


Fig. 2.2: Diagram of the fundamental modes of molecular vibration detectable by mid-IR spectroscopy in the case of a non-linear triatomic molecule, e.g., water.

2.3 FTIR spectrometers

Nowadays, FTIR spectrometers are the most widely used devices to perform IR spectroscopy. A conventional FTIR spectrometer is composed of three main elements: the source, the interferometer and the detector. Although there are many different designs and combinations of these components, their main and most common characteristics will be summarised in the following sections.

2.3.1 Source

The IR light in commercial spectroscopes is commonly generated by *blackbody* sources with broadband spectra. The blackbody radiation is theoretically described by the Planck's law [102, 103]:

$$B_\lambda(\lambda, T_k) = \frac{2hc^2}{\lambda^5} \frac{1}{\exp\left(\frac{hc}{\lambda k_B T_k}\right) - 1} \tag{2.3}$$

where B_λ is the blackbody spectral radiance, as a function of λ and the absolute temperature of the source T_k, and k_B is the Boltzmann's constant. In the International System of Units (SI), B_λ is expressed in $\frac{\mathrm{W}}{\mathrm{sr \cdot m^3}}$. By means of Eq. 2.3, the blackbody spectral radiance emitted by the source can be estimated for some specific values of its absolute temperature, resulting in the curves shown in Fig. 2.3.

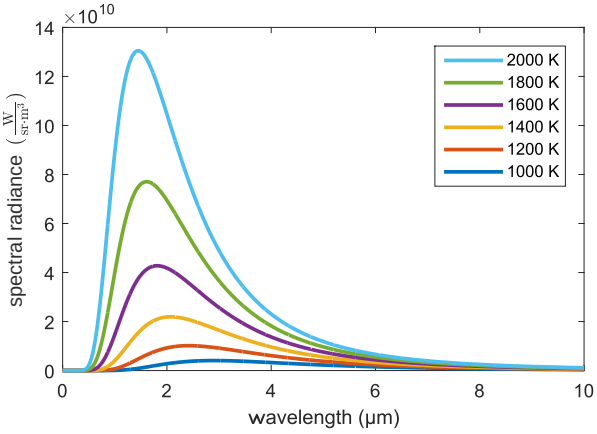


Fig. 2.3: Blackbody spectral radiance at different absolute temperatures.

As can be observed in Fig. 2.3, the overall radiated energy increases with the source's temperature and the maximum peak of the curve moves toward shorter wavelengths (Wien's displacement law). Therefore, although higher spectral radiance is obtained in the IR region at higher temperatures, the proportion of IR radiation compared to the total energy emitted is lower. As a compromise solution, thermal sources operating in the range of 1400 to 2000 K are usually employed [101]. These sources, normally heated by an electric current, consist of a filament or rod made of materials that are relatively inert to atmospheric oxygen at the operating temperature. *Globar* is a patented material of silicon carbide disposed in a diamond lattice, which is widely used as blackbody IR source in FTIR spectrometers. Globar is relatively stable at ∼1300 K (its typical operating temperature [98]) and does not need a special inert atmosphere.

2.3.2 Interferometer

The interferometer is the key element of FTIR spectrometers. Its role is to separate the contributions of the polychromatic spectrum provided by the blackbody source. Commercial FTIR spectrometers typically use Michelson interferometers, whose schematic is shown in Fig. 2.4. The light emitted by the IR source is collimated to get a parallel beam that is directed towards a beam splitter. Half of the radiation is reflected and the other half is transmitted by the beam splitter. The reflected

split beam is reflected again by a fixed mirror, meanwhile the portion of transmitted beam is reflected by a mirror that is continuously moving with a specific frequency and amplitude. Both beams are directed back to the beam splitter where they are recombined. Again, due to the construction of the interferometer, the beam splitter conducts half of the recombined beam towards the source (this part of the light is *lost* for the final measurement and it is not drawn in Fig. 2.4 for clarity) and the other half follows the path towards the sample and the detector.

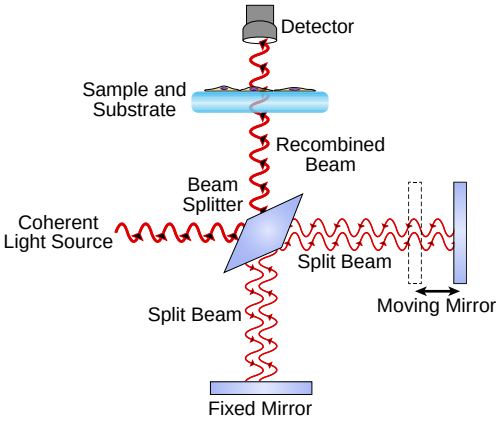


Fig. 2.4: Schematic diagram of a Michelson interferometer configured for FTIR spectroscopy. Adapted from [104].

The split beams travel along paths with different lengths due to the movement of one mirror, making changes in the phases of the beams. Constructive or destructive interference patterns are produced when the beams are recombined, depending on their path difference or phase shift. The final intensity pattern of the recombined beams measured by the detector is called *interferogram* $J(\xi)$, which is a function of the moving mirror displacement ξ. When a broadband source emitting light in a continuous way at all wavenumbers (or wavelengths) is used to illuminate the sample, $J(\xi)$ can be represented by the following integral:

$$J(\xi) = \int_{-\infty}^{\infty} I(\tilde{\nu}) \cos(2\pi\tilde{\nu}\xi) d\tilde{\nu} \tag{2.4}$$

where $I(\tilde{\nu})$ is the spectral intensity passing through the sample and reaching the detector at a specific wavenumber $\tilde{\nu}$. Eq. 2.4 arises when the interference pattern of

the recombined waves is studied [98, 101]. The integral of Eq. 2.4 has the form of a Fourier cosine transform, which is the real part of a full complex Fourier Transform (FT). $I(\tilde{\nu})$ can be computed through the inverse FT, resulting in the integral (apart from normalization constants [105], specified by the symbol $\propto$):

$$I(\tilde{\nu}) \propto \int_{-\infty}^{\infty} J(\xi) \cos(2\pi\tilde{\nu}\xi) d\xi \tag{2.5}$$

Thanks to the duality between different domains granted by the FT, the spectrum $I(\tilde{\nu})$, which is an intensity distribution in the wavenumber domain $\tilde{\nu}$, can be obtained from the interferogram $J(\xi)$, which is the intensity distribution actually recorded by the detector and which depends on the moving mirror position ξ. This is the main property of FTIR spectrometers.

Fig. 2.5 presents an example of interferogram and its associated intensity spectrum measured without any sample placed in the beam path between the interferometer and the detector. Only the values and units of $\tilde{\nu}$ in the intensity spectrum $I(\tilde{\nu})$ have been specified to identify some relevant absorption bands. As the plots correspond to a measurement with no sample, these absorption bands are mostly produced by the water vapour and CO_2 which the IR beam encounters on the air along its travel from the source to the detector (Sec. 2.6.2).

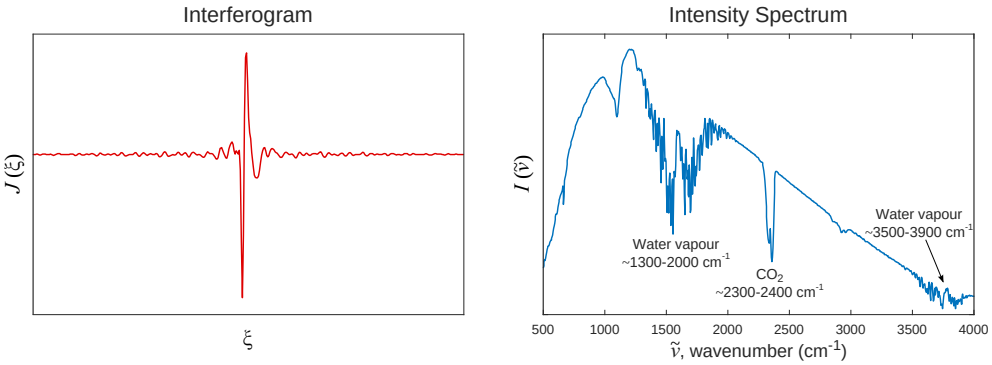


Fig. 2.5: Example of a recorded interferogram (left) and its associated intensity spectrum computed by applying the Fourier transform (right). These graphics correspond to a measurement with no sample. Some relevant absorption bands have been identified in the intensity spectrum. Adapted from [106].

2.3.3 Detector

Detectors for IR region are different from those designed for the visible spectrum or other higher frequency regions because the energy of IR photons is lower. The detectors used in FTIR spectroscopy can be classified in two classical categories [107]: *thermal detectors* and *quantum detectors*.

Thermal detectors measure the heating effect produced in a material when it absorbs IR radiation. Depending on the effect of the temperature change, thermal detectors can also be divided into distinct subcategories: *thermocouples*, which measure an electromotive force or voltage in a junction of different metals; *bolometers* and *thermistors*, which respectively detect changes in the resistance of a conductor or a semiconductor; *pneumatic detectors*, which measure the increase of pressure in an enclosed gas; and *pyroelectric detectors*, which monitor metals behaving as a capacitor whose charge on the surface changes with the temperature. All kinds of thermal detectors have been used for mid-IR spectroscopy in the past. However, the main drawback of thermal detectors is their response time (around several milliseconds), which is too long for the high frequencies of modern FTIR spectrometers [98]. Nowadays, the only thermal detectors that have a practical use are made of a pyroelectric material called Deuterated Triglycine Sulfate (DTGS). DTGS detectors are equipped in low-cost, medium-performance devices and can be operated at room temperature.

Quantum detectors respond to the direct interaction of IR photons with the electrons of the detector material, which can be excited to a higher energy state. This excitation only takes place if the energy of IR photons is higher than a certain threshold E_{min} determined by the material properties. As the photon energy is directly proportional to the wavenumber (Eq. 2.2), the lower the radiation wavenumber, the higher the number of IR travelling photons for a given amount of total energy. Because more electrons are excited by IR photons, the sensitivity of quantum detectors increases when the wavenumber decreases. However, there is a limit $\tilde{\nu}_{min}$ related to E_{min} from which the response drops off abruptly. Quantum detectors can be generally subdivided into *photoemissive detectors* and *semiconductor detectors*.

Photoemissive detectors directly measure the number of electrons released from an illuminated surface. This kind of detectors, which include phototubes and

photomultipliers, cannot be used with IR radiation because they need higher energies to release electrons from the photoemissive surfaces.

Semiconductor detectors are composed of materials with two bands of different energy, namely valence band and conduction band, which are separated by a band gap. If the energy of an incident photon is greater than the band gap, an electron will jump from the valence band into the conduction band and the resistance of the detector will diminish. These devices commonly work as photodiodes and the change in resistance is measured as a variation in the current across the detector. Normally, materials with band gaps between 400 and 800 cm^{-1} are used in FTIR spectroscopy [101]. Mercury Cadmium Telluride (MCT), a mixture of HgTe and CdTe, is the material used in the most common semiconductor mid-IR detectors. MCT has high sensitivity, fast response times ($\sim$ µs) and covers a spectral range from about 700 to over 5000 cm^{-1}. These types of detectors are normally cooled to 77 K with liquid nitrogen because the thermal energy at room temperature (around 208 cm^{-1}) produces unwanted electron promotions to the conduction band, which increase the noise in the measurement.

2.4 FTIR spectra

There are different ways of measuring the interaction of IR light with the sample. Three main sampling strategies or modes are normally applied in the analysis of biological materials by FTIR spectroscopy: transmission, transflection and Attenuated Total Reflection (ATR). Each mode has distinct advantages and disadvantages [108], and its choice depends on the type of study and the available resources. In this thesis, all the described and analysed FTIR measurements were acquired in transmission mode.

In transmission measurements, the detector mainly receives the light which passes through the sample, as was sketched in Fig. 2.4. Transmission is the most straightforward and traditionally used sampling technique. It does not require additional equipment to perform the measurements and allows quicker acquisitions. Moreover, it is exempt from additional artifacts directly caused by reflection but does not avoid scattering effects associated with transmission. Its main drawbacks are the need for thin sections of samples, the complexity of the sample preparation and the cost of the required auxiliary materials.

The biological sample, such as fixed tissues or cells, must be specifically prepared for transmission measurements and laid on a transparent substrate. In particular, the substrate used in transmission commonly consists of *windows* made of CaF_2 (calcium fluoride). CaF_2 is practically transparent from 1000 to 77000 cm^{-1} and its refractive index is relatively low and similar to biological substances (1.4 at 2000 cm^{-1}) [109], which reduces the light loss at the window-sample interface. Moreover, CaF_2 is suitable for direct use in preparations of biological samples, such as cell cultures, because it is chemically inert and water insoluble [84]. The principal inconvenience of CaF_2 is its cost, which can range from around 10 to more than 100 € for each window [110], depending on its size and quality.

The final shape of the intensity spectrum computed from the recorded interferogram is determined by the convolution of the source emission profile, the detector sensitivity profile and the response of any optical component in the beam path [101]. *Transmittance* ($T(\tilde{\nu})$) is the metric used to extract the attenuations caused by the molecules of the sample in the IR beam and (ideally) cancel out the rest of contributions. It is defined as the following ratio [111]:

$$T(\tilde{\nu}) = \frac{I(\tilde{\nu})}{I_0(\tilde{\nu})} \tag{2.6}$$

where $I(\tilde{\nu})$ is the transmitted intensity spectrum passing through the sample (including the substrate) and reaching the detector, and $I_0(\tilde{\nu})$ is the transmitted intensity spectrum reaching the detector in the absence of sample. $I_0(\tilde{\nu})$ is known as the *reference background spectrum* and is commonly measured before $I(\tilde{\nu})$ by selecting a clean region of the substrate, as illustrated in Fig. 2.6.

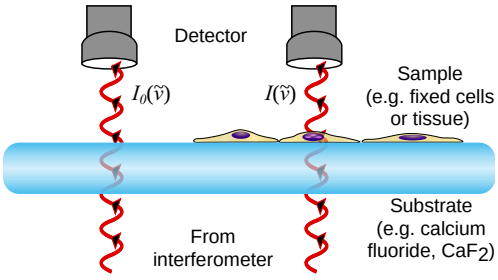


Fig. 2.6: Schematic of the acquisition of the reference background spectrum $I_0(\tilde{\nu})$ and the intensity sample spectrum $I(\tilde{\nu})$ in transmission measurements.

2.4.1 Absorption spectrum and Beer-Lambert law

The higher the amount of radiation absorbed by the sample, the lower the intensity $I(\tilde{\nu})$ reaching the detector and, as can be checked in Eq. 2.6, the lower $T(\tilde{\nu})$. In order to operate with a direct relationship of the absorption level, the metric called *absorbance* ($A(\tilde{\nu})$) is usually employed. For a specific $\tilde{\nu}$, $A(\tilde{\nu})$ is defined as follows [111]:

$$A(\tilde{\nu}) = \log_{10}\left(\frac{1}{T(\tilde{\nu})}\right) = -\log_{10}\left(\frac{I(\tilde{\nu})}{I_0(\tilde{\nu})}\right) \tag{2.7}$$

Both $T(\tilde{\nu})$ and $A(\tilde{\nu})$ are dimensionless, as they are a ratio and the logarithm of a ratio between intensities. The curve of absorbance values for a range of measured wavenumbers is called *absorption spectrum*. As an example, Fig. 2.7 shows $I_0(\tilde{\nu})$, $I(\tilde{\nu})$ and $A(\tilde{\nu})$ spectra acquired from a region of a fixed skin cell as was sketched in Fig. 2.6. These spectra are presented both in the whole range of measured wavenumbers (1000-3800 cm^{-1}) and cropped to the *fingerprint region* (see below).

Absorbance is an experimental quantity that can be ideally related to the theoretical physicochemical properties of the sample by the *Beer-Lambert-Bouguer law* or commonly called the *Beer-Lambert law* [112, 113]:

$$A(\tilde{\nu}) = \ell \cdot C \cdot \varepsilon(\tilde{\nu}) \tag{2.8}$$

where ℓ is the effective optical path length (length which the light beam travels through the sample), C is the concentration of the sample and $\varepsilon(\tilde{\nu})$ is the molar attenuation coefficient or absorptivity. When the sample is a mixture of J chemical components or constituents, which can absorb at the same $\tilde{\nu}$ and have constant absorptivity along the sample thickness, the total absorbance is given by the equation:

$$A(\tilde{\nu}) = \ell \cdot \sum_{j=1}^{J} C_j \cdot \varepsilon_j(\tilde{\nu}) \tag{2.9}$$

where the subscript j identifies each chemical component or constituent. In the previous equations, some complex optical phenomena have been neglected, such as scattering due to inhomogeneous samples, reflections at the substrate-sample interface, chemical changes due to intermolecular interactions and other additional

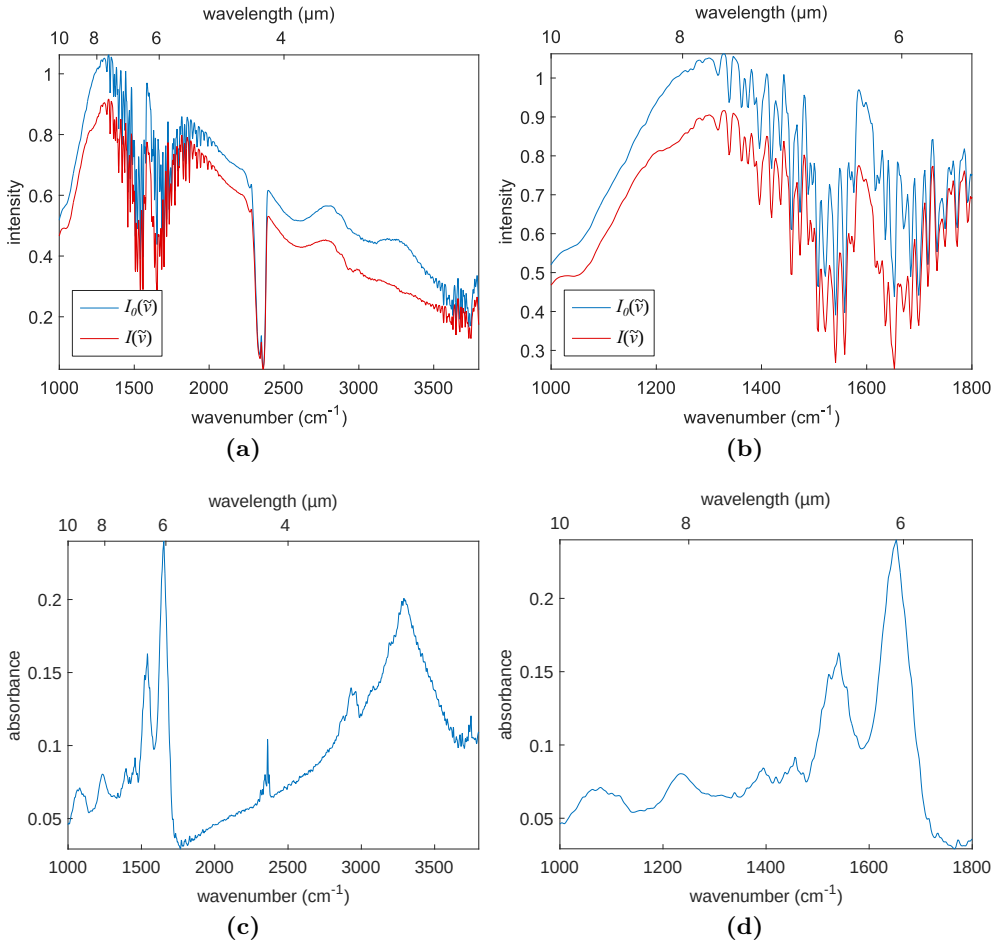


Fig. 2.7: FTIR spectra from a region of a fixed skin cell. Top plots: reference background spectrum $I_0(\tilde{\nu})$ (blue) and intensity sample spectrum $I(\tilde{\nu})$ (red) for (a) the whole range of recorded wavenumbers and (b) the *fingerprint region*. Bottom plots: corresponding absorption spectrum $A(\tilde{\nu})$ for (c) the whole range of recorded wavenumbers and (d) the *fingerprint region*.

effects which may take place in the interaction of the light beam with the sample [113]. However, it highlights that the recorded absorbance is mainly determined by the morphology of the sample (through ℓ), the chemical abundance of its components (through C) and their molecular transitions (through $\varepsilon(\tilde{\nu})$). Note that both ℓ and

C do not depend on $\tilde{\nu}$ and produce *multiplicative effects* in the global shape of the absorption spectrum, whose main peaks are determined by $\varepsilon(\tilde{\nu})$.

Biological materials are made of organic macromolecules, which can be classified as proteins, lipids, carbohydrates and nucleic acids. These macromolecules are composed of different chemical bonds that create structures more complex than non-biological compounds like industrial polymers. Thus, the final absorption spectrum of a biological specimen reflects the overall attenuations of this big mixture of biomolecules, which have strongly overlapping characteristics. Despite this chemical complexity, some relevant peaks can be related to some vibrational modes in a typical biological absorption spectrum, as shown in Fig. 2.8.

Huge efforts have been made in traditional FTIR spectroscopy to identify the most characteristic peak frequencies in biological studies [100, 114]. Most of these relevant peaks are located within the so-called *fingerprint region* [81], which approximately extends from 1000 to 1800 cm^{-1}. This range also contains the strongest peak of a typical biological spectrum, the Amide I peak ($\sim$1640-1680 cm^{-1}), which is associated with the carbonyl stretching of the peptide bond in proteins (Fig. 2.8).

2.4.2 Limitations to the Beer-Lambert law

There are several factors that may cause deviations from the ideal linear relationships with the measured absorbance proposed by the Beer-Lambert law (Eqs. 2.8 and 2.9). Although not exhaustive, a list of some of these factors is presented below [112, 113]:

- *Real limitations*: the Beer-Lambert law was designed to describe the behaviour only of dilute solutions. Therefore, other kinds of samples, such as solid mixtures, may not follow this law.
- *Chemical deviations*: the Beer-Lambert law supposes that the chemical absorbers (molecules, ions, etc.) act independently of each other. On the contrary, intermolecular interactions between chemical species may cause changes in the absorptivity that lead to deviations from the Beer-Lambert law. These effects are likely to occur in biological samples, which have complex chemical compositions and structures.
- *Instrumental and measurement deviations*: several sources of deviations and limitations due to measurement causes can be distinguished:

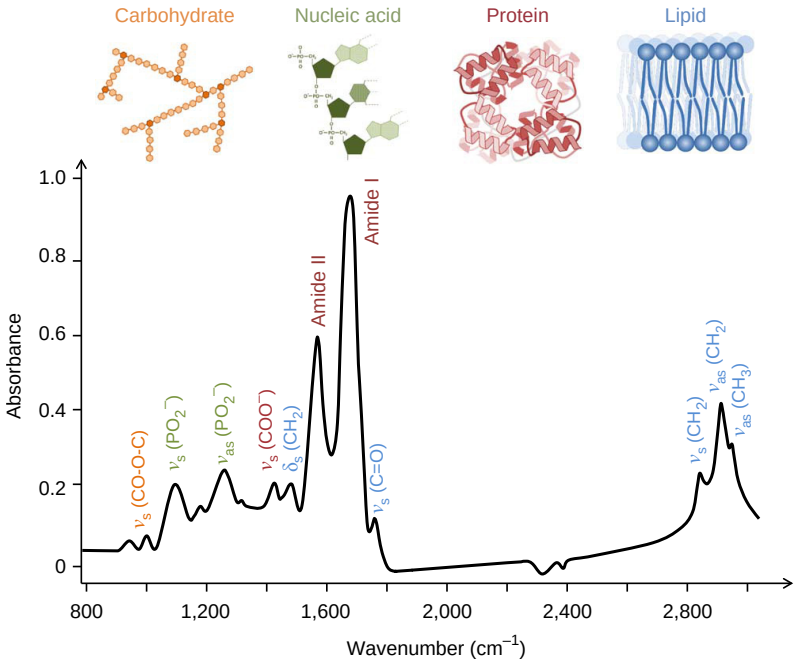


Fig. 2.8: Typical FTIR absorption spectrum of a biological sample in the range 800-3000 cm^{-1}. Some relevant peaks have been identified, which are associated with vibrational modes (ν_s = *symmetric stretching*, ν_{as} = *asymmetric stretching*, δ_s = *symmetric bending*) of different organic macromolecules (see respective colours). Adapted from [108] with permission by courtesy of Macmillan Publishers Ltd.

– *Polychromatic radiation*: strictly speaking, the Beer-Lambert law only applies when the incident light is monochromatic. In the case of using polychromatic radiation, as the one that blackbody sources provide, departures from the Beer-Lambert law will be higher when the differences of absorptivity between wavenumbers increase.

– *Stray light*: it comprehends all the radiation measured by the sensor that was not intended in the original design. This radiation does not pass through the sample and often results from scattering and reflection off the surfaces of the experimental components or the surrounding environment. In this case, the measured absorbance is called *apparent absorbance* ($A_{app}(\tilde{\nu})$) and is lower than the true absorbance of the sample, which limits the dynamic range of the system. These deviations are more significant at

higher absorbance values. In the presence of stray light of intensity $I_s(\tilde{\nu})$, $A_{app}(\tilde{\nu})$ for a specific wavenumber is defined as follows:

$$A_{app}(\tilde{\nu}) = -\log_{10}\left(\frac{I(\tilde{\nu}) + I_s(\tilde{\nu})}{I_0(\tilde{\nu}) + I_s(\tilde{\nu})}\right) \tag{2.10}$$

– *Inhomogeneous or scattering medium*: this is one of the most important factors to disobey the Beer-Lambert law. Uneven optical properties or thickness both in the analysed sample or the substrate, as well as the presence of scattering centres, significantly modify the effective path length and induce severe deviations from the Beer-Lambert law.

– *Non-linearity of the detector*: the non-linearity of the response is more important in MCT detectors. The effect in the recorded spectrum is very similar to that of stray light and the accuracy of the quantitative measurement also decreases with higher absorbance values.

Although the restrictive conditions of the Beer-Lambert law are hardly fulfilled in many situations, the linear relationships with the optical path length and the concentration are used as a guide to develop some preprocessing and analysis methods (Ch. 3).

As last remark, it is worth clarifying the subtle difference between *absorption spectrum* and *absorbance spectrum*, which are sometimes mixed up. The absorption spectrum normally only refers to the physicochemical process that occurs when the molecules of the sample absorb the incident light, which actually would follow the Beer-Lambert law. The absorbance spectrum, on the other hand, includes absorption and others phenomena that attenuate the incident light, such as reflection, scattering, and other physical processes that cause deviations from the Beer-Lambert law. Despite of this difference, both terms are symbolised by $A(\tilde{\nu})$ because the absorbance spectrum is the one actually measured in real situations, whereas the ideal absorption spectrum is commonly unknown.

2.5 Micro-FTIR spectroscopic imaging

FTIR microspectroscopy arises when coupling a spectrometer with a microscope whose optical components are adapted to the mid-IR region [101, 115]. Fig. 2.9a illustrates the configuration of the main elements present in a common microspectrometer and Fig. 2.9b shows a picture of a modern commercial microspectrometer.

Fig. 2.9: (a) Schematic of main components in a common microspectrometer (spectrometer placed at left and microscope at right). Reproduced from [116] with permission by courtesy of SAGE publications Inc. (b) Photograph of a modern commercial microspectrometer (in this case, spectrometer placed at right and microscope at left). Reproduced from [117].

Basically, the microscope is an additional arrangement of optical components located between the interferometer and the detector whose main task is to redirect the IR light leaving the interferometer just into small regions of the sample. The core elements to accomplish this task are a pair of optical components called *objective* and *condenser*. Almost all FTIR microscopes are equipped with a *Schwarzchild* objective and condenser [118], which are a special type of *Cassegrain* objectives. Schwarzchild objective and condenser are composed of a convex and a concave mirror which are concentric. This concentricity grants excellent imaging characteristics over a surprisingly wide Field of View (FOV) [85].

As outlined in Fig. 2.10a, the condenser focuses the IR beam, coming from the interferometer and properly reflected by several mirrors, into a small region of the sample. The sample is mounted on a motorised stage which has a hole to let the IR light pass through freely and whose position can be modified in three directions. The

direction perpendicular to the stage (z axis) can be adjusted to place the sample into the beam focus and the other two directions in the stage plane (x and y axes) allow to choose the specific illuminated region. The objective gathers the light transmitted through the substrate and sample and refocuses it to continue its travel towards the detector. A small part of this light is guided to a traditional optical viewer and a Charge-Coupled Device (CCD) camera, which provide a visible image of the sample. This visible image is used during the acquisition to define the region to measure and to focus the sample correctly. Most of the light leaving the Schwarzchild objective is reflected by a series of mirrors and passes through an adjustable aperture that can crop the Region of Interest (ROI) finally measured by the detector.

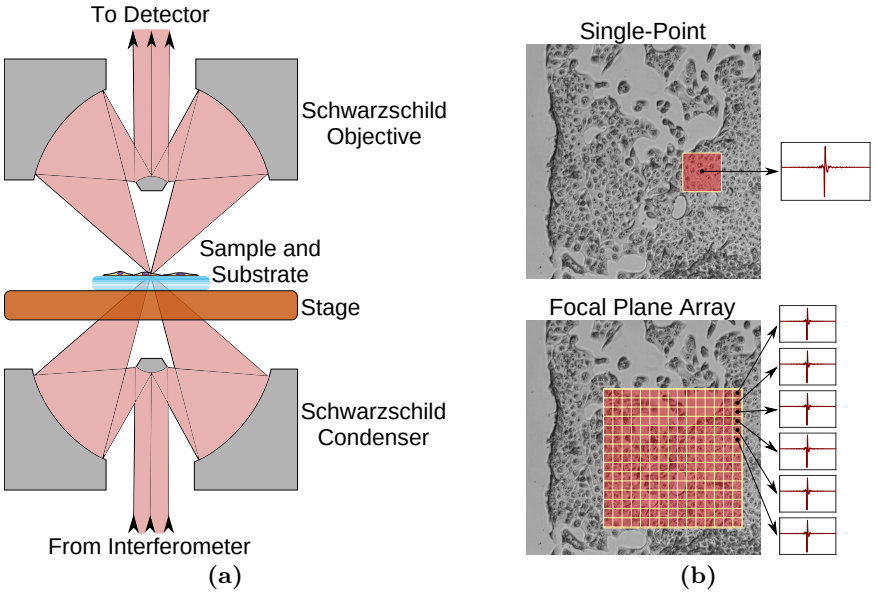


Fig. 2.10: (a) Sketch of the functioning of a Schwarzschild objective and condenser configured in transmission mode. (b) Sketch of the regions covered with a single-point (up) and a Focal Plane Array detector (down) in a cellular sample.

Two main types of detectors are commonly used to perform measurements of IR spectra with micrometric resolution, namely *single-point* and *Focal Plane Array (FPA)* detectors (Fig. 2.9a). Single element detectors have been employed since the beginning of FTIR microspectroscopy by the middle of the last century [119]. The adjustable aperture plays a crucial role with single-point detectors because it

is responsible for refining the spatial localisation of the measured ROI. As sketched in Fig. 2.10b, in single-point measurements only one interferogram is acquired from a squared region of the sample whose side is typically around 25-50 μm.

In the past, single-point detectors were used to create hyperspectral images by mapping the sample point-by-point. However, with the introduction of FPA detectors, originally designed and developed for military applications, the collection time of images drastically decreased [120]. First successful applications of FPA detectors were reported around 20 years ago [121,122]. Since then, FTIR microscopes equipped with multichannel detectors have been the standard to perform IR imaging [116]. Focal Plane Array detectors, also called multichannel detectors [123], are squared grids of multiple small sensing elements, commonly known as *pixels*. Each one of these pixels is capable of capturing an interferogram from a small region of the sample during a single acquisition (Fig. 2.10b). Both single-point and FPA detectors are made of the materials described in Sec. 2.3.3. In particular, MCT sensors require liquid nitrogen for cooling and hence depleting the level of noise (Fig. 2.9a).

FPA detectors can provide spatially-resolved measurements in the form of *hyperspectral images* in relatively short times of acquisition. As illustrated in Fig. 2.11, these hyperspectral images have two spatial dimensions (x, y) that indicate the position of the pixels and a spectral dimension $\tilde{\nu}$ corresponding to the wavenumbers of the measured absorbance spectra.

All the FTIR images analysed in this thesis were captured with an Agilent 620 FTIR microscope coupled to an Agilent 670 FTIR spectrometer with a Globar® light source (Fig. 2.9b). This device has a FPA detector composed of 128×128 (16384) pixels, each one covering around 5.5×5.5 μm^2 in the sample (see next section) when using 15× Cassegrain reflective condenser and objective (NA = 0.62) and jointly covering an effective FOV of 704×704 μm^2 per frame. Normally, the hyperspectral images from contiguous areas of sample, recorded in successive FPA acquisitions, are stitched together in order to augment the total FOV. As a result, FTIR hyperspectral images are information-rich entities containing huge loads of data, easily accounting for several gigabytes.

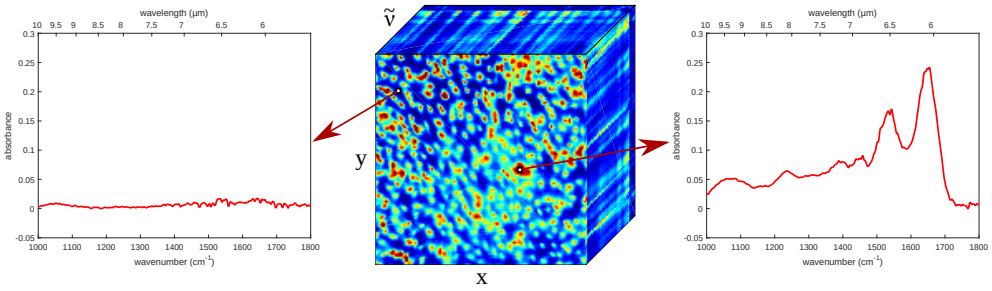


Fig. 2.11: Example of FTIR hyperspectral image (centre), which has two spatial dimensions (x, y) and a spectral dimension $\tilde{\nu}$. The absorbance spectra of two selected pixels are also presented: one corresponding to an empty region of substrate (left) and the another one corresponding to a region of a skin cell (right).

2.5.1 Spatial resolution

An important factor in any microscopy technology is *spatial resolution*, and more specifically *lateral spatial resolution*. Several definitions of this concept exist depending on the specific research community [124]. Roughly speaking, spatial resolution is normally defined as the minimum distance at which two close points of the same intensity can be clearly distinguished as two different points in the image. Mathematically, this minimum distance d is commonly defined by the Rayleigh's criterion as a function of the wavelength λ or the wavenumber $\tilde{\nu}$:

$$d = \frac{0.61 \cdot \lambda}{\mathrm{NA}} = \frac{0.61}{\mathrm{NA} \cdot \tilde{\nu}} \tag{2.11}$$

where NA is the numerical aperture of the optical setup, normally the objective, which is determined by its optical properties. Following this equation, several curves of d as a function of the wavenumber $\tilde{\nu}$ and the wavelength λ for different values of NA have been represented in Fig. 2.12. These NA values correspond to different objectives that can be chosen for the microspectroscope used for the measurements in this thesis [117]. In particular, all the samples that will be analysed were measured with an objective with $\mathrm{NA} = 0.62$.

As can be observed in Fig. 2.12, better resolution (lower d) can be achieved with higher NA and higher wavenumbers (lower wavelengths). This theoretical behaviour is normally followed in practice, although the final practical resolution

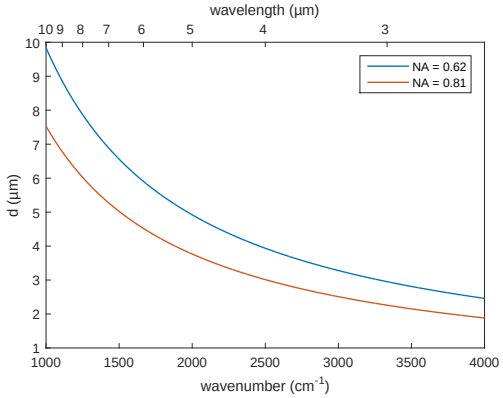


Fig. 2.12: Rayleigh lateral spatial resolution d as a function of the wavenumber and wavelength for several values of numerical aperture NA.

is not only determined by the used objective, but by the total optical system (e.g., internal optical elements). Therefore, the actual resolution of a specific system can only be experimentally measured. A common methodology is using a USAF 1951 resolution target, which has geometrical patterns (e.g. vertical and horizontal bars) at different scales and separated by calibrated distances [117, 124–126]. This target is normally composed of a glass substrate at which a chromium test pattern is vacuum-deposited [124].

Resolution is normally wrongly mixed up with the concept of *effective pixel size* on the sample. The effective pixel size, also called *nominal pixel size* or simply *pixel size*, is mainly determined by the size of the individual elements in the FPA detector and all the optical components of the system (not only the objective), which determine the total system magnification [117, 125]. The pixel size is inversely related to the spatial sampling rate, which is determined by the Nyquist criterion [126]. In order to fulfil this criterion and not lose spatial information, the pixel size must be smaller than the practical resolution. Anyway, the manufacturers of microspectroscopes and objectives normally provide only the information of the pixel size on the sample and favour this confusion between concepts. It must also be remarked that as the actual resolution is not known, the pixel size has been used as a guide for different tasks performed in this thesis. For instance, for the comparison of the spatial dimension between different microscopy modalities (Ch. 4) or for the representation of scale bars in the microscopic images.

2.5.2 Synchrotron light sources

Synchrotron radiation is a very bright broadband light, ranging from the X-ray to the IR region (Fig. 2.1), which is emitted from electrons travelling at relativistic velocities close to the speed of light. These extremely high velocities are achieved by accelerating the electrons into circular trajectories through sequences of magnets. Nowadays, synchrotron sources are built in large expensive facilities and their use is mainly focused on research.

There are several synchrotron facilities around the world, but to access them a well-defined research plan must be justified. Synchrotron radiation has been employed in FTIR microspectroscopy for around two decades [127] and several biomedical applications were reported shortly after [128–130].

The main advantage of synchrotron sources is their high brightness (about two orders of magnitude) compared with common thermal sources [131]. This high flux of photons makes possible to collect spectra with a very high Signal-to-Noise Ratio (SNR) from regions around several microns (e.g., 3×3 μm^2), achieving sub-cellular resolution. However, its main drawback is that synchrotron beams are normally confined into very narrow areas of around 10×10 μm^2 and, typically, they are only used in single-point measurements. Therefore, it takes much time to acquire a hyperspectral FTIR image with a small FOV by sequentially scanning with the single-point detector and the synchrotron beam.

Most of the pioneering work in biomedical FTIR microspectroscopy was accomplished by using synchrotron sources. These works comprehend the reference methodologies and theoretical background that is being extrapolated to thermal sources. Recent advances in optical elements have made possible to develop FTIR microspectrometers, with thermal sources and FPA detectors, capable of collecting high-spatial resolution images with a nominal pixel size up to 1.1×1.1 μm^2, similar to that of a synchrotron [125].

Although the spectral SNR of these commercial devices is still lower than synchrotron's, they allow to measure larger FOV in shorter times and in a common laboratory. In this thesis, all the analysed spectral measurements were collected with commercial devices equipped with thermal sources, but some reference studies (which will be properly cited) used synchrotron sources.

2.6 Artifacts, anomalies and common errors

Artifacts, *anomalies* and *errors* are three nouns normally used to denote those spectral features whose presence implies that the data are imperfect, that the data were derived in an imperfect way or that the data are impure [132]. The word *anomaly* also implies *ignorance* or *lack of knowledge* and normally refers to those unexpected phenomena which cannot be explained with current existing theories [133]. In fact, historically, IR spectroscopy has been mostly an empirical technique whose technology has been always one or several steps ahead of its theoretical background, which has been driven by the experimental findings.

Each new imaging technique introduces a new set of artifacts that may interfere with clinical diagnosis [134]. The underlying mechanisms that generate such artifacts should be identified in order to avoid incorrect interpretation or misleading quantitative analysis of spectra. Nevertheless, the identification of a spectral feature as an artifact is not easy and requires a deep knowledge of the spectroscopic technology and the underlying physicochemical phenomena. In addition, the relevance of an error or artifact will finally depend on the specific application.

The presence of some artifacts can be avoided, or at least minimised, before and during the acquisition of the spectral data, e.g., during the design of the experimental setup or during the preparation of samples. Nevertheless, different types of artifacts appear in the final recorded spectra and they can only be corrected for by some kinds of computerised processes (Sec. 3.2). Several lists of artifacts have been identified in FTIR spectroscopy [132, 135], containing even up to 50 error sources [136]. They range from stochastic measurement noise to various sources of systematic errors, such as non-linear instrument responses, shift problems and interfering effects of undesired chemical and physical variations [137]. Here, only the most relevant artifacts and sources of error encountered during the development of this thesis will be described, which have been grouped under several main topics: instrument, environment, substrate, contamination, scattering and microscopy measurements.

2.6.1 Instrument

The contribution of the instrument is called the *instrument response function* [99]. This function is determined by the convolution of the source emission spectrum (Fig. 2.3), the spectral sensitivity of the detector and the rest of optical components

in the spectrometer, which define the overall or broad shape of the intensity spectra (Figs. 2.5 and 2.7). In the case of the employed spectrometer, this broad shape increases with the wavenumber until reaching a maximum around 1200-1300 cm^{-1} and then decreases monotonically. This suggests that the higher signal levels are recorded approximately in the fingerprint region ($\sim$1000-1800 cm^{-1}).

The source and the rest of optical elements normally have a relatively constant spectral behaviour during time, but their variations can cause unexpected baseline curvatures in $A(\tilde{\nu})$ [99]. In contrast, the detector response may experiment more significant fluctuations during time. In the case of MCT detectors, these fluctuations are constrained as far as the volume of liquid nitrogen in the cooling deposit is enough to maintain a constant cryogenic temperature of 77 K, although some electrons may still jump the band gap and produce random noise (Sec. 2.3.3).

In order to soften the instrumental fluctuations, the interferograms are remeasured in the same region several times and averaged or *co-added* before computing the intensity spectra by Fourier transform. These repeated measurements are called *scans* and are normally carried out in powers of 2, e.g., 32, 64, 128 or 256 scans. As the number of scans increases, the instrument interferences (mainly present as high frequency random noise) are more negligible in $A(\tilde{\nu})$. In some cases, a higher number of scans are co-added when measuring $I_0(\tilde{\nu})$ than when $I(\tilde{\nu})$ is recorded because each $I_0(\tilde{\nu})$ measurement is used as a reference for several regions of the same sample. For instance, in Fig. 2.7, 256 scans were co-added to compute $I_0(\tilde{\nu})$, meanwhile 128 scans were averaged to obtain $I(\tilde{\nu})$.

2.6.2 Environment

Some parts of spectrometers such as the interferometer are normally isolated from the room environment, but other parts such as the analysed sample can be in contact with the room atmosphere. In a laboratory of biospectroscopy, this atmosphere is normally controlled to keep a nearly constant temperature and humidity, but the air existing in the room is regular with no special filtrations. Therefore, the IR beam encounters different gasses whose molecules can absorb mid-IR photons and have influence in the final measurements. In particular, the interferences caused by CO_2, and especially by water vapour, are major problems in FTIR spectroscopy.

Some spectrometers have an isolation chamber to introduce the sample, which is purged with dry air having a very low water content. Fig. 2.13 shows an intensity spectrum recorded after isolating and purging the sample chamber with dry air (blue), which lacks the peaks associated with water vapour that can be observed in the spectrum recorded without purging (red). Fig. 2.14 also presents an absorption spectrum from a different un-purged spectroscope cropped to the range 1300-1800 cm^{-1}. Minimising the presence of water vapour in the isolated atmosphere requires much time (around one hour) and impedes the manipulation of the sample during the isolation. Some spectrometer manufacturers get rid of the isolation chamber and try to alleviate the problem of water vapour by directly purging the environment surrounding the sample with dry air. This system is far from being efficient and the water vapour peaks appear in the recorded spectra, as can be checked in the spectra that were presented in Fig. 2.7.

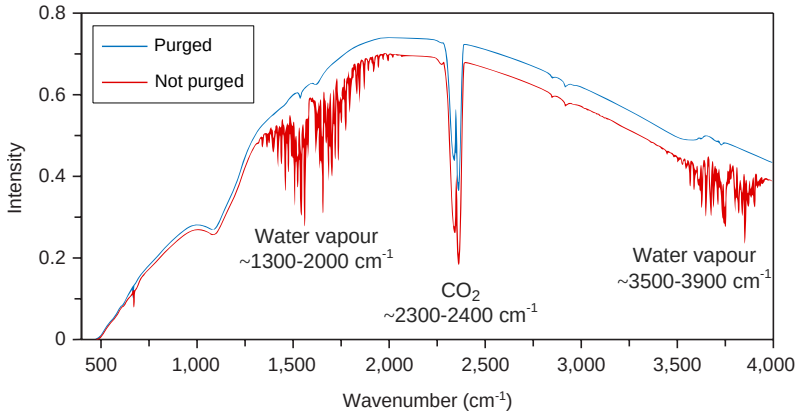


Fig. 2.13: Water vapour intensity spectrum. Intensity spectrum recorded when purging (blue) and not purging (red) the isolated sample compartment with dry air. Adapted from [138] with permission by courtesy of Macmillan Publishers Ltd.

Carbon dioxide (CO_2) has a strong absorption band quite concentrated in the range 2300-2400 cm^{-1}, where no relevant biomolecular absorbers exist. On the contrary, water vapour molecules have a more complex vibrational mechanism and their absorption bands occupy a broad spectral range, from about 1300 to 2000 and from 3500 to 3900 cm^{-1}. The attenuation peaks induced by the water vapour are weaker than those of CO_2 and their pattern resembles high frequency noise. As can be observed in Fig. 2.14 and also in Figs. 2.5, Fig. 2.7 and 2.13, the first range of water

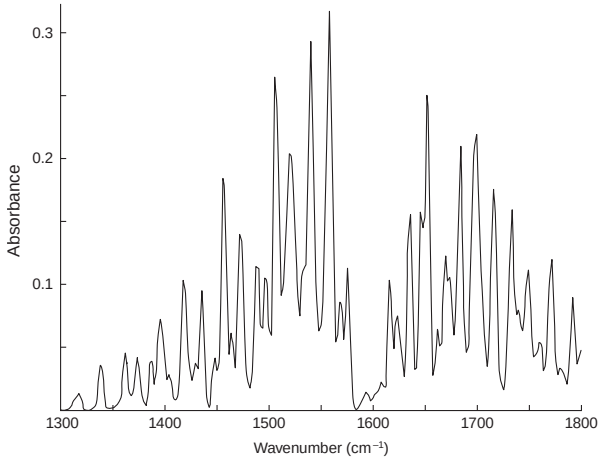


Fig. 2.14: Absorption spectrum of water vapour at room temperature and ambient humidity in the range 1300-1800 cm^{-1}, measured through an un-purged infrared microscope with a path length of about 1 m. Adapted from [101] with permission by courtesy of John Wiley & Sons Inc.

vapour absorption overlaps the fingerprint spectral region and is particularly stronger in the Amide I and Amide II peak regions (Fig. 2.8). Due to the gaseous nature of water vapour, the interaction of its rotational and vibrational transitions create a complex absorption spectrum that strongly depends on the local temperature and humidity, and which is extremely difficult to model during the spectral preprocessing of the recorded data [101].

Biological samples with low concentrations, such as cells, produce relatively low absorption signals. Therefore, their spectra are particularly sensitive to environmental interferences and other sources of noise. Especially in those cases, there is a critical trade-off when selecting the number of scans. A higher number of scans reduces the random noise mainly generated by the detector but involves a longer acquisition time, which produces changes in the concentrations of gases surrounding the sample. As a final consequence, the CO_2 and, more importantly, the water vapour peaks appear superimposed in the absorption spectra and distort the shape of relevant peaks in the fingerprint region (e.g., compare the Amide II peak in Fig. 2.8 and Fig. 2.7). This fact must be considered when comparing spectra acquired with different devices or even with the same spectrometer operating at different conditions (e.g., at different times).

2.6.3 Substrate

As was depicted in Fig. 2.6, the background spectrum $I_0(\tilde{\nu})$ is recorded in a region of substrate containing no biological sample. Although the CaF_2 substrate is normally manufactured with high standards, it may contain some chemical impurities or physical defects (e.g., thickness variability or scratches) that may affect the correct ratioing of $I_0(\tilde{\nu})$ and $I(\tilde{\nu})$. However, the most important differences between $I_0(\tilde{\nu})$ and $I(\tilde{\nu})$ due to the substrate can be caused by the presence of chemical contaminants on its surface. These contaminants may come from a deficient manipulation or preservation, but they can also be introduced during the sample preparation.

For instance, when culturing cell populations on the substrate, different chemical compounds such as cell medium are used, which is a mixture of organic substances. Although these compounds are normally removed by washing with water, relevant concentrations can remain adhered to the substrate. The spectral interferences owing to the substrate may be minimised by selecting a clean region of substrate with no defects. Nevertheless, sometimes this task is not feasible because the contaminants extend all over the *empty* surface of substrate.

Another problem associated with the substrate is the effect of reflection loss in the sample-substrate interface [111]. This effect takes place when the refractive indices of sample and substrate differ. Although the substrate materials are selected so that their refractive indices are very close to those of biological samples (e.g., CaF_2, see Sec. 2.4), they can differ in some regions of the sample. Nevertheless, its influence is normally neglected in the analysis of spectra mainly because the exact optical properties of the biological samples are unknown or very sample-dependent and its quantification is infeasible.

2.6.4 Contamination

The final sample can contain substances or materials different from the biological specimens intended to measure, which create absorption bands that can be incorrectly attributed to the studied specimens. These *contaminants* may come from the sample preparation or may have been accidentally introduced at any phase of the handling. In the first case, the substances may unintentionally remain in the sample, such as residual chemical agents used in cell culture, or left on purpose. For instance, paraffin may not be removed in histological samples to prevent chemical distortions (Ch. 4).

Detecting a contaminant involves checking if its spectrum is *similar* to the *expected* absorption profile. This task is challenging because sometimes the *expected* spectrum may be totally *unknown* or may be too *ideal* for the imperfect recorded spectra, which can contain other kinds of artifacts (Sec. 3.2.1). Moreover, the undesired contaminants may give rise to spectra very similar to the studied specimens, such as cell media, which have a biomolecular composition similar to cells. Finally, in some cases a contaminant may not be detected, mainly because its concentration is not high enough and it gets masked under the specimen's absorption, or it may be confounded with other types of anomalies or artifacts.

Anyway, the most straightforward method to detect a contaminant is applying heuristic rules depending on the problem in hand. For example, one solution is removing the specific spectral regions containing the characteristic absorption peaks of a known contaminant, such as paraffin (Sec. 4.2.3). But in other cases, the features of a reference spectrum can be used to reject contaminated spectra, e.g., the Amide I peak is normally the maximum absorbance in a typical biological spectrum (Fig. 2.8).

2.6.5 Light scattering

Light scattering may be the most distinctive problem in biomedical applications of FTIR spectroscopy. Most problems associated with scattering are generated by the morphological and optical properties of the analysed sample. In applications involving artificial materials, the sample preparation and presentation for FTIR measurements may be adapted in order to avoid or minimise scattering. However, the preparation of biomedical samples is much more restricted and is normally determined by clinical routine or laboratory procedures.

The general phenomenon of scattering in FTIR spectral acquisition is coarsely illustrated in Fig. 2.15. In an ideal homogeneous flat sample (left side of Fig. 2.15), a fraction of the collimated IR light is absorbed by the sample and practically the rest of radiation (neglecting the reflection at interfaces) continues its travel towards the detector without significant deviations. Nevertheless, in a real biological sample with heterogeneous morphological and optical properties, such as a single cell (right side of Fig. 2.15), the incident light is scattered, giving rise to significant deviations in the photon paths. As the scattered rays may not reach the detector, the recorded absorption spectrum may be higher than it would be in the absence of scattering. As

a result, this virtual increment in absorption can be wrongly interpreted as chemical absorptions by the biomolecules of the sample [139].

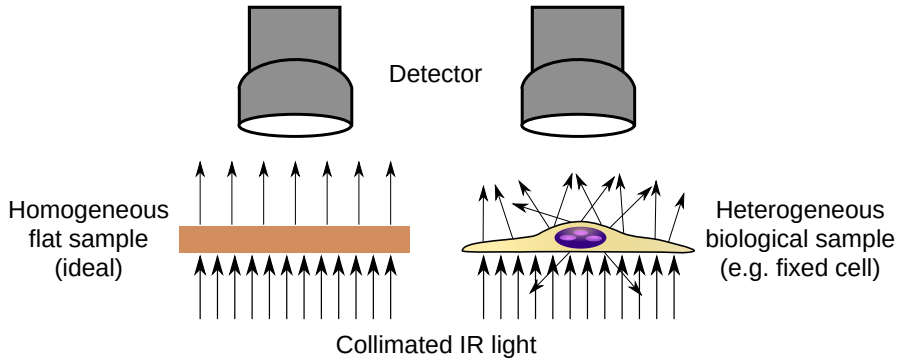


Fig. 2.15: Illustration of scattering phenomena. In an ideal homogeneous flat sample (left) the transmitted light has the same direction of the incident light. However, in a real biological sample with heterogeneous morphological and optical properties (right) the incident light is scattered, giving rise to significant deviations in the photon paths.

In biological samples, many scattering sources are still unknown or require extremely complex mathematical modelling, but few of them have been identified and (at least partially) understood. This last category comprehends the so-called *Mie scattering* and *resonant Mie scattering*, which are mainly related to the morphology and size of the sample.

Mie scattering

In 1908 [140], Mie formulated the theoretical bases of the scattering phenomena produced when dielectric spheres are irradiated with a light whose wavelength is comparable to the size of those spheres. The typical size of animal cells ranges between 8-30 μm and the size of subcellular structures or organelles is between 1-10 μm [86]. These dimensions are comparable and even coincident with the wavelength range commonly used in FTIR spectroscopy, approximately between 2.5-10 μm. Therefore, biological specimens are ideal candidates to experiment the type of scattering formulated by Mie.

Mie-type scattering was firstly identified in cytological samples by Mohlenhoff *et al.* in 2005 [141], although atypical absorbance spectra related to the cell cycle had been

reported several years before [142,143]. The main sources of Mie scattering seem to be the nuclei of cells, especially when the chromatin is more compact and dense. These compact states are experimented when the cell is close to the mitotic phase (Sec. 1.1.2) or when the nucleus permanently shrinks due to cell death (*piknosis*) [141, 144]. Nevertheless, it is still not clear if the real sources of scattering are only the nuclei, the entire cells or their combination. The distortions caused by Mie scattering are more pronounced in cytological samples but they can also be present in spectral histology, particularly at the edges of the tissue [86].

The most prominent features caused by classical Mie scattering in spherical particles are alterations in the baseline of the absorbance spectra. These baseline artifacts are normally modelled as undulating functions with broad sinusoidal shapes, such as the one presented in Fig. 2.16a, which are superimposed on the spectra of biological specimens [101, 145]. Nevertheless, the undulations of the background observed in cells have more maxima and minima than expected from single scattering objects, such as simple spheres, maybe because the size and morphology of cells are not uniform [141].

Resonant Mie scattering

Apart from the *relatively simple* broad baseline oscillations caused by classical Mie scattering, most severe artifacts appear in real measurements of single cells. One of the artifacts identified so far is the so-called *resonant Mie scattering*, which was suggested by Bassan *et al.* in 2009 [146] and tries to explain the anomalies that were previously attributed to the *dispersion artifact* [147,148]. These anomalies consist of significant derivative-like distortions of band shapes, which highly modify the intensity and the positions of the maxima of relevant absorption peaks. As an example, Fig. 2.16b shows an absorbance spectrum contaminated by resonant Mie scattering, which can be compared with the corresponding pure absorption spectrum of the same protein sphere (Fig. 2.16c.). The most prominent distortion can be observed in the Amide I peak, whose relative intensity drastically decreases when it is compared with other peaks such as the Amide II and whose maximum is shifted towards lower wavenumbers.

As commented, Bassan *et al.* [146] was the first in describing that the underlying phenomenon of this type of artifacts is also related to Mie scattering interactions. In classical Mie scattering, the dielectric non-absorbing spherical particles are assumed

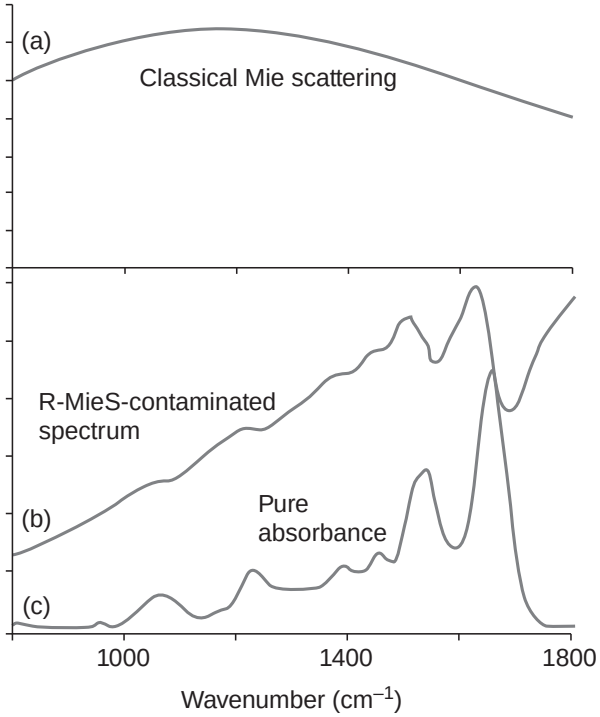


Fig. 2.16: Illustration of Mie and resonant Mie scattering. (a) Example of a classical Mie scattering curve of a spherical particle. (b) Example of absorbance spectrum contaminated with resonant Mie scattering. (c) Pure absorption spectrum of a protein sphere. Adapted from [101] with permission by courtesy of John Wiley & Sons Inc.

to have constant values of the real refractive index, independently of the wavenumber. Nevertheless, in absorbing materials with phenomena of absorption resonance (i.e., the chemical molecules absorb the photons at the illuminated wavenumber), the real refractive index rapidly changes across the wavenumbers of the resonant absorption bands, causing sudden variations in the measured absorbance spectra. In that publication, Bassan *et al.* [146] experimentally demonstrated its theoretical supposition with the absorbance spectra from polymeric microspheres of Poly(methyl methacrylate) (PMMA) measured with synchrotron FTIR microspectroscopy and suggested the potential application to biological samples, especially single cells.

Further details of the scattering phenomena will be described in Sec. 3.2.1.

2.6.6 Micro-FTIR imaging measurements

The use of optical objectives and FPA detectors introduces spatial variability in the recorded spectra. The objectives are not ideal and they introduce a series of aberrations in the final image, especially in the borders of the image. In addition, the elements composing the array of pixels are not exactly equal and do not operate at the same conditions. This is reflected in a difference in the sensitivity and, therefore, the registered intensity may vary between pixels even with homogeneous illumination.

As an example, Fig. 2.17 shows a pseudocolour image constructed with the values extracted from the hyperspectral image of the intensity reference spectrum $I_0(\tilde{\nu})$ at a specific wavenumber ($\tilde{\nu} = 1300$ cm^{-1}, approximately the maximum of intensity). This image also shows the $I_0(\tilde{\nu})$ spectra of four selected pixels in the pseudocolour image.

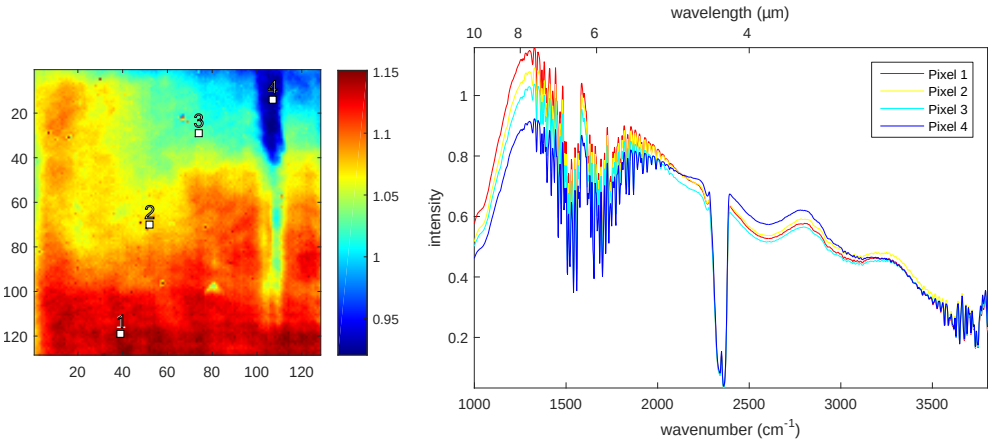


Fig. 2.17: Intensity refererence spectra $I_0(\tilde{\nu})$ of a FPA detector. Left: Pseudocolour image with its colourbar of the intensity values corresponding to a wavenumber $\tilde{\nu} = 1300$ cm^{-1}. Right: Intensity reference spectra $I_0(\tilde{\nu})$ of the pixels selected in the pseudocolour image.

In this case, the region of illuminated substrate is pretty homogeneous and the spatial variability is mainly determined by the sensitivity of the pixels. Nevertheless, the possible spatial inhomogeneity in the substrate is another source of artifacts that introduces undesirable variations in the final measured spectra. This unevenness is *tangled* with other spatial inhomogeneities, for instance, coming from the sensor, the illumination source and the optical components.

As can be observed in the pseudocolour image of Fig. 2.17, the registered intensity values are higher in the lower part of the image, meanwhile there is a small region close to the upper-right corner (pixel 4) where the sensitivity of the sensor is significantly lower. This spatial variability can be confirmed in the spectra of the selected pixels, where the pixel 4 shows lower intensity at lower wavenumbers, but higher sensitivity than the rest of selected pixels at higher wavenumbers.

These spatial differences, although are supposed to be compensated when computing the transmittance and absorbance, introduce variations in the SNR of the measured spectra. In particular, the pixels with lower sensitivity will have a poorer detection limit [101] and will be more prone to different sources of noise. In addition, the sensor can present specific pixels working at saturated levels (*hot pixels*) or not registering any current (*dead pixels*), which normally will provide anomalous measurements that can be identified as *outliers*.

The spatial variability introduced by FTIR microspectrometers can increase the presence of the artifacts described in previous sections and complicate their removal. For instance, some methods exist to reduce the influence of water vapour in the spectrum, which are mainly based on modelling the contributions of a pre-recorded reference spectrum of water vapour [149]. However, that methodology requires to record a battery of reference spectra at different temperatures and concentrations of water vapour in order to model the possible variability. Therefore, if the spatial variability is added to that problem, the model complexity and the risk to introduce additional artifacts highly increase.

Finally, as was commented in Sec. 2.5.1, the use of microscopy instruments is always linked with spatial resolution. Resolution, theoretically determined by the optical power of the instrument and the wavelength of the used light (Fig. 2.12), defines the concept of *diffraction limit* [150], i.e., maximum resolution from which diffraction phenomena occur. Although the advances in optical instrumentation have made progress in the spatial resolution, the diffraction limit will be always present mainly due to the restriction of the wavelength. In addition, practical resolution also depends on the analysed sample.

The main consequence of the diffraction limit is the uncertainty of knowing the spectral *purity* of a pixel [132]. That is, knowing if the recorded spectrum from one pixel only contains contributions from the material within the physical

region covered by that pixel or if there are intrusions from nearby regions due to diffraction phenomena. These undesirable interactions are more critical for lower wavenumbers (Fig. 2.12), which contain more useful biological information (Fig. 2.8). This problem can seriously hamper the analysis of spectral signals, especially in highly heterogeneous samples. Some empirical methods have been proposed to computationally improve the spatial resolution and try to deconvolve the contributions of pixels' spectra [124]. Nevertheless, although these methods can perform an aesthetic improvement in images, their real accuracy and practical improvement for spectral analysis have not been proved.

The theoretical bases of the physicochemical phenomena occurring in modern FTIR microspectroscopes have started to be understood and formulated very recently, both for homogeneous [151] and heterogeneous samples [152]. As stated in [152], the scattering phenomena, sampling geometry, sample morphology and spectral profile in heterogeneous samples (such as biological specimens) are strongly coupled to diffraction effects and jointly perturb the ideal chemical spectrum of the sample. However, these phenomena are not well-understood yet and most practical methods of processing and analysis consider the spectra as obtained with classical spectroscopes. This is the line that will be followed in this thesis, but the advances in the theory of FTIR microspectroscopy will be an important issue to consider in any future research.

Chapter 3

Spectral processing

Contents

3.1 Introduction

FTIR spectra are high-dimensional data structures that contain huge loads of information, part of which can be very useful or, on the contrary, irrelevant and spurious, depending on the specific study or application. This chapter describes several techniques, which can be categorised as *spectral preprocessing* and *feature extraction*, that specifically try to get *better versions* of the FTIR spectra and try to separate the most relevant information in order to improve the performance of later analyses.

3.2 Spectral preprocessing

As was described in Sec. 2.6, FTIR measurements are affected by a number of unwanted phenomena that can hamper their quantitative analysis, which ideally should be based on the chemical information encoded within the absorbance spectra. Although some errors or artifacts may be avoided by good sample preparation and measurement protocols, many of the undesired phenomena remain in the measured spectra and should be corrected for computationally by *preprocessing*.

Preprocessing is normally a determining step in the analysis of FTIR spectra [153–155]. Its main objective is to remove or at least minimise the spectral variability not related to the chemical information of interest. Correct preprocessing can improve the graphical interpretation of the data, as well as reduce the complexity and improve the robustness of the subsequent data modelling (e.g., less biased future predictions) [156]. However, preprocessing of spectral data must be applied with caution and with supporting physical or chemical bases. Incorrect data preprocessing introduces artificial artifacts which can reduce or distort the spectral variance specific of each chemical species and mislead the later analysis. Nevertheless, preprocessing is still an open field with no standard solutions and deep knowledge of spectroscopy theory is normally required to select a specific method, which mainly depends on the final application.

Depending on the bibliographic reference or the application field, spectral preprocessing methods are catalogued into different categories. Here, it will be adopted the classification proposed in [157], where preprocessing methods are coarsely divided into two main groups, namely *filtering methods* and *model-based methods*.

Filtering methods include all those techniques that discard some kind of undesired spectral variation and transform the data into a presumably *better* version. Model-based methods try to get also a *better* version of the data but creating a more explicit mathematical model, which is used both to estimate the useful and the unwanted information. Filtering methods are simpler techniques that have been used traditionally in FTIR spectroscopy, whereas model-based methods have attracted increasing attention in the last few years due to their potential ability to cope with more complex phenomena.

Below, the model-based methods relevant for this thesis will be introduced in the first place. The reasoning and concepts explained in those methods will help to justify the aims of some traditional filtering methods that will be used in later applications.

3.2.1 Model-based methods

Currently, the most advanced model-based techniques used in vibrational spectroscopy are adaptations of the so-called Extended Multiplicative Signal Correction (EMSC). EMSC is a flexible preprocessing method based on linear statistical regression modelling where the influence of unwanted spectral variation is estimated and corrected [137, 157]. The main difficulty of model-based methods in general, and EMSC in particular, is the requirement of knowing the unwanted sources of variation and their theoretical formulation in order to model them in a precise way. The basic principles (which are not trivial) and the versions of EMSC used to correct the FTIR data analysed in this thesis will be discussed in the following paragraphs.

Extended Multiplicative Signal Correction (EMSC)

EMSC allows more selective corrections for various types of unwanted effects, such as scattering, than what is feasible with standard filtering methods. EMSC demonstrated its utility to separate light scattering effects from chemical information in near-IR measurements [158]. Shortly after, this method was also applied to isolate scatter (physical information) from chemical absorbance changes in FTIR images taken from biological samples, in particular from cryo-sections of beef loin [159]. Since then, EMSC has been used as the main preprocessing tool in different problems of FTIR spectroscopy involving biological samples due to its versatility [157].

The starting point for deriving EMSC models is the Beer-Lambert law. As was stated in Sec. 2.4.1, biological samples are complex mixtures of biomolecules whose ideal chemical absorption spectrum would obey the Eq. 2.9. Notwithstanding, as was also described in Secs. 2.4.2 and 2.6, different undesired phenomena *mask* the ideal chemical absorption spectrum within the real measured absorbance spectrum. The key idea of EMSC is approximating the measured absorbance spectrum $A(\tilde{\nu})$ as physical modifications of an idealised chemical spectrum $A_{chem}(\tilde{\nu})$ [137]:

$$A(\tilde{\nu}) = b \cdot A_{chem}(\tilde{\nu}) + A_{phys}(\tilde{\nu}) + e(\tilde{\nu}) \tag{3.1}$$

where scalar b is a multiplicative physical scaling parameter that mimics the effective optical path length ℓ (Eq. 2.9), $A_{phys}(\tilde{\nu})$ represents additive physical contributions to the absorbance spectrum and $e(\tilde{\nu})$ is the residual absorbance which denotes the sum of all other unmodelled effects and measurement errors. This theoretical formulation must be converted into a practical mathematical form, which is accomplished by constructing a linear model with multiplicative and additive terms adapted to the corresponding problem at hand.

First, it is often assumed that the FTIR absorbance spectra obtained from biological samples are very similar [157] and the chemical contributions can be expressed as deviations, $\Delta\varepsilon_j(\tilde{\nu})$, from a (hypothetical or real) reference spectrum $A_{ref}(\tilde{\nu})$ [159]:

$$A_{chem}(\tilde{\nu}) = \sum_{j=1}^{J} C_j \cdot \varepsilon_j(\tilde{\nu}) = \sum_{j=1}^{J} C_j \cdot A_{ref}(\tilde{\nu}) + \sum_{j=1}^{J} C_j \cdot \Delta\varepsilon_j(\tilde{\nu}) \tag{3.2}$$

In some cases, $A_{ref}(\tilde{\nu})$ is computed as the mean spectra of the dataset, but the selection of this ideal reference spectrum depends on the specific problem. In most situations the specific chemical constituents, and particularly their absorptivity values $\varepsilon_j(\tilde{\nu})$, are not known or difficult to obtain. Thus, the last term of Eq. 3.2 is frequently incorporated into the unmodelled residuals $e(\tilde{\nu})$. In addition, considering that the sum of constituents' concentration sums up to 1 ($\sum_{j=1}^{J} C_j = 1$), Eq. 3.1 becomes:

$$A(\tilde{\nu}) = b \cdot A_{ref}(\tilde{\nu}) + A_{phys}(\tilde{\nu}) + e(\tilde{\nu}) \tag{3.3}$$

The physical contributions $A_{phys}(\tilde{\nu})$ are mainly related to scattering effects involved in the measurement. Therefore, $A_{phys}(\tilde{\nu})$ is maybe the most problem-dependent term and its particular modelling has given rise to different versions of EMSC, some of which are later discussed. In some basic versions of EMSC models the scattering effects are represented by an offset parameter a and a polynomial baseline of order N along the wavenumbers $\tilde{\nu}$ with coefficients d_n, that is:

$$A(\tilde{\nu}) = b \cdot A_{ref}(\tilde{\nu}) + a + \sum_{n=1}^{N} d_n \cdot \tilde{\nu}^n + e(\tilde{\nu}) \tag{3.4}$$

The particular case of extending the polynomial baseline up to the quadratic term is normally called the *basic EMSC model*:

$$A(\tilde{\nu}) = b \cdot A_{ref}(\tilde{\nu}) + a + d_1 \cdot \tilde{\nu} + d_2 \cdot \tilde{\nu}^2 + e(\tilde{\nu}) \tag{3.5}$$

As can be observed, the most relevant *simplification* of EMSC is the assumption that only linear relationships exist in the models. Although this linear restriction may oversimplify the physicochemical phenomena in some cases, it facilitates the estimation of the model parameters by statistical techniques. In particular, the unknown model parameters are normally estimated simultaneously by multiple linear regression solved by ordinary or weighted least squares [157, 159]. In the case of weighted least squares, specific weights can be assigned to particular regions of wavenumbers. For example, some ranges containing excessive contaminants, such as the CO_2 region ($\sim$2300-2400 cm^{-1}), can be given a very low weight (e.g., 10^{-6}) in order to reduce their influence in the parameter estimation.

Once the EMSC parameters are estimated, the spectra can be corrected by subtracting the unwanted modelled contributions. For instance, the corrected absorbance spectrum $A_{corr}(\tilde{\nu})$ corresponding to the EMSC model described in Eq. 3.4 would be computed according to:

$$A_{corr}(\tilde{\nu}) = \frac{A(\tilde{\nu}) - a - \sum_{n=1}^{N} d_n \cdot \tilde{\nu}^n}{b} \tag{3.6}$$

Note that in this equation the division by the multiplicative scaling parameter b implies a normalisation of the corrected spectrum.

Mie Scattering-EMSC

In some biological samples with relatively homogeneous spatial properties, such as tissues, the scattering effects can be modelled with relatively *simple* functions, e.g., the polynomial baselines described before. However, biological samples with more complicated optical and morphological characteristics, such as single cells, may produce more complex scattering phenomena, which are reflected in fancier absorbance spectra.

As was described in Sec. 2.6.5, Mie-type scattering is one of the physical artifacts that have been identified and modelled in single cell samples [141, 145]. Mie theory was adapted in 2008 by Kohler *et al.* [160] to develop a version of EMSC that tries to correct Mie scattering artifacts. Full Mie theory [140] describes the scattering produced by an ideal homogeneous non-absorbing sphere in terms of the extinction cross section $Q_{ext}(\tilde{\nu})$ [145]:

$$Q_{ext}(\tilde{\nu}) = \left(\frac{1}{\rho(\tilde{\nu})^2}\right) \sum_{m=1}^{\infty} (2m+1)\, \Re\,(a_m + b_m) \tag{3.7}$$

where a_m and b_m are the scattering coefficients, which are complicated Bessel functions in cylindrical coordinates; $\Re$ symbolises the real part of those functions; and $\rho(\tilde{\nu})$ is the size parameter, which is defined as:

$$\rho(\tilde{\nu}) = 4\pi r(n-1)\tilde{\nu} \tag{3.8}$$

where r denotes the radius of the scattering sphere and n the ratio of the real refractive indices of the particle and the surrounding medium. In the FTIR measurements that will be studied here, the surrounding medium is air (whose real refractive index is essentially 1) and n simplifies to the real refractive index of the scattering particle.

The original full formulation of $Q_{ext}(\tilde{\nu})$ expressed in Eq. 3.7 is normally substituted in many practical applications by an approximation described by van de Hulst in 1957 [161]:

$$Q_{ext}(\tilde{\nu}) \approx 2 - \left(\frac{4}{\rho(\tilde{\nu})}\right) \sin\rho(\tilde{\nu}) - \left(\frac{4}{\rho(\tilde{\nu})^2}\right) (1 - \cos\rho(\tilde{\nu})) \tag{3.9}$$

This formula approximates the original $Q_{ext}(\tilde{\nu})$ predicted by Mie within 1% precision in the case of spheres with a large size parameter and a refractive index close to unity [141, 160]. Fig. 3.1 shows some examples of these curves for a specific real refractive index $n = 1.3$ and a range of radii of the scattering sphere $r = 2 - 14\,\mu\text{m}$. This graphic reveals the undulating nature of the damped sinusoidal functions in Eq. 3.9, which fluctuate around the asymptotic value $Q_{ext}(\infty) = 2$.

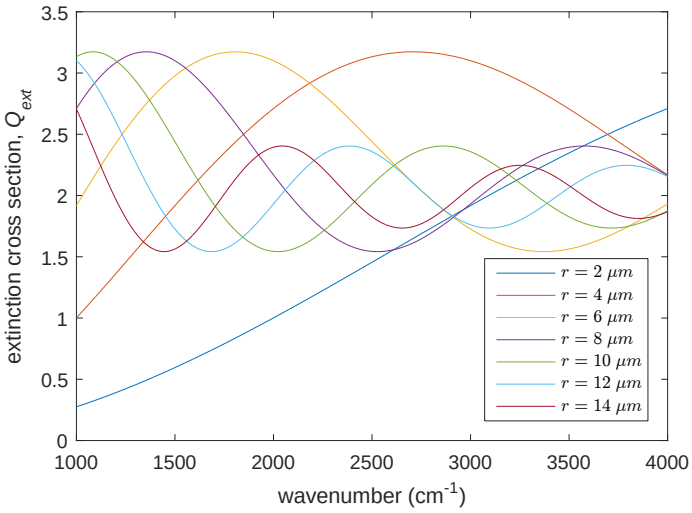


Fig. 3.1: Examples of Mie extinction cross section functions $Q_{ext}(\tilde{\nu})$, according to van de Hulst approximation (Eq. 3.9), for $n = 1.3$ and several values of r (see the legend).

By computing $Q_{ext}(\tilde{\nu})$, the baseline fluctuations attributed to Mie scattering could be estimated through the parameter f, together with an offset a and a multiplicative effect b, in the initial version of the EMSC model proposed by Kohler *et al.* [160]:

$$A(\tilde{\nu}) = b \cdot A_{ref}(\tilde{\nu}) + a + f \cdot Q_{ext}(\tilde{\nu}) + e(\tilde{\nu}) \tag{3.10}$$

Nevertheless, in practical situations the exact values r and n of the studied biological specimen are unknown and the corresponding $Q_{ext}(\tilde{\nu})$ cannot be computed. One possible solution is to consider a set of Mie extinction functions $Q_{ext}(\tilde{\nu})$, covering a range of feasible values of r and n, and try to find the combination of those functions that best fits the baseline deviations in the measured absorbance spectrum. The main problem of this approach is the non-linear nature of $Q_{ext}(\tilde{\nu})$ functions, which can cause

problems of stability and speed in the correction algorithm.

In order to simplify this problem, Kohler *et al.* [160] proposed to apply multivariate metamodelling [162, 163]. This technique consists of computing the set of $Q_{ext}(\tilde{\nu})$ curves for a range of possible values of r and n, which are placed (e.g., by rows) inside the matrix $\mathbf{Q}$, and then decomposing this matrix by non-centred Principal Component Analysis (PCA) (Sec. 3.3.1) into the corresponding matrices of scores $\mathbf{T}$ and loadings or eigenvectors $\mathbf{P}$:

$$\mathbf{Q} = \mathbf{T} \cdot \mathbf{P}^{\mathsf{T}} + \mathbf{E} \tag{3.11}$$

Thus, the original set of functions can be approximated up to a certain degree (see Sec. 3.3.1) by the subspace determined by the first eigenvectors placed by columns in $\mathbf{P}$ and the errors of this approximation are gathered within the residual matrix $\mathbf{E}$. As an example, with the permutations of 10 equidistant values in the range $r = [2, 8]$ µm and 10 equidistant values in the range $n = [1.1, 1.5]$, a total of 100 Mie extinction curves $Q_{ext}(\tilde{\nu})$ can be computed to create the matrix $\mathbf{Q}$. When this matrix is decomposed by non-centred PCA according to Eq. 3.11, the first six loadings or eigenvectors p_i (first six columns of matrix $\mathbf{P}$) account for nearly 100% of the explained variance. As can be observed in Fig. 3.2, the loadings of this example have different characteristics of frequency and damping with the wavenumbers.

By using this new compressed subspace, the $Q_{ext}(\tilde{\nu})$ curve within the EMSC model formulated in Eq. 3.10 is approximated by the first K loadings or eigenvectors $p_i(\tilde{\nu})$ and the original factor f is divided into K new scalar coefficients g_i. The errors due to this approximation are accumulated in the unmodelled residual $e(\tilde{\nu})$, resulting the equation:

$$A(\tilde{\nu}) = b \cdot A_{ref}(\tilde{\nu}) + a + \sum_{i=1}^{K} g_i \cdot p_i(\tilde{\nu}) + e(\tilde{\nu}) \tag{3.12}$$

Finally, as in any EMSC model, the coefficients can be estimated by ordinary or weighted least squares regression and the corrected absorbance spectrum can be computed as:

$$A_{corr}(\tilde{\nu}) = \frac{A(\tilde{\nu}) - a - \sum_{i=1}^{K} g_i \cdot p_i(\tilde{\nu})}{b} \tag{3.13}$$

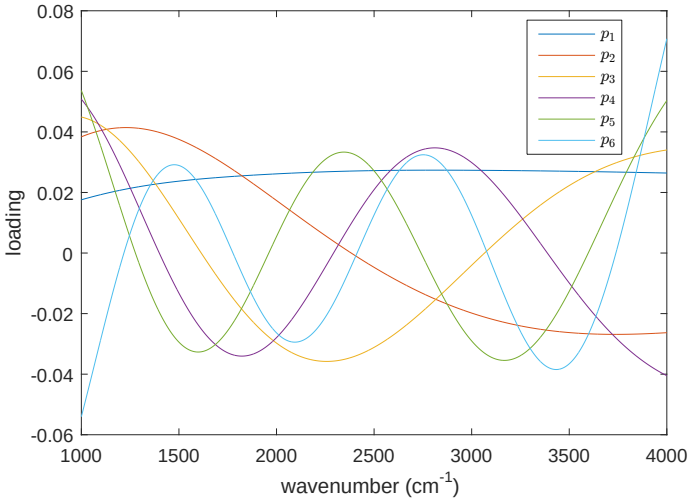


Fig. 3.2: Example of the first six loadings (accumulating nearly 100% of explained variance) of the non-centred PCA metamodel of matrix **Q**, which contains 100 Mie extinction curves $Q_{ext}(\tilde{\nu})$ covering the range of parameters $r = [2, 8]$ and $n = [1.1, 1.5]$

Resonant Mie Scattering-EMSC

As was commented in Sec. 2.6.5, the phenomenon termed as *resonant Mie scattering* was firstly identified by Bassan *et al.* in 2009 [146] as the responsible for the sudden changes in intensity peaks and shifts in the maxima of relevant absorption bands in the measured spectrum. Shortly after, the researchers from the same group published a version of EMSC which tries to correct the artifacts caused by resonant Mie scattering and which was called Resonant Mie Scattering (RMieS)-EMSC [164]. In order to understand the main novelties introduced by the RMieS-EMSC algorithm, some concepts of optics must be described before.

Optical parameters and anomalous dispersion

The absorption of radiation and the interaction of a sample with a medium are governed by two optical parameters, $n(\tilde{\nu})$ and $k(\tilde{\nu})$, which respectively constitute the real and the imaginary parts of the *complex refractive index*, $\tilde{n}(\tilde{\nu})$ [98]:

$$\tilde{n}(\tilde{\nu}) = n(\tilde{\nu}) + ik(\tilde{\nu}) \tag{3.14}$$

where $i = \sqrt{-1}$. $n(\tilde{\nu})$ is often simply called *refractive index* and $k(\tilde{\nu})$ is known as

the *absorption index*. For pure materials containing a single chemical component, the absorption index $k(\tilde{\nu})$ is also related to the absorptivity of the material $\varepsilon(\tilde{\nu})$ by the following equation [98, 101, 165]:

$$k(\tilde{\nu}) = \frac{\ln 10 \cdot C \cdot \varepsilon(\tilde{\nu})}{4\pi\tilde{\nu}} \tag{3.15}$$

In cases where the Beer-Lambert law applies, considering the Eq. 2.8 for a pure substance, the Eq. 3.15 can be rewritten as:

$$k(\tilde{\nu}) = \frac{\ln 10 \cdot A(\tilde{\nu})}{4\pi\tilde{\nu} \cdot \ell} \tag{3.16}$$

This last equation suggests that (under the ideal conditions when the Beer-Lambert law applies) the absorption index $k(\tilde{\nu})$ essentially follows the behaviour of the measured absorbance spectrum $A(\tilde{\nu})$. And particularly, the function $k(\tilde{\nu})$ will have equivalent peaks to the absorption bands in $A(\tilde{\nu})$.

Eq. 3.14 states that the refraction and the absorption of light are coupled processes and the absorption of light is always accompanied by changes in the refractive index [101]. This relationship between $n(\tilde{\nu})$ and $k(\tilde{\nu})$ is mathematically expressed by the Kramers-Kronig transforms, which are based on the principle of causality and connect the real and imaginary parts of many complex quantities in physics [103, 165]:

$$n_{KK}(\tilde{\nu}) = n(\tilde{\nu}) - n_\infty = \frac{2}{\pi}\mathcal{P}\int_0^\infty \frac{s \cdot k(s)}{s^2 - \tilde{\nu}^2} ds \tag{3.17a}$$

$$k(\tilde{\nu}) = -\frac{2\tilde{\nu}}{\pi}\mathcal{P}\int_0^\infty \frac{n(s) - n_\infty}{s^2 - \tilde{\nu}^2} ds \tag{3.17b}$$

where n_∞ is the average real refractive index and $\mathcal{P}$ symbolises the Cauchy principal value of the integral, which tries to resolve the singularity of the denominator at $s = \tilde{\nu}$. The output of the first Kramer-Kronig transform is termed as n_{KK} for convenience in later deductions.

As commented before (Eq. 3.16), $k(\tilde{\nu})$ will present absorption bands at the corresponding wavenumbers (*resonant wavenumbers*) where the molecules of the sample absorb photons. Consequently, following the Eq. 3.17a, the refractive index

$n(\tilde{\nu})$ changes across the absorption bands in $k(\tilde{\nu})$. As an example, Fig. 3.3 shows the refractive index (Fig. 3.3a) and the absorption index (Fig. 3.3b) for a measurement of the polymer Poly(methyl methacrylate) (PMMA). As can be seen, the transitions experimented by $n(\tilde{\nu})$ are sharper when the corresponding absorption bands in $k(\tilde{\nu})$ are stronger and narrower. The most representative example of this behaviour is the absorption band in $k(\tilde{\nu})$ associated with the carbonyl stretching bond at $\sim$1730 cm^{-1}, which creates a sharp *derivarive-like* transition in the refractive index $n(\tilde{\nu})$. As can also be observed, outside the regions containing absorption bands, the refractive index $n(\tilde{\nu})$ is approximately constant. The change of n with the wavenumber is called *dispersion of the refractive index* and the sudden change experimented across an absorption band is normally known as *anomalous dispersion* [98, 101, 165].

RMieS-EMSC algorithm

Bassan *et al.* [164] merged in 2010 the described phenomenon of anomalous dispersion with classical Mie scattering theory to create an improved version of the EMSC algorithm proposed by Kohler *et al.* in 2008 [160] in order to correct for resonant Mie scattering. The rest of this section will explain the main ideas behind the RMieS-EMSC algorithm described in [164].

As already commented, classical Mie scattering assumes a constant refractive index to compute the broadband baseline oscillations approximated by the curves $Q_{ext}(\tilde{\nu})$ (Eq. 3.9) because it was originally formulated for dielectric *non-absorbing* spheres. The main novelty suggested in [146] to explain resonant Mie scattering is considering that the scattering spheres can also absorb light and, therefore, the refractive index can vary with the wavenumber in order to compute $Q_{ext}(\tilde{\nu})$. The main problem of this approach is that the exact curves $n(\tilde{\nu})$ and $k(\tilde{\nu})$ of a sample are normally unknown and they must be approximated.

The key step of RMieS-EMSC is computing the evolution of the refractive index $n(\tilde{\nu})$ by means of the first Kramers-Kronig transform. To do so, an ideal reference spectrum $A_{ref}(\tilde{\nu})$, free of scattering artifacts, is used to approximate the absorption bands within the absorption index $k(\tilde{\nu})$, whose actual value is unknown. In fact, in analogy with Eq. 3.16, $k(\tilde{\nu})$ is supposed to be nearly proportional to $A_{ref}(\tilde{\nu})$:

$$k(\tilde{\nu}) \propto A_{ref}(\tilde{\nu}) \tag{3.18}$$

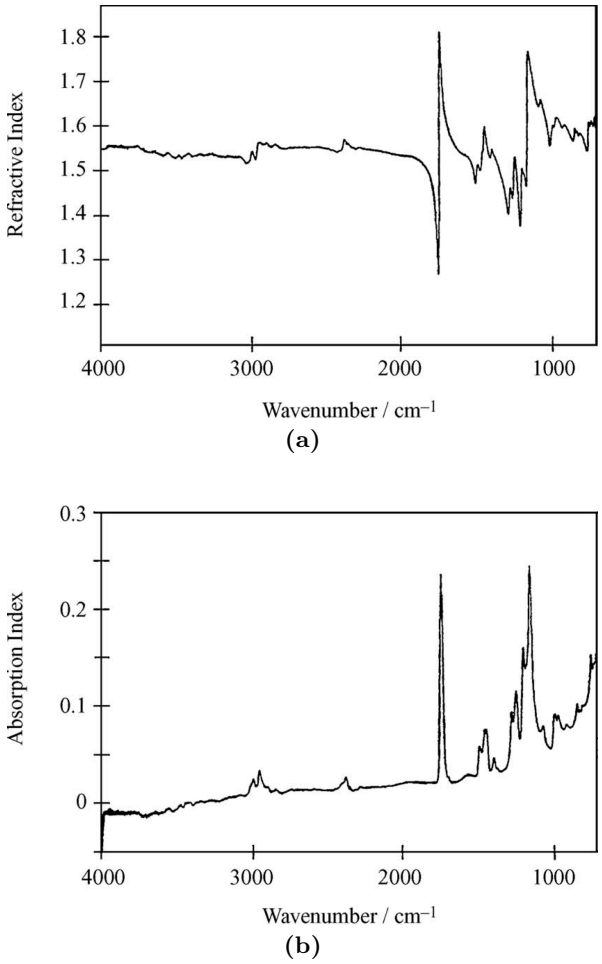


Fig. 3.3: Spectra of optical indices of PMMA. (a) Real refractive index spectrum $n(\tilde{\nu})$. (b) Absorption index or imaginary refractive index spectrum $k(\tilde{\nu})$. Reproduced from [98] with permission by courtesy of John Wiley & Sons, Inc.

Considering this equation and omitting the $\frac{2}{\pi}$ factor in Eq. 3.17a, the following proportional relationship can be stated for the output n_{KK} of the first Kramer-Kronig transform:

$$n_{KK}(\tilde{\nu}) \propto \mathcal{P} \int_0^\infty \frac{s \cdot k(s)}{s^2 - \tilde{\nu}^2} ds \propto \mathcal{P} \int_0^\infty \frac{s \cdot A_{ref}(s)}{s^2 - \tilde{\nu}^2} ds \tag{3.19}$$

As the last proportionality is obtained after several approximations, two new parameters, a_n and b_n, are introduced to estimate the refractive index:

$$n(\tilde{\nu}) = a_n + b_n \cdot n_{KK} \tag{3.20}$$

where a_n is an approximation of the average real refractive index n_∞ and b_n is an amplification factor for n_{KK}, which mainly tries to correct the approximation of $k(\tilde{\nu})$ by $A_{ref}(\tilde{\nu})$ in Eq. 3.19. With this definition of $n(\tilde{\nu})$, Eq. 3.8 can be modified to compute the size parameter $\rho(\tilde{\nu})$:

$$\rho(\tilde{\nu}) = 4\pi r(a_n + b_n \cdot n_{KK} - 1)\tilde{\nu} \tag{3.21}$$

Following the same multivariate metamodelling defined by Koler *et al.* [160] to estimate the particular value of $Q_{ext}(\tilde{\nu})$, a range of possible values for the model parameters can be considered in order to compute a set of scattering curves $Q_{ext}(\tilde{\nu})$ and create the matrix $\mathbf{Q}$. In this case, instead of two, a total of three parameters must be explored: r, a_n and b_n. The ranges of values that will be considered here were suggested by Bassan *et al.* in another publication [166], where the RMieS-EMSC algorithm was applied to a small set of synchrotron FTIR spectra taken from fixed cultured cells of a human prostate adenocarcinoma cell line. These ranges of values, specially for r (which is supposed to be mainly related to the radii of compact nuclei), are compatible with the type of cells that will be analysed in this thesis. In particular, 1000 Mie extinction curves $Q_{ext}(\tilde{\nu})$ can be computed with the permutations of 10 equidistant values in the range $r = [2, 8]$ μm, 10 equidistant values in the range $a_n = [1.1, 1.5]$ and 10 equidistant values in the range $b_n = [0, a_n - 1]$ (see [164,166] for further details in the selection of these ranges). This set of 1000 $Q_{ext}(\tilde{\nu})$ scattering curves can be approximated (up to 99.9% of explained variance [164]) by the first 7 loadings computed via the non-centred PCA model (Eq. 3.11).

Finally, the RMieS-EMSC algorithm assumes that the measured spectrum $A(\tilde{\nu})$ is a linear combination of a scaled reference spectrum, an offset, a linear baseline (this term was not considered by Koler *et al.* [160]), a scattering curve (defined by the PCA metamodelling) and a spectrum of unmodelled features:

$$A(\tilde{\nu}) = b \cdot A_{ref}(\tilde{\nu}) + a + d_1 \cdot \tilde{\nu} + \sum_{i=1}^{7} g_i \cdot p_i(\tilde{\nu}) + e(\tilde{\nu}) \tag{3.22}$$

And as in other EMSC models, the coefficients can be estimated by ordinary or weighted least squares regression and the corrected absorbance spectrum can be computed as:

$$A_{corr}(\tilde{\nu}) = \frac{A(\tilde{\nu}) - a - d_1 \cdot \tilde{\nu} - \sum_{i=1}^{7} g_i \cdot p_i(\tilde{\nu})}{b} \tag{3.23}$$

The weakest part of RMieS-EMSC is the selection of a correct reference spectrum $A_{ref}(\tilde{\nu})$ free of scattering to perform the estimation of both $n(\tilde{\nu})$ (and consequently the scattering $Q_{ext}(\tilde{\nu})$ curves) and the multiplicative parameter b in the final linear regression model. The main problem is that the *perfect* reference spectrum would be the pure absorption spectrum of the sample, which is precisely the ideal output of the algorithm. The practical solution adopted in [164] is converting the described method into an iterative process where an *approximately* correct reference spectrum is used in the first iteration and the corrected spectrum $A_{corr}(\tilde{\nu})$ for each following iteration is considered as the reference spectrum of the next iteration.

The pipeline with the main steps of the iterative RMieS-EMSC algorithm is presented in Fig. 3.4.

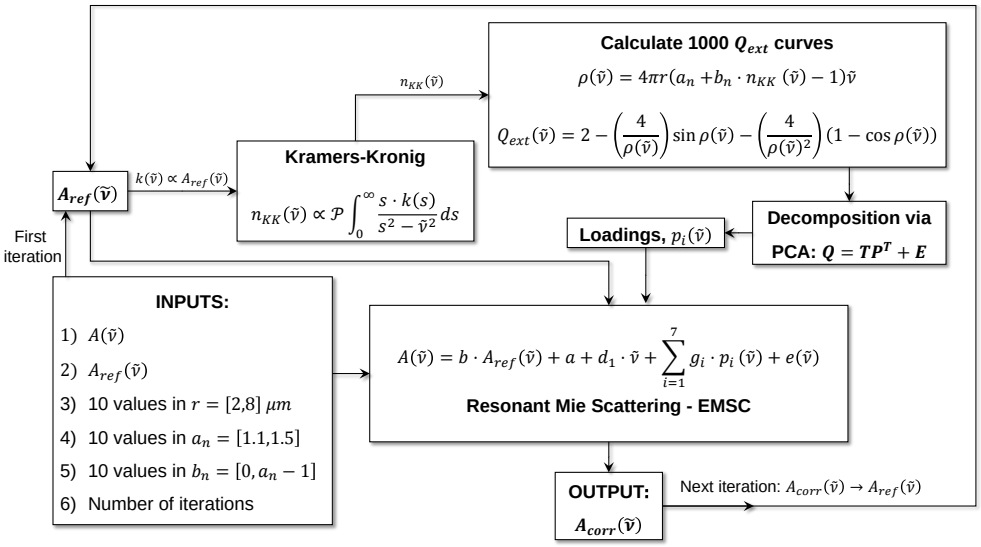


Fig. 3.4: Pipeline of the RMieS-EMSC algorithm. Adapted from [164] with permission by courtesy of The Royal Society of Chemistry.

The selection of the initial reference spectrum $A_{ref}(\tilde{\nu})$ has been a great source of discrepancies in the spectroscopic community [167] mainly because it can potentially introduce serious chemical deviations from the original spectrum. It has been argued that this selection is not critical and the corrected solution finally converges for different reference spectra after a *sufficient* number of iterations [168]. Nevertheless, this number of iterations can be very high and in the order of hundreds of iterations [168]. A spectrum of a thin layer of Matrigel (an artificial basement membrane consisting mainly of proteins), which does not present fluctuations due to scattering and whose absorption peaks are similar to biological samples, is normally used as the initial reference spectrum [164, 166]. Following this practice, Matrigel is the initial $A_{ref}(\tilde{\nu})$ that has been used here in all the applications of RMieS-EMSC.

Examples of corrections by EMSC models

In order to illustrate the outputs obtained by Mie-EMSC [160] and RMieS-EMSC [164], two spectra corresponding to two different pixels from hyperspectral images of A-375 skin cells have been selected. As can be observed in the first row of Fig. 3.5, the selected raw spectra present clear differences although both of them cover approximately the same range of absorbance values.

The presence of resonant Mie scattering (compare with Fig. 2.16) is less pronounced in the first spectrum (Fig. 3.5a), which only shows a slight *depression* beyond the Amide I peak ($\sim$1650 cm^{-1}) and whose characteristic peaks are well defined. Nevertheless, resonant Mie scattering is much more present in the second spectrum (Fig. 3.5b), which is mainly reflected in a fancier baseline, more distorted peaks and a stronger derivative-like depression beyond the Amide I peak. Another distinguishing factor is the presence of random noise, which is higher in the second spectrum and more pronounced at higher wavenumbers (possibly due to the lower sensitivity of the detector at those wavenumbers). In addition, a range enclosing the CO_2 region ($\sim$2250-2450 cm^{-1}) has been discarded to avoid additional artifacts.

In the rest of subfigures of Fig. 3.5, the Matrigel reference spectrum has been plotted for comparison. It can be observed that the Matrigel spectrum presents *ideal* characteristics regarding scattering, specially evident in the null absorption region approximately between 1900 and 2500 cm^{-1}, and the position of its characteristic peaks are very similar to the raw cell spectra.

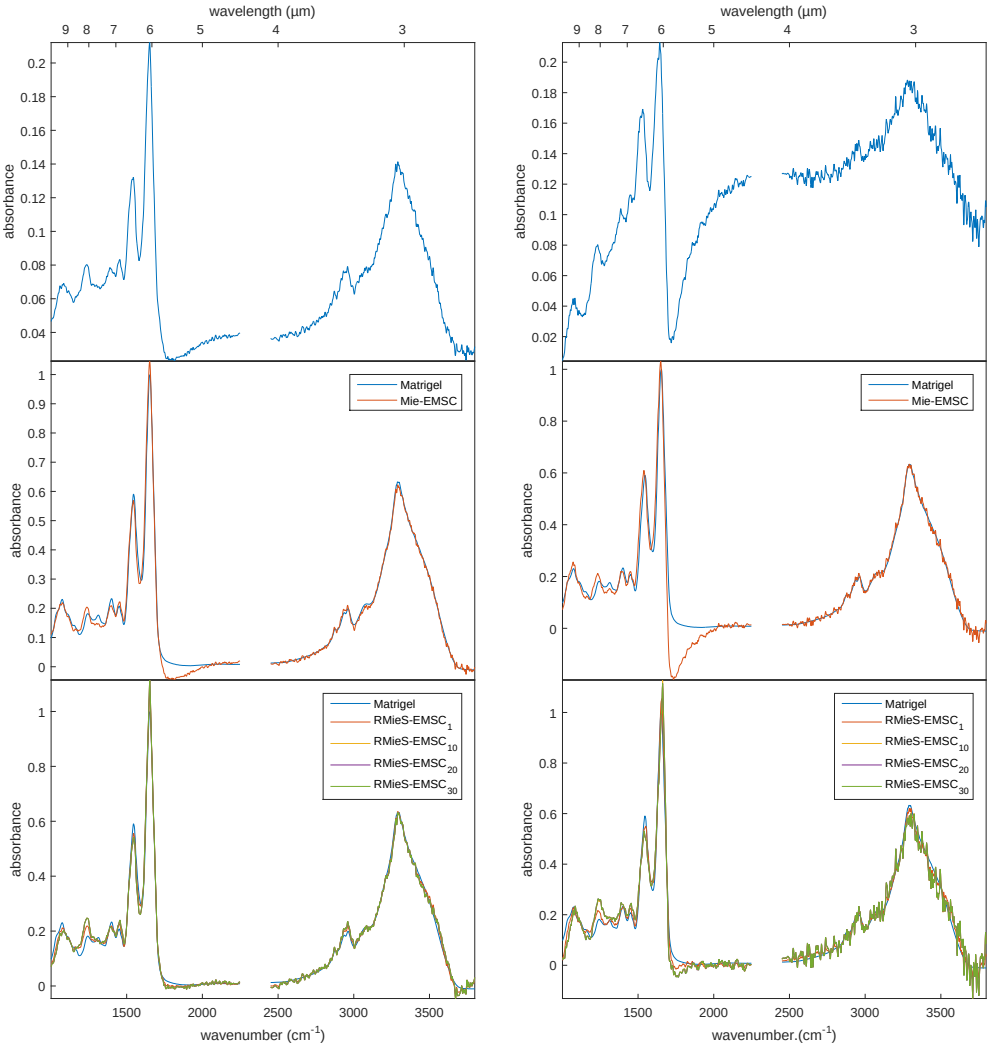


Fig. 3.5: Example of corrections by EMSC models in two different pixels' spectra (a) and (b) extracted from hyperspectral images of A-375 skin cells. First row: Raw spectra. Second row: Matrigel reference spectrum and corrected spectrum by Mie-EMSC [160]. Third row: Matrigel reference spectrum and spectra corrected by RMieS-EMSC [164] after different iterations (specified by the subscripts in the legends).

The results of the spectra corrected by Mie-EMSC can be seen in the second row of Fig. 3.5. When comparing with the raw spectra, Mie-EMSC seems to decrease the baseline artifacts in most parts of the wavenumber range. Unfortunately, the depression beyond the Amide I peak is even accentuated, suggesting a suboptimal correction or amplification of the resonant Mie scattering artifacts. The implicit normalisation performed by the model can be appreciated in the change of covered absorbance values. With respect to the random noise, Mie-EMSC seems to slightly smooth the spectra.

The outputs for different iterations of RMieS-EMSC are shown in the last row of Fig. 3.5. As described before, the iterative RMieS-EMSC algorithm considers the Matrigel spectrum as the initial reference $A_{ref}(\tilde{\nu})$ in the first iteration and the corrected spectrum becomes the $A_{ref}(\tilde{\nu})$ in subsequent iterations. This mechanism gradually makes the corrected spectra more different from the Matrigel spectrum and hypothetically more similar to the corresponding ideal absorption spectra. Moreover, the random noise present in the raw spectra is also recovered in increasing iterations. In view of the represented spectra, the corrected spectra seem to become very stable after the 10th iteration and no differences can be appreciated between the 20th and 30th iterations (totally overlapped in the figures). Apart from the commented characteristics, the most relevant difference between the corrections of Mie-EMSC and RMieS-EMSC is the better compensation for the *derivative-like* artifacts supposedly created by resonant Mie scattering, clearly appreciated in the depression beyond the Amide I peak.

The stability of the iterative RMieS-EMSC algorithm and the increasing presence of the random noise with the number of iterations have been further studied in a dataset of 300 skin cell spectra. Fig. 3.6a displays the 300 raw spectra extracted from pixels of hyperspectral images of four different cell lines (A-375, HaCaT, NIH-3T3, SK-MEL-28). All those pixels contained different regions of cellular material and were randomly selected so that each cell line was equally represented (i.e., 75 spectra/cell line). The heterogeneity introduced by the physical phenomena can be clearly appreciated in the variations of offsets, baselines and ranges of absorbance values in the raw spectra. The 300 corrected spectra after 30 iterations of RMieS-EMSC are shown in Fig. 3.6b. It can be observed that the corrected spectra are much more similar between them and the remaining variability is supposed to be mainly

governed by the chemical contributions and the unmodelled effects, which include the random noise. This random noise is more evident in the spectra with lower signal levels after the normalisation, which amplifies the absorbance values equally in the whole range of wavenumbers.

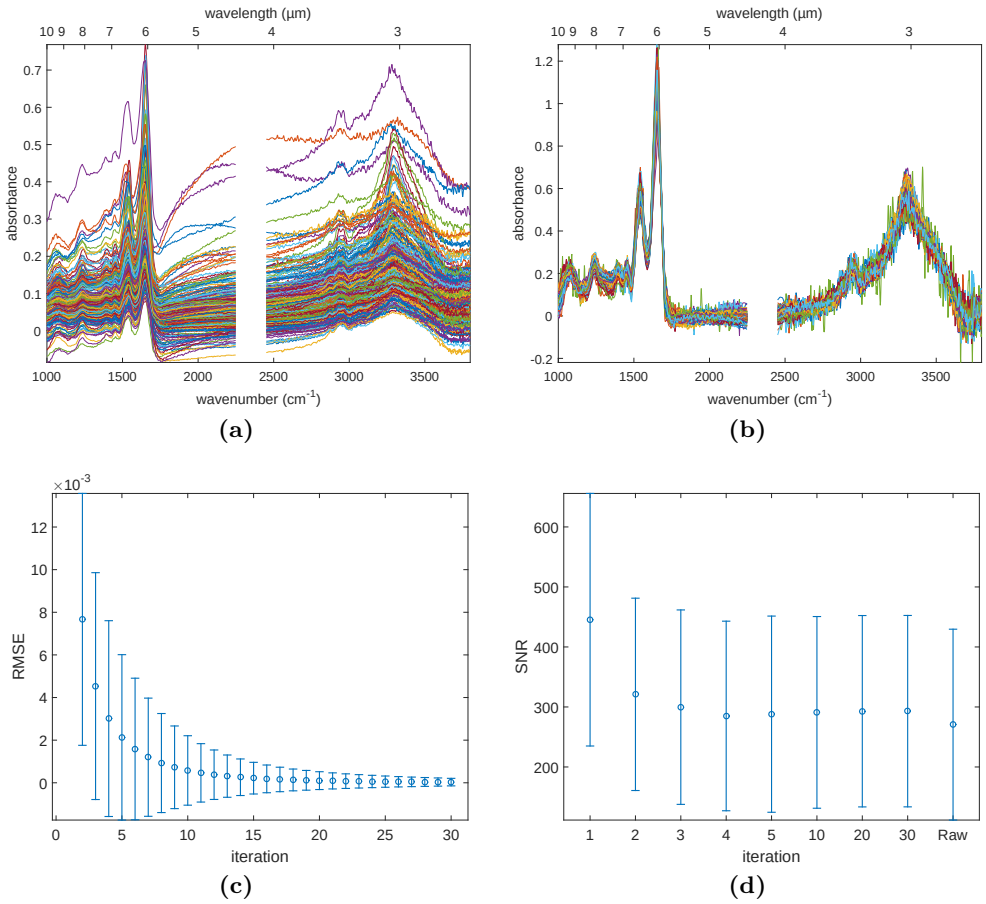


Fig. 3.6: Example of corrections by the RMieS-EMSC algorithm in a small dataset of 300 pixels' spectra from hyperspectral images of four types of skin cells (A-375, HaCaT, NIH-3T3, SK-MEL-28). (a) Raw spectra. (b) Spectra corrected by RMieS-EMSC after 30 iterations. Error bars (mean values symbolised by a circle and standard deviation by bars) for different iterations of RMieS-EMSC of (c) Root-Mean-Square Error (RMSE) and (d) Signal-to-Noise Ratio (SNR).

The stability of RMieS-EMSC has been quantified by the Root-Mean-Square Error (RMSE), which basically accounts for the differences between the corrected spectra at a specific iteration and at the previous iteration. That is, for a generic iteration j, RMSE is computed as:

$$RMSE_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[A^{j}_{corr}(\tilde{\nu}_i) - A^{j-1}_{corr}(\tilde{\nu}_i)\right]^2} \tag{3.24}$$

where N here indicates the total number of discrete wavenumbers measured in the absorbance spectra. The values of RMSE were computed for all the iterations of the 300 spectra of the selected dataset. The obtained results are summarised in terms of the mean and the standard deviation with the error bars depicted in Fig. 3.6c. As can be seen, the mean value of RMSE gradually decreases with the number of iterations, being lower than 10^{-3} from the 8th iteration. In addition, the standard deviation also diminishes with the number of iterations, denoting a stabilisation even for the most problematic spectra from around 20 iterations.

The level of random noise can be assessed in the corrected spectra after each iteration by computing the Signal-to-Noise Ratio (SNR) with the methodology explained in section Sec. 3.2.2. Similarly to RMSE, the values of SNR for the 300 selected spectra are summarised in terms of mean and standard deviation in Fig. 3.6d. As was already observed in the individual spectra of Fig. 3.5, the presence of random noise is lower after the first iteration giving the highest SNR. In this case, the mean and standard deviation values of SNR become approximately steady after around 4 iterations, suggesting that the random noise is quickly recovered. It must be highlighted the high dispersion existing in the SNR values, mainly caused by the relevant differences in the signal measured at different regions of the cell. In particular, pixels containing only cytoplasm generate lower absorbance values than pixels containing nuclear regions and have lower values of SNR. Finally, it can be seen that the SNR is slightly higher in the corrected spectra than in the raw spectra.

Comments and limitations of RMieS-EMSC

Some comments should be made regarding the RMieS-EMSC algorithm:

- The selection of the number of iterations is not a trivial task. On the one hand, the more iterations, the closer the corrected spectrum should be to the

ideal absorption spectrum. On the other hand, the computational cost can be prohibitive for real practical solutions, specially with FTIR images containing thousands of spectra. As a reference, the current implementation of the RMieS-EMSC algorithm takes around 1-2 seconds per iteration and per spectra in an average up-to-date computer. It has been argued that few iterations (<10) could be enough for classification purposes, at least for histological images [168], although a reliable interpretation of the biochemical information would require more iterations. As has been shown here in a small dataset of 300 skin cell spectra, most corrected spectra seem to converge from the 10th iteration, but more stable solutions are obtained with 20 iterations.

- The noise in the measurements can be amplified in the corrected spectra with poorer SNR due to the intermediate computations of the method, specially in the Kramer-Kronig transform [166]. For example, when the corrected spectrum is used as a reference in the second and successive iterations, the water vapour peaks and the random noise of the detector can mislead the estimation of the scattering curves, which are supposed to be affected only by the absorbance of the sample. These problems may be minimised in practice in synchrotron measurements, where ideal measurement conditions are highly controlled. In fact, all the algorithms described so far to correct scattering in FTIR spectra were mainly supported by measurements performed in synchrotron facilities. Nevertheless, in a common laboratory of biospectroscopy, the measurement conditions may experiment fluctuations that may be more critical for some final applications, such as the analysis of single cells.

- Its effectiveness has mainly been demonstrated in polymeric (PMMA) spheres [146]; in simulated spectra of cells with resonant Mie scattering artifacts, which were artificially introduced [164]; and reduced dataset of cells, mostly measured with synchrotron microspectroscopes [166]. In addition, this effectiveness has only been supported by qualitative or visual graphics, such as representations of the corrected spectra and PCA score plots.

- Despite the relatively complex formulations considered by RMieS-EMSC, it still assumes severe simplifications, e.g., in the Mie scattering theory or in the computation of the real refractive index. In the last years, more advanced theoretical background has been formulated to cope with the physical

phenomena in FTIR microspectroscopy [169, 170]. Even very recently a revised and faster version of RMieS-EMSC, which seems to incorporate some improvements (still to demonstrate), has been described [171, 172]. Unfortunately, no practical implementations of these improved methodologies were available during the development of this thesis.

Despite all these limitations, the RMieS-EMSC algorithm described here is the most advanced practical implementation currently available and has been used as the reference preprocessing method in the applications involving single cells.

3.2.2 Filtering methods

As previously commented, filtering methods are simpler techniques than model-based methods which transform the spectra with the aim of discarding the unwanted information and retaining or enhancing the useful information. The main filtering techniques that were used in this thesis are: *smoothing* or *denoising*, *baseline correction*, *differentiation* and *normalisation*.

Smoothing/Denoising

Smoothing filters basically try to reduce the presence of random noise in the spectrum. The main challenge is to remove this noise without degrading the underlying information within the signal. By far, the most used smoothing filter in spectroscopy is the Savitzky-Golay (SG) filter [173]. This filter works by moving a *sliding window* which is sequentially centred in each point of the spectrum and spans a specified number of points. The absorbance values of those points are used to fit a polynomial by least squares (computed with convolution functions previously calculated) and the absorbance value of the central point is substituted by its corresponding value in the polynomial.

Two parameters must be defined in Savitzky-Golay (SG) filters: the *order* of the fitting polynomial and the *points* that define the size of the sliding window. The least squares fitting is normally performed with a 2nd or 3rd order polynomial to avoid excessive distortions. The number of points must be odd and has a major impact in the final solution. Fig. 3.7a illustrates the influence of the points used in the sliding window when using a SG filter with 2nd order fitting polynomial to smooth a skin

cell spectrum. As can be observed, the higher the number of points, the more high frequency random noise is removed. However, the characteristic peaks also lose more details and become more distorted. This smoothing procedure was applied to the small dataset of 300 raw spectra that were plotted in Fig. 3.6a and their corresponding SNRs were computed. Fig. 3.7b summarises the variations of SNR with the number of points of the sliding window in terms of the mean and standard deviation. As expected, the SNR gradually increases with the number of fitting points, but this increment is not homogeneous as demonstrated by the simultaneous growth of the standard deviation.

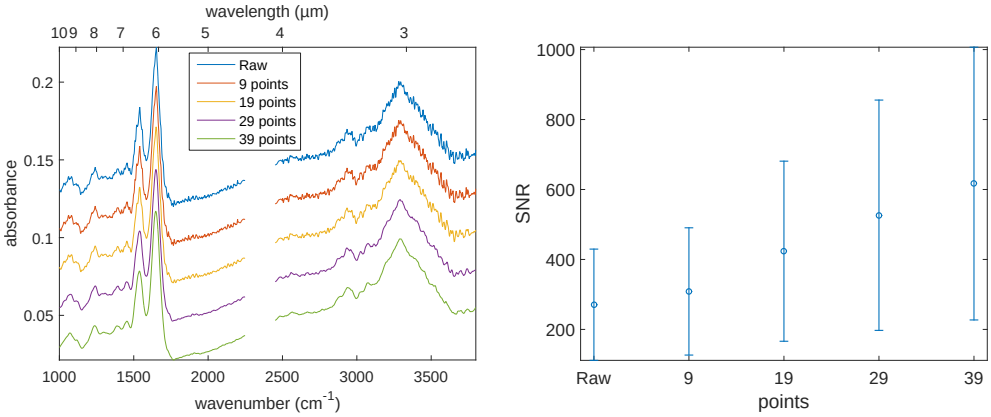


Fig. 3.7: Example of smoothing by Savitzky-Golay filtering with a 2nd order fitting polynomial and different window sizes (points). (a) Raw and smoothed spectra of a single pixel from a skin cell. Artificial offsets have been added for clarity. (b) Error bars (mean values symbolised by a circle and standard deviation by bars) of the Signal-to-Noise Ratio (SNR) of the raw and smoothed spectra of Fig. 3.6a.

Smoothing must be applied with extreme caution. Moderate smoothing does not have a relevant influence but excessive smoothing can introduce critical artifacts. In general, smoothing provides only an aesthetic improvement in the visualization of spectra although it is sometimes used to reduce noise in the generation of 2D images from the hyperspectral data cube (e.g., Sec. 4.2.3). As an alternative, in quantitative analysis the huge information contained in the multidimensional spectra can be used to discard the influence of random noise by means of multivariate techniques, e.g., PCA (Sec. 3.3.1).

Signal-to-noise ratio

Signal-to-Noise Ratio (SNR) is the variable commonly used to assess the quality of a spectrum with respect to the presence of noise. The generic definition of SNR is pretty simple:

$$SNR = \frac{Signal}{Noise} \tag{3.25}$$

Nevertheless, there is not a general consensus about the specific features of the spectrum which must be considered as *Signal* and *Noise*, giving rise to different practical implementations of the SNR in spectroscopy [99, 174–176]. Fig. 3.8 sketches the elements used in this thesis to compute the SNR of individual spectra:

- *Signal*: following the general trend, it is chosen as the amplitude of the strongest peak in the spectrum, that is the Amide I peak. This amplitude is computed with respect to the absorbance minimum inside a range of wavenumbers enclosing the Amide I peak (1480-1780 cm^{-1}).

- *Noise*: this is the greatest source of discrepancies. As a rule, it must be computed in a region without absorption bands so that only *random noise* should be present. Here, the noise is computed as the root-mean-square level (rms) in the region 1950-2150 cm^{-1}, where no absorption peaks should exist. However, as shown in Fig. 3.8, the presence of other types of artifacts that generate complex baselines, such as Mie scattering, can mislead the contributions of noise. Therefore, a broad baseline (computed with an *aggressive* SG smoothing filter of 51 points) is removed from that spectral region before calculating the rms.

Baseline correction

As was described in Sec. 3.2.1, spectral baselines can be distorted due to different physical contributions, such as scattering, changes in measurement conditions or other instrumental factors. The complexity of the baseline distortions can be high, such as in cell samples affected by strong Mie scattering. In those cases, model-based techniques may be the proper solution to remove baseline artifacts. Nevertheless, in applications where those baseline distortions are not critical, simpler techniques may

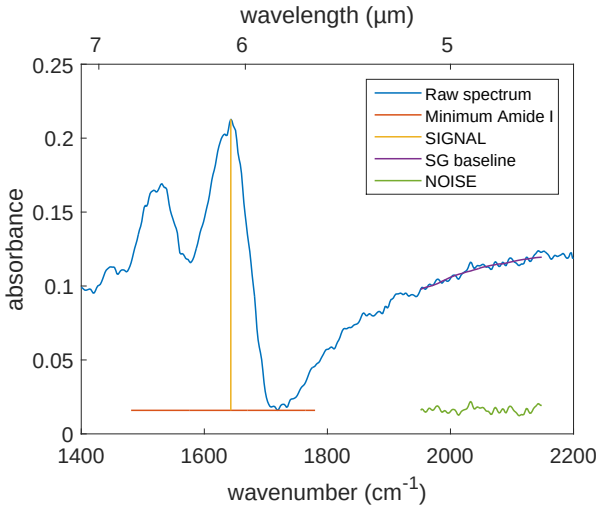


Fig. 3.8: Sketch of the elements used to compute the Signal-to-Noise Ratio (SNR).

provide acceptable results. Different baseline correction methods can be applied in FTIR spectra [153], but two methods have been mainly applied in this thesis are:

- *Offset correction*: it is the simplest baseline correction method. It consists in subtracting a constant value of absorbance (i.e., an horizontal line) to the whole spectrum. The subtracted value is normally the minimum or the mean absorbance value of the spectrum. This method is equivalent to estimate and subtract the parameter a in EMSC models (Sec. 3.2.1).

- *Rubberband baseline correction*: it consists in automatically finding a convex polygonal line which does not intersect the spectrum and whose edges are relative minima within the spectrum [177]. Although being automatic is the main advantage of this method, it can introduce severe artifacts specially in spectra with complex baselines, such as cell spectra.

Fig. 3.9 illustrates the offset and baseline estimated by the previous methods in a raw cell spectrum cropped to the fingerprint region and the resulting corrected spectra after their removal. As can be observed, the relative heights of characteristic peaks slightly change with the rubberband method, but these distortions may be higher in cases with more complex baselines. Finally, as was sketched in Fig. 3.8, SG

smoothing filter with a high number of fitting points can also be used to estimate a broad baseline, but its use is merely restricted to regions without broad absorption peaks that can interfere with the baseline estimation.

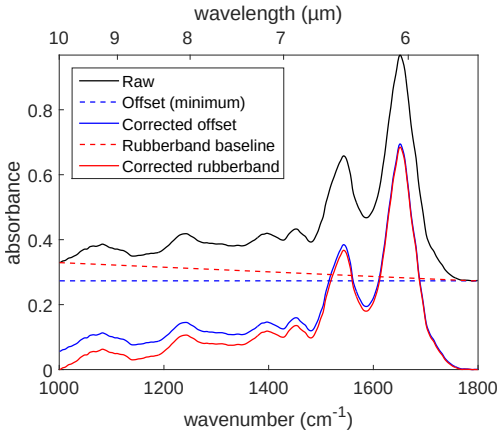


Fig. 3.9: Example of baseline correction methods.

Differentiation

Derivative filters are widely used in spectroscopy to enhance spectral features and identify overlapping absorption bands in complex spectral profiles [153, 177–179]. In addition, derivative filters reduce baseline artifacts: first differentiation removes an additive offset, meanwhile second differentiation also removes a linear baseline.

The main drawback of differentiation is the simultaneous enhancement of noise and the consequent degradation of the SNR. Not only random noise can be enhanced, uncompensated absorption bands from water vapour will also be enhanced by the derivative filters [138]. In order to slightly alleviate this problem, a version of Savitzky-Golay filters exists to compute the derivatives by least squares fitting of a polynomial [173]. In fact, SG differentiation is normally the default technique to compute derivatives in spectroscopy.

As an example, Fig. 3.10 shows the outputs after applying the 1st (DiffSG1) and 2nd (DiffSG2) order differentiation by SG filters with different window sizes (points) to a raw spectrum of skin cell. The spectral information contained in the absorption

peaks of the original spectrum augments with derivative filters, as can be seen with the increase of the number of spectral peaks. However, noise also grows, specially at higher wavenumbers, and almost completely hides the signal in the 2nd order differentiation. The noise is reduced with higher numbers of fitting polynomial points, but the useful signal is also distorted. Traditionally, 2nd order differentiation has been preferred over 1st order derivatives, but its use may be detrimental in applications involving low-SNR spectra.

Normalisation

According to the Beer-Lambert law (Eqs. 2.8 and 2.9), the absorption spectra depend on the thickness (optical path length) and concentration of the sample. Depending on the application, this information may be useful whereas it can mislead the analyses that focus on the biochemical information. In the last cases, several normalisation methods exist that mitigate the influence of the sample thickness and concentration, which are individually applied to each pixel's absorbance spectrum:

- *Min-Max normalisation*: it consists in firstly subtracting the minimum absorbance value of the spectrum (similar to the simple offset correction) and then dividing the whole spectrum by the maximum absorbance. In cases where the spectrum is cropped to the fingerprint, this maximum value is normally the Amide I peak. It is the simplest normalisation method but it may provide acceptable results in situations where the scattering phenomena are not critical.

- *Standard Normal Variate*: the mean absorbance value is subtracted from the whole spectrum and the resulting spectrum is divided by the standard deviation of the whole absorbance spectrum. Originally applied in 1989 to remove the multiplicative interferences of scatter and particle size in near-IR diffuse reflectance spectra [180], this approach has been historically used in different spectroscopic problems to reduce scattering effects.

- *Vector normalisation*: the whole absorbance spectrum is divided or scaled by its Euclidean or L_2 norm. In some cases, the mean absorbance value of the whole spectrum is previously subtracted [153]. This type of normalisation is normally used after differentiation [177].

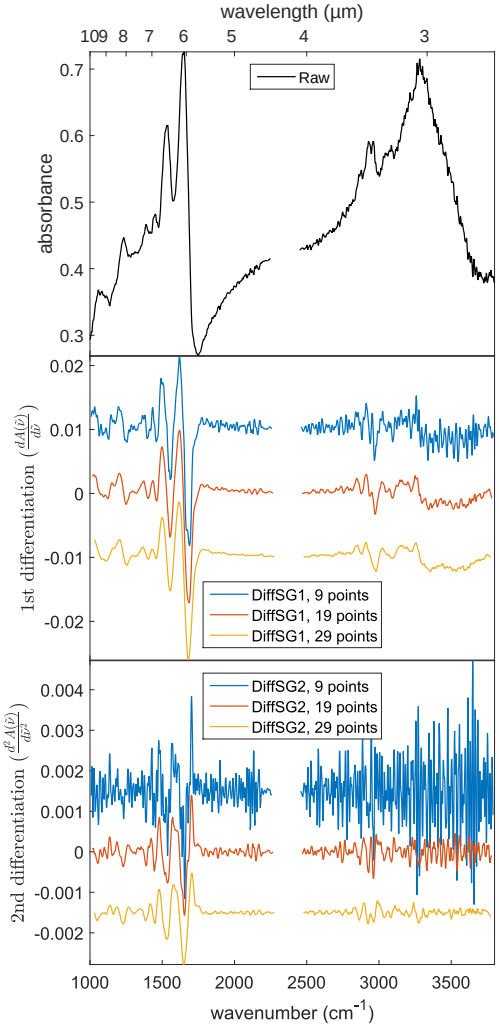


Fig. 3.10: Example of spectral differentiation. Top subfigure: raw spectrum of a skin cell (A-375). Middle and bottom subfigure: 1st and 2nd order differentiation by Savitzky-Golay (SG) filtering with a 2nd order fitting polynomial and different window sizes (points) specified in the legends. Artificial offsets have been added for clarity.

The importance of normalisation can be better understood with the two raw spectra from pixels containing structures of the same skin cell line (HaCaT) presented in Fig. 3.11. The first impression is that each spectrum spans along different ranges

of absorbance values and that there is an offset between them. In spite of that, both spectra seem to have similar shape (characteristic peaks) but at different scale.

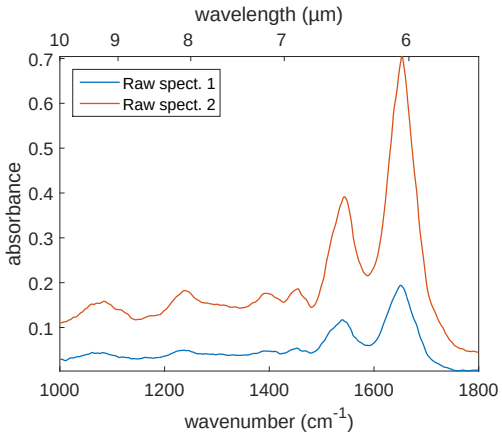


Fig. 3.11: Example of raw spectra from pixels containing structures of the same skin cell line (HaCaT), which are normalised in Fig. 3.12.

The coarse differences observed in the raw spectra of Fig. 3.11 are presumably caused by variations in thickness and concentration of the biological material within the areas covered by the pixels. Fig. 3.12 depicts the same spectra of Fig. 3.11 after applying the normalisation methods described above, together with the resulting corrected spectra after 30 iterations of RMieS-EMSC, which also performs a normalisation of the spectra. As can be observed, the normalised spectra are much more similar, presenting almost the same spectral features and mainly differing in random noise content. In spite of that, the results of the normalisation methods differ even in this *simple* example without relevant distortions due to scattering and other complex perturbations.

3.3 Feature extraction

Chemometrics [174, 181–183] is an interdisciplinary field that includes all those methods related to the mathematical manipulation and interpretation of chemical data [182], which are normally characterised by having multiple variables. Chemometrics shares many techniques with other disciplines, such as *pattern recogni-*

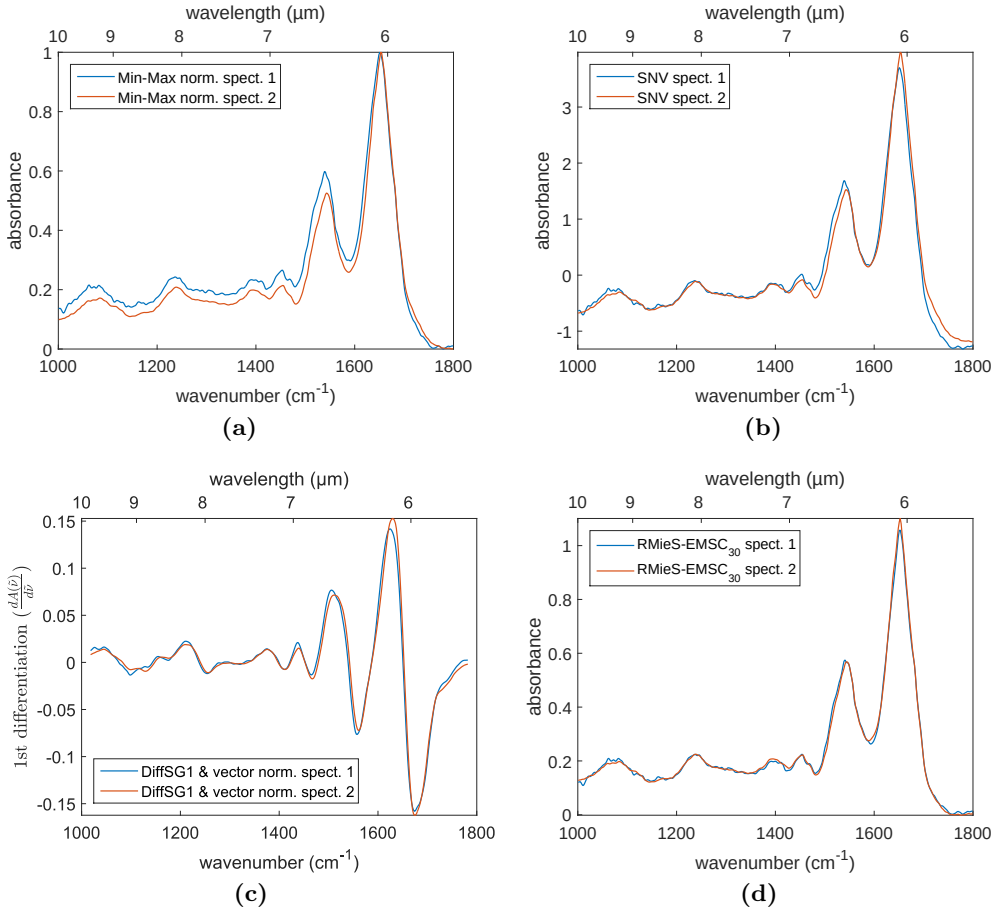


Fig. 3.12: Example of spectral normalisations of the raw spectra shown in Fig.3.11. (a) Min-Max normalisation. (b) 2nd order differentiation by Savitzky-Golay (SG) filter (2nd order polynomial and 19 fitting points) and vector normalisation. (c) Standard Normal Variate (SNV). (d) RMieS-EMSC after 30 iterations.

tion [184, 185], *machine learning* [186] or *data mining* [187]. Nevertheless, chemometric methodologies are specially oriented toward the analysis of data with multiple variables, normally called *multivariate data*.

As FTIR spectra are multivariate data structures with hundreds or even thousands of variables (i.e., the measured absorbance values at different wavenumbers), they have been traditionally treated by chemometric methods associated with multivariate analysis. Among these methodologies, feature extraction techniques have special relevance. They treat to extract the most relevant characteristics within the huge loads of information contained in the FTIR spectra. Two of the most important feature extraction techniques in chemometrics have been used during the development of this thesis: Principal Component Analysis (PCA) and Partial Least Squares (PLS).

3.3.1 Principal component analysis

Principal Component Analysis (PCA) [188–190] is the most widespread technique in multivariate analysis. Multivariate data consist of a large number of variables which are normally highly correlated, giving rise to a huge amount of redundant information. The main aim of PCA is to reduce the dimensionality of multivariate datasets while retaining as much of their variation as possible.

PCA looks for a linear transformation of the initial variables so that the new variables, called Principal Components (PCs), are aligned with the directions of maximum variation of the data. This new set of variables has two distinguishing characteristics: they are uncorrelated and they are ordered in terms of explained variance. Mathematically, the original multivariate dataset, which consists of I samples or objects (e.g., spectra) with J variables (e.g., absorbance values at specific wavenumbers), can be used to construct the original data matrix $\mathbf{X}$. The transformation computed by PCA can be expressed as follows:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^{\mathsf{T}} + \mathbf{E} \tag{3.26}$$

where $\mathbf{T}$ is the matrix of *scores*, $\mathbf{P}$ is the matrix of *loadings*, the superscript T indicates *transpose* and $\mathbf{E}$ is the *residual* matrix. The corresponding dimensions of the matrices involved in this relationship are sketched in Fig. 3.13.

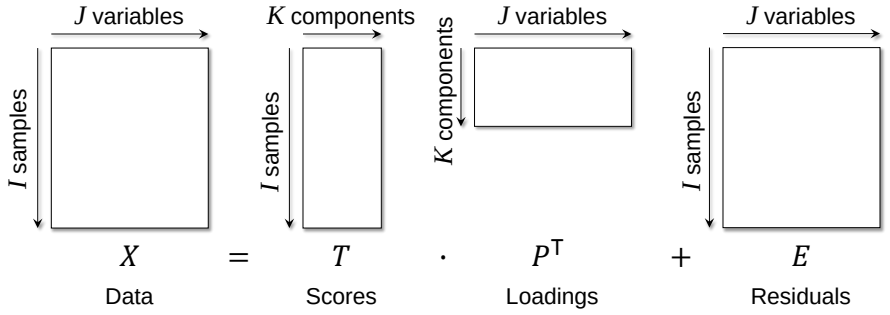


Fig. 3.13: Sketch of the matrices involved in the decomposition by PCA.

Each row of the matrix $\mathbf{T}$ contains the *scores* of each sample or object of the dataset. These scores are the coordinates of that object in the new space of K PCs. Each column of $\mathbf{P}$ (note the transpose in Eq. 3.26) corresponds to the *loadings* or *eigenvectors* of the new space of PCs. These eigenvectors form a basis of the new vector space of PCs. The number of computed PCs, K, is a parameter that must be selected and is limited by the minimum dimension of the original data matrix $\mathbf{X}$. That is, $K \leq \{I, J\}$. Depending on the selected K, the original dataset is approximated up to certain degree of intrinsic variation by the product $\mathbf{T} \cdot \mathbf{P}^{\mathsf{T}}$ and the errors of this approximation are accumulated in the residual matrix $\mathbf{E}$.

In order to compute this transformation, the following steps are normally applied:

1. Translation of the origin of the initial axes to the mean point of the dataset. This step, normally called *mean-centring*, is not a requirement but it is often applied because it removes offset effects [191]. When it is not applied, the global method is called *non-centred* PCA. Unless explicitly stated, mean-centring will be applied when performing PCA.

2. Computation of the eigenvalues of the covariance matrix of the dataset. The sum of all these eigenvalues accounts for the total variance within the dataset. The magnitude of each eigenvalue is related to the explained variance along a specific direction in the centred space.

3. Sorting of the computed eigenvalues in decreasing order so that the first eigenvalues explain the highest variance in the data.

4. Calculation of the eigenvectors corresponding to each ordered eigenvalue in a

sequential way so that they form an orthonormal basis capable of generating a vector subspace within the original vector space.

5. Finally, the scores are computed as the projections of the variables in the original space into the new subspace of eigenvectors.

The result of the PCA process is sketched in Fig. 3.14. In this figure, an artificial two-dimensional dataset (blue stars) defined by the variables (x_1, x_2) is used to compute the new variables or principal components $(\mathrm{PC}_1, \mathrm{PC}_2)$, which are centred in the mean point of the whole dataset $(\overline{x}_1, \overline{x}_2)$. As can be observed, in the new feature space the first axis PC_1 is aligned with the direction of highest variation or dispersion of the original dataset, whereas PC_2 is orthogonal to PC_1.

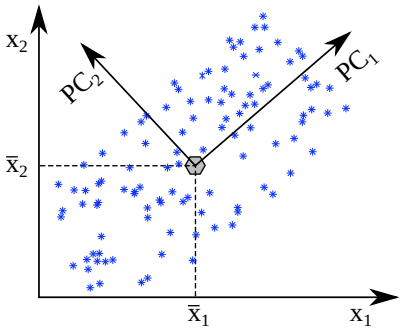


Fig. 3.14: Sketch of the Principal Component Analysis (PCA) transformation for a hypothetical two-dimensional dataset.

The applications of PCA include the reduction of dimensionality, denoising and the detection of outliers.

Dimensionality reduction

As described above, when the first PCs are retained and the rest are discarded, a new feature space of lower dimensionality is obtained, which keeps the main information (in terms of variation) of the original dataset. The level of kept information is determined by the percentage of cumulative explained variance, which is related to the corresponding eigenvalues of the retained components.

The benefits of dimensionality reduction include, for instance, the visualisation of the data, the simplification of the modelling and later analysis of the dataset or

the compression of the data. Among these applications, the visualisation of high-dimensional data in 2D or 3D spaces has special interest to perform an exploratory analysis of the intrinsic structure of the dataset. The information obtained with the exploratory analysis can be very valuable as a complement or as a guide for the later processing and analysis.

As an example, the dataset of 300 pixels' spectra of four types of skin cells that was presented in Fig. 3.6 has been decomposed by PCA. Fig. 3.15 shows the main results both for the raw spectra and the spectra corrected by RMieS-EMSC after 30 iterations. The score plots of the two first PCs (Figs. 3.15a and 3.15d) allow to appreciate an approximation of the *proximity* of the spectra in the original multidimensional space. Attending to the colours of the score plots, the spectra belonging to the same cell type tend to form more clear groups or clusters in the raw spectra, meanwhile there is a higher spread when they are corrected by RMieS-EMSC. This behaviour suggests that the initial differences between cell lines in the raw spectra may be mainly driven by physical effects, which are mostly removed during the preprocessing by RMieS-EMSC. It must be highlighted that the information of the cell types (labels) has only been used *a posteriori* to colour the points of each spectrum but not during the processing by PCA (see Sec. 3.3.2).

The plots of the loadings or eigenvectors (Figs. 3.15b and 3.15e) give information about the relative *weights* that each measured wavenumber has in the corresponding PC. Hence, they inform of the relative *importance* of each absorption band in the existing variations within the dataset. For instance, the two first PCs in the raw spectra (Fig. 3.15b), give higher weights to the Amide I ($\sim$1650 cm^{-1}) and Amide II ($\sim$1550 cm^{-1}) peaks. The loadings of the spectra corrected by RMieS-EMSC (Fig. 3.15e) also assign higher weights to the regions around the Amide I peak, but the position of the maximum weight is slightly shifted depending on the PC. This behaviour warns about suboptimal corrections of resonant Mie scattering, which is supposed to be the main responsible for the position shifts of the strongest absorption peaks (Sec. 3.2.1).

The evolution of the cumulative explained variance with the number of retained PCs (Figs. 3.15c and 3.15f) also gives important information regarding the *complexity* of the relationships between the spectra of the datasets. In this case, there is a clear difference between the raw and the corrected datasets. Meanwhile in the raw spectra

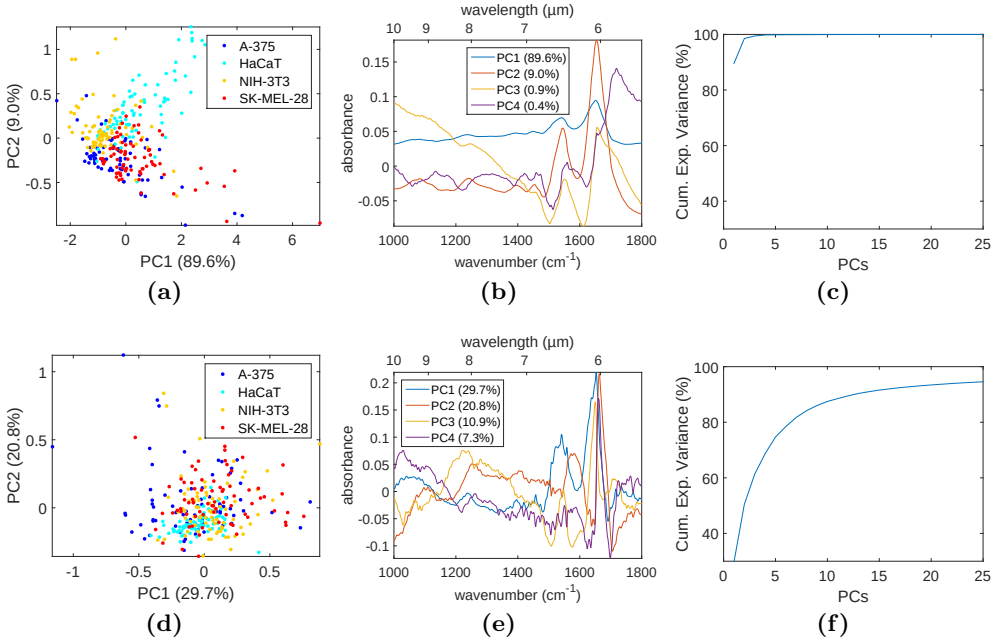


Fig. 3.15: Example of dimensionality reduction and visualisation by PCA of the dataset of skin cell spectra that was presented in Fig. 3.6, both for raw spectra (upper row) and spectra corrected by RMieS-EMSC after 30 iterations (bottom row). (a),(d) Score plots of the two first PCs with the corresponding percentage of explained variance in parentheses. Colours indicate the specific cell type corresponding to each spectrum. (b),(e) Loadings or eigenvectors of the 4 first PCs with the corresponding percentage of explained variance in parentheses. (c),(f) Evolution of the cumulative explained variance for the first 25 PCs.

the two first PCs accounts for more than 98% of the total variance, in the corrected spectra the 95% of explained variance is not reached even retaining the first 25 PCs and the two first PCs roughly explain 50% of the total variance. This difference, for instance, has two consequences:

- Simpler models can be constructed for the raw spectra because most of the information can be condensed by fewer PCs.

- The score plot of the raw spectra (Fig. 3.15a) provides more reliable and further information of the intrinsic structure of the multidimensional dataset than the spectra corrected by RMieS-EMSC (Fig. 3.15d).

Denoising

Another application or *consequence* of PCA is the reduction of noise when the data are approximated by the first PCs. This is because the first PCs are mainly associated with genuine variations within the dataset, whereas the last PCs just measure uncorrelated *noise* [189, 192]. Nevertheless, the number of components from which the explained variation is essentially noise depends on the specific problem or dataset. In a practical way, PCA is normally used as an indirect denoising method, rather than an explicit smoothing methodology. It is important to realise that with this methodology the amount of random noise present in an individual spectrum is reduced by using the total information provided by all the spectra from the dataset. In contrast, common filtering methods only use the information of individual spectra.

In order to illustrate the denoising functionality of PCA, Fig. 3.16a displays the spectra obtained when different numbers of PCs are retained for the same spectrum, which belongs to the dataset of 300 spectra corrected by RMieS-EMSC that was presented in Fig. 3.6b. As can be observed, the reconstructed spectra do not show the high-frequency fluctuations present in the original spectrum even when 20 PCs are kept. In addition, Fig. 3.16b shows the error bars of the SNR for different numbers of retained PCs after applying PCA to the aforementioned dataset of 300 spectra. The reduction of SNR with the number of PCs again demonstrates the increase of noise content when more PCs are retained.

Detection of outliers

There is no clear definition of outliers. In general, outliers are samples that are somehow *anomalous* or *unusual*. However, outliers do not necessarily mean *wrong samples*, but they may belong to sub-populations that are not sufficiently represented in the studied dataset [174, 190]. What to do with the outliers depends on the application, but having a way to automatically detect such *anomalous* samples is very useful.

The ambiguous definition of outliers is coupled with the lack of a universal method for their detection. Here, the Mahalanobis distance is employed as a tool to assess how far each object or spectrum is from the centre of the reduced space of PCs. Mahalanobis distance [193, 194] is a distance widely used in multivariate analysis that

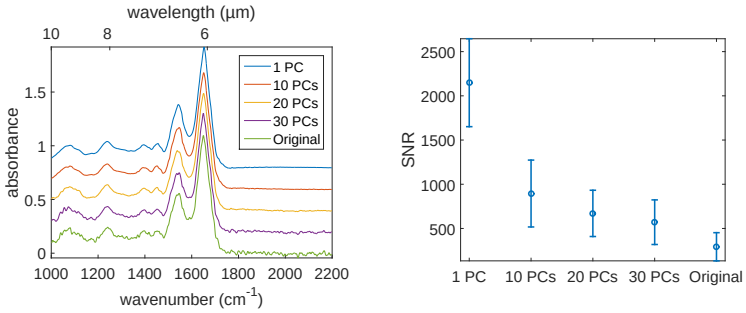


Fig. 3.16: Example of spectral smoothing or denoising by PCA. (a) Original and reconstructed spectra of the same spectrum after retaining the number of PCs specified in the legend. Artificial offsets have been added for clarity. (b) Error bars (mean values symbolised by a circle and standard deviation by bars) of the Signal-to-Noise Ratio (SNR) of the original and reconstructed spectra after applying PCA to the dataset of 300 spectra corrected by RMieS-EMSC that was presented in Fig. 3.6b.

takes into account the intrinsic variations of the dataset by applying a normalisation with the covariance matrix $\mathbf{S}$. Formally, the Mahalanobis distance is defined as:

$$d_{Mah}(\mathbf{x}_i) = \sqrt{(\mathbf{x}_i - \overline{\mathbf{x}})\mathbf{S}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})^{\mathsf{T}}} \tag{3.27}$$

where $\mathbf{x}_i$ is the coordinate vector of the specific object (e.g., the vector of scores), $\overline{\mathbf{x}}$ is the vector of the mean point of the dataset (in the case of centred PCA it is the zero vector) and $\mathbf{S}$ is the covariance matrix of the dataset. As the PCs are uncorrelated, $\mathbf{S}$ is a diagonal matrix whose values are the eigenvalues of each PC.

Once computed the values of d_{Mah} for all spectra, the highest values will be associated with extreme samples. Again, the number of PCs used to construct the reduced space also has influence in the value of d_{Mah}. Fig. 3.17 treats to illustrate the procedure and results of the detection of outliers when retaining 2 and 25 PCs in the dataset of 300 spectra corrected by RMieS-EMSC that was presented in Fig. 3.6b.

Fig. 3.17a shows the box plots of the values of the Mahalanobis distance for the whole dataset when retaining 2 and 25 PCs. In these box plots, whiskers extend to the most extreme data point that is no more than 1.5 times the Interquartile Range (IQR) from the edge of the box. All the values outside the whiskers' range are considered

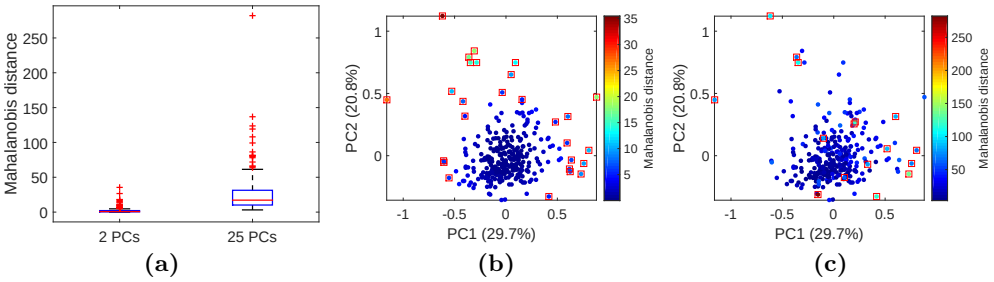


Fig. 3.17: Example of outlier detection by PCA and Mahalanobis distance in the dataset of 300 spectra corrected by RMieS-EMSC that was presented in Fig. 3.6b. (a) Box plot of the Mahalanobis distance when retaining 2 and 25 PCs (outliers are symbolised by red plus signs). Score plots of the two first PCs with the outliers marked by red squares when retaining 2 PCs(b) and 25 PCs (c). The points of each score plot are coloured according to the corresponding colour bar of Mahalanobis distance values.

as outliers and are symbolised by red plus signs. The corresponding points designed as outliers in the box plots can be identified (red squares) in the score plots of the first two PCs. When 2 PCs are retained to find the outliers (Fig. 3.17b), the points marked as outliers are approximately located outside an elliptical region centred in the origin of the reduced space of two PCs. However, the position of the outliers is not so evident when 25 PCs are retained to detect the outliers (Fig. 3.17c) and even some of them are *surprisingly* close to the centre of the reduced space of two PCs. In this dataset, the last option for detecting outliers when 25 PCs are retained seems more logical as they account for a higher level of explained variance (Fig. 3.15f).

3.3.2 Partial least squares

Partial Least Squares (PLS) is another methodology very popular in the chemometric field [174, 181, 183]. Whilst PCA is an *unsupervised* method of dimensionality reduction, as it only takes into account the information contained in the data matrix $\mathbf{X}$, Partial Least Squares (PLS) can be considered as a *supervised* methodology. That is because it also uses the information of the response vector $\mathbf{y}$ to compute a reduced space of components, also referred to as *latent variables*, from the original input variables. This is performed by defining the latent variables through the covariance between the response and the input variables [195].

PLS essentially tries to relate two types of variables: the measurements or *independent variables* contained in the data matrix $\mathbf{X}$ and the response or *dependent variable* $\mathbf{y}$. This is performed by linearly decomposing those variables in a similar way as in PCA:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^{\mathsf{T}} + \mathbf{E} \tag{3.28a}$$

$$\mathbf{y} = \mathbf{T} \cdot \mathbf{q}^{\mathsf{T}} + \mathbf{f} \tag{3.28b}$$

where the first equation has the same form as Eq. 3.26 (but the values obtained by PLS are different), and $\mathbf{q}$ and $\mathbf{f}$ are respectively analogue to a vector of loadings and a vector of residuals for the response vector $\mathbf{y}$. Note that the vector of scores $\mathbf{T}$ is the same in both decompositions of variables $\mathbf{X}$ and $\mathbf{y}$. The matrices involved in Eq. 3.28a have the same structure that was sketched in Fig. 3.13, meanwhile the matrices and vectors of the new equation Eq. 3.28b are sketched in Fig. 3.18.

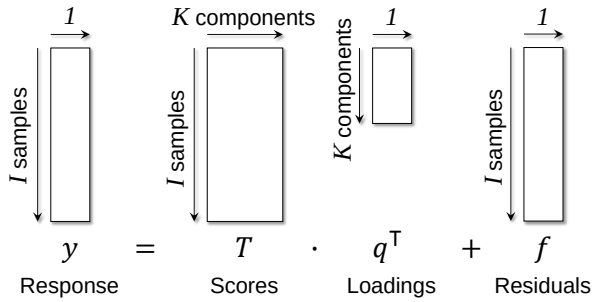


Fig. 3.18: Sketch of the matrices and vectors involved in the decomposition of the response vector by PLS.

Several algorithms exist in the literature to compute the described decomposition by PLS. Here, a variant of the algorithm called *PLS1* [181,183] is applied to establish a relationship between both types of variables ($\mathbf{X}$ and $\mathbf{y}$). PLS1 is an iterative algorithm where the desired number of PLS components K are computed in each iteration. Each iteration of PLS1 consists of the following steps:

1. Compute the weight vector $\mathbf{w}$ of the current PLS component:

$$\mathbf{w} = \mathbf{X}^{\mathsf{T}} \cdot \mathbf{y} \tag{3.29a}$$

2. Compute the vector of scores (corresponding column of the matrix $\mathbf{T}$) for the current PLS component:

$$\mathbf{t} = \frac{\mathbf{X} \cdot \mathbf{w}}{\|\mathbf{w}\|} \tag{3.29b}$$

3. Compute the vector of loadings of data (corresponding column of the matrix $\mathbf{P}$) for the current PLS component:

$$\mathbf{p} = \frac{\mathbf{X}^\mathsf{T} \cdot \mathbf{t}}{\|\mathbf{t}\|^2} \tag{3.29c}$$

4. Compute the loading of the response (corresponding scalar element of the vector $\mathbf{q}$) for the current PLS component:

$$q = \frac{\mathbf{t}^\mathsf{T} \cdot \mathbf{y}}{\|\mathbf{t}\|^2} \tag{3.29d}$$

5. Calculate the residuals of the data matrix $\mathbf{X}_{res}$ and the response vector $\mathbf{y}_{res}$ by subtracting the effect of the current PLS component:

$$\mathbf{X}_{res} = \mathbf{X} - \mathbf{t} \cdot \mathbf{p}^\mathsf{T} \tag{3.29e}$$

$$\mathbf{y}_{res} = \mathbf{y} - \mathbf{t} \cdot q \tag{3.29f}$$

6. If more PLS components are needed, $\mathbf{X}$ and $\mathbf{y}$ are respectively replaced by $\mathbf{X}_{res}$ and $\mathbf{y}_{res}$ and the process is repeated from step 1.

In the above equations, $\|\cdot\|$ symbolises the Euclidean norm of a vector. In PLS1 algorithm, the computed scores $\mathbf{T}$ are orthogonal (as in PCA), but the loadings $\mathbf{P}$ are neither orthogonal nor normalised. Moreover, an additional matrix $\mathbf{W}$ of dimensions $J \times K$, whose columns correspond to the weight vectors $\mathbf{w}$ of each iteration, is also obtained.

Originally, the PLS method was designed for *regression* problems, that is, those cases where the response vector $\mathbf{y}$ is a continuous variable [196]. Nevertheless, it has been successfully adapted to classification and discrimination applications, where it is normally called PLS-Discriminant Analysis (DA) (Sec. 5.2.8). Indeed, in discrimination problems with high number of correlated variables (e.g., FTIR data) where dimensionality reduction is needed, PLS normally outperforms PCA. This is

because *unsupervised* PCA basically computes the new reduced space of variables based on the *gross* variability of the dataset but it is not capable of distinguishing *among-groups* and *within-groups* variability [197]. Nevertheless, PCA is an important preliminary step of purely discriminative algorithms, which otherwise would have a poor performance with high-dimensional multivariate data.

Chapter 4

Multimodal registration of histopathological samples

Contents

4.1 Introduction

In this chapter, the first application of this thesis, aimed at improving the diagnosis of histopathological samples, will be thoroughly described. It consists of an automated framework to spatially align different sections of tissue measured by two distinct imaging modalities: FTIR microspectroscopic imaging and conventional optical microscopy. In addition, the usefulness of its development and application will be supported by data that demonstrate the improvement in classification accuracy of different pathological states of colorectal cancer when it is combined with other approaches to extract information from the tissue.

As was described in Sec. 1.1.6, the final diagnosis of colorectal cancer is performed by expert pathologists, who examine histological sections extracted from suspicious biopsies acquired during colonoscopy. These slices of tissue are commonly stained with H&E to colour the tissue structures (nuclei in blue or purple, cytoplasm and connective tissue in pink) and observed with an optical microscope illuminated with visible light. However, this assessment is still done in a qualitative manner, relying heavily on the judgement and experience of the pathologist. Indeed, several studies [74, 75] have revealed a suboptimal inter-observer variability in the differentiation and reporting of colorectal polyps, suggesting that more objective criteria should be applied for risk stratification in screening and surveillance guidelines.

FTIR images are information-rich data structures that can be analysed, alone or together with other imaging modalities, to provide objective pathological diagnoses. In order to develop new diagnostic algorithms, the different regions of tissue should be correctly labelled and must spatially match between images of distinct modalities. As commented, the H&E stained samples are the *gold standard* where the pathologists can distinguish and label different anatomical and pathological structures within the tissue. Therefore, H&E images may be the most interesting imaging modality to be fused with FTIR images. However, the H&E stain irreversibly changes the chemical composition of the tissue and introduces confounding artifacts in the FTIR spectra [198]. In some cases, the tissue is measured by FTIR spectroscopy before applying the H&E stain. Nevertheless, this option is technically challenging and it impedes performing posterior FTIR measurements and retrospective studies. Therefore, a common solution is to measure different slices of tissue with each imaging modality.

4.1.1 Objective

The main objective of this application is to establish a methodology where the two types of images (H&E and FTIR) from different tissue sections are aligned or registered. The aim of this alignment is to compute the spatial transformation that communicates the coordinate systems of the two images. This transformation makes possible to translate or fuse the spatial information contained within each imaging modality. In particular, the position and auxiliary information (e.g. pathological labels) of different regions of interest, identified by an expert pathologist in the H&E images, can be transferred to the FTIR coordinates with the calculated spatial transformation. In doing so, objective diagnostic algorithms may be created in future studies by employing the fused pixels' information from each imaging modality.

The developed registration method must be robust against the different sources of variability intrinsic to the problem (Sec. 4.1.2). Therefore, another important objective is to assess the robustness and effectiveness of the proposed methodology in a real-world dataset.

4.1.2 Problem overview

Fig. 4.1 outlines the problem faced in this application. Different histological sections are extracted with a microtome from a colon biopsy that has been chemically treated and fixed in a paraffin block for preservation purposes. Some sections follow the H&E staining process and one of them is carried to an optical microscope to take Red-Green-Blue (RGB) images in the visible spectrum range. This colour image is composed of three channels (red, green and blue) so that a three-valued intensity vector is associated with each pixel.

Another section is directly analysed by means of FTIR spectroscopy without any further chemical processing. In the ideal case, this last section would be adjacent to the one used in the optical microscope. However, this ideal case cannot be guaranteed in clinical routine due to problems in the handling and cutting of the biopsy sample. FTIR sensors can measure a large quantity of wavelengths in the near and mid-IR range, providing hyperspectral images that may have hundreds or even thousands of channels. As can be seen in Fig. 4.1, FTIR images are three-dimensional data structures with two spatial dimensions (x,y) that define the position of each pixel and a third spectral dimension ($\tilde{\nu}$), which consists of the recorded wavenumbers. Thus,

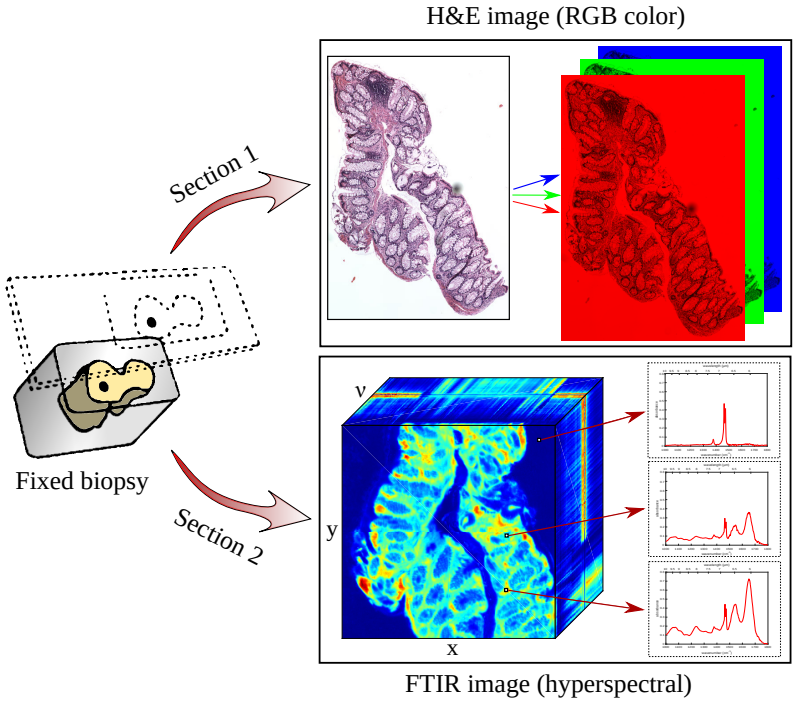


Fig. 4.1: Registration problem overview. Several histological sections are extracted from a paraffin-fixed colon biopsy. One of the sections is stained with H&E and analysed with an optical microscope to take a three-channel RGB image. A different section is directly measured with a FTIR spectrometer to obtain a hyperspectral image, where each pixel has an associated spectrum of hundreds or thousands of channels.

each pixel has an associated absorption spectrum that represents the overall chemical composition of a tissue portion of several squared microns.

The variability between the images of both modalities can be divided into four main sources:

- *Intensity values:* due to the distinct range of wavelengths covered by each image detector. Finding a correct relationship between intensities of corresponding pixels is one of the challenges of this *multimodal* problem.

- *Pixel size:* owing to the characteristics of the optics and the detectors. This can be solved by applying a proper scaling if the exact resolutions in both images are known or by introducing a scaling factor in the spatial transformations.

- *Coarse spatial differences:* because each image device has its own spatial reference. Therefore, large differences in spatial positions and orientation may exist between corresponding anatomical structures.

- *Local spatial differences:* as a consequence of not using the same tissue sections. These dissimilarities may be divided into anatomical differences naturally present in the tissue and local deformations introduced by the physical manipulation, such as cutting with the microtome. If the sections are not adjacent, these local differences are further exaggerated.

The process of finding the alignment between images acquired with different types of sensors and matching their spatial references is called *multimodal registration*. Classical reviews of image registration can be found in [199–202] and a recent comprehensive overview in [203, 204], where the most recent advances in this field are described as well as the techniques applied to medical images. In the problem addressed in this paper, the use of images from different modalities along with the fact that they come from different sections of tissue complicate this task.

4.1.3 Related work

Many approaches exist to deal with the problem of multimodal registration, mainly because the decisions made in each step of the process depend on the application and on the characteristics of the involved images. To the author's knowledge, only two recent studies [205, 206] tackle the problem of registering H&E and FTIR images. Nevertheless, in both of them the same histological sections were firstly measured by FTIR spectroscopy and later stained with H&E to be analysed by optical microscopy.

In [205], the images were taken from Tissue Microarrays (TMAs) of prostate cancer and converted into binary form by trying to separate the pixels that contained tissue from the background. As stated in [205], this binarisation bypasses the multimodal problem because the only reliable features to establish a spatial correspondence between these kinds of images are macroscopic sample shapes and empty spaces (glandular lumens and breaks) inside the tissue.

A more general problem is treated in [206], where images of small regions of tissue obtained using FTIR or other spectroscopic modalities are registered with whole-slide

H&E images by a *template matching* procedure. In that study a pre-segmentation step is performed on each image through *k-means* clustering, taking the FTIR spectrum and the RGB vector as input features for each pixel. The spatial arrangement of the segmented pixels is then matched through an exhaustive sparse search with geometrical restrictions (only translations or little rotations up to ±30 degrees) due to its computational complexity. However, the pre-segmentation step diminishes the spatial local information and can also introduce severe mistakes because the natural clusters may not reflect the same spatial structures in both images due to the different information contained in the RGB vectors and the FTIR spectra.

A different multimodal registration problem involving microscopic images from adjacent sections of prostatic tissue is addressed in [207]. In that study, several adjacent sections with different grades of cancer are stained with either H&E, immunohistochemical or fluorescence dyes. The spatial correspondence between adjacent sections is obtained by means of a rigid registration based on the features automatically provided by the Scale Invariant Feature Transform (SIFT) algorithm [208, 209] and filtered using the Random Sample Consensus (RANSAC) algorithm [210]. This approach is similar to the one used in the first step of the registration framework proposed here (Sec. 4.2.5).

4.1.4 Proposed framework

As stated in Sec. 4.1.1, the objective of this application is to establish a complete pipeline where the two types of images of the different sections of tissue are aligned. After setting some parameters, this alignment is performed in an automatic way through two registration steps which analyse different characteristics of the grayscale images obtained from the initial images. The key step of the suggested method is to obtain these grayscale images from the FTIR hyperspectral data cube.

The proposed framework starts by obtaining grayscale images from the two imaging modalities, which are the inputs of the registration method. The main aim of this preliminary step is to get images with similar local contrast where analogous anatomical structures are easily distinguishable.

The registration process consists of two steps. The first step produces a fast coarse alignment that offers good initial conditions for the finer registration performed in the

second step. The first registration step is a *feature-based registration* that uses the SIFT algorithm [208, 209] to automatically find and match relevant keypoints in both images. These matches are filtered with the RANSAC algorithm [210] to estimate a coarse rigid transformation. Several combinations of SIFT parameters were considered due to the wide spatial variability between the studied sections of tissue, which may not be immediately adjacent. The best combination of parameters in each sample was chosen through the maximisation of a multimodal similarity measure between the aligned images.

The second registration step consists of an *intensity-based registration* that seeks to refine the alignment and to compensate for the local spatial differences between the tissue sections of the two imaging modalities. In this last stage, a non-rigid transformation is iteratively estimated to maximise again the same multimodal similarity measure. Several parameters must be adjusted in the intensity-based algorithm, but this time a global combination of parameters was empirically fixed to reach a reasonable fine alignment.

The obtained results for the available dataset were qualitatively and quantitatively assessed. The rigid transformation obtained with the first registration step was evaluated by comparison with a *gold standard* rigid transformation that was estimated based on manually selected landmarks. The evaluation of the second non-rigid registration step is more difficult because no *gold standard* can be practically established for comparison. Therefore, the results were assessed by studying the modification of a different multimodal similarity measure before and after this last registration step.

4.2 Materials and methods

4.2.1 Dataset

The available dataset consists of 47 colon samples from different pathological groups: 16 normal (non-cancerous), 16 intermediate (comprising adenoma and hyperplastic tissue) and 15 tumoral (cancerous). All samples were fixed and embedded in paraffin blocks. One slice of 7 µm thickness and several contiguous slices of 3 µm were extracted from each block with a microtome. The thicker slice was not further treated

and was measured with the micro-FTIR imaging system that was described in Sec. 2.5. The samples were measured in transmission mode and the absorption spectra were acquired between 1000-3900 cm^{-1} with a wavenumber interval of 4 cm^{-1}. Multiple frames were acquired and combined to allow a larger overall FOV, which varied depending on the size of the sample and the particular ROI.

The rest of sections were chemically deparaffinated and stained with H&E. One of these sections, ideally contiguous to the one measured by FTIR spectroscopy, was chosen to acquire RGB images with an optical microscope. A Philip Harris DMSK211 microscope with a coupled digital camera of 1280×1024 pixels was used to record sequential images that covered the full H&E sample. An objective lens of $20\times$ was used, giving a pixel size of 0.4×0.4 μm^2. The movement of the microscope stage was automatised with a custom hardware based on Raspberry Pi. The tiled images were stitched with Microsoft Image Composite Editor (ICE) in order to obtain the final RGB image of the whole sample.

4.2.2 Registration pipeline

A block diagram with the main steps of the proposed registration pipeline is shown in Fig. 4.2. In order to illustrate the outputs of each step, a representative image from a specific sample is shown next to each block.

The inputs of the process are the two raw data structures of each imaging modality (RGB and FTIR), which were described in Sec. 4.1.2. The first stage of the framework consists of obtaining *representative grayscale images* from the two raw inputs. This is the key step in the pipeline and allows to apply registration methods that have proven to be effective in other medical areas that operate with multimodal grayscale images [203,204,211,212]. The aim is to obtain two images with similar local contrast where analogous anatomical structures can be easily observed. These two grayscale images, named T_1 (RGB) and R (FTIR), are the inputs of the next step.

A *feature-based registration* with a rigid spatial transformation is performed in the second level of the pipeline. Its main objectives are to compensate for the coarse spatial differences between images and to provide an initial alignment based on the correspondence of keypoints, which are automatically detected and matched in both images using SIFT and RANSAC as described in Sec. 4.2.5. In this initial alignment,

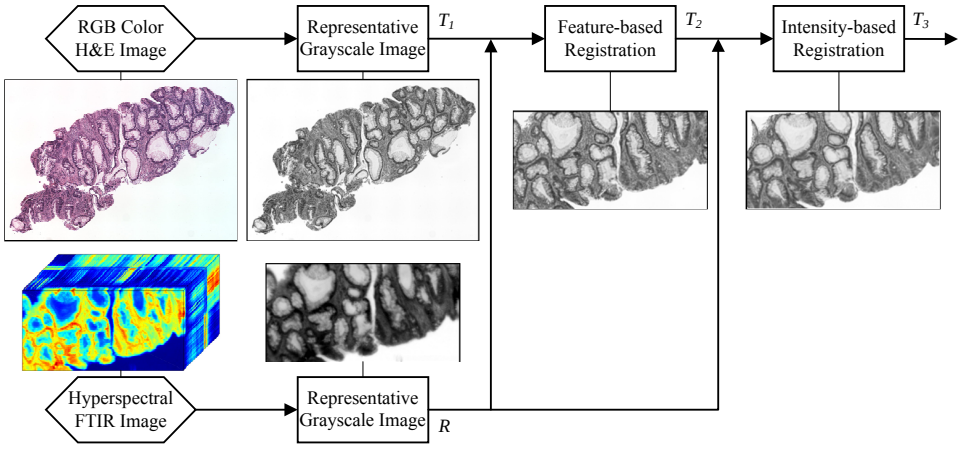


Fig. 4.2: Block diagram of the proposed registration framework. The outputs of each step are illustrated with a representative image from a single sample.

the representative FTIR grayscale image R acts as the *reference* image and the H&E grayscale image T_1 is the *target* image that is shifted through a rigid transformation to produce image T_2.

In the third and final step of the framework, an *intensity-based registration* is applied to perform a non-rigid spatial transformation. It takes as inputs the unaltered *reference* image R and the *target* image T_2 to produce the final output image T_3. The aim of this step is twofold: on the one hand, to refine any remaining misalignment and on the other hand, to diminish local spatial differences due to the use of different slices of tissue.

4.2.3 Representative images

The critical step in this work is converting the data structures of each modality into grayscale images such that registration can be performed in the grayscale domain. The goal is to create two images with similar local contrast where corresponding anatomical structures can be easily observed. The process to obtain these grayscale images is different for each imaging modality.

H&E grayscale image

Two steps were applied to construct the H&E grayscale image:

1. *Resizing:* A bi-cubic interpolation with an antialiasing filter was performed to downscale the original H&E colour images to the same pixel size as FTIR images.

2. *Transformation to grayscale:* The RGB values of the resized images were converted to grayscale by computing the luma $\mathcal{Y}$, which is the achromatic component representing luminance in the YIQ colour space [213]. In the studied dataset, luma component demonstrated to have a high rate of convergence to an optimal solution, especially in the first registration step, as described in Sec. 4.3.2. In addition, it demonstrated to be more robust against different illumination conditions (some samples had uneven spatial illumination) and staining variability (e.g. cancerous regions stain darker than normal ones) than other chromatic components. In a theoretical way, human visual system is more sensitive to luminance differences rather than chromatic differences. The luminance is closely related to the perceptual attribute called *brightness*, which is the visual sensation according to which a source appears to emit more light or less than another does [213]. On the other hand, absorbance values of the FTIR images may be interpreted as the capability of an object to absorb more light or less than another does. Therefore, theoretically luma component may be a good candidate to be inversely related to a transformation that condenses the absorbance values of the FTIR images (see next section). As commented before, this theoretical intuition was empirically supported by the results obtained and compared with a manual gold standard registration. The specific conversion to obtain $\mathcal{Y}$ is a weighted combination of the three-valued vector ($\mathcal{R}$,$\mathcal{G}$,$\mathcal{B}$):

$$\mathcal{Y} = 0.299 \cdot \mathcal{R} + 0.587 \cdot \mathcal{G} + 0.114 \cdot \mathcal{B} \tag{4.1}$$

FTIR grayscale image

Before transforming the FTIR hyperspectral data into a single grayscale image, the spectra must be preprocessed. As was extensively described in Sec. 3.2, preprocessing is a relevant stage in FTIR data analysis and interpretation [153, 177]. It is

essential to remove the unwanted effects involved in the spectral data acquisition and to highlight specific information within the spectra. No universal preprocessing method exists [108]; the specific combination of preprocessing steps depends on the application.

In order to illustrate the applied preprocessing, the final representative FTIR grayscale image of one sample is shown in Fig. 4.3d. In this image five pixels have been coarsely marked with different colours and numbers. Pixels 1 and 2 exemplify two regions without tissue (substrate), pixel 3 corresponds to a region where there is a low amount of sample (interior of lumen), pixel 4 represents a region where the presence of organic material is higher than in pixel 3 (outer region of lumen) and, finally, pixel 5 illustrates a region with the highest content of tissue of the selected pixels. The same colours and numbers are used to represent the corresponding raw spectra in Fig. 4.3a. Strong peaks due to contaminants such as CO_2 ($\sim$2300-2400 cm^{-1}) and paraffin ($\sim$2800-3000 cm^{-1}) can be easily observed in the raw spectra because they are even present in the substrate regions (pixels 1 and 2).

The preprocessing steps applied in these images, with the reasoning behind their use, were:

1. *Spectral window selection:* The raw absorption spectra were cropped to the so-called *fingerprint region* (1000-1800 cm^{-1}), which contains the vibrational frequencies of the chemical bonds from the most relevant biomolecules [81]. Thus, attention is focused on the most informative wavenumbers and the stronger contaminant bands are rejected (Fig. 4.3b).

2. *Removal of remaining spectral contaminants:* Relatively strong peaks due to paraffin still appear in the *fingerprint region* (1360-1390 and 1430-1490 cm^{-1}). The absorbance values corresponding to those ranges of wavenumbers are removed from the spectra to reduce the noise produced in the final images by the interference of these contaminants and the rest of meaningful biochemical peaks.

3. *Denoising:* A Savitzky-Golay filter was applied to smooth the spectra and reduce the random noise. Its parameters were fixed to a window size of 15 points and a 2nd order fitting polynomial, which were a good trade-off between noise attenuation and signal distortion in the studied dataset.

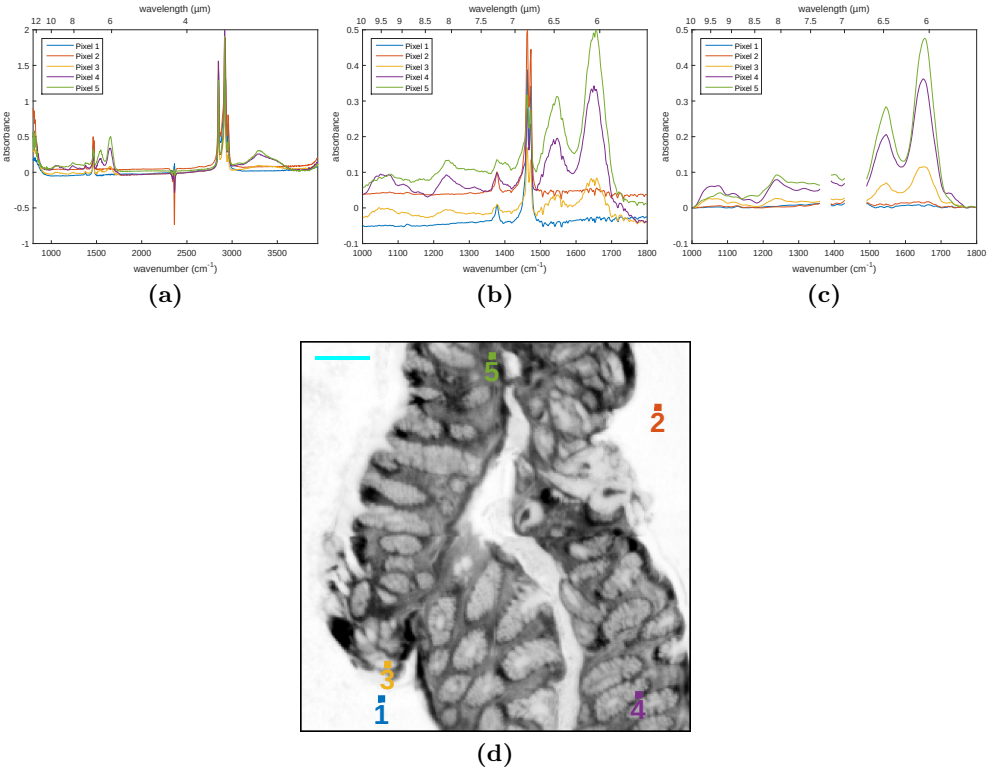


Fig. 4.3: (a) Full-range raw spectra drawn in the same colour of their corresponding pixels. (b) Raw spectra cropped to *fingerprint region*. (c) Preprocessed spectra after cropping to the fingerprint region (1000-1800 cm^{-1}), smoothing by Savitzky-Golay filtering, rejecting the paraffin ranges (1360-1390 and 1430-1490 cm^{-1}) and applying rubberband baseline correction. (d) Representative FTIR grayscale image of a colon sample where five pixels have been marked and numbered with different colours. Cyan scale bar represents 200 μm.

4. *Baseline correction:* It is crucial to diminish the spectral baseline effects due to scattering, heterogeneity in external illumination, supporting substrate or sensor's sensitivity and other changing conditions during data collection [153]. There is a large variety of baseline correction algorithms. Here the *rubberband baseline correction* method was employed [177], which subtracts a convex polygonal line whose edges are minima within the spectrum. In particular, the parameterless implementation of the *rubberband baseline correction* included in the IRootLab toolbox [214] was used.

The final preprocessed spectra of the selected pixels are shown in Fig. 4.3c. The selected pixels have been numbered in increasing order of maximum absorbance values. Absorbance is related to the concentration and thickness of the biological content as stated by the Beer-Lambert law (Sec. 2.4.1). There are some metrics commonly used in vibrational spectroscopy to condense the absorbance measurements, such as peak height or peak integral [215]. In this study, different combinations of these metrics were explored, such as finding the maximum or computing the integral of the complete *fingerprint region* or only the strongest peak (Amide I peak, between 1630-1670 cm^{-1}). However, it was found that the standard deviation (*std*) of each spectrum was the most robust metric (in terms of noise, sharpness and independence of previous preprocessing steps) to produce an image with a local contrast similar to the corresponding H&E grayscale image. In particular, the intrinsic removal of the mean absorbance value performed by *std* reduces the influence of spectral baseline artifacts (e.g. produced by scattering) inefficiently removed in the baseline correction preprocessing step. More specifically, if $\mathbf{A} = \{A_1, \ldots, A_{N_w}\}$ is the associated absorption spectrum of a pixel with N_w wavenumbers, then the assigned value to that pixel was computed as:

$$std(\mathbf{A}) = \sqrt{\frac{1}{N_w - 1} \sum_{i=1}^{N_w} \left(A_i - \overline{A}\right)^2}, \tag{4.2}$$

where $\overline{A} = \frac{1}{N_w} \sum_{i=1}^{N_w} A_i$ is the mean absorbance value of $\mathbf{A}$. The obtained *std* values were linearly mapped to get a grayscale image with intensities between 0 and 255. However, this gives the substrate low gray values, as opposed to the H&E grayscale image where the substrate has high values and appears white. For this reason the gray values were inverted to obtain the reference image R in the proposed registration method.

4.2.4 Similarity measures

Images from different imaging modalities have different intensity characteristics due to the specific properties of each sensor. Thus, multimodal registration problems require similarity measures based on statistical relationships between the pixels of the images rather than direct intensity relationships as in monomodal scenarios [216].

Two similarity measures widely used in multimodal problems were employed: the correlation ratio and the mutual information.

Correlation ratio

The correlation ratio (CR) assumes a functional relationship between the intensities of the registered images [217, 218]. CR does not consider any specific relationship, whereas the correlation coefficient does. The correlation coefficient assumes a linear relationship between intensities, making it more appropriate for monomodal registration problems. CR can take values from 0 (no functional dependence) to 1 (purely deterministic dependence) [218]. The closer CR is to 1, the more similar R and T are and, in the extreme case (CR = 1), it would inform of a hypothetical perfect alignment. To compute CR, the reference image R and the target image T are viewed as random variables. Let $\mathbf{x}$ denote a pixel that has an intensity $T(\mathbf{x})$ in image T; Ω the set of pixels in the overlapping region between R and T; N_p the total number of pixels in Ω; Ω_i the subset of Ω whose pixels have the same intensity level i in R, that is, $\Omega_i = \{\mathbf{x} \in \Omega,\, R(\mathbf{x}) = i\}$; $N_{p,i}$ the number of pixels in Ω_i; then, CR is defined as:

$$\mathrm{CR}(R,T) = 1 - \frac{1}{N_p \sigma^2} \sum_i N_{p,i} \sigma_i^2, \tag{4.3}$$

where

$$\begin{aligned} \sigma^2 &= \frac{1}{N_p} \sum_{\mathbf{x}\in\Omega} T(\mathbf{x})^2 - m^2, \quad m = \frac{1}{N_p} \sum_{\mathbf{x}\in\Omega} T(\mathbf{x}), \\ \sigma_i^2 &= \frac{1}{N_{p,i}} \sum_{\mathbf{x}\in\Omega_i} T(\mathbf{x})^2 - m_i^2, \quad m_i = \frac{1}{N_{p,i}} \sum_{\mathbf{x}\in\Omega_i} T(\mathbf{x}). \end{aligned} \tag{4.4}$$

Mutual information

Mutual information (MI) is a measure from information theory that characterises the amount of shared information between the registered images [219]. It assumes a probabilistic relationship between the intensities of the registered images R and T.

MI can range from 0 (statistical independence) to any positive quantity and, as in CR, the higher MI, the more similar R and T are. However, interpretation of MI is difficult because there is not a maximal value that can be used as a reference for a hypothetical *perfect* alignment. MI is computed in terms of image entropies:

$$\mathrm{MI}(R,T) = H_R + H_T - H_{RT}. \tag{4.5}$$

The three involved entropies are defined as:

$$\begin{aligned} H_R &= -\sum_i P_i \log_2(P_i), \\ H_T &= -\sum_j P_j \log_2(P_j), \\ H_{RT} &= -\sum_{i,j} P_{ij} \log_2(P_{ij}), \end{aligned}$$

where P_i is the probability of an intensity level i occurring in image R; P_j is the probability of intensity level j occurring in image T; P_{ij} is the joint probability of both intensity levels i in image R and j in image T occurring at the same position.

4.2.5 Feature-based registration

Scale Invariant Feature Transform (SIFT) is an algorithm to detect and describe local features in images [208, 209]. The main characteristic of SIFT is its ability to find distinctive keypoints that are invariant to location, scale and rotation, and robust to affine transformations (changes in scale, rotation, shear and position) and changes in illumination, which makes it usable for object and pattern recognition. SIFT is the core algorithm employed for keypoint selection and feature extraction. In the SIFT algorithm, a series of keypoints that are invariant to scale and orientation are firstly detected by seeking extrema in a Difference-of-Gaussian (DoG) transformation. At each candidate keypoint, a local descriptor relative to scale-invariant coordinates is computed based on local image gradients. The feature descriptors of the candidate keypoints in both images are matched by a nearest-neighbour strategy through minimum Euclidean distance.

The values of four relevant SIFT parameters have been explored to get different rigid transformations for each sample:

- *Number of orientations* (r) *and width of the descriptor* (n)*:* these two parameters change the size (rn^2) of the local descriptor vector, which is computed in a $n \times n$ array of histograms of r orientations. When these values increase, the descriptor complexity and discriminative properties grow. The highest rates of convergence were achieved with values of $r = \{4, 8\}$ and $n = \{4, 6\}$.

- *Number of scale samples per octave* (s)*:* this parameter modifies the number of extrema detected in the DoG transformation. The higher its value, the more candidate keypoints are detected. Although Lowe [209] experimentally found an optimal value of 3, more robust results (higher rates of convergence) were obtained with values of $s = \{6, 10\}$ in the studied dataset.

- *Ratio of distances closest/next closest* (*rod*)*:* each candidate keypoint is only retained if the ratio of distances between the first and the second nearest matched neighbour is below *rod*. When this threshold decreases, a higher number of false matches is discarded although good matches can also be rejected. The highest rates of convergence were obtained with $rod = \{0.8, 0.9\}$.

Afterwards, the Random Sample Consensus (RANSAC) algorithm [210] is used to filter the candidate keypoints matched by means of the SIFT descriptors. In RANSAC, a spatial transformation model between both images must be imposed in order to estimate its parameters. In this case, a rigid transformation $\mathcal{T}$ is considered, which follows the equation:

$$\mathcal{T}(\mathbf{x}) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix}, \tag{4.6}$$

where the original spatial coordinates $\mathbf{x} = (x, y)$ are converted into the new ones by applying a global translation (t_x, t_y) and a rotation of angle θ.

RANSAC computes the parameters t_x, t_y and θ by considering a minimum percentage of inlier matched keypoints (5% in this case) and a maximal alignment

error between the transformed keypoints and their corresponding matches. For this work the alignment error was limited to a maximum of 15 pixels.

Due to the spatial variability between samples, a rigid transformation may not be constructed for some combinations of parameters because no inlier correspondences are detected. The selection of the best combination of parameters for each sample was therefore based on the maximisation of a similarity measure (Sec. 4.2.4) between the output target image T_2 and the reference image R. In particular, the correlation ratio (CR) was the variable to maximise.

4.2.6 Intensity-based registration

The second registration step is based on a variational approach, which has been formulated in the frequency domain [220, 221] and also implemented in the frequency domain [222], providing a fast and efficient registration method. This method produces a non-rigid displacement field $\mathbf{u}$ that compensates for the remaining differences between the reference R and the target T_2 images caused by an imperfect alignment in the first registration step and the local spatial differences between tissues. The obtained non-rigid displacement field $\mathbf{u} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ will make the transformed target image similar to the reference image, $T_2(\mathbf{x} - \mathbf{u}(\mathbf{x})) \approx R(\mathbf{x})$, where $\mathbf{u}(\mathbf{x}) = (u_x(\mathbf{x}), u_y(\mathbf{x}))^\top$ and $\mathbf{x}$ is the spatial position $\mathbf{x} = (x, y) \in \mathbb{R}^2$.

The non-parametric registration can be approached in terms of calculus of variations by defining the following energy functional to be minimised:

$$\mathcal{J}[\mathbf{u}] = \mathcal{D}[R, T_2; \mathbf{u}] + \alpha \mathcal{S}[\mathbf{u}]. \tag{4.7}$$

The energy term $\mathcal{D}$ measures the distance between the deformed target and the reference images; $\mathcal{S}$ is a penalty term which acts as a regulariser and determines the smoothness of the displacement field; and $\alpha > 0$ weighs the influence of the regularisation.

The distance measure $\mathcal{D}$ is chosen depending on the datasets to be registered. In this application, since the H&E and FTIR datasets do not share the same intensity range (*multimodal* problem), statistical-based measures are more appropriate. Among the similarity measures described in Sec. 4.2.4, the correlation ratio (CR) was used in this case because it provided more accurate alignments and had a better behaviour

against wrong global minima during its optimisation. The regularisation term $\mathcal{S}$ gives the smoothness characteristics to the displacement field [223]. In this problem, the diffusion term was used, which is given by the energy of the first derivatives of $\mathbf{u}$ [220].

As described in [220], the energy functional (Eq. 4.7) can be translated into the frequency domain by means of Parseval's theorem. Then $\mathcal{J}[\mathbf{u}] = \tilde{\mathcal{J}}[\tilde{\mathbf{u}}]$, with $\tilde{\mathbf{u}}(\boldsymbol{\omega}) = (\tilde{u}_x(\boldsymbol{\omega}), \tilde{u}_y(\boldsymbol{\omega}))^\top$ being the frequency counterpart of the displacement field, $\boldsymbol{\omega} = (\omega_x, \omega_y)$ is the two-dimensional variable in the frequency domain. The minimisation of the energy functional $\tilde{\mathcal{J}}[\tilde{\mathbf{u}}]$ leads to the translation of the Euler-Lagrange equation into the frequency domain and provides the following iteration for the Fourier transform of the l-th component of the displacement field:

$$\tilde{u}_l^{(\eta)}(\boldsymbol{\omega}) = H(\boldsymbol{\omega})\big(\tilde{u}_l^{(\eta-1)}(\boldsymbol{\omega}) - \alpha\, \tilde{f}_l^{(\eta-1)}(\boldsymbol{\omega})\big), \tag{4.8}$$

where $\eta \in \mathbb{N}$ is the iteration index, $l = \{1, 2\}$ in this 2D problem, $H(\boldsymbol{\omega})$ is a low-pass filter in the frequency domain and $\tilde{\mathbf{f}}(\boldsymbol{\omega}) = (\tilde{f}_x(\boldsymbol{\omega}), \tilde{f}_y(\boldsymbol{\omega}))^\top$ is the Fourier transform of the external forces field. For further details, please refer to [220, 221].

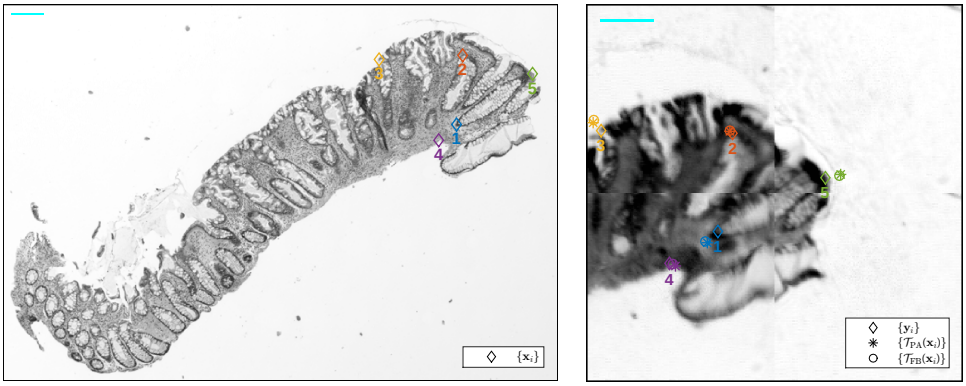
From Eq. 4.8, the target image is iteratively modified by $\mathbf{u}$ and goes through different intermediate states T_i until the final image $T_3(\mathbf{x}) = T_2(\mathbf{x} - \mathbf{u}(\mathbf{x}))$ is obtained. Eq. 4.8 provides a stable implementation for the computation of a numerical solution for the displacement field in a more efficient way than existing approaches if the two-dimensional fast Fourier transform is used [221].

The values of the most relevant parameters in this step were: a maximum number of iterations (η_{max}) equal to 400 and $\alpha = 500$. The images R and T_2 were preprocessed with a 3×3 and a 5×5 median filter, respectively, before applying this last registration step to avoid excessive local distortions that may be induced by noise artifacts.

4.2.7 Evaluation

No *ground truth* for alignment evaluation exists in this problem mainly for two reasons: placing external fiducial marks in the microscopic tissue is not trivial; the deformations produced during the processing of the different slices of tissue can be highly variable. Even so, special care was taken to evaluate the obtained results in both registration stages.

The rigid registration step (Sec. 4.2.5) was assessed by a procedure similar to the one used in [207]: an artificial rigid transformation was computed in each sample as a *gold standard* for comparison. This *gold standard* transformation was estimated by applying Procrustes analysis [224–226] to a set of manually selected landmarks or control points. These sets of landmarks were manually chosen by trying to select anatomical structures shared by both images to register (R and T_1), such as distinctive points in the glands or in the tissue border. Formally, a set of N_p pixels or spatial points $\{\mathbf{x}_i\}$ were selected in the original target image T_1 and their corresponding set of points $\{\mathbf{y}_i\}$ were marked in the reference image R, where $i = 1, \ldots, N_p$. For each sample, five corresponding points ($N_p = 5$) were selected in both images. As an illustration, Fig. 4.4 shows the images T_1 and R of one tissue sample where the two corresponding sets of points $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$ are marked with different coloured diamonds.



(a) T_1 **(b)** R

Fig. 4.4: Illustration of the sets of manual landmarks used in the evaluation of the feature-based registration step. Cyan scale bars represent 200 µm. (a) Target image T_1 with five manually selected points $\{\mathbf{x}_i\}$ marked with diamonds. (b) Reference image R with the corresponding set of manually selected points $\{\mathbf{y}_i\}$ (diamonds), the transformed set $\{\mathcal{T}_{\mathrm{PA}}(\mathbf{x}_i)\}$ (asterisks) obtained by Procrustes analysis and the transformed set $\{\mathcal{T}_{\mathrm{FB}}(\mathbf{x}_i)\}$ (circles) obtained by the feature-based registration step.

The set $\{\mathbf{x}_i\}$ can be mapped into the reference space of image R with a generic rigid transformation (Eq. 4.6) to give a set of transformed points $\{\mathcal{T}(\mathbf{x}_i)\}$. The root-mean-square of the distances between the points of a generic transformed set $\{\mathcal{T}(\mathbf{x}_i)\}$ and their corresponding $\{\mathbf{y}_i\}$ points is the Fiducial Registration Error (FRE) [226]:

$$\mathrm{FRE}(\mathbf{x}_i, \mathbf{y}_i, \mathcal{T}) = \sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} \left(\{\mathcal{T}(\mathbf{x}_i)\} - \{\mathbf{y}_i\}\right)^2}, \tag{4.9}$$

Two FRE values can be computed for each sample by applying each rigid transformation: one for the *gold standard* transformation computed with Procrustes analysis ($\mathrm{FRE_{PA}}$) and one for the transformation obtained with the proposed feature-based registration ($\mathrm{FRE_{FB}}$). These values respectively inform of the mean alignment error in pixel units of the transformed sets $\{\mathcal{T}_{\mathrm{PA}}(\mathbf{x}_i)\}$ and $\{\mathcal{T}_{\mathrm{FB}}(\mathbf{x}_i)\}$. Procrustes analysis [224] computes the optimal linear transformation (rigid in this case) by least squares minimisation of the distances between the cloud of landmarks selected in the reference image R ($\{\mathbf{y}_i\}$) and their corresponding transformed landmarks from the target image T_1 ($\{\mathcal{T}(\mathbf{x}_i)\}$), i.e., by minimising FRE [225, 226]. Thus, $\mathrm{FRE_{PA}}$ represents the lowest possible error (in terms of minimum least squares) for the manually selected sets of landmarks, what justifies the use of Procrustes analysis as the *gold standard* rigid transformation. The difference error ($\Delta\mathrm{FRE} = \mathrm{FRE_{FB}} - \mathrm{FRE_{PA}}$) was also computed for each sample to show explicitly the disagreement between both rigid registration methods. The transformed sets $\{\mathcal{T}_{\mathrm{PA}}(\mathbf{x}_i)\}$ and $\{\mathcal{T}_{\mathrm{FB}}(\mathbf{x}_i)\}$ for one tissue sample have been represented in Fig. 4.4b with asterisks and circles, respectively. As can be seen, the agreement between the two rigid registrations in the chosen sample is very high.

The evaluation of the non-rigid transformation performed in the intensity-based registration step (Sec. 4.2.6) is not an easy task and no well-established validation methods exist [201,216]. The comparison with a manual *gold standard* transformation based on selected landmarks is not appropriate due to the local deformations produced by the non-rigid transformation. These spatially uneven deformations may lead to a deceptive estimation of the errors in the positioning of the manually selected landmarks, which are actually very difficult to pick due to the use of different slices of tissue. A usual validation method for non-rigid intensity-based approaches is the computation of a similarity measure (Sec. 4.2.4) between the target and the reference images before and after applying the non-rigid registration [216]. In order to avoid false conclusions, the similarity measure used for this evaluation must be different from the one (CR) maximised during the registration. In this case, MI was computed for each sample and its improvement was statistically analysed by pathological group

and over the whole dataset. In each group, the right-tailed Wilcoxon signed-rank test [227] was applied, whose null hypothesis states that the differences before and after applying the intensity-based registration step have zero median, meanwhile the alternative hypothesis states that the median of these differences is positive. A p-value was obtained for each group to assess the statistical significance of rejecting the null hypothesis in favour of the alternative hypothesis of this non-parametric test.

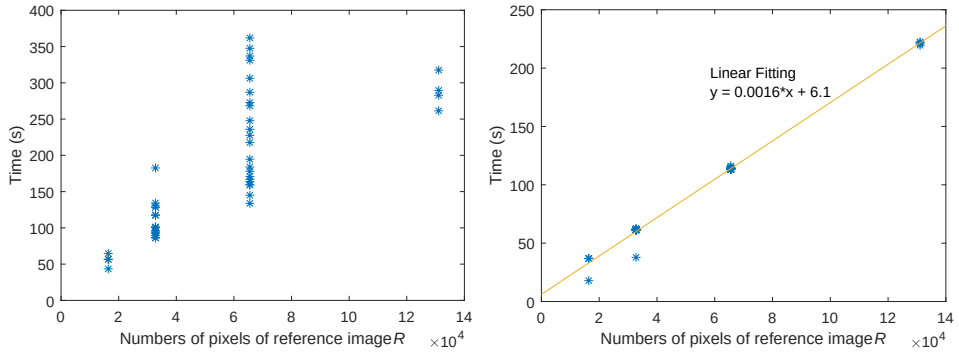
Finally, careful visual assessment remains the first and most important validation check [201,216] especially in this multimodal registration problem which involves non-rigid deformations. The best visual results were obtained when CR was considered as the similarity measure to maximise through the complete registration pipeline. It must be remarked that the subjectivity in this last assessment is unavoidable due to the high variability introduced by the non-rigid distortions.

4.2.8 Implementation

All the steps of the suggested registration framework, including its evaluation, were implemented in MATLAB. As external programs, the IRootLab toolbox [214] was used to perform the preprocessing of FTIR images and the implementation of SIFT and RANSAC algorithms included in the software Fiji/ImageJ [228, 229] were linked with in-house MATLAB algorithms through the MIJ Java package [230].

The implemented code was applied to the available dataset (Sec. 4.2.1) in a computer with a processor Intel® Core™ i7-4790K @ 4.00GHz and 16GB of RAM memory, running under Windows 8.1 and with the version R2016b of MATLAB installed. Fig. 4.5 shows the running times in this computer platform versus the size of the reference image R, in terms of total number of pixels. As described in Sec. 4.2.1, the FTIR reference image R is composed of different frames of 128×128 pixels which cover the corresponding region of interest in each sample. As can be seen in Fig. 4.5a, the computation time for the complete proposed framework varies from around 40 to 360 seconds, with an increasing but not clear relationship with the size of the reference image R. However, as presented in Fig. 4.5b the running time in the intensity-based registration step does follow a linear relationship with the size of R with an estimated slope by least-squares (coefficient of determination $R^2 = 0.99$) of 1.6 miliseconds/pixel. The variability in the total running time comes from the

feature-based registration step, whose computation not only depends on the size of the images to be registered but also on their content. Images with higher number of relevant structures increases the detection of candidate keypoints and the time for matching and filtering them with SIFT and RANSAC algorithms (Sec. 4.2.5).



(a) **(b)**

Fig. 4.5: Running times (in the computer platform described in the main text) vs. number of pixels of reference image R using the samples of the available dataset for (a) the complete proposed framework and (b) the intensity-based registration step.

Tab. 4.1 summarises the most relevant parameters of both registration steps, including the applied values in the available dataset and the risks of using too low or too high values. The relevance of SIFT parameters in the feature-based registration step were described in Sec. 4.2.5. Choosing incorrect values of these parameters would create incorrect or insufficient matches between keypoints, leading to the computation of wrong rigid transformations or, even worse, the failure of finding any transformation. In the intensity-based registration step, the main parameter that can be tuned is α, which controls the smoothness of the applied deformation. The higher α, the smoother the deformation, at the risk of not compensating sufficiently for the local spatial differences. On the contrary, the lower α, the more unrealistic excessive deformations may be applied. Finally, η_{max} is mainly related to the chosen α. In this case, a relatively high value of η_{max} was selected to be conservative.

As last remarks, the suggested parameter values should be a good starting point for other datasets with different sources of variability between images. In the feature-based registration step, the indicated methodology of exploring different combinations

Tab. 4.1: Summary of the most relevant parameters of each registration step

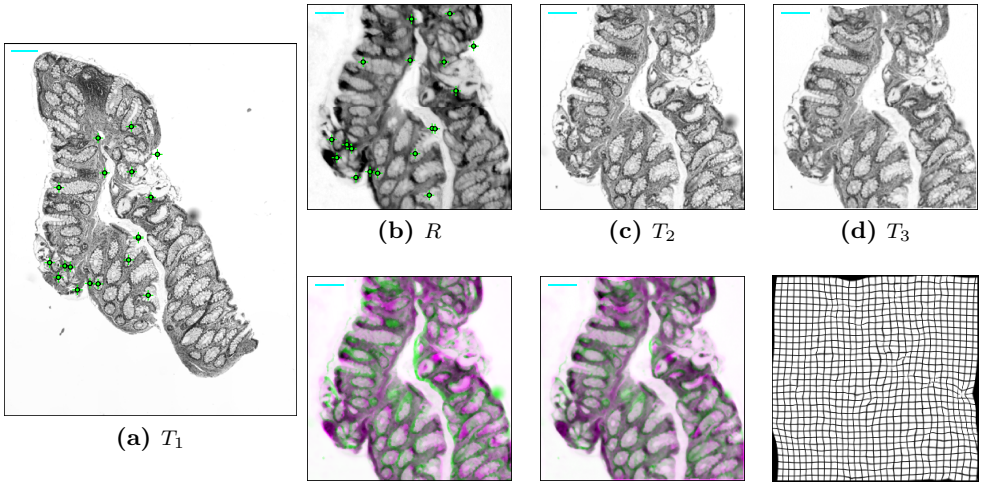
Registration step	Parameter	Applied values	Risk of low values	Risk of high values
Feature-based	r	$\{4, 8\}$	Descriptor with lacking information	Descriptor with excessive noise
	n	$\{4, 6\}$		
	s	$\{6, 10\}$	Not enough keypoints	Many confounding keypoints
	rod	$\{0.8, 0.9\}$	Right matches discarded	Wrong matches retained
Intensity-based	α	500	Excessive local deformation	Insufficient local deformation
	η_{max}	400	Optimal solution not reached	Excessive computation time

of parameters and choosing the one that maximises a similarity measure between images (CR in this case) proved to be able to handle satisfactorily different scenarios. In the intensity-based registration step, no significant changes in the results were observed in the range between a double and a half of the suggested values.

4.3 Results

4.3.1 Qualitative results

One sample per pathological group has been chosen to present the most relevant images obtained in the different steps of the proposed registration framework. Fig. 4.6 shows the images of the chosen intermediate sample, Fig. 4.7 the normal sample and Fig. 4.8 the tumoral sample. In these figures, the automatically-detected inlier keypoints of the feature-based registration step (Sec. 4.2.5) have been marked in the original target T_1 and reference R images. The output target images of the feature-based T_2 and the intensity-based T_3 registrations are also shown. These output images have been independently overlaid with the reference image R to create composite RGB images whose gray levels denote regions with similar intensities in the two overlaid images, whereas magenta and green regions correspond to different intensities. Although the presence of magenta or green in some cases does not mean a bad alignment (due to the multimodality of the overlaid images), these representations allow a better visualisation of the alignment produced by each registration step.

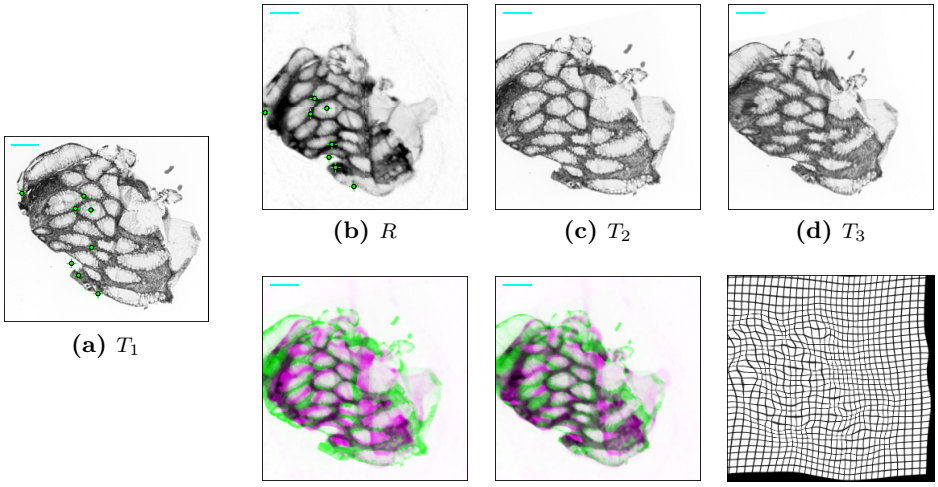


(a) T_1 **(b)** R **(c)** T_2 **(d)** T_3 **(e)** R-T_2 composite **(f)** R-T_3 composite **(g)** Deformation

Fig. 4.6: Registration results for an intermediate sample. Cyan scale bars represent 200 µm. (a) Target image T_1 with automatically-detected inlier keypoints. (b) Reference image R with automatically-detected inlier keypoints. (c) Feature-based registration output image T_2. (d) Intensity-based registration output image T_3. (e) Composite RGB image of overlaid images R and T_2. (f) Composite RGB image of overlaid images R and T_3. (g) Artificial grid deformed with the non-rigid displacement field computed in the intensity-based registration.

Finally, and also for better visualisation of the applied deformation, the figures show synthetic grids with the same dimensions as T_2 and T_3, which were deformed with the non-rigid displacement fields $\mathbf{u}$ computed in the intensity-based registration step. Each squared interval of the undeformed synthetic grids was chosen to occupy 16 pixels both in horizontal and vertical directions.

As can be seen in figures 4.6 to 4.8, the feature-based registration step is capable of compensating for coarse misalignment (very relevant in the chosen intermediate sample) and cropping the initial target image to the region of interest of the reference image. It can be observed that the detected inlier keypoints of the tumoral sample are placed near the edge of the tissue. In the intermediate and the normal samples, the keypoints are also identified in the inner region of the tissue, mainly close to the glandular regions. As the SIFT algorithm detects keypoints at different resolutions, the ones which appear outside the tissue at this resolution belong to the border of the tissue at a lower resolution.



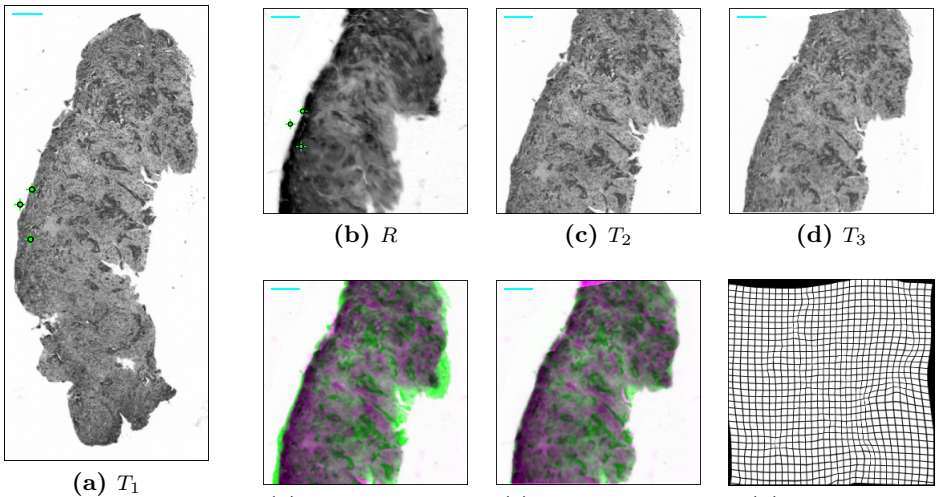
(a) T_1 **(b)** R **(c)** T_2 **(d)** T_3 **(e)** R-T_2 composite **(f)** R-T_3 composite **(g)** Deformation

Fig. 4.7: Registration results for a normal sample. Cyan scale bars represent 200 µm. (a) Target image T_1 with automatically-detected inlier keypoints. (b) Reference image R with automatically-detected inlier keypoints. (c) Feature-based registration output image T_2. (d) Intensity-based registration output image T_3. (e) Composite RGB image of overlaid images R and T_2. (f) Composite RGB image of overlaid images R and T_3. (g) Artificial grid deformed with the non-rigid displacement field computed in the intensity-based registration.

It can also be seen that the intensity-based registration step refines the general misalignment (better appreciated in the chosen normal sample) and also produces local deformations that improve the correspondences inside the tissue. These local modifications can be better viewed in the glands of the intermediate and normal tissues.

4.3.2 Quantitative evaluation

Tab. 4.2 summarises the quantitative evaluation of the results for the 16 Intermediate (INT), 16 Normal (NOR) and 15 Tumoral (TUM) samples of the dataset, identified by the code IDsample. The feature-based registration results show: the combination of SIFT parameters that has been chosen as optimal (IDcomb; see Tab. 4.3); the fiducial registration error for the Procrustes analysis (FRE_{PA}) *gold standard* registration, for the proposed feature-based registration (FRE_{FB}) and the difference between them



(a) T_1 **(b)** R **(c)** T_2 **(d)** T_3 **(e)** R-T_2 composite **(f)** R-T_3 composite **(g)** Deformation

Fig. 4.8: Registration results for a tumoral sample. Cyan scale bars represent 200 µm. (a) Target image T_1 with automatically-detected inlier keypoints. (b) Reference image R with automatically-detected inlier keypoints. (c) Feature-based registration output image T_2. (d) Intensity-based registration output image T_3. (e) Composite RGB image of overlaid images R and T_2. (f) Composite RGB image of overlaid images R and T_3. (g) Artificial grid deformed with the non-rigid displacement field computed in the intensity-based registration.

($\Delta\text{FRE} = \text{FRE}_{\text{FB}} - \text{FRE}_{\text{PA}}$); all FRE values are expressed in *pixels*. The intensity-based registration results show the mutual information between the reference and the target images before the non-rigid registration ($\text{MI}_{\text{before}}$), after the non-rigid registration (MI_{after}) and the difference between them ($\Delta\text{MI} = \text{MI}_{\text{after}} - \text{MI}_{\text{before}}$); all MI values are expressed in *bits*. The raw results presented in Tab. 4.2 will be analysed in the next sections.

Feature-based registration

Due to the variability between tissue sections in the studied samples, 16 combinations of the SIFT parameters (Sec. 4.2.5) were considered in order to increase the probabilities of convergence and optimisation of the first registration step. Tab. 4.3 shows the 16 combinations (IDcomb) of the four SIFT parameters (r, n, s and *rod*) explored for each sample. The table also presents the number (#) and the

Tab. 4.2: Quantitative evaluation of all the samples in the dataset.

IDsample	Feature-based registration				Intensity-based registration		
	IDcomb	FRE_{PA}	FRE_{FB}	ΔFRE	MI_{before}	MI_{after}	ΔMI
INT1	3	5.6	6.2	0.6	1.666	1.743	0.077
INT2	1	8.2	10.4	2.2	1.249	1.410	0.161
INT3	8	6.0	6.5	0.5	1.300	1.701	0.400
INT4	9	3.6	10.8	7.2	1.413	1.377	-0.037
INT5	13	1.8	2.3	0.5	2.310	2.312	0.002
INT6	1	1.2	1.4	0.2	1.870	2.195	0.325
INT7	16	2.8	3.9	1.1	2.273	2.342	0.069
INT8	14	4.7	5.4	0.7	1.713	1.790	0.078
INT9	6	18.4	46.1	27.7	0.948	1.324	0.376
INT10	5	3.1	3.5	0.4	1.478	1.709	0.232
INT11	8	10.1	23.0	13.0	1.470	1.632	0.163
INT12	16	5.2	5.9	0.7	1.322	1.389	0.067
INT13	6	6.3	31.5	25.2	1.174	1.446	0.272
INT14	11	3.2	3.4	0.2	1.065	1.127	0.062
INT15	7	5.2	6.0	0.8	1.435	1.594	0.159
INT16	11	7.8	7.9	0.1	1.418	1.508	0.090
NOR1	8	11.8	15.5	3.7	1.748	1.832	0.085
NOR2	16	10.4	10.7	0.3	1.532	1.759	0.227
NOR3	1	6.5	14.6	8.1	1.549	1.470	-0.079
NOR4	7	4.1	9.6	5.5	1.767	1.856	0.089
NOR5	1	11.8	25.8	14.0	1.245	1.405	0.160
NOR6	13	2.0	2.3	0.3	1.711	1.748	0.037
NOR7	4	1.6	1.6	0.1	1.864	1.894	0.030
NOR8	8	4.1	7.5	3.3	1.575	1.731	0.157
NOR9	6	6.4	93.3	86.9	1.327	1.398	0.070
NOR10	4	3.2	3.3	0.1	1.249	1.289	0.040
NOR11	3	5.5	6.7	1.1	1.674	1.803	0.130
NOR12	1	7.3	9.2	1.9	1.357	1.465	0.109
NOR13	6	16.0	24.4	8.5	1.734	1.658	-0.075
NOR14	5	9.3	11.1	1.7	0.985	1.110	0.125
NOR15	3	6.5	9.0	2.5	1.073	1.243	0.170
NOR16	8	2.7	3.2	0.5	1.635	1.772	0.138
TUM1	8	4.8	5.1	0.3	1.511	1.622	0.110
TUM2	8	6.4	6.6	0.1	1.258	1.407	0.148
TUM3	8	7.7	15.3	7.6	1.073	1.334	0.261
TUM4	8	2.7	3.5	0.8	1.086	1.192	0.106
TUM5	15	5.4	7.6	2.2	1.097	1.166	0.069
TUM6	15	16.2	20.4	4.3	0.844	0.865	0.021
TUM7	15	3.4	3.8	0.3	1.133	1.293	0.159
TUM8	13	10.4	12.4	2.1	1.304	1.473	0.169
TUM9	3	5.1	5.9	0.8	1.367	1.501	0.134
TUM10	15	4.1	4.3	0.1	2.303	2.350	0.047
TUM11	7	7.5	7.8	0.3	1.295	1.542	0.247
TUM12	14	10.6	11.0	0.4	1.710	1.940	0.229
TUM13	-	-	-	-	1.134	1.153	0.019
TUM14	1	6.6	10.0	3.5	1.437	1.576	0.139
TUM15	6	15.5	17.0	1.4	1.049	1.325	0.276

percentage (%) of samples which each combination has chosen as optimal (in terms of maximal CR) in the feature-based registration step. As can be seen, although three combinations of parameters have not been selected as optimal, there is not a clearly prevailing combination. This fact justifies the exploration of different SIFT parameters in this dataset. The exact combination (IDcomb) of SIFT parameters that each sample has chosen as optimal is shown in Tab. 4.2.

Tab. 4.3: Explored combinations, identified by a number (IDcomb), of the four SIFT parameters (r, n, s and rod) with the number (#) and the percentage (%) of samples which has chosen each combination as optimal in the feature-based registration step.

IDcomb	**SIFT parameters**				#	%
	r	n	s	rod		
1	4	4	6	0.8	6	13.0
2	4	4	6	0.9	0	0.0
3	4	4	10	0.8	4	8.7
4	4	4	10	0.9	2	4.3
5	4	6	6	0.8	2	4.3
6	4	6	6	0.9	5	10.9
7	4	6	10	0.8	3	6.5
8	4	6	10	0.9	9	19.6
9	8	4	6	0.8	1	2.2
10	8	4	6	0.9	0	0.0
11	8	4	10	0.8	2	4.3
12	8	4	10	0.9	0	0.0
13	8	6	6	0.8	3	6.5
14	8	6	6	0.9	2	4.3
15	8	6	10	0.8	4	8.7
16	8	6	10	0.9	3	6.5

By using the manually selected landmarks, the fiducial registration error for the Procrustes analysis $\mathrm{FRE_{PA}}$ considered as the *gold standard* registration, the proposed feature-based registration $\mathrm{FRE_{FB}}$ and the differences between them ($\Delta\mathrm{FRE} = \mathrm{FRE_{FB}} - \mathrm{FRE_{PA}}$) were computed for each sample (see Tab. 4.2). Fig. 4.9 presents the box plots [231] which summarise the distributions of these error metrics if the samples are divided into the three pathological groups or if all samples are considered. In these box plots, whiskers extend to the most extreme data point that is no more than 1.5 times the IQR from the edge of the box. All the values outside the whiskers range are considered as outliers (red plus signs).

As can be seen in Fig. 4.9a, the $\mathrm{FRE_{PA}}$ values (blue box plots) are very similar between pathological groups, being their medians around 6 pixels. Only one outlier

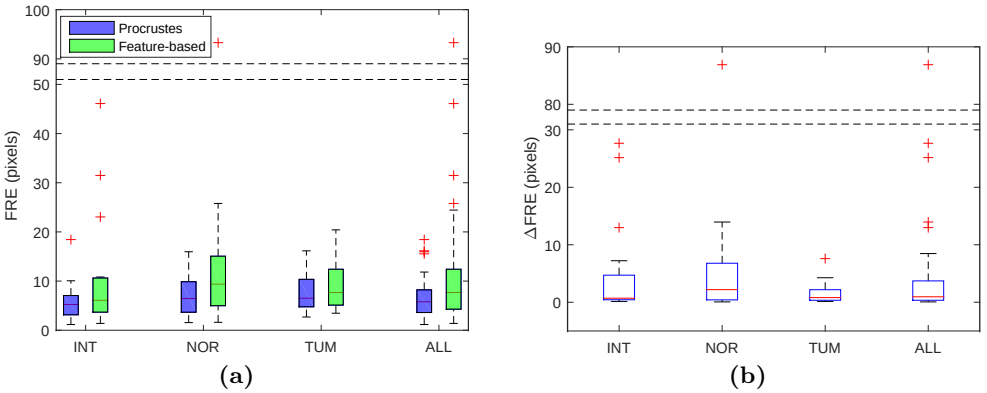


Fig. 4.9: Evaluation of the feature-based registration step. Box plots for the Intermediate (INT), Normal (NOR), Tumoral (TUM) and ALL samples groups representing (a) the fiducial registration error for the Procrustes analysis $\mathrm{FRE_{PA}}$ *gold standard* registration (blue), the proposed feature-based $\mathrm{FRE_{FB}}$ registration (green) and (b) their differences ($\Delta\mathrm{FRE} = \mathrm{FRE_{FB}} - \mathrm{FRE_{PA}}$). In both subfigures, the empty regions of the vertical axes have been cropped to improve the visualisation.

exists for the intermediate group, although three samples (with $\mathrm{FRE_{PA}}$ values above 15 pixels) are considered as outliers in the group of all samples. A higher dispersion exists in $\mathrm{FRE_{FB}}$ values (green box plots) although their distributions are also similar between pathologies and their median values remain under 10 pixels for all the groups. Three outliers are present in the intermediate group with values between 20 and 50 pixels and one extreme outlier with a $\mathrm{FRE_{FB}}$ value over 90 pixels exists in the normal group.

In Fig. 4.9b, the distributions of the paired differences of errors (ΔFRE) show a bit higher dispersion in the normal group with a median of 2.2 pixels meanwhile the rest of pathological groups have a median error difference under 1 pixel. If all the samples are considered, the median of (ΔFRE) is also under 1 pixel with a low dispersion, excluding the 5 outlier samples with (ΔFRE) values above 10 pixels. It must be remarked that there was one tumoral sample (TUM13; see Tab. 4.2) where no evaluation was performed because no corresponding points could be manually identified.

Intensity-based registration

In order to assess the improvement due to the non-rigid transformation provided by the intensity-based registration, the reference image R was compared with the initial target image T_2 and with the final target image T_3. To that end, the mutual information before ($\text{MI}_{\text{before}}$; computed between R and T_2) and after (MI_{after}; computed between R and T_3) the last registration step as well as their difference ($\Delta\text{MI} = \text{MI}_{\text{after}} - \text{MI}_{\text{before}}$) were computed for each sample (see Tab. 4.2). Fig. 4.10 shows the box plots of the distributions of these MI values divided by pathological groups and considering all the samples. The rules regarding the notches, whiskers and outliers of these box plots are the same as in Fig. 4.9.

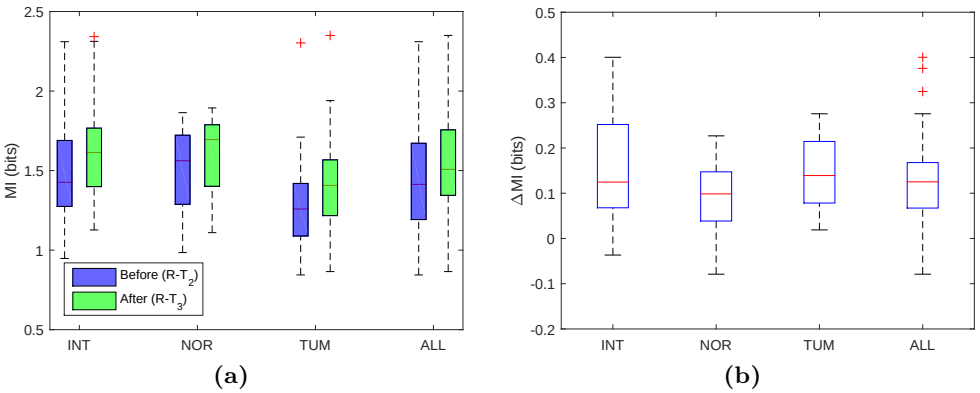


Fig. 4.10: Evaluation of the intensity-based registration step. Box plots for the Intermediate (INT), Normal (NOR), Tumoral (TUM) and ALL samples groups representing (a) the mutual information before ($\text{MI}_{\text{before}}$; computed between R and T_2) the intensity-based registration (blue), after (MI_{after}; computed between R and T_3) the intensity-based registration (green) and (b) their differences ($\Delta\text{MI} = \text{MI}_{\text{after}} - \text{MI}_{\text{before}}$).

As observed in Fig. 4.10a, the distributions of MI_{after} are above their respective distributions of $\text{MI}_{\text{before}}$ for all the groups. Although these tendencies suggest an improvement in the alignment of the images, it is not possible to state confidently such an improvement due to the partial overlap caused by the spread of the corresponding distributions. The analysis of the paired differences ΔMI removes the within-sample variability and reduces the source of uncertainty to the spread in these differences [232]. The distributions of ΔMI (Fig. 4.10b) present predominant positive values

(with the exception of one intermediate and two normal samples; see Tab. 4.2) with medians around 0.1 bits in the three pathological groups and a median of 0.125 bits in the group of all samples. The right-tailed Wilcoxon signed-rank test [227] returned p-values under 0.001 for all the groups, supporting the alternative hypothesis that the medians of all ΔMI distributions are greater than 0. This result suggests that there is a statistically significant improvement in the alignment of the images.

4.4 Discussion

The main novelty of this work is the use of a representative grayscale image extracted from the FTIR hyperspectral data cube which condenses the most relevant information of the biological structures of the tissue. The preprocessing to extract this grayscale image is crucial to get a spatial contrast similar to the grayscale image obtained from the H&E image. This grayscale image provides higher spatial information than the binary or the clustered images which were used as inputs in other previous approaches that treated to solve similar multimodal registration problems [205, 206]. This increase in local information is essential in this work in order to tackle the additional problem of aligning different sections of tissue.

In the feature-based registration step a medium level of information at different levels of resolution is explored by the SIFT algorithm to automatically detect relevant landmarks. No optimal combination of SIFT parameters clearly prevailed in the studied samples (Tab. 4.3), which justifies the use of different SIFT values and an optimisation procedure in terms of CR. This fact also confirms the wide morphological variability between samples, whose sections of tissue may not even be adjacent. Regarding the differences in pathology, normal and intermediate samples have more distinctive references inside the tissue due to the presence of glandular structures, whilst tumoral tissue contains more heterogeneous structures as the severity of cancer increases. As a consequence, the automatically detected inlier keypoints in the tumoral samples are mainly located at the tissue borders, which may be a problem if not enough border regions are captured in the image. Apart from that, similar results were obtained for all the pathological groups in the comparison of the feature-based registration step with a *gold standard* manual registration. Most of samples obtained a difference in positioning error ΔFRE under 10 pixels and the median of the

distribution of all samples was under 1 pixel. These differences may be considered as quite satisfactory taking into account that the main purpose of the first registration step is to provide a coarse initial alignment for the second registration step.

The final intensity-based registration step is essential to refine the initial coarse alignment provided by the detected keypoints and to compensate for the spatial unevenness between the different sections of tissue. This step is a complement of the first registration step because it considers the images at the lowest level of information given by their intensities. The values of the parameters used in this step were empirically tuned in the studied dataset as a good global trade-off between achieving a finer alignment and avoiding excessive local deformations. The statistical significant increase of the mutual information shared between the registered images confirmed the alignment improvement that had been already visually observed, independently of the pathological state.

4.4.1 Advantages and limitations

Taking into account the previous considerations, the main advantages of the proposed method may be divided in the following points:

- *Simplification of the multimodal problem*: the multimodal problem of finding the correspondence between the three-valued H&E pixels and the hyper-valued FTIR pixels is simplified by condensing their information in grayscale images. It allows the use of registration methods in the grayscale domain, which have been successfully and efficiently applied in other medical areas.

- *Gain of spatial information*: using grayscale images augments the spatial information, compared to other related approaches recently published. In particular, a simple binarisation is applied in [205], meanwhile previous independent segmentations by clustering within each image are proposed in [206]. It must be remarked that in those studies the same tissue section was measured by FTIR microspectroscopic imaging and then stained to get the H&E image. Therefore, relatively simple spatial transformations had to be calculated; specially restricted in [206], where only translations and rotations up to ± 30 degrees were considered due to its computational complexity and

probably to the uncertainty introduced by the pre-segmentation step. In this problem, the gain in spatial information provided by the grayscale image is crucial to compensate for the local spatial differences caused by employing distinct tissue sections for H&E and FTIR images. In addition, the suggested approach avoids pre-segmentation steps, which would increase the complexity of the problem (segmentation is maybe the most challenging task in medical image analysis) and may accumulate errors coming from the incorrect partition of corresponding anatomical structures.

- *Estimation of local deformations*: the computation of non-rigid spatial transformations by the intensity-based registration step is another improvement compared to [207], where distinct tissue sections treated with different stains are aligned only with rigid transformations. The computed non-rigid displacements are decisive to improve the alignment of more distant sections and compensate for the deformations due to the handling and cutting of the biopsies.
- *Robustness against pathological variability*: the methodology is versatile and robust against distinct sources of variability, including different morphological configurations depending on the pathological state. This robustness is increased in the first registration step by considering different combinations of parameters and maximising a similarity measure (CR).

On the other hand, the main limitations of the framework are:

- *Presence of distinctive anatomical structures*: the feature-based registration step tries to find relevant keypoints normally present inside the tissue in normal or intermediate pathological states, such as glandular structures. However, the tissue heterogeneity increases with the pathological state and the distance between tissue sections. Therefore, in those cases the presence of tissue borders in the recorded images may be crucial to find an optimal spatial transformation.
- *Global regularisation in non-rigid registration step*: the smoothness of the displacement field is equally defined in the whole image by the parameter α. This characteristic may be problematic in samples with uneven local deformations, for example, due to the presence of different pathological regions. Therefore, there may be zones where under- or over-deformations may be computed.

- *Possible unrealistic deformations*: because the intensity-based registration method only considers the low-level information of pixels' intensities. A possible solution would be to create a model of tissue deformation with higher levels of information, which may require challenging tasks such as pre-segmenting inner tissue structures (e.g. nuclei or glands) or estimating mechanical properties of the biological material.

- *Automatic parameter optimisation in non-rigid registration step*: single values for the parameter α and the related η_{max} were fixed for all the samples as a good global trade-off, but better results may be obtained by individually tuning these parameters. However, the main challenge is to pick the optimal values automatically without applying over-deformations. Opposing to the feature-based registration step, where the best combination of parameters is chosen by the maximisation of CR, there is no auxiliary reliable metric that may indicate the optimal configuration for each sample.

4.5 Application

This multimodal registration framework was employed for a preliminary assessment of the capabilities of FTIR spectra to discriminate normal and cancerous regions in colorectal tissue samples. Some images and the quantitative results presented here were kindly provided by collaborators from UoE and GHFT.

This preliminary binary classification of spectra tried to discriminate spectra that were extracted from glandular regions containing epithelial cells. As was explained in Sec. 1.1.3, those regions are responsible for the onset of the development of malignant patterns and their spectra are likely to contain more differential features from the diagnostic point of view. Nevertheless, the identification of those glandular regions in the FTIR images is not trivial and two different approaches were explored, as sketched in Fig. 4.11.

Pathologists are used to interpreting the patterns in the H&E stained images and they are able to better delineate the glandular regions in those kinds of images (green boundaries in Fig. 4.11). This first approach uses the developed registration method to transfer the regions annotated by the pathologist in the H&E images to the FTIR

space. Hence, a binary mask is obtained, which isolates the pixels (black pixels in Fig. 4.11) whose spectra must be retained in each image for the later classification. The main problem of this approach is that several errors are introduced, mainly due to the use of distinct tissue sections for both imaging modalities.

The second approach to identify the glandular regions is a direct segmentation based on unsupervised classification or *clustering* of individual FTIR spectra. An example of the results obtained by *K-means* clustering can be seen in Fig. 4.11. As observed, clustering groups pixels into different categories symbolised by different colours. Pixels from *clusters* likely associated with glandular regions are finally retained for the later classification. Although this alternative offers an objective delineation, it has several drawbacks as the dependence of the number of groups (very difficult to state for each sample) and the dependence on the *natural* tendency of FTIR spectra from glandular regions to form isolated groups in the spectral feature space.

Finally, the third approach merges the previous approaches and only the pixels retained both by registration and clustering are used in the classification.

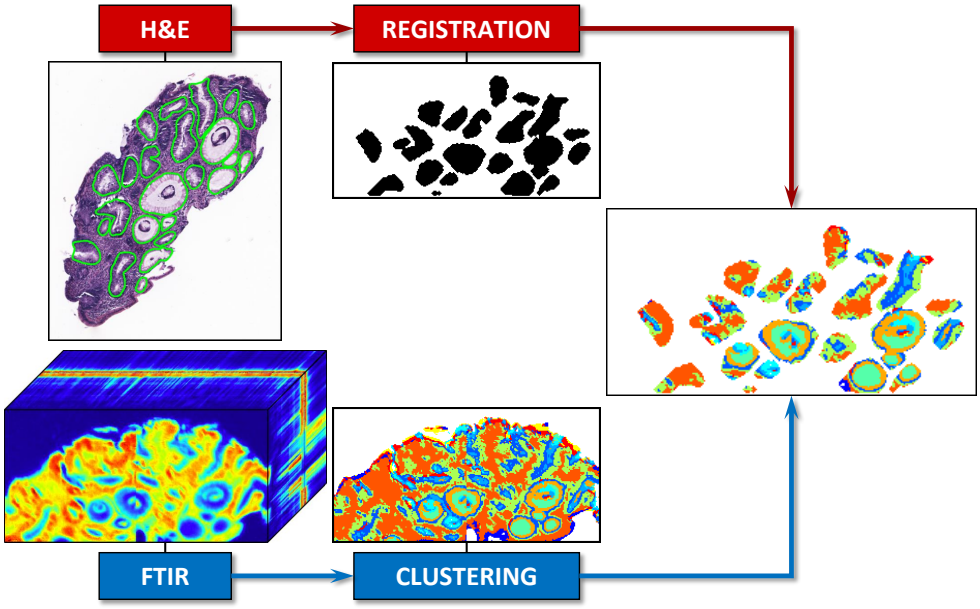


Fig. 4.11: Sketch of the fusion of spatial information obtained by registration and clustering.

The binary classification used *PCA-LDA* (Sec. 5.2.8) to discriminate the individual extracted pixels into 2 groups (Normal vs Cancer). In order to better assess the generalisation capabilities of the discrimination framework, a *Leave-one-map-out cross-validation* was performed [186]. That is, alternatively, the spectra belonging to each tissue sample were considered as the test set, and the classification algorithms were trained with the rest of spectra. The final assessment metrics were computed with the predicted labels for those test sets. *Sensitivity* (Sn) and *Specificity* (Sp) were the indices used to evaluate the results from a diagnostic point of view. They are defined as:

$$Sn(\%) = \frac{TP}{TP + FN} \cdot 100, \qquad Sp(\%) = \frac{TN}{TN + FP} \cdot 100 \tag{4.10}$$

where TP are the *True Positives* (pixels from cancerous regions correctly classified), FN are the *False Negatives* (pixels from cancerous regions incorrectly classified), TN are the *True Negatives* (pixels from normal regions correctly classified) and FP are the *False Positives* (pixels from normal regions incorrectly classified).

Tab. 4.4 presents the final quantitative results in terms of Sn and Sp together with the total number of analysed spectra and tissue samples used for each approach to extract the previous spatial information of the glandular regions. As can be observed, clustering offered better Sn and Sp values than registration, but their combination clearly outperformed the individual approaches. Further applications to larger datasets of tissue samples are currently under development. Those classification models will also try to discriminate intermediate states of pathology.

Tab. 4.4: Quantitative results of the preliminary discrimination of normal and cancerous regions in colorectal tissue samples.

Spatial Information	Sn	Sp	**Number of spectra**	**Number of samples**
Registration	79%	83%	770K	25
Clustering	84%	86%	260K	26
Registration + Clustering	**90%**	**93%**	81K	24

Chapter 5

Discrimination of skin cancer cells

Contents

5.1 Introduction

This chapter will present the second main application of this thesis. It seeks to contribute to the findings about the use of FTIR microspectroscopic imaging as a tool to improve cancer diagnosis by means of cytopathological samples. The followed methodology, which fuses techniques described in previous chapters with new methods from the areas of *machine learning* and *pattern recognition*, will be detailed together with the main reasoning behind each applied step. In addition, the final results will be quantitatively assessed and thoroughly discussed.

A limiting factor for the use of FTIR spectroscopes in biomedical problems has been the sensitivity of the systems, which is related to the relatively high acquisition times needed to obtain measurements of sufficient quality. This problem is of greater concern for single layers and/or individual cells, which contain less biological material than tissues and provide lower signals. Therefore, to obtain spectra from cells with signal-to-noise ratio comparable to tissues, the acquisition time is normally on the order of two or even four times longer for cells than for tissue. Moreover, the measurement of cells has normally been a very tedious task because they are frequently spread out in the sample preparations. These facts have hampered the proper analysis of cell spectra and prevented the systematic assessment of their discriminative power.

Recently, modern FTIR microspectroscopes have increased their acquisition speed mainly thanks to the development of larger and more sensitive imaging sensors. The present work constitutes a proof-of-concept to assess if a modern benchtop FTIR microspectroscope, together with the existing protocols of sample preparation and spectral analysis, are ready to provide a reliable diagnostic system using cytological samples. In particular, the ability of FTIR spectra to differentiate between cells from cultures of four different skin cell lines, including two melanoma cell lines with malignant phenotypes, was studied.

5.1.1 Related work

Because of the aforementioned difficulties, the application of FTIR microspectroscopic imaging to cytopathological problems is considerably less than the large number of studies and significant advances accomplished in histopathology. Nevertheless, several research groups around the world have pushed, and still push, forward with its application and development.

Until the advent of FTIR microscopes, cytological studies analysed average FTIR spectra recorded from large samples of *cell pellets*, which lacked enough spatial resolution to accurately distinguish cell subpopulations [233]. The first relevant studies of cells by combining FTIR spectroscopes and optical microscopy were performed with synchrotron sources around the beginning of this century [128,143]. In fact, these kinds of facilities (Sec. 2.5.2) were used in pioneering studies that revealed one of the main problems in FTIR cytology: Mie and Resonant Mie Scattering (RMieS) [141,146,160]. Improvements in optical components and sensor sensitivities extended the use of FTIR microspectrometers with thermal sources to laboratories. In addition, the introduction of FPA sensors enabled faster measurements of cell preparations [234].

The most relevant bibliography involving FTIR cell analyses can be found in recent reviews [114, 235]. Regarding the discrimination of cells from the diagnostic point of view, commonly referred to as Spectral Cytopathology (SCP), the most important works using FTIR microscopy are mainly related to Diem's collaborations [236–243]. All these studies are focused on smear cells directly extracted from different parts of the patients rather than in cell cultures. They have covered different types of cancer pathologies: urine [239], cervix [240] or upper respiratory and digestive tract [241–243]. In most of them, cells were deposited on low-e slides and measured in transflection. Due to the preparation of the extracted samples, cells were relatively spread and isolated. A patented method called PapMap [234], which computes a mean spectrum per isolated cell and discards clumped cells, is used in those studies.

As stated above, most existing studies related to cancer discrimination analyse cells directly extracted from the patient. In principle, this is the ideal methodology to develop decision support systems to assist the clinician in the diagnosis of cytopathological samples. Nevertheless, one of the problems of using these kinds of samples is the need for an explicit labelling by expert pathologists, which is considered the *ground truth*, or *gold standard*. This process is very time-consuming for the experts and is also subject to their interpretation. In addition, most of the extracted cell samples contain debris and undesirable heterogeneities that may diminish the quality and reliability of the cell datasets. These problems, added to the intrinsic difficulties when recording FTIR spectra of cells, have resulted in most existing studies being based on reduced datasets of spectra (e.g., less than 1000 cell spectra). As

a consequence, their results are generally founded on just qualitative assessments (e.g., by means of PCA scores plots) or quantitative classifications where no clear separations of the training and test sets (e.g., by cross-validation) are stated. Hence, their findings offer limited statistical significance and generalisation capabilities of the models are very difficult to guarantee, especially given the high-dimensional nature of FTIR spectra.

5.1.2 Objective

The main objective of this study is to assess the capabilities of FTIR spectra to discriminate different skin cell lines, comprising two non-tumoral and two tumoral types. These cell lines were cultured and fixed in a controlled environment before being measured with a modern state-of-the-art commercial FTIR microspectrometer. The selected cellular models are rather simple, stable and have significant differences in tumorigenicity. Cultured cells belong to catalogued cell lines with approximately *constant* characteristics within the same populations, which reduces the uncertainty linked to the establishment of a reliable *ground truth*. In addition, cell cultures provide higher spatial densities of cells, which increases the efficiency for recording larger amounts of data. Therefore, these cultured cells should potentially constitute a suitable standard model for the evaluation of the processing and classification of FTIR cell spectra with higher levels of statistical significance than current related studies.

Nevertheless, some technical problems and additional difficulties arise when dealing with cell cultures. Some of these particularities introduce confounding artifacts that may mislead the discrimination. The aim of this work is to apply different data analysis methodologies that diminish those critical biases and promote a discrimination based on the genuine biochemical information of the cell lines. Special efforts will be made to evaluate the generalisation capabilities and robustness of those methodologies against possible fluctuations in experimental conditions. This variability, introduced both during sample preparation and spectral acquisition, will also be considered during the discrimination analysis. Therefore, another goal of this work is to identify the possible limitations of the current measurement and sample preparation protocols.

5.2 Materials and methods

5.2.1 Discrimination pipeline

Fig. 5.1 shows the flow diagram with the main steps that were followed to discriminate skin cell lines, including melanoma and non-melanoma cells, based on their FTIR spectra. This process starts with the preparation of cultured samples of catalogued skin cell lines and their measurement with a modern FTIR microspectrometer in order to obtain the hyperspectral images. From these images, the pixels associated with cells were separated from the non-cell pixels so that only useful spectra were retained for subsequent steps.

FTIR spectra extracted from the retained pixels were individually preprocessed by different techniques to normalise their values and remove unwanted variations which may mislead the posterior quantitative analysis. A mean spectrum was computed for each cell in order to reduce the complexity of the dataset and try to mitigate remaining undesirable effects in the preprocessed spectra of individual pixels, such as random noise. The anomalous or extreme values in the dataset of mean cell spectra were filtered out to promote stability and reduce bias in the dimensionality reduction and classification algorithms.

An exploratory analysis was performed in the resulting dataset in order to study the main trends within the data. Finally, the kept mean cell spectra went through a process of feature extraction and supervised classification where different alternatives were explored, too. This final process of classification was subject to a methodology called *nested Cross-Validation (CV)* consisting of two *loops*, which separates the training and testing subsets of spectra in order to avoid over-fitting and give a reliable measurement of the performance of the constructed classification algorithms. All of these steps will be described in detail in the following sections.

5.2.2 Hyperspectral images

Cell culture and sample preparation

Tab. 5.1 summarises the cell lines used in this study (A-375, HaCaT, NIH-3T3, SK-MEL-28). With the selection of these cell lines the two major cellular skin

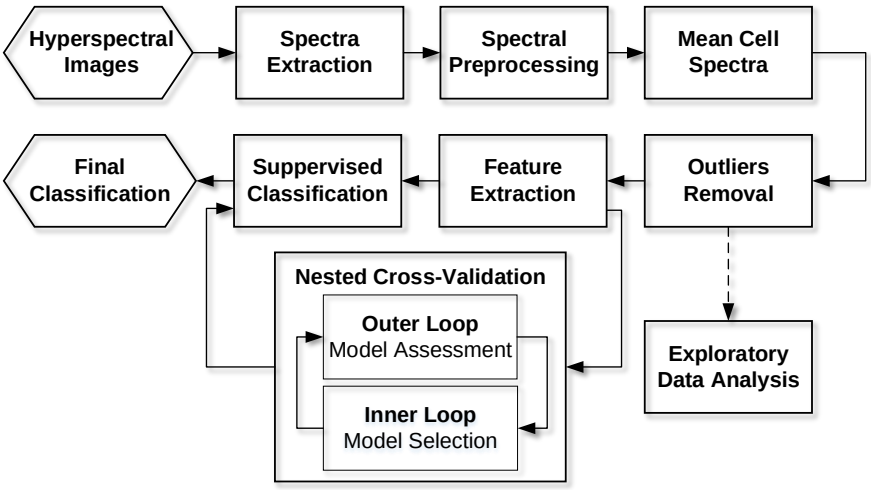


Fig. 5.1: Flow diagram of the main steps applied for the discrimination of skin cells.

constituents, keratinocytes and fibroblasts, together with two skin cancer cell types are represented (Sec. 1.1.6). These cell lines were obtained from CLS Cell Lines Service GmbH (CLS) [244] and Leibniz Institute-German Collection of Microorganisms and Cell Cultures (DSMZ) [245]. All cell lines were individually cultured in Dulbecco's Modified Eagle Medium (DMEM - high glucose, Sigma-Aldrich) supplemented with 10% Fetal Bovine Serum (FBS Good, PAN-Biotech) and 2 mM L-Glutamine (Lonza) at 37 °C and 5% CO_2. Cells were grown to near confluence within two to three days (A-375, NIH-3T3) or five to seven days (HaCaT, SK-MEL-28) and then transferred to new culture plates. All cell lines were regularly tested for mycoplasma infection by Polymerase Chain Reaction (PCR).

Tab. 5.1: Information about the cell lines used in this study.

Name	A-375	HaCaT	NIH-3T3	SK-MEL-28
Species	Homo sapiens	Homo sapiens	Mus musculus	Homo sapiens
Origin	Skin	Skin	Embryo	Skin
Type	Melanoma	Keratinocytes	Fibroblasts	Melanoma
Provider	CLS	CLS	DSMZ	CLS
Reference	300110	300493	ACC 59	300337
Cells/well	$1.5 \cdot 10^5$	$2 \cdot 10^5$	$1 \cdot 10^5$	$1 \cdot 10^5$
Incubation time	one day	two days	one day	one day

CaF_2 windows (grade VUV, 12.5 × 12.5 mm in size and 1.5 mm thick) were obtained from CRYSTAL GmbH [246]. Both sides of the windows were optically polished. The windows were decontaminated and stored in pure alcohol before use. Before seeding cells, CaF_2 windows were placed in a 12-well plate, washed twice with Phosphate Buffered Saline (PBS) and kept in cell media. Cells were counted and the cell suspensions were dropped onto the windows and incubated at 37 °C and 5% CO_2. Densities (in cells/well) and incubation times are shown in Tab. 5.1.

Once the cells had built a confluent layer on the CaF_2 window, they were washed twice with PBS, fixed with 1% glutaraldehyde/PBS for 30 minutes and washed again. The fixed samples were dehydrated by carrying out an ascending ethanol series. Finally, the samples were air-dried and stored in order to perform posterior measurements. More detailed information about the cell culture and the sample preparation can be found in [247–249].

FTIR measurements

The measurements were performed with the micro-FTIR spectroscopic imaging system described in Sec. 2.5. Six frames per sample were taken and combined, covering a horizontal and vertical area of around 2.1 × 1.4 mm^2 that included the whole cell culture. The measurements were carried out in transmission mode and absorption spectra were acquired between 1000-3900 cm^{-1} with a wavenumber interval of 4 cm^{-1}. The images of the reference backgrounds (taken from regions of empty substrate) and the images of the cell samples were created by co-adding 256 and 128 scans, respectively.

Batches definition

Cells from each cell line were independently cultured and later seeded and fixed separately in specific regions of a CaF_2 window forming a *cell sample*. The separation of these regions were delimited by silicone inserts or *moulds* with four rectangular holes that were attached to the CaF_2 window before cell seeding and were removed after fixation. Fig. 5.2 shows a picture of one of these inserts attached to a square CaF_2 window.

Fig. 5.3 shows the representative grayscale images obtained from the FTIR

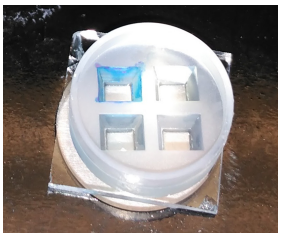


Fig. 5.2: Photography of the inserts or *moulds* used to separate different cell cultures in the same CaF_2 window.

hyperspectral images of all the cell samples that were prepared and measured for this study. As can be observed, three samples per cell line were measured giving a total of twelve cell samples. These samples have been arranged in different *batches* according to preparation and measurement criteria:

- The samples within *Batch 1* were cultured at the same time and later seeded and fixed in the same CaF_2 window. They were also measured during the same day in the FTIR microspectrometer.

- The samples within *Batch 2* and *Batch 3*, with the exception of the samples of HaCaT cell line, were cultured at the same time (around one year and a half later than cells from *Batch 1*) and were also seeded and fixed in two separate CaF_2 windows. HaCaT samples of *Batch 2* and *Batch 3* were later cultured (3 months later) and seeded and fixed in a separate CaF_2 window due to problems during the preparation of that cell line in the original CaF_2 windows of those batches. Finally, all the samples from *Batch 2* and *Batch 3* (including HaCaT cells) were measured during the same day in the FTIR microspectrometer, around nine months after acquiring the images of *Batch 1*.

Both preparation and measurement factors introduce variations in the FTIR spectra. Those variations will determine the strategies of the quantitative analysis and will also have influence in the obtained results.

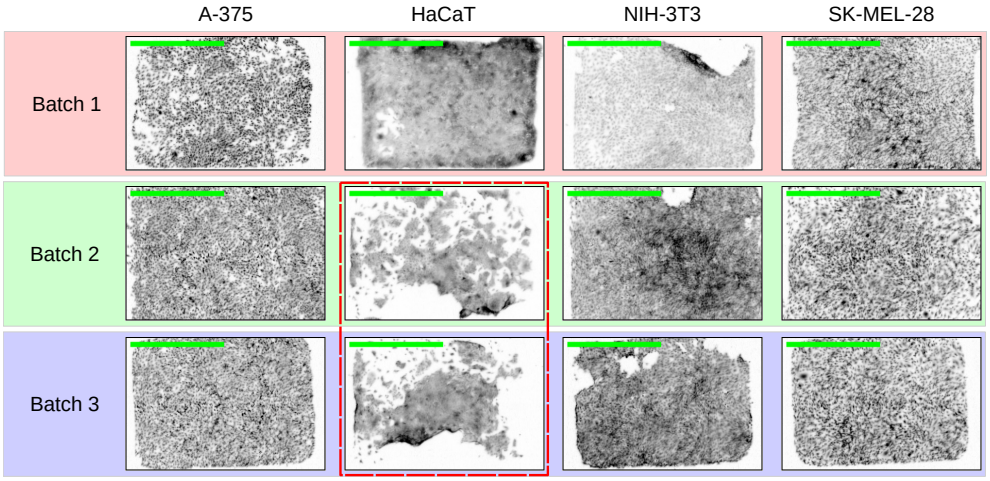


Fig. 5.3: FTIR grayscale images of the measured samples of skin cultured cell lines. Images from different cell lines are arranged by columns and from different batches by rows. The red dotted line reminds that images of HaCaT cells from *Batch 2* and *Batch 3* were cultured in a separate CaF_2 window. Green scale bars represent 1 mm.

5.2.3 Spectra extraction

The first step in the analysis of the FTIR hyperspectral images is to separate pixels containing cell structures from those containing only substrate. Those different regions can be distinguished when a representative grayscale image is computed from the FTIR data cube in a similar way that was described in Sec. 4.2.3. In this case, the spectral window was cropped to the region 1610-1690 cm^{-1}, which contains the strongest peak of the spectrum (Amide I) and, therefore, the influence of noise in the final image is reduced. The standard deviation (Eq. 4.2) of the cropped spectral region was computed for each pixel and the range of values obtained in the whole image was linearly transferred to the intensity grayscale range of 0-255, allowing 0.5% of saturated pixels in both intensity endpoints to consider the impulsive noise and dead pixels. Finally, the intensity range was inverted in order to mimic the intensity pattern of the white light images (light in the substrate and dark in cells). This methodology was used to create the images presented in Figs. 5.3 and 5.4a.

In these grayscale images, substrate pixels have the highest (lightest) grayscale levels, nuclear regions have the lowest (darkest) ones and pixels mainly associated

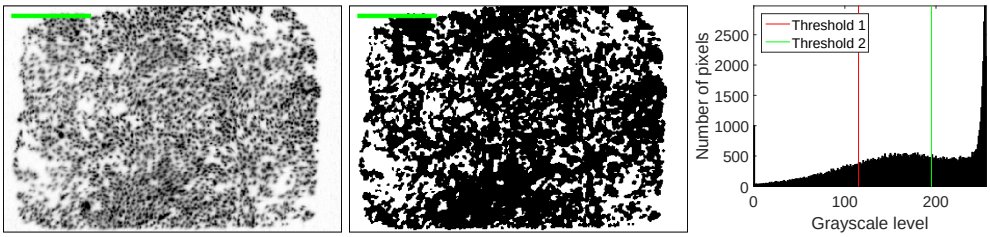


Fig. 5.4: Illustration of automatic binarisation by Otsu's method. (a) FTIR grayscale image. (b) Binary mask. (c) Histogram of the grayscale levels of image (a) and the automatic thresholds computed by Otsu's method. The threshold 2 (green vertical line) was used to create the binary mask (b). Green scale bars of images (a) and (b) represent 500 µm.

with cytoplasm have intermediate grayscale levels. This behaviour is reflected in the histogram of the grayscale images (Fig. 5.4c). Attending to the characteristics of the histogram, the Otsu's method [250] can be used to automatically compute two thresholds that separate the three classes of pixels (nucleus, cytoplasm and substrate). The highest threshold (threshold 2 in Fig. 5.4c) can be used to compute a binary mask that separates the regions mainly containing substrate from cell structures (Fig. 5.4b). These binary masks were computed individually for each cultured sample and the spectra of cell structures (black pixels in Fig. 5.4b) were retained for next steps of the discrimination pipeline.

5.2.4 Spectral preprocessing

As was explained in Sec. 3.2, spectral preprocessing is recognised as a key step in quantitative analysis of FTIR spectra, especially for cytological studies. In order to better understand the need of spectral preprocessing, Fig. 5.5 shows several examples of raw spectra corresponding to pixels from different regions of the four studied cell lines. Those pixels, whose effective size is 5.5×5.5 µm^2, have been overlaid on independent white light images of higher resolution aligned with the FTIR hyperspectral images.

As can be observed, even spectra from nearby pixels in the same image can be highly variable. Most of the *gross* variations are mainly determined by differences in concentrations and optical path lengths (thickness), which commonly result in offsets

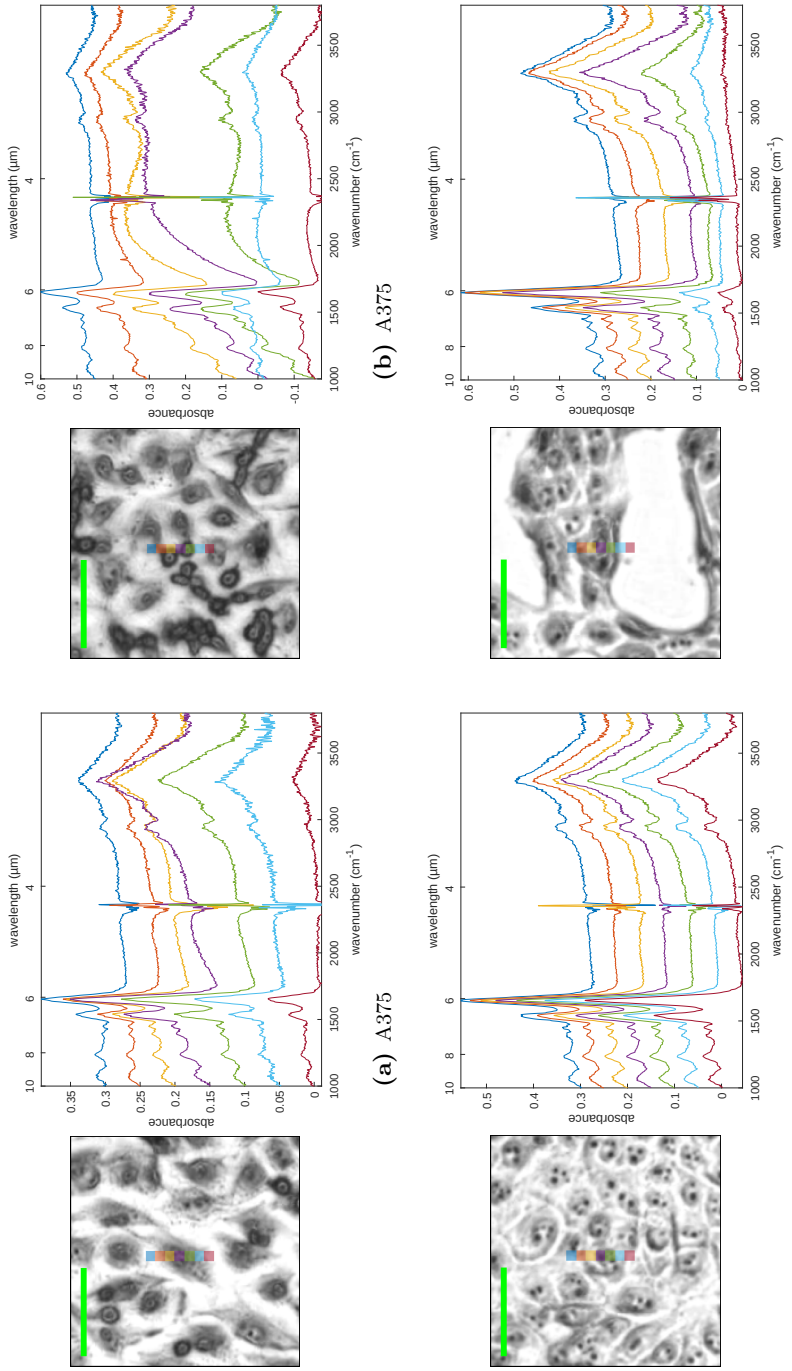
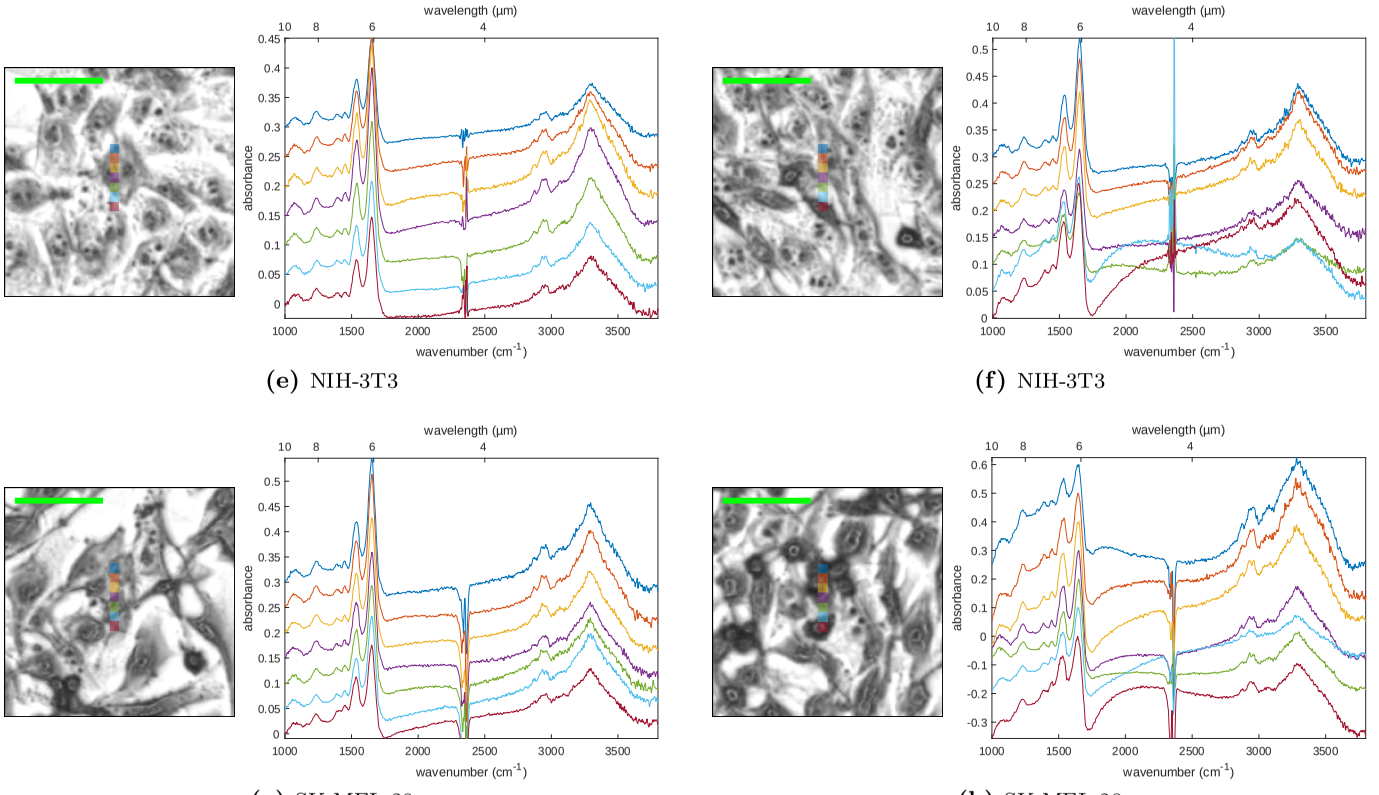


Fig. 5.5: Examples of raw spectra extracted from FTIR hyperspectral images of skin cells. Each subfigure shows at left a white light image with a vertical line of selected pixels overlaid in different colours and at right the absorbance spectra corresponding to each pixel with the same colour. A proper offset has been added to each spectrum for clarity. (a) and (b) A375 cell line. (c) and (d) HaCaT cell line. Green scale bars of white light images represent 50 µm.



(e) NIH-3T3 **(f)** NIH-3T3

(g) SK-MEL-28 **(h)** SK-MEL-28

Fig. 5.5: (cont.) Examples of raw spectra extracted from FTIR hyperspectral images of skin cells. Each subfigure shows at left a white light image with a vertical line of selected pixels overlaid in different colours and at right the absorbance spectra corresponding to each pixel with the same colour. A proper offset has been added to each spectrum for clarity. (e) and (f) NIH-3T3 cell line. (g) and (h) SK-MEL-28 cell line. Green scale bars of white light images represent 50 μm.

and multiplicative factors. Nevertheless, more complex distortions can be observed in some cases, showing the pattern distinctive of Mie and Resonant Mie Scattering (RMieS) (e.g., complex baselines, lower ratio Amide I-Amide II peaks, derivative-like depressions beyond the Amide I peak), which were thoroughly described in Secs. 2.6 and 3.2. By looking at the white light images, those complex spectra can be associated with regions where the cells are more compact and rounded. Hence, they have characteristics that produce Mie scattering phenomena.

This morphological heterogeneity, which gives rise to critical variations in the FTIR spectra, can be related to different stages of the cell cycle, as was demonstrated in previous studies [143, 251–253]. Those studies showed that cells in stages G_2 and close to mitotic phase M (Sec. 1.1.2) have more compact structures that increase the presence of RMieS artifacts. They also demonstrated that FTIR spectra from cells of the same cell type could be discriminated attending to their stage in the cell cycle. In addition, some other cells may have entered into apoptosis, due to the high levels of cell stress that can appear locally in the cell culture, acquiring compact and rounded configurations too.

The main aim of spectral preprocessing is to normalise and standardise the spectra so that they can be categorised based on biochemical properties, rather than unwanted physical properties (thickness, concentrations, morphology, etc) that increase the spectral variability and presumably can confound the discrimination. The raw spectra together with four different preprocessing methods that were described in Sec. 3.2 have been considered in order to study their possible influence in the discrimination of cell lines:

- *Raw*: the spectra were not further preprocessed.
- *Min-Max*: Min-Max normalisation.
- *SNV*: Standard Normal Variate [180].
- *DiffSG1*: 1st order differentiation by Savitzky-Golay (SG) filter (2nd order polynomial and 19 fitting points) and vector normalisation.
- *RMieS-EMSC*: Resonant Mie Scattering-Extended Multiplicative Signal Correction algorithm after 20 iterations. As was shown in Sec. 3.2.1, 20 iterations

is a conservative number that generally gives a stable correction. Nevertheless, it took 4 weeks to individually preprocess all the initial extracted spectra (∼480000) in the computer platform that was described in Sec. 4.2.8, which simultaneously ran the current version of the RMieS-EMSC algorithm in 4 parallel MATLAB sessions.

Traditionally, the first three preprocessing methods (Min-Max, SNV and DiffSG1) have been mainly applied in FTIR spectroscopic measurements of tissues, although they have been also employed for cell spectra [177]. Nevertheless, since the characterisation of RMieS artifacts and the publication of RMieS-EMSC algorithm, it has generally become the default preprocessing method in cytological studies.

In all cases, the spectral range was cropped to the *fingerprint region* (1000-1800 cm^{-1}). This restriction was applied to the raw spectra and before preprocessing except for the case of RMieS-EMSC, whose spectra were cropped after preprocessing so that the RMieS baselines could be modelled and corrected more accurately with the information of higher wavenumbers (Sec. 3.2.1).

5.2.5 Mean cell spectra

The vast majority of cytological studies based on FTIR hyperspectral images compute a mean or average spectrum per cell in order to reduce complexity and increase the robustness of later analyses. These mean spectra are computed after (instead of before) preprocessing the raw spectra of individual pixels independently in order to compensate for uneven spectral distortions due to the heterogeneous spatial properties of cells [252].

This approach is in line with *object-based* or *object-oriented* classification approaches, which are very popular in the analysis of hyperspectral images in remote sensing [254]. Object-based approaches incorporate a certain level of spatial information that generally improves the performance of *pixel-wise* or *pixel-by-pixel* classification.

In order to compute mean cell spectra, it is essential to delimit the regions of pixels belonging to each individual cell. The binarisation method described in Sec. 5.2.3 follows a similar philosophy to the so-called PapMap method [234], which was

designed to identify isolated cells from smears preparations and discard clumped cells. Nevertheless, unlike smears of cells, which are normally scattered and isolated, cultured cells tend to grow into clusters due to spatial constrictions. Therefore, segmentation methods [255,256] more advanced than a simple binarisation are needed to separate the regions which belong to single cells. The segmentation of cells was performed with a methodology similar to the one proposed by Filik *et al.* [257], which applies the marker-controlled watershed transformation.

Watershed transformation

The watershed transformation is one of the most used techniques of mathematical morphology [255, 258, 259] and also one of the most used methods to segment cells and nuclei in different microscopy modalities [260, 261]. It was firstly defined in [262] and the most efficient algorithmic implementation based on hierarchical queues was found in [263].

The watershed transformation considers a grayscale image as a topographical surface whose height is determined by the grayscale intensity level of each pixel and which is composed of different local maxima and minima. Fig. 5.6a shows the main elements of this analogy. The local minima of the image are surrounded by lighter pixels that form *catchment basins*. The watershed transformation supposes that a *hole* exists at each local minimum and that the surface is *flooded* from these *holes* Fig. 5.6b. As the water level increases, the catchment basins get filled and the water tries to be transferred between them. In order to prevent that merging, a kind of *dams* are built at each contact point. When the water level reaches the global maximum height, the union of all constructed dams constitute the *watershed lines*, which indicate the limits of the catchment basins.

In order to illustrate this segmentation method, Fig. 5.7 shows an example where the input grayscale image (Fig. 5.7a) can be considered as a topographical surface (Fig. 5.7b). The watershed transformation segments the input image in different connected objects limited by the watershed lines (Fig. 5.7c). The number of final segmented objects (four in the considered example) is defined by the number of catchment basins existing in the input image, which in turn is determined by the number of connected local minima.

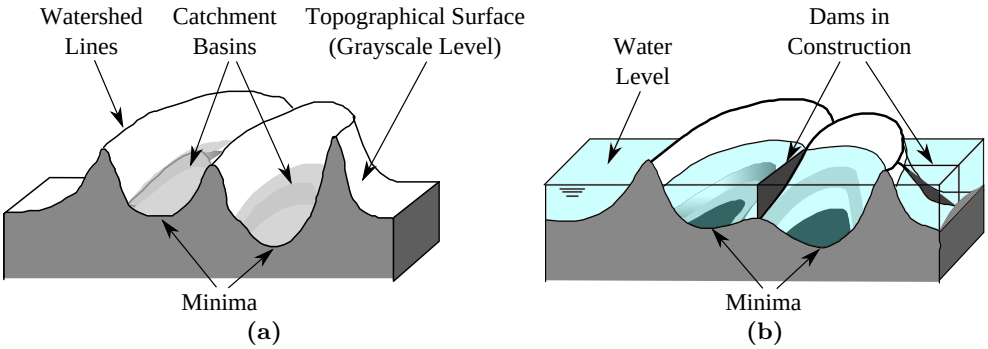


Fig. 5.6: Analogy of the watershed transformation. (a) Topographical patterns of a grayscale image. (b) Construction of the watersheds by flooding. Adapted from [264].

(a) **(b)** **(c)**

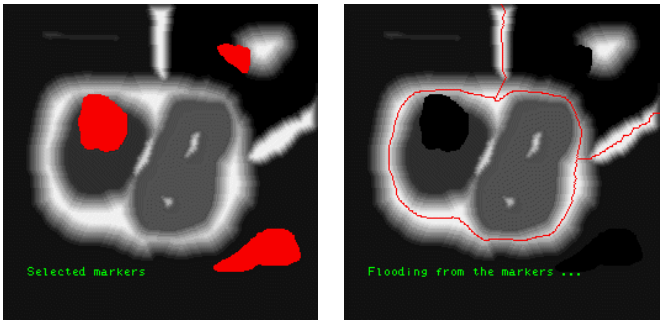
Fig. 5.7: Example of segmentation with the watershed transformation. (a) Input grayscale image. (b) Representation of the grayscale image as a topographical surface. (c) Final segmented image with the watershed lines in red, delimiting four independent regions. Reproduced from [265] with permission by courtesy of Serge Beucher (Center of Mathematical Morphology).

Marker-controlled watershed

One of the major problems of the watershed transformation is over-segmentation. As described before, each local minimum of the input image will give rise to a segmented region. Apart from being sensitive to local irregularities and minima due to noise, the final segmentation may produce excessive undesired divisions. The marker-controlled watershed, firstly defined in [258], is the major enhancement of

the watershed transformation and tries to prevent the over-segmentation problem by introducing the notion of *markers* of the objects to be segmented. Each marker is a set of connected pixels that defines a specific region in the image. The modification consists of imposing the marker regions as global minima in the input image. During the *flooding* of this modified image, the water springs up only from these new global minima and the constructed dams enclose the modified catchment basins. Thus, a segmented region arises from each selected marker.

As an example, Fig. 5.8a shows the same image that was segmented in Fig. 5.7 with three regions of connected pixels selected as markers in red. When comparing the final segmented image (Fig. 5.8b) with the result of the regular watershed transformation (Fig. 5.7c), it can be observed that the vertical watershed line that divided the two central objects has disappeared. It can also be checked that the number of final segmented objects in the marker-controlled watershed is determined by the number of defined markers (three in this example).



(a) **(b)**

Fig. 5.8: Example of the marker-controlled watershed. (a) Input grayscale image with three regions of connected pixels selected as markers in red. (b) Final segmented image with the watershed lines in red. Reproduced from [265] with permission by courtesy of Serge Beucher (Center of Mathematical Morphology).

In many situations, the aim of the segmentation task is to isolate specific objects in the images. In those cases, the concept of *foreground* and *background* markers arises, which are regions defined inside and outside the objects of interest, respectively. The methodology to find or define these two types of markers may be different, but all of them finally act as global minima in the modified input image to segment.

Often, the objects of interest have relatively uniform intensities and their borders, identified as regions of high local contrast, become local maxima in the gradient image. In those cases, the gradient image is used as the input image of the watershed transformation and the final watershed lines tend to match the borders of the uniform objects. However, the use of the gradient image is not mandatory and the selection of the input image will depend on the specific problem. The same reasoning applies to the markers, whose definition is also problem-dependent.

Cell segmentation

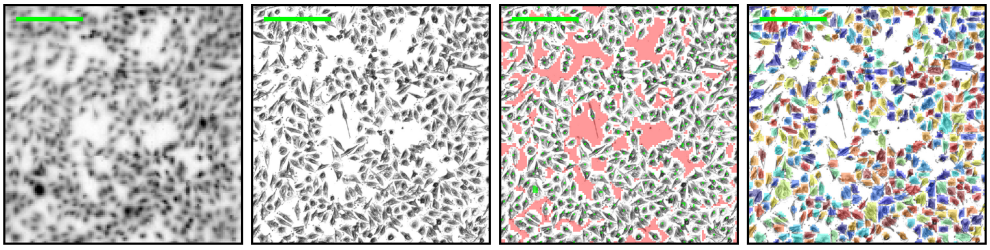
As described above, three elements must be defined to segment cells with the marker-controlled watershed, which in this case are:

- *Input image*: the FTIR grayscale image that was described in Sec. 5.2.3.
- *Outer markers*: the binary mask of substrate, also described in Sec. 5.2.3, after being processed by morphological erosion with a disk of radius 1 as structuring element [255, 259].
- *Inner markers*: the regional minima of the FTIR grayscale image outside the binary mask of substrate, which are mainly associated with the nuclear regions that have higher absorbance than their neighbourhood.

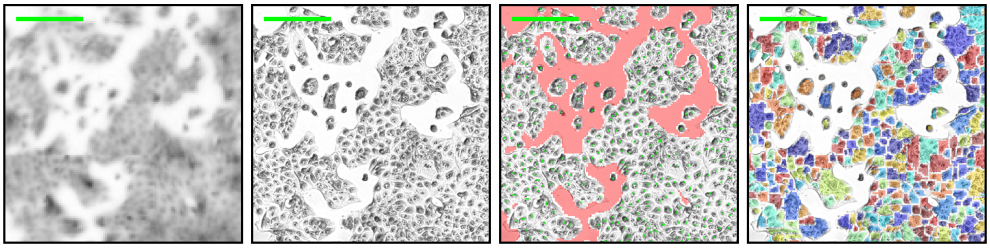
A cropped image per cell line has been selected to illustrate the elements and results of the cell segmentation in Fig. 5.9. The described segmentation method gave good overall results for the four studied cell lines, which have different growth behaviour. HaCaT cells presented the major difficulties because they tend to form a layer with very close cells, which are not well resolved in some regions at the measured spatial resolution (effective pixel size of 5.5×5.5 μm^2). Finally, regions with less than 5 and more than 100 pixels were rejected in order to correct over- and under-segmentation respectively.

5.2.6 Outliers removal

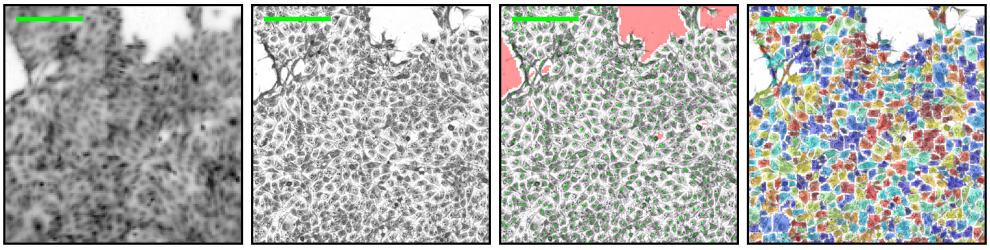
Outliers can seriously bias and deteriorate the performance of the learning framework. The correct way of proceeding with outliers is attempting to determine and address their causes in order to detect possible failures in the followed methodologies [174].



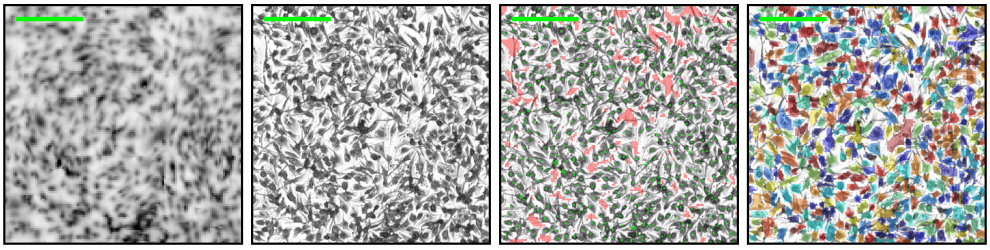
(a) A-375



(b) HaCaT



(c) NIH-3T3



(d) SK-MEL-28

Fig. 5.9: Illustration of cell segmentation. An example for each cell line is shown by rows: (a) A-375, (b) HaCaT, (c) NIH-3T3 and (d) SK-MEL-28. Column 1: FTIR representative grayscale image. Column 2: White light image aligned with the FTIR image. Column 3: Markers of the marker-controlled watershed segmentation overlaid on the white light image: inner markers coloured in green and outer markers coloured in red. Column 4: Final segmentation mask overlaid on the white light images: isolated cellular regions are defined by different pseudo-colours. Green scale bars represent 200 µm.

In this problem, there are many sources that may cause anomalous samples, starting from the cell culture preparation, the acquisition of spectra and finally in the previous steps of the classification pipeline (e.g., spectral artifacts introduced or not corrected during preprocessing). Due to the complexity of the problem, the limited technological maturity of FTIR microspectroscopic imaging and the lack of standardised and universally-recognised analysing protocols, it is reasonable to remove all possible perturbations that may be caused by outliers and leave the (difficult) task of analysing outliers for future studies.

The whole dataset of mean spectra per segmented cell region was studied to detect outliers. For this task, the preprocessing option RMieS-EMSC was taken as a reference to define those cell regions whose mean spectra are anomalous. This process was performed by using PCA decomposition (retaining the first PCs that accounted for 99% of total variance) and Mahalanobis distance in the same way that was described in Sec. 3.3.1. The cell regions with mean spectra preprocessed by RMieS-EMSC and detected as outliers were discarded in all the preprocessing alternatives described in Sec. 5.2.4 in order to keep the same set of cells for the final analysis.

5.2.7 Feature extraction

Feature extraction techniques aim to reduce the number of variables included in the classification model. It is essential to reduce the complexity and increase the stability of discrimination models. The problem of high dimensionality in hyperspectral images is well-known in the field of remote sensing [266,267], commonly receiving the name of *Hughes phenomenon* [268]. In summary, this phenomenon states that as the number of dimensions increases, the effectiveness of the classifier decreases. The main reason is that the number of parameters involved in the classification model increases with the number of dimensions and, for a fixed sample size, the uncertainty in the estimation of those parameters become wider [266]. PCA and PLS (Sec. 3.3) were the methods used to reduce the dimensionality of FTIR spectra and try to prevent this problem.

5.2.8 Supervised classification

Supervised classification is the task of *pattern recognition* [184, 185] or *machine learning* [186] whose goal is to create models capable of predicting or inferring the

labels, *classes* or *groups* of the samples from a dataset, taking as inputs a set of measured *features*. These models must be constructed or *trained* with a set of samples whose labels or classes are known. In this problem, the *features* are the *scores* of the components retained in the feature extraction step. These scores are used to predict the *label*, i.e., the *cell line*, of each sample or cell. As there are four different cell lines, this study can be categorised as a multiclass classification problem. Two main aspects must be defined in the supervised classification step, namely the specific classification algorithm and the measurement used to assess its performance.

Classification algorithms

Three classification algorithms or techniques, very popular in the chemometrics field [182, 183], have been explored: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Partial Least Squares Discriminant Analysis (PLS-DA). The main difference is in the definition of the class boundary; hence, these three alternatives have been considered in order to diminish their possible influence in the final discrimination.

Linear discriminant analysis

LDA assumes an underlying multivariate Gaussian distribution for each class or group, which is mainly defined by a mean vector and a covariance matrix. Those two parameters, estimated with the training samples, are used to create discriminant rules based on posterior probabilities [186]. Hence, a new object or sample will be assigned to that class with the largest posterior probability based on the values of its features [182]. The feature space is divided into different regions separated by *decision boundaries*, which correspond to those feature values where the posterior probabilities of distinct classes are equal.

In the special case of LDA, all the classes are considered to have the *same covariance matrix*. As a consequence of this assumption or *simplification*, the boundaries between classes created in the feature space are *linear*. That is, the specific regions of the feature space assigned to each cell line are separated by hyperplanes.

PCA will be used as the method for dimensionality reduction in the feature extraction step and the scores of the retained PCs will be the features used by LDA. This combination will be denoted by *PCA-LDA*.

Quadratic discriminant analysis

QDA follows the same methodology of LDA, but allows each class to have its *own covariance matrix*. Therefore, QDA can potentially represent classes with different variance or spread better than LDA. The main consequence is that decision boundaries are no longer linear but *quadratic*. This increment in complexity increases the risk of over-fitting but in some problems, as in biological applications, this additional level of sophistication may provide better results [183]. This classifier will also be used in conjunction with PCA and their combination will be referred to as *PCA-QDA*.

Partial least squares discriminant analysis

PLS-DA [197, 269] is a technique extensively used in chemometrics, which directly employs the latent variables computed by PLS (Sec. 3.3.2) to discriminate different classes. Following the reasoning and notation that was described in Sec. 3.3.2, once the PLS model is built, it can be used to predict the values of the response variable $\mathbf{y}$ of new samples, which in principle is a continuous variable. In particular, the independent variables $\mathbf{X}$ and the response variables $\mathbf{y}$ can be related by the following expression [183]:

$$\mathbf{y} = \mathbf{X} \cdot \mathbf{b} + \mathbf{f} = \mathbf{T} \cdot \mathbf{q}^{\mathsf{T}} + \mathbf{f} \tag{5.1}$$

where $\mathbf{b}$ is a regression coefficient vector of dimensions $J \times 1$. This vector can be used to predict the response vector $\hat{\mathbf{y}}$ of new samples from their measurements matrix or independent variables $\hat{\mathbf{X}}$ by:

$$\hat{\mathbf{y}} = \hat{\mathbf{X}} \cdot \mathbf{b} \tag{5.2}$$

The estimation of $\mathbf{b}$ can be obtained through the pseudo-inverse of $\mathbf{X}$ (denoted $\mathbf{X}^{+}$) as follows:

$$\mathbf{b} = \mathbf{X}^{+} \cdot \mathbf{y} = \mathbf{W} \cdot \mathbf{q}^{\mathsf{T}} \tag{5.3}$$

where $\mathbf{W}$ is the matrix whose columns contains the weight vectors of the PLS components (Sec. 3.3.2).

In PLS-DA, the response or dependent variable $\mathbf{y}$ of the training samples only contains discrete values that codify different classes or labels. In particular, the implementation based on PLS1 only allows to discriminate two classes, which are normally coded as +1 and -1 in $\mathbf{y}$. Therefore, PLS-DA based on PLS1 is a pure two-

class or binary classifier and its extension to multiclass problems is less straightforward than in LDA and QDA. Here, the *one-versus-all* strategy was adopted, which consists of creating as many binary classifiers as the existing classes (4 in this case). During the training of each classifier, alternatively, the samples belonging to one of the classes are considered as the *positive* class (coded by +1) and the samples of the rest of classes are grouped together to form the *negative* class (coded by -1). When a new test sample arrives, a specific continuous value of the response variable $\hat{y}$ will be computed for each classifier by using Eq. 5.2. The predicted label is finally assigned to that class which was considered as *positive* in the classifier that provided the highest predicted value $\hat{y}$. The continuous nature of the $\hat{y}$ variable allows to perform this judgement of superiority between binary classifiers.

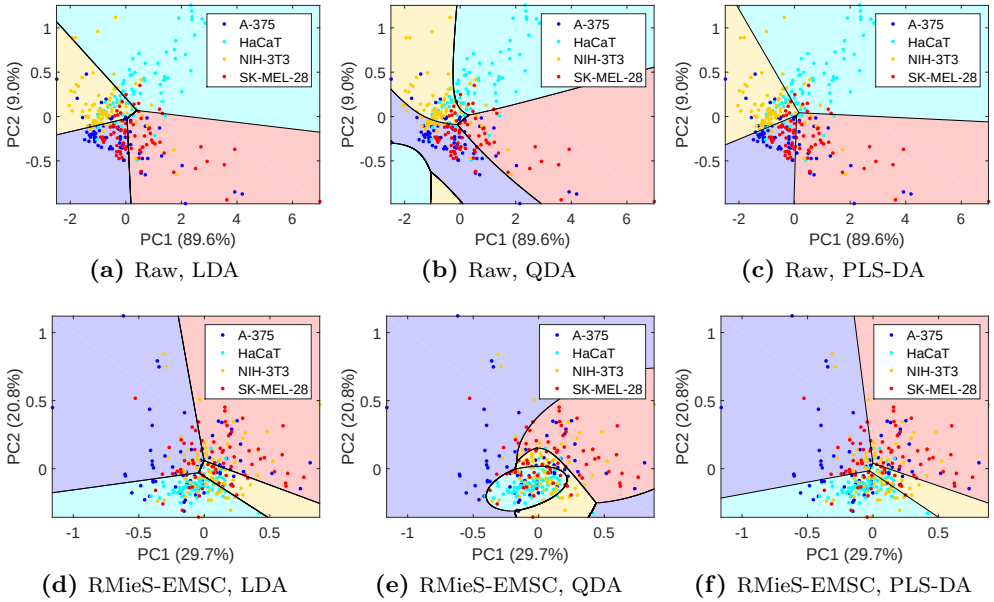
Classification boundaries

The described classification algorithms produce different decision or classification boundaries by using the features of the training dataset. In order to illustrate these differences, Fig. 5.10 shows the areas assigned to each cell line and their limits created with the dataset of 300 skin cell spectra that was presented in Fig. 3.6 and Fig. 3.15, both for raw spectra and spectra corrected by RMieS-EMSC. In these examples, the inputs or independent variables have been simplified to only two features (the two first PCs) so that the boundaries can be easily visualised.

As can be checked in the figures, both LDA and PLS-DA create linear boundaries (straight lines in the input space) meanwhile QDA generates more complex limits with quadratic nature (e.g., elliptic, parabolic or hyperbolic curves). The boundaries created by LDA and PLS-DA are very similar in these simplified cases, but their differences increase when a higher number of PC is retained. In those input spaces of higher dimensions (e.g., more than 3), the linear boundaries created by LDA and PLS-DA become hyperplanes and the quadratic boundaries of QDA generalise to quadric hypersurfaces.

Assessment metric

Another important component in classification is the metric for its evaluation. As the problem in hand is a multiclass problem, a suitable assessment metric is the *overall*



(a) Raw, LDA **(b)** Raw, QDA **(c)** Raw, PLS-DA

(d) RMieS-EMSC, LDA **(e)** RMieS-EMSC, QDA **(f)** RMieS-EMSC, PLS-DA

Fig. 5.10: Example of boundaries computed by different classification algorithms when retaining the two first PCs in the dataset of 300 skin cell spectra that was presented in Fig. 3.6 and Fig. 3.15, both for raw spectra (upper row) and spectra corrected by RMieS-EMSC (bottom row). (a),(d) LDA. (b),(e) QDA. (c),(f) PLS-DA.

accuracy [266,270], which condenses the global classification performance of all classes. As will be presented later (Tab. 5.2), a small imbalance exists between the four classes of the dataset, reaching almost a 3:1 ratio in some cases (e.g., NIH-3T3/HaCaT in *Batch 2*, see Tab. 5.2). To avoid favouring the larger classes [271, 272], the *Balanced Accuracy* (BA) was used instead of the overall accuracy to select the optimal models and to assess the final multiclass classifications. BA is the mean of the accuracies for each class and is defined as:

$$\mathrm{BA} = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{c_{ii}}{\sum_{j=1}^{N_c} c_{ij}} \tag{5.4}$$

where N_c is the number of classes (4 in this case) and c_{ij} is the number of spectra of class i classified as class j. Hence, c_{ii} is the number of cell spectra correctly classified for the class i.

5.2.9 Nested cross-validation

One of the main concerns in classification is to obtain a reliable measure of the performance of the developed models. In order to avoid, or at least diminish, over-fitting and assess generalization ability of the classifier, the test and training sets of spectra must be very well defined and separated during the whole process of training, optimisation and assessment of the learning framework [183, 186]. This task was accomplished by applying a hierarchical Cross-Validation (CV) approach, called *nested CV*, which consists of two *loops*: an *outer loop* for assessing the constructed classification models and an *inner loop* for training and optimising those learning models. Following this structure, two cross-validation alternatives were applied with the aim of checking the dependency of the discrimination capabilities of FTIR spectra on sample preparation and measurement conditions.

Outer loop: model assessment

In classification problems it is relatively easy to obtain high classification success rates even starting with random data [183]. This phenomenon, normally referred to as *over-fitting*, makes possible to get very good or even perfect discrimination on the dataset used to train the classification models, but generally offers poor prediction capabilities for new *unseen* samples [273]. In order to prevent this, in the *outer loop* a group of mean cell spectra called *test set* is separated from the rest of the dataset and does not take part during the construction of the learning models performed in the *inner loop*. Once the models are finished, the spectra from the test set act as *unseen* samples and they are used to assess the models constructed in the *inner loop*. This technique tries to give an estimate of the generalisation capabilities of those models.

Inner loop: model selection

The remaining samples not considered as test set in the *outer loop* are normally referred to as *training set*. The classification models have a series of parameters which must be tuned with that training set. These parameters can be divided into those tuned with the statistics of the samples, e.g., the specific weight given to a variable, and those whose values must be explicitly imposed by the expert, which are sometimes called *hyperparameters*. As the hyperparameters also influence the performance of the learning model, the selection of their values must be optimised.

In this case, the only hyperparameter is the number of components K retained during the feature extraction step, which determines the complexity of the model. In order to optimise its value and, again, reduce the risk of over-fitting, the original training set is alternatively subdivided into two subsets inside the *inner loop* of CV. The first subset of samples, normally called *validation set* [183, 186], plays the role of independent set, which assesses the performance of the classification models trained with the samples of the second subset. In this process, all the samples are considered once inside a validation set and a global measure of performance in terms of BA can be computed for each value of K. As an illustration, the left plot of Fig. 5.11 presents two curves with the evolution of BA with K for different validation sets.

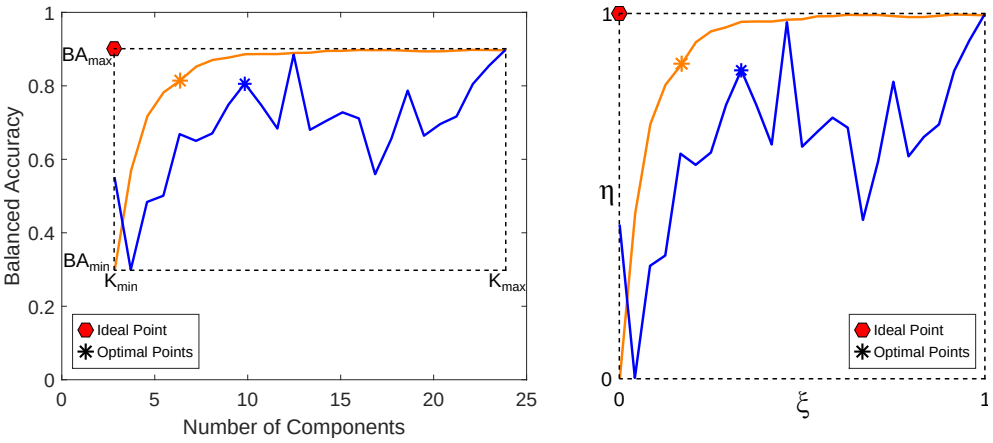


Fig. 5.11: Sketch of the optimisation of the number of retained components K. The left plot presents two curves (yellow and blue) with the evolution of the Balanced Accuracy (BA) with K for different validation sets. The right plot presents the same curves after applying the transformation to the normalised space where the distances to the ideal performance point are computed. In both plots, the ideal performance point and the selected optimal points for each curve have been symbolised as shown in the legends.

One straightforward way to optimise K would be to take the value corresponding to the maximum of those computed curves. However, in cases like the curves presented in Fig. 5.11, taking the maximum value would naïvely increase the complexity of the model. The *Hughes phenomenon* (Sec. 5.2.7), together with the fact that the presence of noise increases with the number of retained components (Sec. 3.3), recommend to keep the value of K as low as possible. Therefore, an optimisation procedure was developed to try to find automatically a compromise solution between the

maximisation of BA and the minimisation of K. The key concept of this procedure is that the ideal point, for a given maximum performance, would be the point $(K_{min}, \mathrm{BA}_{max})$. Therefore, the optimal point inside the computed curve would be the closest one to that ideal point. As the variables of the two original axes (K, BA) have different *nature*, they must be normalised in order to give them a fair weight. Mathematically, the coordinates (ξ, η) of this referenced and normalised space are computed by using the maximum and minimum along each original axis:

$$\xi = \frac{K - K_{min}}{K_{max} - K_{min}} \tag{5.5a}$$

$$\eta = \frac{\mathrm{BA} - \mathrm{BA}_{min}}{\mathrm{BA}_{max} - \mathrm{BA}_{min}} \tag{5.5b}$$

As can be seen in the right plot of Fig. 5.11, the performance curves are *distorted* in the normalised space and the ideal performance point becomes the point $(\xi_{ideal}, \eta_{ideal}) = (0, 1)$. In this new normalised space, the optimal number of retained components K_{opt} is finally computed as the value of the curve that minimises the distance to the ideal point d_{opt}. That is:

$$K_{opt} = \arg\min_{K} \{d_{opt}\} \tag{5.6}$$

where

$$d_{opt} = \sqrt{(\xi - \xi_{ideal})^2 + (\eta - \eta_{ideal})^2} \tag{5.7}$$

The optimal points for each curve of Fig. 5.11 are symbolised by asterisks. In the final applications, the values of minimum (K_{min}) and maximum (K_{max}) number of retained components were respectively fixed to 1 and 25. The final step in the *inner loop* consists of constructing the classification models by using the computed K_{opt} values and the whole original training set. Those optimised models are taken to the *outer loop* to assess their generalisation performance with the isolated test set.

Cross-validation alternatives

Once the CV method has been chosen, it is important to decide how the test, training and validation sets will be created from the whole dataset of mean cell spectra that were retained after the removal of outliers. The prediction capabilities of the classification models are determined by the similarities between the spectra from the

test sets and the training sets. In order to assess those similarities, two approaches were followed to separate the mean cell spectra of each kind of set:

- *One-Batch-Out CV*: consecutively, the spectra extracted from the images of one batch (Fig. 5.3) are considered as the test set in the *outer loop* and the classification algorithms are trained with the spectra of the other batches in the *inner loop*. The process is repeated so that each batch is considered once as the test set. The results for each test batch are combined to provide the final performance measurement (a single value of BA). The idea behind this approach is to assess the uniformity of cell lines between batches and to check if there may be some critical factors for correct discrimination (e.g., the sample preparation procedure or the measurement conditions).

- *In-Batch CV*: as shown in Fig. 5.12 for *Batch 1*, each image is split in 5 vertical stripes with an equal number of segmented cells. Each vertical stripe of the same colour is considered as the test set and the remaining stripes form the training set. This is repeated until all stripes have been in the test set once. The main reason to group cells in vertical stripes, instead of randomly choosing them for the test sets, is to try to construct synthetic subimages where the spatial variability within the same cell culture is also assessed. When this process is finished for one batch, it is repeated for the remaining batches independently. The results for each test stripe and batch are finally combined to provide the final performance measurement (a single value of BA). The aim of this approach is to assess the discrimination between cell lines inside each batch and to compare the performance with the *One-Batch-Out CV* approach.

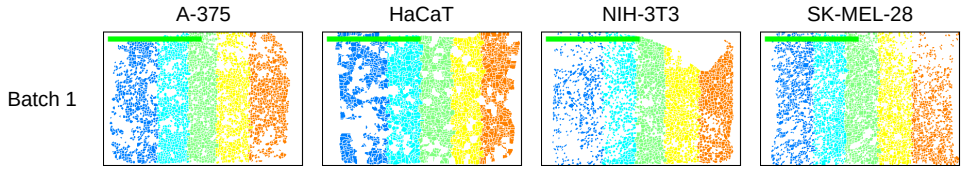


Fig. 5.12: Sketch of the *In-Batch CV* approach. Each image is split in 5 vertical stripes with an equal number of segmented cells; CV is performed independently within each batch by alternately considering one of the stripes from each cell line as the test set and the rest of stripes as the training set. Green scale bars represent 1 mm.

5.3 Results

5.3.1 Exploratory data analysis

Before presenting the final classification results, it is useful and even advisable to perform an exploratory analysis of the data in search for trends or patterns that may justify the outputs of the classification. This analysis is centred on the inputs of the final classification module, i.e., the mean cell spectra after outliers removal (Fig. 5.1). This task consists of visualising the main outputs of descriptive statistics and PCA.

Descriptive statistics

Morphological information

The information about the final number of segmented cell regions and their corresponding number of pixels retained after outliers removal for each cell line and batch are presented in Tab. 5.2. A total of 22700 cells were retained for the final analysis. As can be observed, the number of cells varies between cell cultures.

Tab. 5.2: Information about the retained cells after outliers removal. Number of segmented cellular regions (first number) and number of pixels inside them (second number) for each cell line (columns) and batch (rows). The last row and column present the corresponding total marginal values.

	A-375	**HaCaT**	**NIH-3T3**	**SK-MEL-28**	**Total**
Batch 1	2247 / 35473	1638 / 41525	1750 / 31524	2495 / 34610	8130 / 143132
Batch 2	2438 / 31314	966 / 23367	2656 / 45348	1931 / 35875	7991 / 135904
Batch 3	1867 / 27906	875 / 20609	2318 / 45377	1519 / 32022	6579 / 125914
Total	6552 / 94693	3479 / 85501	6724 / 122249	5945 / 102507	22700 / 404950

The information of the total number of retained cells and pixels can be complemented by the histograms of pixels per cell shown in Fig. 5.13. The majority of cells cover less than 20 pixels but there are slight differences between samples and cell lines. As already pointed out (e.g., see Figs. 1.13 and 5.9), HaCaT cells tend to form more homogeneous monolayers and adopt flatter configurations than the other cell lines, which is reflected in more uniform histograms. On the contrary, malignant cell lines (A-375 and SK-MEL-28) and in a less extent NIH-3T3 cells normally grow in a more proliferative and disordered way. As a result, those cell lines present higher proportions of smaller cells, which also tend to be more rounded and compact and, hence, more liable to produce scattering artifacts.

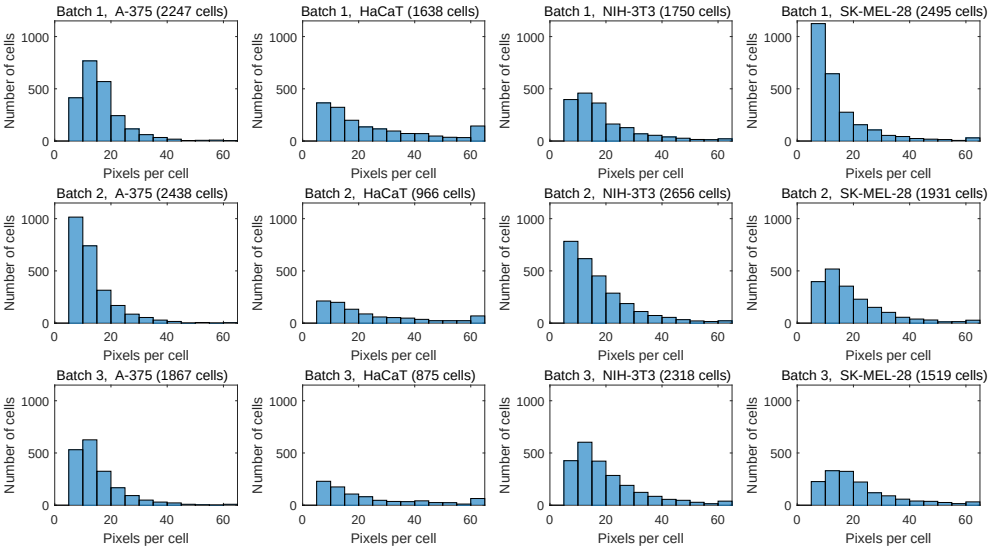


Fig. 5.13: Histograms of pixels per cell in the final retained cells of each sample.

Spectral information

The mean and the standard deviation (std) of the retained mean cell spectra were computed for each hyperspectral image as measurements of central tendency and dispersion. Fig. 5.14 displays these values for the raw spectra without preprocessing in the form of shaded error bars for each cell culture or hyperspectral image. In addition, the mean spectrum of each cell line has been represented in order to disclose the differences between batches for each cell line. In these plots, the gross variations (offsets, maximum absorbance, etc.) of raw spectra that have been extensively described in Secs. 3.2 and 5.2.4 become evident. This variability between mean cell spectra within the same hyperspectral image and also between batches can seriously hamper the prediction capabilities of the learning models.

Spectral preprocessing techniques aim to reduce the unwanted variability of raw measurements by standardising or normalising the spectra. As an illustration of the effects of spectral preprocessing, Fig. 5.15 shows the equivalent shaded error bars of the mean cell spectra preprocessed by the RMieS-EMSC alternative. The most distinguishing effect of this preprocessing alternative is the drastic reduction of the standard deviation of spectra, i.e., differences between spectra become much more

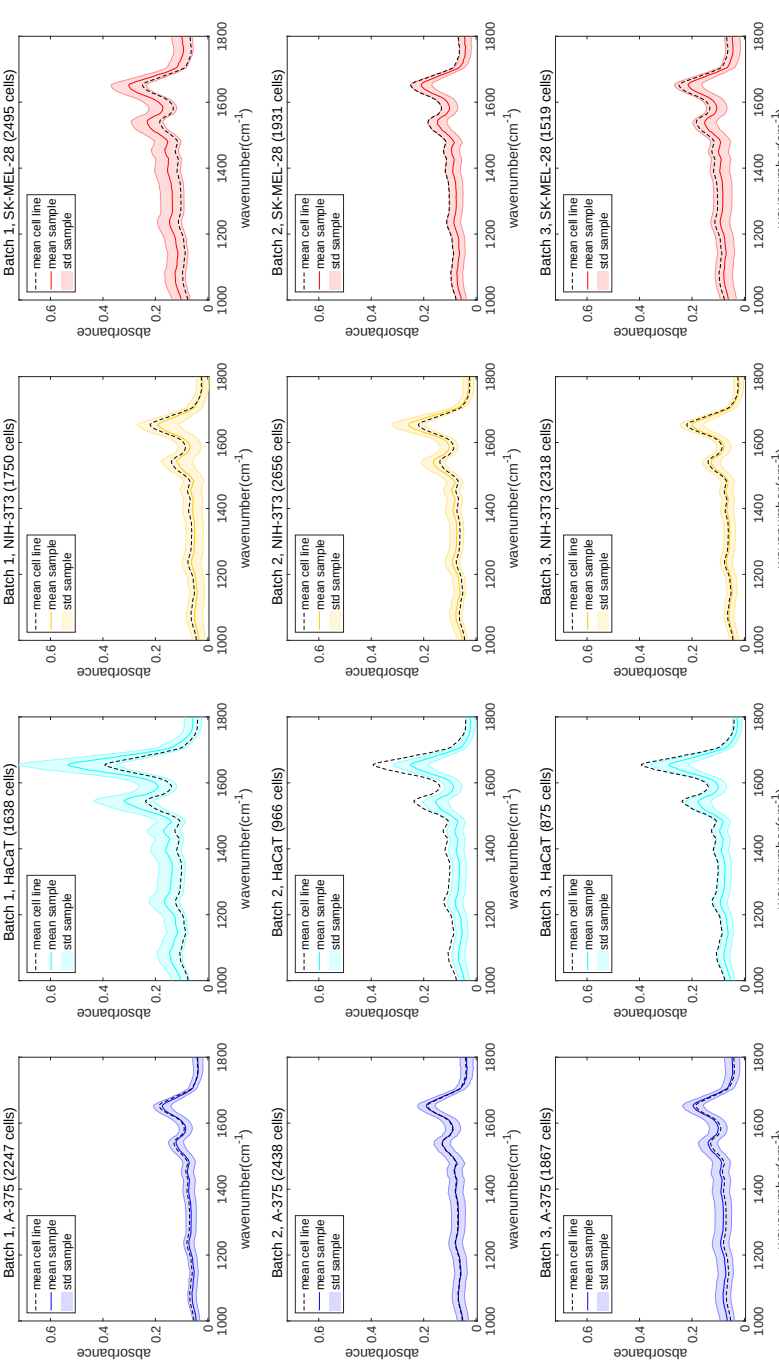


Fig. 5.14: Shaded error bars of the mean cell spectra without preprocessing (Raw) retained after outliers removal for each batch (rows) and cell line (columns). As specified in the legends, continuous lines of the corresponding colours represent the mean spectra of each cell culture (hyperspectral image) and the shaded regions the standard deviation (std) for each measured wavenumber. The black discontinous lines are the mean spectra of all the mean cell spectra of each cell line (column).

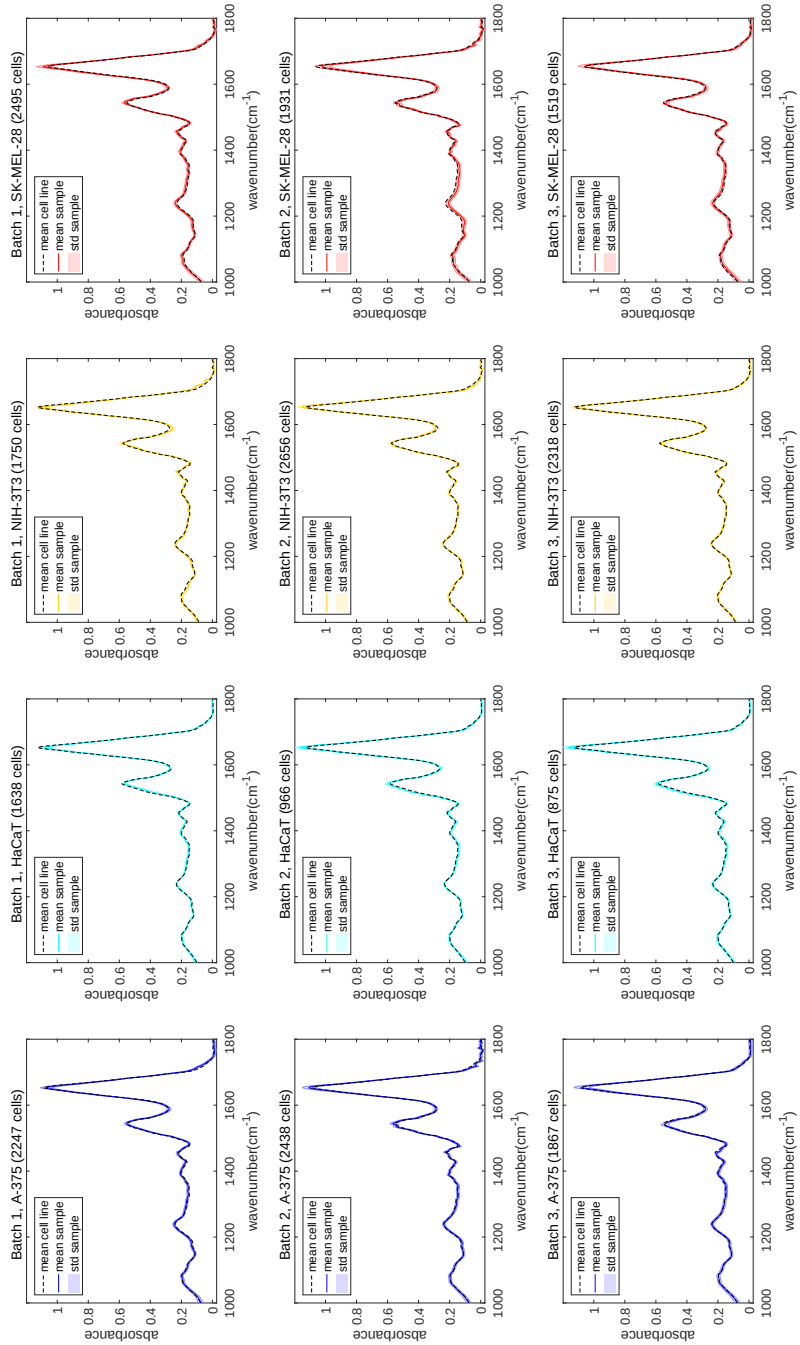


Fig. 5.15: Shaded error bars of the mean cell spectra preprocessed by RMieS-EMSC and retained after outliers removal for each batch (rows) and cell line (columns). As specified in the legends, continuous lines of the corresponding colours represent the mean spectra of each cell culture (hyperspectral image) and the shaded regions the standard deviation (std) for each measured wavenumber. The black discontinous lines are the mean spectra of all the mean cell spectra of each cell line (column).

subtle. Depending on the preprocessing alternative this variability will be mainly caused by different sources and, in the case of RMieS-EMSC (the most complex alternative), the main source is supposed to be the biochemistry of the cells. If the drawn preprocessed spectra are closer studied, those subtle differences can be identified. As an example, in A-375 and SK-MEL-28 cell lines of *Batch 2* high *frequency* variations in the range 1300-1800 cm^{-1} are more relevant than in the other batches, which may suggest a higher content of water vapour artifacts.

For brevity, the shaded error bars of the rest of alternatives are not shown but in exchange the mean spectra for each cell line and preprocessing alternative are presented in Fig. 5.16. As can be observed, the differences between cell lines change depending on the preprocessing. In fact, the order of presentation of the preprocessing alternatives corresponds with the level of differences observed. This order also agrees with the level of transformation that individual spectra experiment during the corresponding preprocessing. In this sense, Min-Max alternative broadly removes the offsets and applies a basic normalisation. Therefore, with this preprocessing the differences observed between cell lines are likely due to Mie scattering artifacts and, by extension, to cell morphology as evidenced by the uneven ratios of Amide I-Amide II peaks and remaining baselines. Either way, it must be demonstrated if the reduction of undesirable artifacts with the most complex preprocessing alternatives is necessary and effective enough to discriminate, or improve the discrimination of, cell lines.

Principal component analysis

As was described in Sec. 3.3.1, the capabilities of PCA for dimensionality reduction can be used to visualise the intrinsic structures of hyperspectral datasets. Because PCA is an unsupervised method (no information of the cell class is used), their plots provide useful details about the *proximity* or *similarity* between spectra and the possible *natural* groups that they may form. Following this line, PCA can be used to extract preliminary information from the different datasets which each cross-validation alternative (Sec. 5.2.9) will have to deal with.

In the case of *One-Batch-Out CV*, it is interesting to explore the whole dataset of mean cell spectra in order to study if there are relevant differences between the batches. The most important plots after applying PCA to the whole dataset of retained mean cell spectra are shown in Fig. 5.17. Again, RMieS-EMSC has been

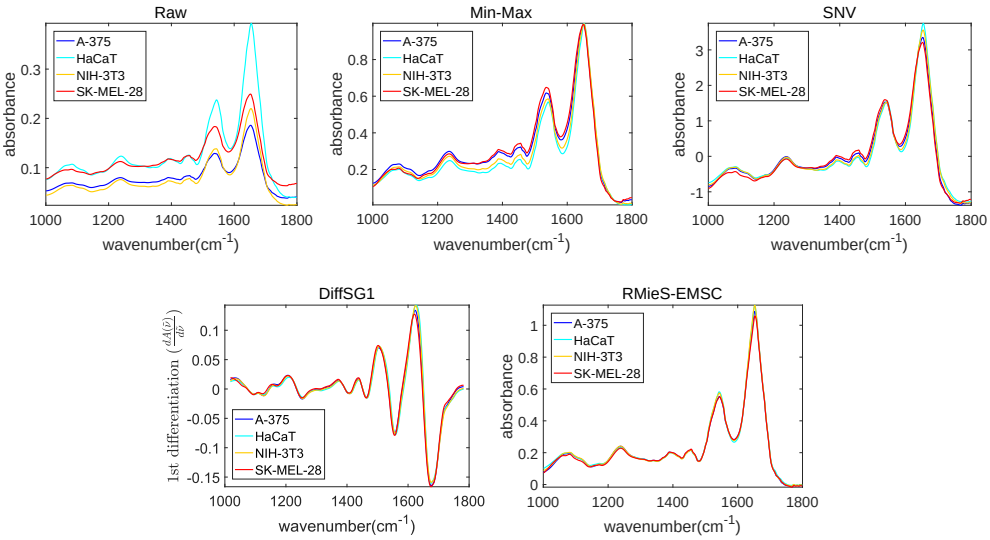


Fig. 5.16: Mean spectra of the mean cell spectra retained after outliers removal for each cell line and preprocessing alternative.

chosen as the reference preprocessing because it is the most advanced alternative. In Fig. 5.17a, 100 mean cell spectra from each cell line and batch have been randomly selected to construct the score plots of the 2 first PCs. In that plot, points from the same cell line have been drawn with the same colour and same symbols have been used to identify points from each batch. As can be observed, some subgroups of the same cell lines seem to appear mainly from non-tumoral HaCaT and NIH-3T3, which spread less than melanoma cell lines A-375 and SK-MEL-28. Despite those local subgroups, a global separation between cell lines is not so evident.

In Fig. 5.17b the points belonging to each cell line have been isolated in distinct subplots in order to better distinguish differences between batches within each cell line. In those subplots, batches have been denoted by the same symbols used in Fig. 5.17b but assigned different colours. As can be seen, spectra from *Batch 1* clearly tend to cluster separately from the rest of batches in all cell lines. However, points from *Batch 2* and *3*, with the exception of NIH-3T3, are much more overlapping. If these subplots are compared with Fig. 5.17a, now spectra from *Batch 1* can be better identified in the lower part of the graphs. In that batch, NIH-3T3 spectra seem more isolated than the other cell lines. In addition, HaCaT cells from *Batches 2* and *3* can

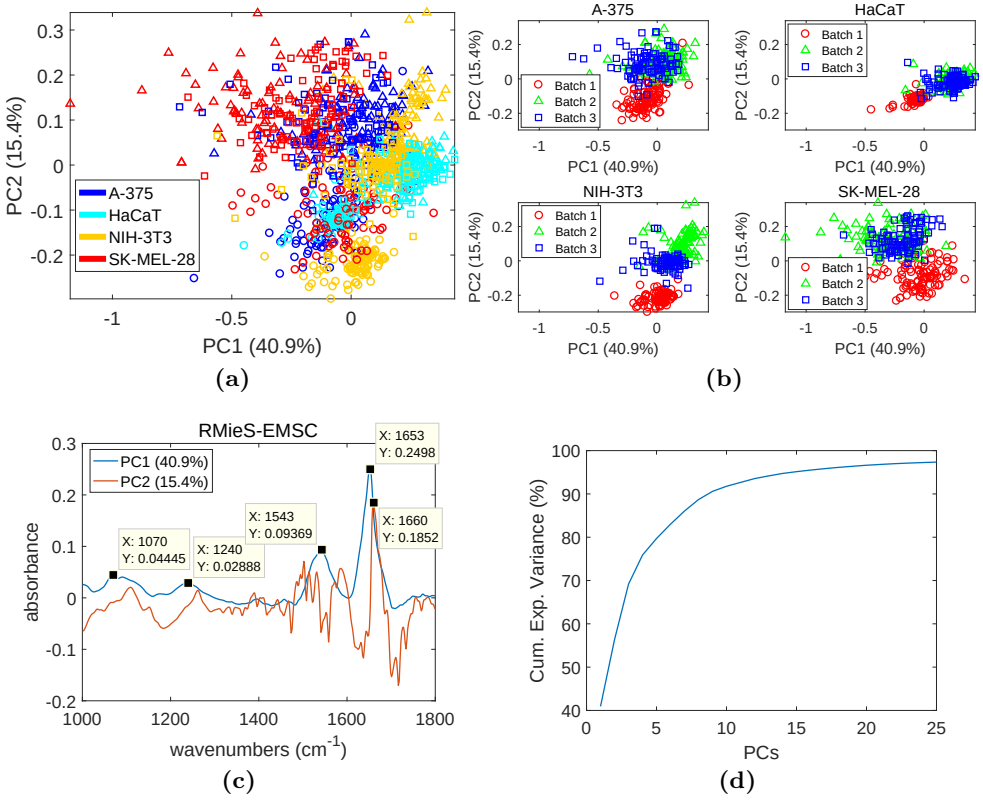


Fig. 5.17: Principal component analysis of the whole dataset of retained mean cell spectra preprocessed by RMieS-EMSC. (a) Score plots of the two first PCs with the corresponding percentage of explained variance in parentheses where 100 mean cell spectra for each cell line (depicted with different colours) and batch (depicted with different symbols) have been randomly chosen from the whole dataset. (b) Same score plots than in (a) showing only the randomly selected spectra of each cell line; symbols used to denote the batch are the same as in figure (a) but colours have been changed to better identify differences between batches. (c) Loadings of the 2 first PCs with the corresponding percentage of explained variance in parentheses. (d) Evolution of the cumulative explained variance for the first 25 PCs.

be better distinguished in the right central region more or less mixed with NIH-3T3 spectra. Finally, spectra from melanoma cell lines A-375 and SK-MEL-28 of *Batches 2* and *3* are much more tangled in the upper-central region of the chart.

Fig. 5.17c displays the corresponding loading vectors of the 2 first PCs. As observed, the highest weights of the first PC (PC1), which explains 40.9% of the total

variance, are located around the Amide I peak (∼1600-1700 cm^{-1}, with maximum at 1653 cm^{-1}) and Amide II peak (∼1500-1600 cm^{-1}, with maximum at 1543 cm^{-1}). Other stronger weights occur around 1070 cm^{-1} and 1240 cm^{-1}. The second PC (PC2) explains 15.4% of the total variance and it gives important clues about the presence of two spectral artifacts (Sec. 2.6). The first artifact is Resonant Mie Scattering, which causes shifts in the maxima's position of the strongest absorption peaks and may be responsible for the difference in the position of the maximum weight (1660 cm^{-1}) with respect to PC1. The second artifact is likely water vapour, which may be responsible for the *high-frequency* fluctuations throughout the range 1300-1800 cm^{-1}.

Finally, Fig. 5.17d shows the evolution of cumulative explained variance with the number of retained PCs. As can be observed, the first 5 PCs account for 80% of the total variance and 15 PCs must be retained to explain a 95% percentage.

In the case of *In-Batch CV*, the similarities between cell lines within each batch may be inferred from the described study of the whole dataset. However, it may be more convenient to analyse each batch independently in order to reduce the interferences of other batches during the computation of the main directions of variation. Fig. 5.18 shows the most relevant plots when applying PCA to the subsets of retained mean cell spectra preprocessed by RMieS-EMSC from each batch.

The score plots of the 2 first PCs of each batch (first column of Fig. 5.18) reveal congruent information with the study of the whole dataset about the relative dispersion of each cell line: HaCaT and NIH-3T3 tend to form more compact subgroups, meanwhile A-375 and SK-MEL-28 are relatively more spread. Concerning the separation between cell lines, slight differences can be observed within each batch: in *Batch 1*, NIH-3T3 seems to be the most isolated cell line followed by HaCaT, which has partial overlaps with the other cell lines; in *Batch 2*, HaCaT and SK-MEL-28 spectra seem well separated but there is a significant overlap between A-375 and NIH-3T3; in *Batch 3*, HaCaT and NIH-3T3 seem to overlap only partially with the rest of cell lines, meanwhile A-375 and SK-MEL-28 spectra appear more mixed, as in *Batch 1*.

The loading vectors of the 2 first PCs (second column of Fig. 5.18) again highlight the problems of RMieS, which may be responsible for the shifts in the maximum weights around the Amide I peak, and water vapour. In particular, fluctuations due

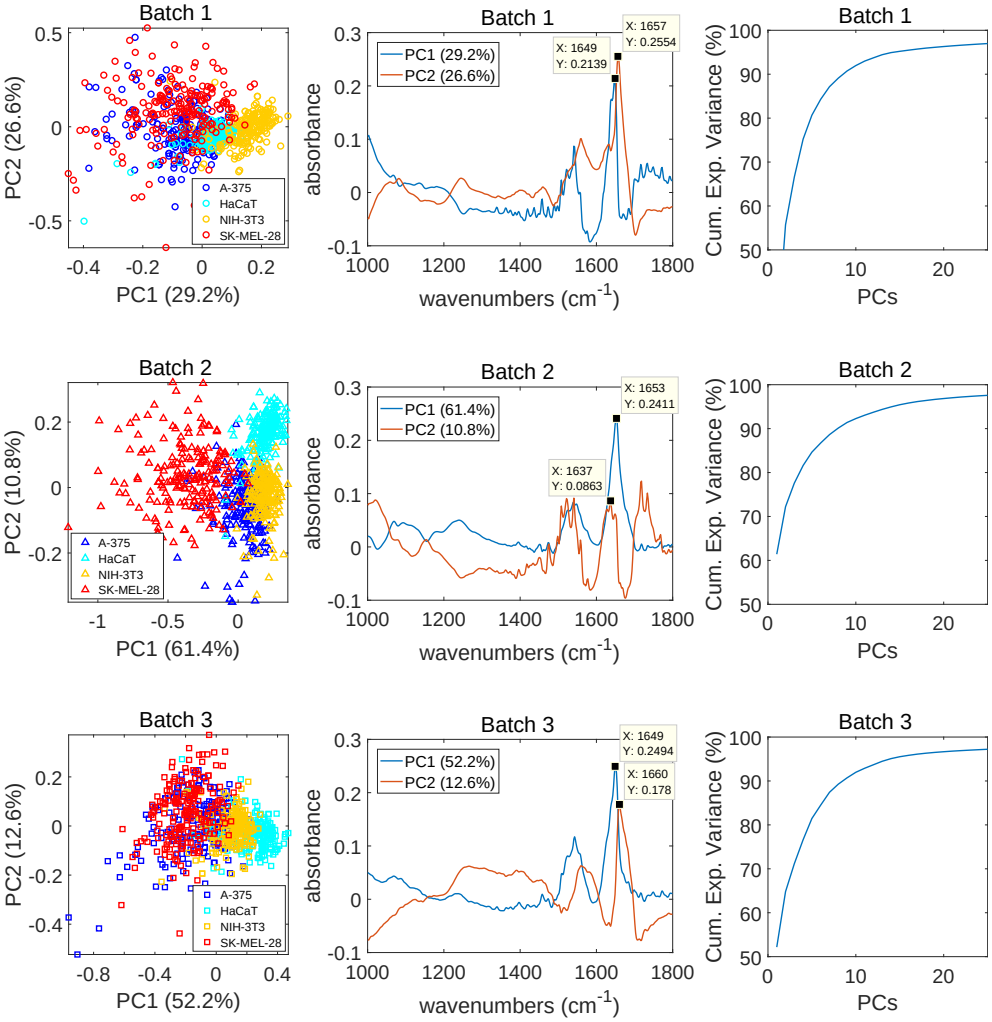


Fig. 5.18: Principal component analysis of the subsets of retained mean cell spectra from each batch (rows) preprocessed by RMieS-EMSC. First column: Score plots of the two first PCs, with the corresponding percentage of explained variance in parentheses, where 200 mean cell spectra for each cell line (depicted with different colours) have been randomly chosen from the corresponding batch's subset. Second column: Loadings of the 2 first PCs with the corresponding percentage of explained variance in parentheses. Third column: Evolution of the cumulative explained variance for the first 25 PCs.

to water vapour seem stronger in *Batch 2*, clearly distorting the maximum weights. Finally, the plots of the cumulative explained variance with the number of retained PCs (third column of Fig. 5.18) inform that, although in *Batch 1* the first PC is less relevant than in the rest of batches, more than 80% and 95% of the total variance can be explained by respectively retaining 5 and 15 PCs in all batches.

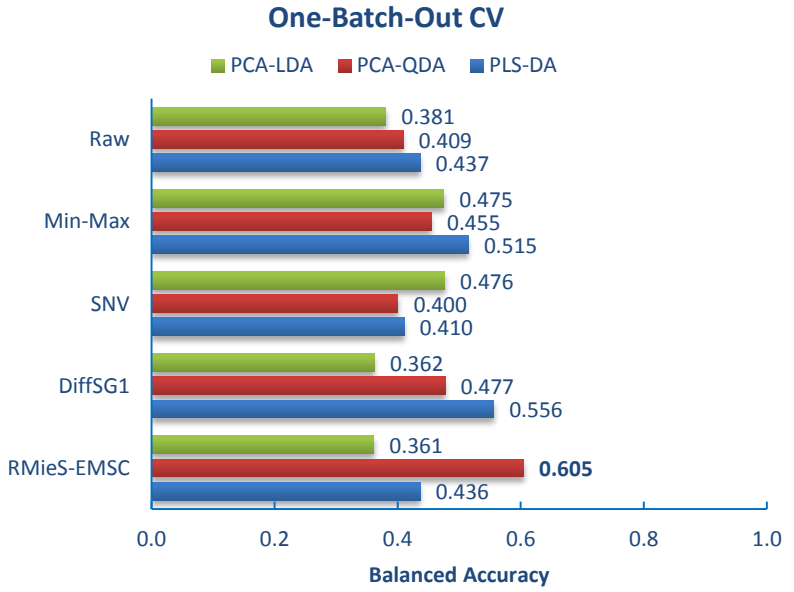
5.3.2 Classification results

The preliminary descriptive information of the most relevant trends are based on two principal components and must be contrasted with objectively-assessed classification models that consider the whole dataset of cells and a larger number of components. Fig. 5.19 shows graphically the final classification results in terms of Balanced Accuracy (BA) for the combinations of explored preprocessing options (Sec. 5.2.4), classification algorithms (Sec. 5.2.8) and cross-validation alternatives (Sec. 5.2.9). Each final BA was computed by combining the predicted labels of the corresponding test sets of mean cell spectra.

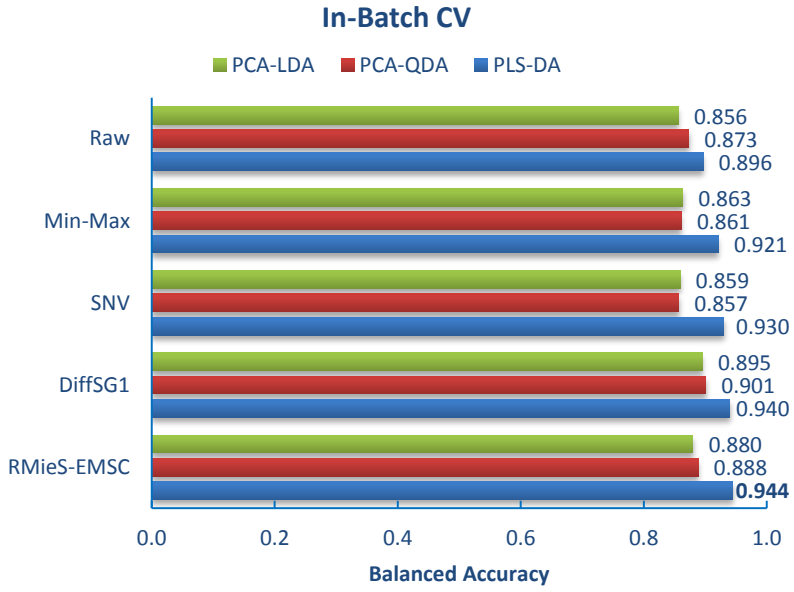
At first glance, there is an evident difference between the ranges of BA values obtained by the two CV approaches. Meanwhile in *One-Batch-Out CV* BA values range from around 0.35 to approximately 0.6 (combination RMieS-EMSC and PCA-QDA), in *In-Batch CV* all the combinations provide a final BA between 0.85 and 0.95 (combination RMieS-EMSC and PLS-DA). Regarding preprocessing, there is no clearly prevailing option and the final performance also depends on the employed classification algorithm. Although the maximum BA values in both CV alternatives have been reached by a combination including RMieS-EMSC.

If now we focus on the best combinations of each CV alternative, further details may be obtained about the underlying reasons to get those final BA values. The aim is to gain deeper information about why the misclassification may occur even in the best scenario.

Firstly, it is worth studying the optimisation curves of the number of retained components, computed with the corresponding validation sets as was described in Sec. 5.2.9. Figs. 5.20 and 5.21 show the optimisation curves for the best combinations of *One-Batch-Out CV* and *In-Batch CV*, respectively. In *One-Batch-Out CV*, a remarkably higher performance (around double BA) is reached when the *Batch 1* is left out as test set and algorithms are trained and validated only with the spectra



(a)



(b)

Fig. 5.19: Classification results in terms of Balanced Accuracy for the different preprocessing options (grouped by rows), classification algorithms (colours) and cross-validation alternatives: (a) *One-Batch-Out CV* and (b) *In-Batch CV*.

from *Batch 2* and *3* than when one of those batches is left out and the other one forms the initial training set with *Batch 1*. This fact confirms that there are more similarities between spectra from *Batch 2* and *3* than with those from *Batch 1*, which is in line with the preliminary studies by PCA (Fig. 5.17). On the other hand, the optimisation curves of *In-Batch CV*, individually constructed for each batch by leaving out vertical stripes of cells in the hyperspectral images, present much more similar characteristics between them. Even so, *Batch 2* provides a slightly lower performance than the other batches but in all cases BA is above 0.9. This suggests a high degree of similarity between the spectra from the same hyperspectral image and, hence, from the same cell line, as well as relevant differences with other cell lines within the same batch.

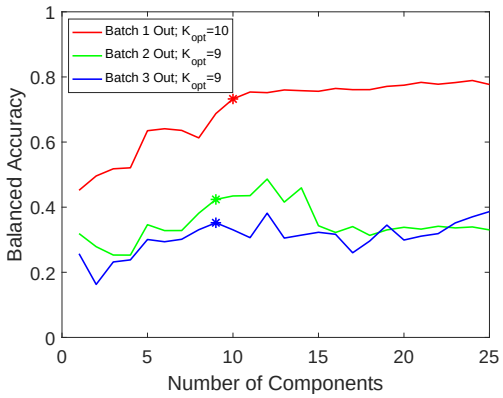


Fig. 5.20: Optimisation curves for the best combination (RMieS-EMSC and PCA-QDA) of *One-Batch-Out CV*.

Finally, again for the best combinations (highest BA) of each CV alternative, the predicted cell labels in the test sets were joined to create pseudo-colour images by using the colour code shown in Fig. 5.22. Figs. 5.23 and 5.24 respectively present those images for *One-Batch-Out CV* and *In-Batch CV* with the same distribution of batches (rows) and cell lines (columns) as in Fig. 5.3. Note that cells within each image should be coloured with the colour specified outside each column for a correct classification. These images provide more detailed information about the position of the misclassified spectra.

When analysing Fig. 5.23, some observations can be stated about the classification performed by the best combination of *One-Batch-Out CV*:

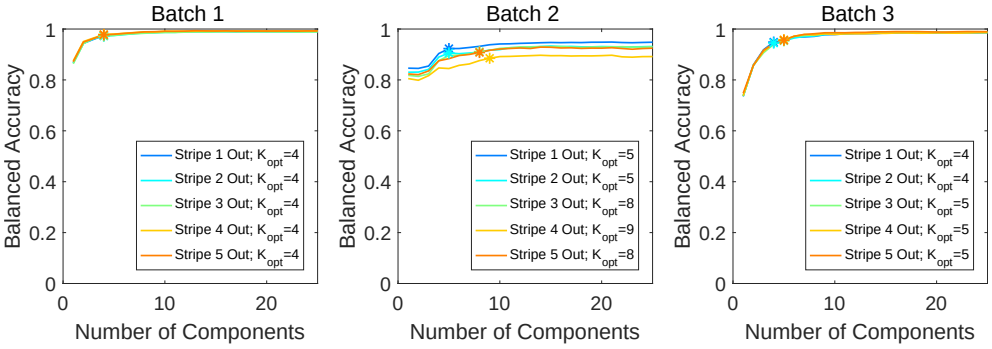


Fig. 5.21: Optimisation curves for the best combination (RMieS-EMSC and PLS-DA) of *In-Batch CV*.

Fig. 5.22: Colour code for the predicted labels of cells in the images of qualitative results.

- *Batch 1*: it presents the highest rate of misclassification. A significant amount of malignant A-375 cells are incorrectly labelled as the benign NIH-3T3. The worst case is the non-tumoral HaCaT cell line, whose cells are mostly *identified* as the tumoral SK-MEL-28. A small number of NIH-3T3 cells are confounded with cancerous A-375 cells. Lastly, almost half of SK-MEL-28 cells are wrongly classified as A-375.

- *Batch 2*: the malignant A-375 cell line of this batch is another case with extreme misclassification, again with the benign NIH-3T3. However, the rest of cell lines are very well categorised.

- *Batch 3*: A-375 cells are also mistaken with NIH-3T3 but in a lower degree than in Batch 2. The rest of cell lines present high rates of correct classification, although some cells from NIH-3T3 and SK-MEL-28 are mainly confused with the same cell types as in *Batch 1*.

Finally, the images of the best combination of *In-Batch CV* (Fig. 5.24) confirm the good general classification of most cells. The only remarkable inaccuracies are mainly located in the malignant A-375 and the benign NIH-3T3 cell lines of *Batch 2*, whose cells are mutually confused with the other cell line.

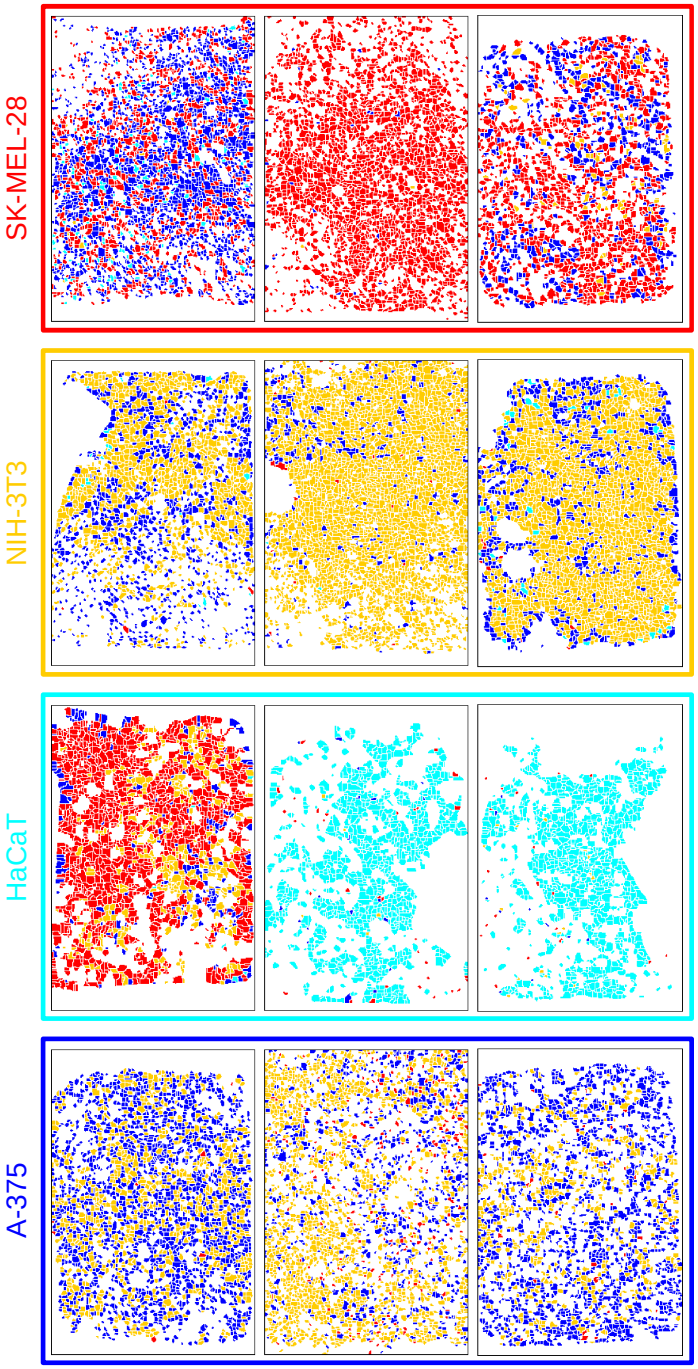


Fig. 5.23: Qualitative results for the best combination (RMieS-EMSC and PCA-QDA) of *One-Batch-Out CV*.

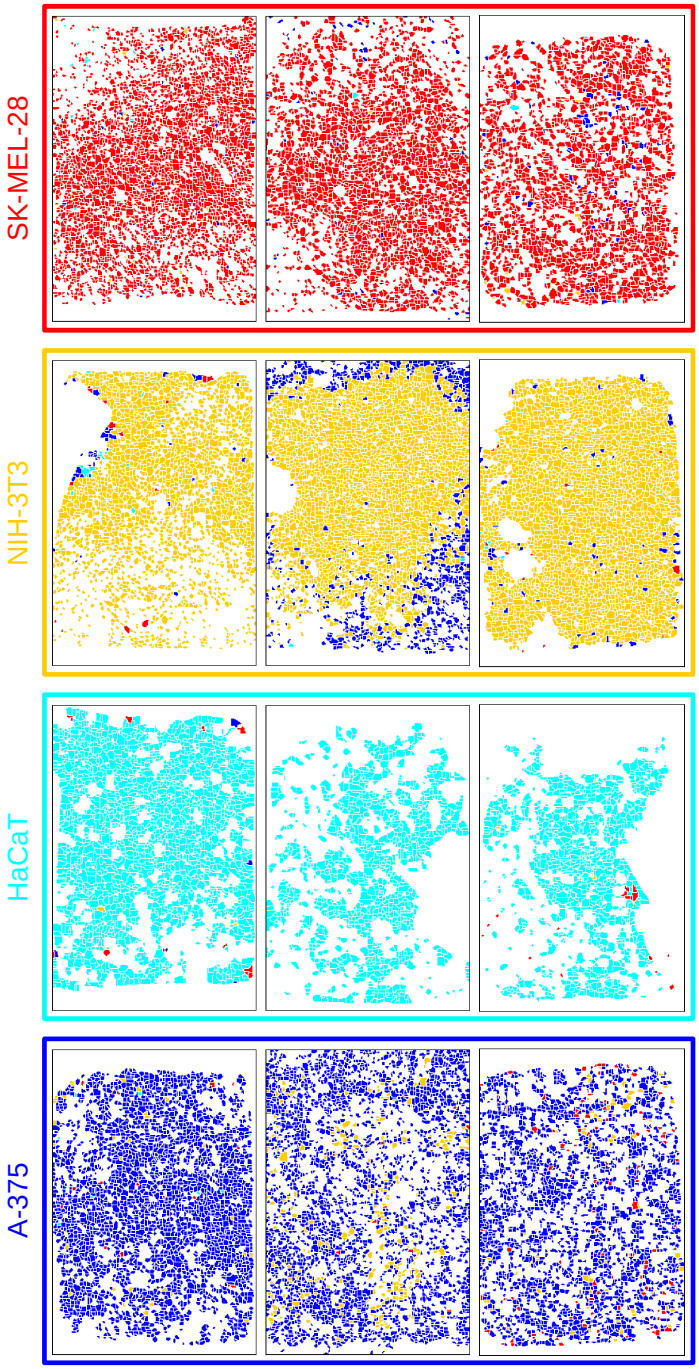


Fig. 5.24: Qualitative results for the best combination (RMieS-EMSC and PLS-DA) of *In-Batch CV*.

5.4 Discussion

In principle, the excellent classification results obtained in the *In-Batch CV* suggest a high potential of FTIR spectra to discriminate the skin cell lines, at least inside each batch. Nevertheless, those high rates of success are obtained independently of the preprocessing technique and even for the raw spectra. This fact warns that the results of the *In-Batch CV* may be too optimistic and some level of over-fitting may be playing a critical role. Those suspicions are supported by the less favourable results obtained in the *One-Batch-Out CV* alternative, which better assesses the generalisation capabilities of the discrimination framework.

Indeed, a deeper study of the most optimistic discriminative option reveals that critical differences exist, especially between cells from *Batch 1* and the other batches. These findings are also supported by the unsupervised PCA score plots of the exploratory analyses. Taking into account that cells from *Batch 1* were prepared and measured at time points different from *Batch 2* and *3* (Sec. 5.2.2) and, although similar sample preparation and measurement protocols were followed for all batches, they are likely to introduce confounding factors in FTIR spectra which are critical for the correct discrimination of cell lines. Fortunately, when the experimental conditions are more comparable, as in *Batch 2* and *3*, the extrapolation of the discrimination capabilities is more satisfactory.

These findings put into doubt the actual potential of FTIR spectra to detect subtle biochemical changes between pathological phenotypes and raises the question of whether the discrimination of cells is critically biased by the experimental factors. These factors are mixed with the rest of physicochemical information in the final hyperspectral images and discovering the exact sources of these perturbations is not easy. Despite this, the exploratory analysis of preprocessed spectra has revealed the undesirable presence of water vapour effects even in the loadings of first PCs. These water vapour fluctuations, which highly depend on the environmental conditions during the spectral acquisition, could not be avoided even when purging the sample area with dry air. Moreover, those interferences, whose relative importance is higher in spectra with lower absorbance, increase their impact in the global dataset due to the spectral normalisation.

The exploratory analyses also offered interesting information about the efficiency

and consequences of the RMieS-EMSC algorithm, which is currently the most advanced preprocessing method available practical with a implementation. In this study, 20 iterations of the algorithm were used, which took around 4 weeks to process the whole dataset of extracted individual spectra (around 400 thousand in total). However, shifts in the most relevant weights of the PCA loadings around the Amide I peak suggest a suboptimal correction of Resonant Mie Scattering artifacts. These artifacts, associated with morphological differences, are accentuated by cell culture conditions, where cells have sustained nutrients, support and enough space to proliferate, favouring the presence of mitotic stages. Differences in growing and proliferation properties between cell lines are responsible for uneven sizes and shapes, with tumoral cell lines more prone to being smaller and more compact. Therefore, Mie and Resonant Mie Scattering artifacts may have also played determinant confounding roles in the discrimination of cell lines. For example, this factor may be responsible for the misidentification of A-375 cells as NIH-3T3.

Clinically, the confusion of cancerous cells as normal (e.g., A-375 cells classified as NIH-3T3) is critical from a diagnostic point of view and a hypothetical decision support system would fail in the detection of the correct pathology for samples in a different batch, which is the likely clinical scenario. On the other hand, the designation of normal cells as cancerous (e.g., HaCaT in *Batch 2* classified as SK-MEL-28) would give rise to over-diagnosis, over-treatment and waste of resources. Hence, the transfer of the current technologies and methodologies to clinical practice is not recommended at this point in time.

Chapter 6

Conclusions

An extensive review of the current literature concerning the application of FTIR spectroscopy for cancer diagnosis has been accomplished throughout the development of this thesis. The most relevant references have been properly cited along this final manuscript. The most important biomedical notions have been synthesised and considered during the whole technical process.

Thorough reviews and understanding of the current knowledge of FTIR technology, ranging from the instrumentation to the involved physicochemical factors, have been also carried out. Special attention has been given to the possible spectral artifacts introduced in the different experimental steps. These artifacts are mixed with the biochemical information of the biological samples and can potentially mislead their later analysis. Huge efforts have been made to reduce the presence of those artifacts both during the acquisition of spectra and during their computational processing.

The most advanced existing methodologies for the correct handling, processing and analysis of hyperspectral images and signals have been also identified and adapted to the specific problems and real applications tackled in this thesis. The most relevant contributions of this thesis to the global knowledge in the spectroscopic and clinical fields have been condensed into two different applications.

In the first main application, a multimodal registration framework for the automatic alignment of FTIR spectroscopic and H&E stained images from different

histological sections has been presented. This methodology was used to register 47 colon samples from three different pathological groups (16 normal, 16 intermediate and 15 tumoral) with good overall qualitative and quantitative results. The proposed method exploits the information of concentration and thickness contained in the absorbance FTIR spectra to generate a grayscale image with a contrast similar to the one obtained from the H&E image. Thus, the morphological structures are highlighted and can be used as a reference for the alignment of the two multimodal images. This approach does not need a prior segmentation step that may introduce errors and reduce the spatial information content.

This automatic method can be easily extrapolated to other kinds of pathologies, such as prostate or breast cancer, where the inner part of the tissue contains relevant morphological structures. Moreover, it can also be applied to more heterogeneous tissues if their borders are also captured in the image. The values of the parameters used in the two steps of the registration framework may be tuned depending on the intrinsic morphological variability of the tissue. In particular, the required deformations computed in the second registration step should be higher as the distance of the sections of tissue to register increases.

The proposed method can improve the accuracy to combine the spatial information extracted from both the traditional H&E stained images and the emerging FTIR spectroscopy, even if different sections of tissue are used. These combinations can result in richer diagnostic algorithms which may consider complementary aspects of the pathological tissue, following the same philosophy as other approaches that fuse different medical imaging modalities.

In the second main application of this thesis, the potential application of the current state-of-the-art FTIR technology to cytopathological diagnosis has been assessed. A dataset of approximately 22700 cultured cells, derived from two tumoral and two non-tumoral skin cell lines, has been analysed to evaluate the discrimination capabilities of current FTIR technology in cytopathological problems. This number of analysed cells is relatively high when compared with existing studies. The analysed cells were distributed across three different batches, which were prepared and measured at different time points by following similar protocols to maximise uniformity between batches. However, these controlled protocols are limited by current technological and procedural restrictions, e.g., the lack of an isolation chamber

for the measured samples, or the need of a particular growing and preparation protocol for each cell line. Therefore, external perturbations were inevitably introduced both during sample preparation and spectral acquisition.

Different methodologies and approaches have been explored to process FTIR hyperspectral images and correctly analyse cell spectra throughout the whole discrimination pipeline. Excellent discrimination results are obtained when the algorithms are trained and tested with cell spectra from the same batch. However, those results are not confirmed when cells from different batches are mixed to construct the algorithms and they are finally applied to a different batch. This disappointing fact questions the real generalisation capabilities of the discriminative properties of FTIR spectra, which seems to be critically influenced by the differential factors between batches, such as the sample preparation protocol or the measurement conditions (e.g., different water vapour content).

Experimental design of future studies should assess if the optimistic results observed between cell lines of the same batches are really driven by genuine biochemical differences rather than artificial perturbations introduced by sample preparation and measurement protocols. In any case, the high rates or misclassification for some cell lines in different batches warn about the need of a better standardisation of the aforementioned protocols, together with the specific analysis methodologies. This is fundamental to develop and establish a reliable diagnostic technology of cancerous cells based on FTIR microspectroscopy in the near future.

Further development of standardisation protocols is needed, which will also have to incorporate and adapt to the existing routine procedures of anatomical pathology laboratories. Additional improvements can be introduced in the different steps of the proposed discrimination pipeline apart from the experimental improvements. Regarding spectral analysis, it is expected that the most relevant factors may be the preprocessing and the supervised classification framework. In the case of preprocessing, more complex techniques for the removal of interferences such as water vapour should be explored in case that measurements could not be performed in isolated environments. In addition, new advances in algorithms for the correction of RMieS, with higher computing speed, more stability and robustness against perturbations (e.g., again water vapour), should be incorporated.

Finally, the biomedical relevance and the extrapolation of the results obtained with cultured cells are limited because they are grown in artificial conditions. Therefore, the development of a reliable decision support system will necessarily have to use FTIR spectra of cells extracted from the patient, with a reference *ground truth* provided by pathologists, such as in histopathological studies. This is a question that may not be answered in the short term, at least until more urgent technical problems are addressed and more standardised protocols for sample preparation, measurement and analysis are developed and widely used.

References

[1] World Health Organization, "Cancer," http://www.who.int/cancer/en/, 2017, [Last access: 09/04/2017].

[2] P. Anand, A. B. Kunnumakara, C. Sundaram, K. B. Harikumar, S. T. Tharakan, O. S. Lai, B. Sung, and B. B. Aggarwal, "Cancer is a preventable disease that requires major lifestyle changes," *Pharmaceutical Research*, vol. 25, no. 9, pp. 2097–2116, 2008.

[3] V. Kumar, A. Abbas, and J. Aster, "Neoplasia," in *Robbins & Cotran Pathologic Basis of Disease*, 9th ed. Elsevier, 2015, ch. 7, pp. 265–340.

[4] R. Weinberg, *The biology of cancer*, 2nd ed. Garland Science, Taylor & Francis Group, 2013.

[5] M. Greaves and C. C. Maley, "Clonal evolution in cancer," *Nature*, vol. 481, no. 7381, pp. 306–313, Jan. 2012.

[6] A. J. Minn and J. Massagué, "Invasion and metastasis," in *Devita, Hellman, and Rosenberg's Cancer: Principles & Practice of Oncology*, 9th ed. Wolters Kluwer, 2011, ch. 10, pp. 113–127.

[7] R. A. Weinberg, "How cancer arises," *Scientific American*, vol. 275, no. 3, pp. 62–71, 1996.

[8] G. Cooper, "The Cell Cycle," in *The Cell : A Molecular Approach*, 4th ed. ASM Press Sinauer Associates, 2007, ch. 16, pp. 649–688.

[9] A. B. Pardee, "A Restriction Point for Control of Normal Animal Cell Proliferation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 71, no. 4, pp. 1286–1290, Apr. 1974.

[10] M. B. Kastan and J. Bartek, "Cell-cycle checkpoints and cancer," *Nature*, vol. 432, no. 7015, pp. 316–323, Nov. 2004.

[11] C. J. Sherr, "Cancer cell cycles," *Science*, vol. 274, no. 5293, pp. 1672–1677, 1996.

[12] G. I. Evan and K. H. Vousden, "Proliferation, cell cycle and apoptosis in cancer," *Nature*, vol. 411, no. 6835, pp. 342–348, May 2001.

[13] K. Vermeulen, D. R. Van Bockstaele, and Z. N. Berneman, "The cell cycle: a review of regulation, deregulation and therapeutic targets in cancer," *Cell Proliferation*, vol. 36, no. 3, pp. 131–149, 2003.

[14] B. Vogelstein and K. W. Kinzler, "Cancer genes and the pathways they control," *Nature Medicine*, vol. 10, no. 8, pp. 789–799, Aug. 2004.

[15] S. Maddika, S. R. Ande, S. Panigrahi, T. Paranjothy, K. Weglarczyk, A. Zuse, M. Eshraghi, K. D. Manda, E. Wiechec, and M. Los, "Cell survival, cell death and cell cycle pathways are interconnected: Implications for cancer therapy," *Drug Resistance Updates*, vol. 10, no. 1, pp. 13 – 29, 2007.

[16] D. Hanahan and R. Weinberg, "Hallmarks of Cancer: The Next Generation," *Cell*, vol. 144, no. 5, pp. 646 – 674, 2011.

[17] D. Hanahan and R. A. Weinberg, "The Hallmarks of Cancer," *Cell*, vol. 100, no. 1, pp. 57 – 70, 2000.

[18] D. Merino and D. Malkin, "p53 and hereditary cancer," in *Mutant p53 and MDM2 in Cancer*, S. P. Deb and S. Deb, Eds. Springer Netherlands, 2014, pp. 1–16.

[19] K. H. Vousden and X. Lu, "Live or let die: the cell's response to p53," *Nature Reviews Cancer*, vol. 2, no. 8, pp. 594–604, 2002.

[20] A. H. Bild, G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Chasse, M.-B. Joshi, D. Harpole, J. M. Lancaster, A. Berchuck, J. A. Olson, J. R. Marks, H. K. Dressman, M. West, and J. R. Nevins, "Oncogenic pathway signatures in human cancers as a guide to targeted therapies," *Nature*, vol. 439, no. 7074, pp. 353–357, Jan. 2006.

[21] A. Waugh and A. Grant, *Ross and Wilson Anatomy and Physiology in Health and Illness*, 9th ed. Churchill Livinstone, 2001.

[22] Encyclopdia Britannica, "Structures of the human large intestine, rectum, and anus," https://www.britannica.com/science/large-intestine, 2003, [Last access: 06/07/2017].

[23] L. P. Gartner, J. L. Hiatt, and J. M. Strum, *Cell biology and histology*, 6th ed. Wolters Kluwer Health/Lippincott Williams & Wilkins, 2011.

[24] S. V. Schaeybroeck, M. Lawler, B. Johnston, M. Salto-Tellez, J. Lee, P. Loughlin, R. Wilson, and P. G. Johnston, "Colorectal Cancer," in *Abeloff's Clinical Oncology*, 5th ed. Elsevier Saunders, 2014, ch. 77, pp. 1278–1335.

[25] J. Tobias and D. Hochhauser, *Cancer and its Management*, 7th ed. John Wiley & Sons, Ltd, 2015.

[26] S. K. Libutti, L. B. Saltz, and C. G. Willett, "Cancer of the Colon," in *Devita, Hellman, and Rosenberg's Cancer: Principles & Practice of Oncology*, 9th ed. Wolters Kluwer, 2011, ch. 89, pp. 1084–1126.

[27] S. K. Libutti, C. G. Willett, and L. B. Saltz, "Cancer of the Rectum," in *Devita, Hellman, and Rosenberg's Cancer: Principles & Practice of Oncology*, 9th ed. Wolters Kluwer, 2011, ch. 90, pp. 1127–1141.

[28] F. T. Bosman, F. Carneiro, R. H. Hruban, and N. D. Theise, *WHO classification of tumours of the digestive system*, 4th ed. Lyon: International Agency for Research on Cancer, 2010.

[29] T. Winslow, "Skin With Melanocyte Anatomy-HP," National Cancer Insititute, https://visualsonline.cancer.gov/details.cfm?imageid=8284, 2008, [Last access: 04/07/2017].

[30] A. L. Kierszenbaum and L. Tres, *Histology and Cell Biology: An Introduction to Pathology*, 4th ed. Saunders, 2014.

[31] A. Reszko, S. Z. Aasi, L. D. Wilson, and D. J. Leffel, "Cancer of the Skin," in *Devita, Hellman, and Rosenberg's Cancer: Principles & Practice of Oncology*, 9th ed. Wolters Kluwer, 2011, ch. 117, pp. 1610–1633.

[32] Mayo Clinic, "Where skin cancer develops," http://www.mayoclinic.org/diseases-conditions/skin-cancer/multimedia/where-skin-cancer-develops/img-20007623, 2015, [Last access: 04/07/2017].

[33] National Cancer Institute, "Common Moles," https://visualsonline.cancer.gov/details.cfm?imageid=9187, 2011, [Last access: 02/07/2017].

[34] ——, "Melanoma," https://visualsonline.cancer.gov/details.cfm?imageid=9186, 2011, [Last access: 02/07/2017].

[35] C. L. Slingluff, K. Flaherty, S. A. Rosenberg, and P. W. Read, "Cutaneous Melanoma," in *Devita, Hellman, and Rosenberg's Cancer: Principles & Practice of Oncology*, 9th ed. Wolters Kluwer, 2011, ch. 119, pp. 1643–1691.

[36] T. C. Gangadhar, L. A. Fecher, C. J. Miller, G. Karakousis, R. Vonderheide, G. Xu, and L. M. Schuchter, "Melanoma," in *Abeloff's Clinical Oncology*, 5th ed. Elsevier Saunders, 2014, ch. 69, pp. 1071–1091.

[37] H. L. Kaufman and J. M. Mehnert, Eds., *Melanoma*. Springer International Publishing, 2016.

[38] D. Schadendorf, D. E. Fisher, C. Garbe, J. E. Gershenwald, J.-J. Grob, A. Halpern, M. Herlyn, M. A. Marchetti, G. McArthur, A. Ribas, A. Roesch, and A. Hauschild, "Melanoma," vol. 1, p. 15003, Apr. 2015.

[39] R. J. Friedman, D. S. Rigel, and A. W. Kopf, "Early detection of malignant melanoma: The role of physician examination and self-examination of the skin," *CA: A Cancer Journal for Clinicians*, vol. 35, no. 3, pp. 130–151, 1985.

[40] D. S. Rigel, J. Russak, and R. Friedman, "The Evolution of Melanoma Diagnosis: 25 Years Beyond the ABCDs," *CA: A Cancer Journal for Clinicians*, vol. 60, no. 5, pp. 301–316, 2010.

[41] International Agency for Research on Cancer, "GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012." http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx, 2017, [Last access: 06/07/2017].

[42] M. Ervik, F. Lam, J. Ferlay, L. Mery, I. Soerjomataram, and F. Bray, "Cancer Today, Global Cancer Observatory," Lyon, France: International Agency for Research on Cancer, http://gco.iarc.fr/today, 2016, [Last access: 06/07/2017].

[43] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, "Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012," *International Journal of Cancer*, vol. 136, no. 5, pp. E359–E386, 2015.

[44] F. Bray, J.-S. Ren, E. Masuyer, and J. Ferlay, "Global estimates of cancer prevalence for 27 sites in the adult population in 2008," *International Journal of Cancer*, vol. 132, no. 5, pp. 1133–1145, 2013.

[45] International Agency for Research on Cancer, "GLOBOCAN - Glossary of Terms," http://globocan.iarc.fr/Pages/glossary.aspx, 2017, [Last access: 06/07/2017].

[46] P. Pisani, F. Bray, and D. M. Parkin, "Estimates of the world-wide prevalence of cancer for 25 sites in the adult population," *International Journal of Cancer*, vol. 97, no. 1, pp. 72–81, 2002.

[47] World Health Organization, "Cancer, fact sheet," http://www.who.int/mediacentre/factsheets/fs297/en/, 2017, [Last access: 06/07/2017].

[48] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistics, 2012," *CA: A Cancer Journal for Clinicians*, vol. 65, no. 2, pp. 87–108, 2015.

[49] M. B. Amin, S. Edge, F. Greene, D. R. Byrd, R. K. Brookland, M. K. Washington, J. E. Gershenwald, C. C. Compton, K. R. Hess, D. C. Sullivan, J. M. Jessup, J. D. Brierley, L. E. Gaspar, R. L. Schilsky, C. M. Balch, D. P. Winchester, E. A. Asare, M. Madera, D. M. Gress, and L. R. Meyer, Eds., *AJCC Cancer Staging Manual*, 8th ed. Springer International Publishing, 2017.

[50] J. D. Brierley, M. K. Gospodarowicz, and C. Wittekind, *TNM Classification of Malignant Tumours*, 8th ed. Wiley Blackwell/John Wiley & Sons, Inc, 2016.

[51] J. L. Young, S. D. Roffers, L. A. G. Ries, A. G. Fritz, and A. A. Hurlbut, Eds., *SEER Summary Staging Manual - 2000: Codes and Coding Instructions*. National Cancer Institute, NIH Pub. No. 01-4969, Bethesda, MD, 2001.

[52] T. Winslow, "Colorectal Cancer: Tumor Sizes," For the National Cancer Institute, http://www.teresewinslow.com/port.asp, 2005, [Last access: 04/07/2017].

[53] Cancer Research UK, "Statistics by cancer type," http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type, 2016, [Last access: 09/07/2017].

[54] Surveillance, Epidemiology, and End Results (SEER) Program, "Cancer Stat Facts," https://seer.cancer.gov/statfacts/, 2016, [Last access: 09/07/2017].

[55] K. S. Suvarna, C. Layton, and J. D. Bancroft, *Bancroft's theory and practice of histological techniques*, 7th ed. Churchill Livingstone Elsevier, 2013.

[56] G. Rolls, "An introduction to the preparation of specimens for microscopy in the histopathology laboratory," Leica Biosystems, http://www.leicabiosystems.com/pathologyleaders/an-introduction-to-specimen-preparation/, 2017, [Last access: 24/06/2017].

[57] E. Uthman, "Tissue processing," Flickr, https://flic.kr/s/aHsiDrC9Ek, 2006, [Last access: 24/06/2017].

[58] M. Fleming, S. Ravula, S. F. Tatishchev, and H. L. Wang, "Colorectal carcinoma: Pathologic aspects," *Journal of Gastrointestinal Oncology*, vol. 3, no. 3, 2012.

[59] M. T. Sheaff and N. Singh, *Cytopathology: an introduction*. London New York: Springer-Verlag London, 2013.

[60] R. Nayar, Ed., *Cytopathology in Oncology*. Springer-Verlag Berlin Heidelberg, 2014.

[61] F. W. Abdul-Karim, J. A. Brainard, S. I. Odronic, and C. W. Michael, "Cytopathology," in *Essentials of Anatomic Pathology*, 4th ed., L. Cheng and D. G. Bostwick, Eds. Springer International Publishing, 2016, ch. 1, pp. 1–118.

[62] G. Gill, *Cytopreparation: principles & practice*. Springer New York, 2013.

[63] J. M. Davis, Ed., *Animal Cell Culture: Essential Methods*, 1st ed. John Wiley & Sons, Inc., mar 2011.

[64] R. I. Freshney, *Culture of Animal Cells: A Manual of Basic Technique and Specialized Applications*, 7th ed. John Wiley & Sons, Inc., 2016.

[65] European Collection of Cell Cultures (ECACC)/Sigma-Aldrich, "Fundamental Techniques in Cell Culture - Laboratory Handbook," Health Protection Agency, http://www.sigmaaldrich.com/life-science/cell-culture/learning-center/ecacc-handbook.html, 2010, [Last access: 28/06/2017].

[66] Thermo Fisher Scientific, "Gibco Cell Culture Basics," http://www.thermofisher.com/es/es/home/references/gibco-cell-culture-basics.html, 2017, [Last access: 28/06/2017].

[67] R. J. Geraghty, A. Capes-Davis, J. M. Davis, J. Downward, R. I. Freshney, I. Knezevic, R. Lovell-Badge, J. R. W. Masters, J. Meredith, G. N. Stacey, P. Thraves, and M. Vias, "Guidelines for the use of cell lines in biomedical research," *British Journal of Cancer*, vol. 111, no. 6, pp. 1021–1046, Sep. 2014.

[68] D. J. Giard, S. A. Aaronson, G. J. Todaro, P. Arnstein, J. H. Kersey, H. Dosik, and W. P. Parks, "In Vitro Cultivation of Human Tumors: Establishment of Cell Lines Derived From a Series of Solid Tumors," *JNCI: Journal of the National Cancer Institute*, vol. 51, no. 5, pp. 1417–1423, 1973.

[69] P. Boukamp, R. T. Petrussevska, D. Breitkreutz, J. Hornung, A. Markham, and N. E. Fusenig, "Normal keratinization in a spontaneously immortalized aneuploid human keratinocyte cell line," *The Journal of Cell Biology*, vol. 106, no. 3, pp. 761–771, 1988.

[70] J. L. Jainchill, S. A. Aaronson, and G. J. Todaro, "Murine Sarcoma and Leukemia Viruses: Assay Using Clonal Lines of Contact-Inhibited Mouse Cells," *Journal of Virology*, vol. 4, no. 5, pp. 549–553, 1969.

[71] T. E. Carey, T. Takahashi, L. A. Resnick, H. F. Oettgen, and L. J. Old, "Cell surface antigens of human malignant melanoma: mixed hemadsorption assays for humoral immunity to cultured autologous melanoma cells." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 73, no. 9, pp. 3278–3282, 1976.

[72] N. A. C. S. Wong, L. P. Hunt, M. R. Novelli, N. A. Shepherd, and B. F. Warren, "Observer agreement in the diagnosis of serrated polyps of the large bowel," *Histopathology*, vol. 55, no. 1, pp. 63–66, 2009.

[73] B. Denis, C. Peters, C. Chapelain, I. Kleinclaus, A. Fricker, R. Wild, B. Augé, I. Gendre, P. Perrin, D. Chatelain, and J. F. Fléjou, "Diagnostic accuracy of community pathologists in the interpretation of colorectal polyps," *European Journal of Gastroenterology & Hepatology*, vol. 21, no. 10, 2009.

[74] P. G. van Putten, L. Hol, H. van Dekken, J. Han van Krieken, M. van Ballegooijen, E. J. Kuipers, and M. E. van Leerdam, "Inter-observer variation in the histological diagnosis of polyps in colorectal cancer screening," *Histopathology*, vol. 58, no. 6, pp. 974–981, 2011.

[75] J. K. Turner, G. T. Williams, M. Morgan, M. Wright, and S. Dolwani, "Interobserver agreement in the reporting of colorectal polyp pathology among bowel cancer screening pathologists in Wales," *Histopathology*, vol. 62, no. 6, pp. 916–924, 2013.

[76] L. Brochez, E. Verhaeghe, E. Grosshans, E. Haneke, G. Piérard, D. Ruiter, and J.-M. Naeyaert, "Inter-observer variation in the histopathological diagnosis of clinically suspicious pigmented skin lesions," *The Journal of Pathology*, vol. 196, no. 4, pp. 459–466, 2002.

[77] S. Lodha, S. Saggar, J. T. Celebi, and D. N. Silvers, "Discordance in the histopathologic diagnosis of difficult melanocytic neoplasms in the clinical setting," *Journal of Cutaneous Pathology*, vol. 35, no. 4, pp. 349–352, 2008.

[78] E. B. Hawryluk, A. J. Sober, A. Piris, R. M. Nazarian, M. P. Hoang, H. Tsao, M. C. Mihm, and L. M. Duncan, "Histologically challenging melanocytic tumors referred to a tertiary care pigmented lesion clinic," *Journal of the American Academy of Dermatology*, vol. 67, no. 4, pp. 727 – 735, 2012.

[79] S. Patrawala, A. Maley, C. Greskovich, L. Stuart, D. Parker, R. Swerlick, and B. Stoff, "Discordance of histopathologic parameters in cutaneous melanoma: Clinical implications," *Journal of the American Academy of Dermatology*, vol. 74, no. 1, pp. 75 – 80, 2016.

[80] Biomarkers Definitions Working Group, "Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework," *Clinical Pharmacology & Therapeutics*, vol. 69, no. 3, pp. 89–95, 2001.

[81] G. Bellisola and C. Sorio, "Infrared spectroscopy and microscopy in cancer research and diagnosis," *American Journal of Cancer Research*, vol. 2, no. 1, pp. 1–21, 2012.

[82] J. T. Bushberg, J. M. Boone, E. M. Leidholdt, and J. A. Seibert, *The Essential Physics of Medical Imaging*, 3rd ed. Wolters Kluwer Health/Lippincott Williams & Wilkins, 2011.

[83] J. L. Prince and J. M. Links, *Medical Imaging Signals and Systems*, 2nd ed. Pearson Prentice Hall, 2014.

[84] B. H. Stuart, *Infrared Spectroscopy: Fundamentals and Applications*. John Wiley & Sons, Ltd, 2004.

[85] R. Salzer and H. W. Siesler, Eds., *Infrared and Raman Spectroscopic Imaging*. John Wiley & Sons, 2014.

[86] M. Pilling and P. Gardner, "Fundamental developments in infrared spectroscopic imaging for biomedical applications," *Chemical Society Reviews*, vol. 45, pp. 1935–1957, 2016.

[87] C. Kendall, M. Isabelle, F. Bazant-Hegemark, J. Hutchings, L. Orr, J. Babrah, R. Baker, and N. Stone, "Vibrational spectroscopy: a clinical tool for cancer diagnostics," *Analyst*, vol. 134, pp. 1029–1045, 2009.

[88] M. Diem, A. Mazur, K. Lenau, J. Schubert, B. Bird, M. Miljković, C. Krafft, and J. Popp, "Molecular pathology via IR and Raman spectral imaging," *Journal of Biophotonics*, vol. 6, no. 11-12, pp. 855–886, 2013.

[89] O. J. Old, L. M. Fullwood, R. Scott, G. R. Lloyd, L. M. Almond, N. A. Shepherd, N. Stone, H. Barr, and C. Kendall, "Vibrational spectroscopy for cancer diagnostics," *Analytical Methods*, vol. 6, pp. 3901–3917, 2014.

[90] A. B. Seddon, "Mid-infrared (IR) A hot topic: The potential for using mid-IR light for non-invasive early detection of skin cancer in vivo," *Physica Status Solidi B*, vol. 250, no. 5, pp. 1020–1027, 2013.

[91] H. J. Byrne, M. Baranska, G. J. Puppels, N. Stone, B. Wood, K. M. Gough, P. Lasch, P. Heraud, J. Sule-Suso, and G. D. Sockalingum, "Spectropathology for the next generation: Quo vadis?" *Analyst*, vol. 140, pp. 2066–2073, 2015.

[92] V. Naranjo, F. Peñaranda, M. Alcañiz, B. Napier, G. S. Mark Farries and, J. Ward, C. Barta, R. Hasal, A. Seddon, S. Sujecki, U. M. Samir Lamrini and, O. Bang, P. M. Moselund, M. Abdalla, D. D. Gaspari, R. M. Vinella, H. Malm, G. R. Lloyd, N. Stone, J. Nallala, L. K. Juergen Schnekenburger and, and B. Kemper, "MINERVA Project, mid- To near Infrared Spectroscopy for Improved Medical Diagnostics," in *European Project Space on Intelligent Systems, Pattern Recognition and Biomedical Systems - EPS Lisbon,*. ScitePress, 2015, pp. 53–69.

[93] Vivid Components Ltd, "MINERVA project," http://minerva-project.eu/, 2017, [Last access: 19/07/2017].

[94] J. M. Chalmers and P. R. Griffiths, Eds., *Handbook of Vibrational Spectroscopy*. John Wiley & Sons, Ltd, 2002, vol. 4.

[95] ISO, "ISO 20473:2007(E): Optics and photonics – Spectral bands," International Organization for Standardization, Geneva, CH, Standard, Apr. 2007.

[96] P. Ronan, "Em spectrum," Wikimedia Commons, https://commons.wikimedia.org/wiki/File:EM_spectrum.svg, 2007, [Last access: 08/04/2017].

[97] NASA, "Em spectrum properties," Wikimedia Commons, https://commons.wikimedia.org/wiki/File:EM_Spectrum_Properties_reflected.svg, 2013, [Last access: 08/04/2017].

[98] P. R. Griffiths and J. A. De Haseth, *Fourier transform infrared spectrometry*. John Wiley & Sons, Inc, 2007.

[99] B. Smith, *Fundamentals of Fourier transform infrared spectroscopy*, 2nd ed. Boca Raton, FL: CRC Press, 2011.

[100] I. u. Rehman, Z. Movasaghi, and S. Rehman, *Vibrational Spectroscopy for Tissue Analysis*. Boca Raton, FL: CRC Press, 2012.

[101] M. Diem, *Modern Vibrational Spectroscopy and Micro-Spectroscopy: Theory, Instrumentation and Biomedical Applications*, 1st ed. John Wiley & Sons, Inc, 2015.

[102] E. A. Sharkov, "Black-body radiation," in *Passive Microwave Remote Sensing of the Earth: Physical Foundations*, 1st ed. Springer-Verlag Berlin Heidelberg, 2003, pp. 203–221.

[103] E. Hecht, *Optics*, 5th ed. Boston: Pearson Education, Inc, 2017.

[104] Sanchonx, "FTIR Interferometer," Wikimedia Commons, https://commons.wikimedia.org/wiki/File:FTIR_Interferometer.png, 2011, [Last access: 16/04/2017].

[105] P. Yip, "Sine and cosine transforms," in *Transforms and Applications Handbook, Third Edition*, ser. Electrical Engineering Handbook. CRC Press, 2010, ch. 3, pp. 1–34.

[106] Midac Corporation, "The Fundamentals of Infrared Spectroscopy," Midac Corporation, http://www.midac.com/files/Tn-100.PDF, 2012, [Last access: 30/04/2017].

[107] F. L. Pedrotti, L. M. Pedrotti, and L. S. Pedrotti, *Introduction to optics*. Pearson Prentice Hall, 2007.

[108] M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood, K. A. Heys, C. Hughes, P. Lasch, P. L. Martin-Hirsch, B. Obinaju, G. D. Sockalingum, J. Sulé-Suso, R. J. Strong, M. J. Walsh, B. R. Wood, P. Gardner, and F. L. Martin, "Using Fourier transform IR spectroscopy to analyze biological materials," *Nature Protocols*, vol. 9, no. 8, pp. 1771–1791, Aug. 2014.

[109] Crystran Ltd., "Calcium Fluoride (CaF2)," Crystran Ltd., https://www.crystran.co.uk/optical-materials/calcium-fluoride-caf2, 2012, [Last access: 27/04/2017].

[110] ——, "Calcium Fluoride - Circular Windows," Crystran Ltd., https://www.crystran.co.uk/windows/calcium-fluoride-windows/calcium-fluoride-circular-windows, 2012, [Last access: 27/04/2017].

[111] P. R. Griffiths, "Introduction to Vibrational Spectroscopy," in *Handbook of Vibrational Spectroscopy*, 1st ed., J. M. Chalmers and P. R. Griffiths, Eds. John Wiley & Sons, Ltd, 2002.

[112] S. Crouch, D. West, F. Holler, and D. A. Skoog, *Fundamentals of Analytical Chemistry*, 9th ed. Brooks/Cole, Cengage Learning, 2014.

[113] P. R. Griffiths, "Beer's Law," in *Handbook of Vibrational Spectroscopy*, 1st ed., J. M. Chalmers and P. R. Griffiths, Eds. John Wiley & Sons, Ltd, 2002.

[114] G. Clemens, J. R. Hands, K. M. Dorling, and M. J. Baker, "Vibrational spectroscopic methods for cytology and cellular research," *Analyst*, vol. 139, no. 18, pp. 4411–4444, 2014.

[115] H. J. Humecki, *Practical Guide to Infrared Microspectroscopy.* CRC Press, Taylor & Francis Group, 1995.

[116] R. Bhargava, "Infrared Spectroscopic Imaging: The Next Generation," *Applied Spectroscopy*, vol. 66, no. 10, pp. 1091–1120, 2012.

[117] L. Tisinger, "Industrial applications of FTIR imaging microscopy," Agilent Technologies, http://www.agilent.com/cs/library/eseminars/Public/Chemical_Imaging_for_Industrial_Laboratories.pdf, 2016, [Last access: 04/05/2017].

[118] M. J. Riedl, *Optical design fundamentals for infrared systems.* Bellingham, Wash: SPIE, 2001.

[119] K. Yeh, R. Reddy, and R. Bhargava, "Fourier Transform Infrared Spectroscopic Imaging: An Emerging Label-Free Approach for Molecular Imaging," in *Emerging Imaging Technologies in Medicine*, ser. Imaging in Medical Diagnosis and Therapy. Taylor & Francis, 2012, pp. 303–320.

[120] R. Bhargava, B. G. Wall, and J. L. Koenig, "Comparison of the ft-ir mapping and imaging techniques applied to polymeric systems," *Applied Spectroscopy*, vol. 54, no. 4, pp. 470–479, 2000.

[121] E. N. Lewis, P. J. Treado, R. C. Reeder, G. M. Story, A. E. Dowrey, C. Marcott, and I. W. Levin, "Fourier Transform Spectroscopic Imaging Using an Infrared Focal-Plane Array Detector," *Analytical Chemistry*, vol. 67, no. 19, pp. 3377–3381, 1995.

[122] L. H. Kidder, I. W. Levin, E. N. Lewis, V. D. Kleiman, and E. J. Heilweil, "Mercury cadmium telluride focal-plane array detection for mid-infrared Fourier-transform spectroscopic imaging," *Optics Letters*, vol. 22, no. 10, pp. 742–744, 1997.

[123] R. Bhargava and I. W. Levin, Eds., *Spectrochemical Analysis Using Infrared Multichannel Detectors*. Blackwell Publishing Ltd, 2005.

[124] P. Lasch and D. Naumann, "Spatial resolution in infrared microspectroscopic imaging of tissues," *Biochimica et Biophysica Acta (BBA) - Biomembranes*, vol. 1758, no. 7, pp. 814 – 829, 2006.

[125] M. Kansiz, "Advancing FTIR Imaging - Synchrotron IR Imaging in Your Lab," Agilent Technologies, http://www.agilent.com/cs/library/eseminars/Public/Advancing_FTIR_Imaging-Synchrotron_IR_Imaging_in_Your_Lab.pdf, 2016, [Last access: 13/05/2017].

[126] R. K. Reddy, M. J. Walsh, M. V. Schulmerich, P. S. Carney, and R. Bhargava, "High-Definition Infrared Spectroscopic Imaging," *Applied Spectroscopy*, vol. 67, no. 1, pp. 93–105, 2013.

[127] G. L. Carr and G. P. Williams, "Infrared microspectroscopy with synchrotron radiation," pp. 51–58, 1997.

[128] N. Jamin, P. Dumas, J. Moncuit, W.-H. Fridman, J.-L. Teillaud, G. L. Carr, and G. P. Williams, "Highly resolved chemical imaging of living cells by using synchrotron infrared microspectrometry," *Proceedings of the National Academy of Sciences*, vol. 95, no. 9, pp. 4837–4840, 1998.

[129] P. Dumas and L. Miller, "The use of synchrotron infrared microspectroscopy in biological and biomedical investigations," *Vibrational Spectroscopy*, vol. 32, no. 1, pp. 3 – 21, 2003.

[130] L. M. Miller and P. Dumas, "Chemical imaging of biological tissue with synchrotron infrared light," *Biochimica et Biophysica Acta (BBA) - Biomembranes*, vol. 1758, no. 7, pp. 846 – 857, 2006, vibrational Microscopic Imaging: Towards Molecular Pathology.

[131] P. Dumas, G. D. Sockalingum, and J. Sul-Suso, "Adding synchrotron radiation to infrared microspectroscopy: what's new in biomedical applications?" *Trends in Biotechnology*, vol. 25, no. 1, pp. 40 – 44, 2007.

[132] J. M. Chalmers, "Mid-infrared Spectroscopy: Anomalies, Artifacts and Common Errors," in *Handbook of Vibrational Spectroscopy*, 1st ed., J. M. Chalmers and P. R. Griffiths, Eds. John Wiley & Sons, Ltd, 2002.

[133] E. W. Ciurczak, "Anomalies in Near-infrared Spectroscopy," in *Handbook of Vibrational Spectroscopy*, 1st ed., J. M. Chalmers and P. R. Griffiths, Eds. John Wiley & Sons, Ltd, 2002.

[134] E. Pusey, R. B. Lufkin, R. K. Brown, M. A. Solomon, D. D. Stark, R. W. Tarr, and W. N. Hanafee, "Magnetic resonance imaging artifacts: mechanism and clinical significance." *RadioGraphics*, vol. 6, no. 5, pp. 891–911, 1986.

[135] S. G. Kaplan and M. A. Quijada, "Fourier Transform Methods," in *Spectrophotometry: Accurate Measurement of Optical Properties of Materials. Experimental Methods in the Physical Sciences*, 2014, vol. 46, ch. 4, pp. 97 – 141.

[136] J. R. Birch and F. J. J. Clarke, "Fifty categories of ordinate error in Fourier transform spectroscopy," *Spectroscopy Europe*, vol. 7, no. 4, pp. 16–22, 1995.

[137] A. Kohler, M. Zimonja, V. Segtnan, and H. Martens, "2.09 - Standard Normal Variate, Multiplicative Signal Correction and Extended Multiplicative Signal Correction Preprocessing in Biospectroscopy," in *Comprehensive Chemometrics*, S. D. Brown, R. Tauler, and B. Walczak, Eds. Oxford: Elsevier, 2009, pp. 139 – 162.

[138] H. Yang, S. Yang, J. Kong, A. Dong, and S. Yu, "Obtaining information about protein secondary structures in aqueous solution using Fourier transform IR spectroscopy," *Nature Protocols*, vol. 10, no. 3, pp. 382–396, 2015.

[139] P. Bassan, "Light scattering during infrared spectroscopic measurements of biomedical samples," Ph.D. dissertation, Faculty of Engineering and Physical Sciences, University of Manchester, 2011.

[140] G. Mie, "Beiträge zur Optik trüber Medien, speziell kolloidaler Metallösungen (Contributions to the optics of turbid media, particularly of colloidal metal solutions)," *Annalen der Physik*, vol. 330, no. 3, pp. 377–445, 1908.

[141] B. Mohlenhoff, M. Romeo, M. Diem, and B. R. Wood, "Mie-Type Scattering and Non-Beer-Lambert Absorption Behavior of Human Cells in Infrared Microspectroscopy," *Biophysical Journal*, vol. 88, no. 5, pp. 3635 – 3640, 2005.

[142] S. Boydston-White, T. Gopen, S. Houser, J. Bargonetti, and M. Diem, "Infrared spectroscopy of human tissue. V. Infrared spectroscopic studies of myeloid leukemia (ML-1) cells at different phases of the cell cycle," *Biospectroscopy*, vol. 5, no. 4, pp. 219–227, 1999.

[143] H.-Y. N. Holman, M. C. Martin, E. A. Blakely, K. Bjornstad, and W. R. Mckinney, "IR spectroscopic characteristics of cell cycle and cell death probed by synchrotron radiation based Fourier transform IR spectromicroscopy," *Biopolymers*, vol. 57, no. 6, pp. 329–335, 2000.

[144] J. K. Pijanka, A. Kohler, Y. Yang, P. Dumas, S. Chio-Srichan, M. Manfait, G. D. Sockalingum, and J. Sule-Suso, "Spectroscopic signatures of single, isolated cancer cell nuclei using synchrotron infrared microscopy," *Analyst*, vol. 134, pp. 1176–1181, 2009.

[145] M. Romeo, B. Mohlenhoff, and M. Diem, "Infrared micro-spectroscopy of human cells: Causes for the spectral variance of oral mucosa (buccal) cells," *Vibrational Spectroscopy*, vol. 42, no. 1, pp. 9–14, 2006.

[146] P. Bassan, H. J. Byrne, F. Bonnier, J. Lee, P. Dumas, and P. Gardner, "Resonant Mie scattering in infrared spectroscopy of biological materials - understanding the 'dispersion artefact'," *Analyst*, vol. 134, pp. 1586–1593, 2009.

[147] M. Romeo and M. Diem, "Correction of dispersive line shape artifact observed in diffuse reflection infrared spectroscopy and absorption/reflection (transflection) infrared micro-spectroscopy," *Vibrational Spectroscopy*, vol. 38, no. 1, pp. 129 – 132, 2005.

[148] P. Bassan, H. J. Byrne, J. Lee, F. Bonnier, C. Clarke, P. Dumas, E. Gazi, M. D. Brown, N. W. Clarke, and P. Gardner, "Reflection contributions to the dispersion artefact in FTIR spectra of single biological cells," *Analyst*, vol. 134, pp. 1171–1175, 2009.

[149] S. W. Bruun, A. Kohler, I. Adt, G. D. Sockalingum, M. Manfait, and H. Martens, "Correcting Attenuated Total ReflectionFourier Transform Infrared Spectra for Water Vapor and Carbon Dioxide," *Applied Spectroscopy*, vol. 60, no. 9, pp. 1029–1039, 2006.

[150] G. L. Carr, "Resolution limits for infrared microspectroscopy explored with synchrotron radiation," *Review of Scientific Instruments*, vol. 72, no. 3, pp. 1613–1619, 2001.

[151] B. J. Davis, P. S. Carney, and R. Bhargava, "Theory of Midinfrared Absorption Microspectroscopy: I. Homogeneous Samples," *Analytical Chemistry*, vol. 82, no. 9, pp. 3474–3486, 2010.

[152] ——, "Theory of Mid-infrared Absorption Microspectroscopy: II. Heterogeneous Samples," *Analytical Chemistry*, vol. 82, no. 9, pp. 3487–3499, 2010.

[153] P. Lasch, "Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging," *Chemometrics and Intelligent Laboratory Systems*, vol. 117, pp. 100 – 114, 2012.

[154] R. Gautam, S. Vanga, F. Ariese, and S. Umapathy, "Review of multidimensional data processing approaches for raman and infrared spectroscopy," *EPJ Techniques and Instrumentation*, vol. 2, no. 1, p. 8, 2015.

[155] H. J. Byrne, P. Knief, M. E. Keating, and F. Bonnier, "Spectral pre and post processing for infrared and Raman spectroscopy of biological tissues and cells," *Chemical Society Reviews*, vol. 45, pp. 1865–1878, 2016.

[156] J. E. Franke, "Inverse Least Squares and Classical Least Squares Methods for Quantitative Vibrational Spectroscopy," in *Handbook of Vibrational Spectroscopy*, 1st ed., J. M. Chalmers and P. R. Griffiths, Eds. John Wiley & Sons, Ltd, 2002.

[157] N. K. Afseth and A. Kohler, "Extended multiplicative signal correction in vibrational spectroscopy, a tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 117, pp. 92 – 99, 2012.

[158] H. Martens, J. P. Nielsen, and S. B. Engelsen, "Light Scattering and Light Absorbance Separated by Extended Multiplicative Signal Correction. Application to Near-Infrared Transmission Analysis of Powder Mixtures," *Analytical Chemistry*, vol. 75, no. 3, pp. 394–404, 2003.

[159] A. Kohler, C. Kirschner, A. Oust, and H. Martens, "Extended Multiplicative Signal Correction as a Tool for Separation and Characterization of Physical and Chemical Information in Fourier Transform Infrared Microscopy Images of Cryo-Sections of Beef Loin," *Applied Spectroscopy*, vol. 59, no. 6, pp. 707–716, 2005.

[160] A. Kohler, J. Sulé-Suso, G. D. Sockalingum, M. Tobin, F. Bahrami, Y. Yang, J. Pijanka, P. Dumas, M. Cotte, D. G. van Pittius, G. Parkes, and H. Martens, "Estimating and Correcting Mie Scattering in Synchrotron-Based Microscopic Fourier Transform Infrared Spectra by Extended Multiplicative Signal Correction," *Applied Spectroscopy*, vol. 62, no. 3, pp. 259–266, 2008.

[161] H. C. van de Hulst, *Light Scattering by Small Particles*. John Wiley & Sons, Inc, 1957.

[162] K. Tondel and H. Martens, "Analyzing complex mathematical model behavior by partial least squares regression-based multivariate metamodeling," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 6, no. 6, pp. 440–475, 2014.

[163] H. Martens, "Quantitative Big Data: where chemometrics can contribute," *Journal of Chemometrics*, vol. 29, no. 11, pp. 563–581, 2015.

[164] P. Bassan, A. Kohler, H. Martens, J. Lee, H. J. Byrne, P. Dumas, E. Gazi, M. Brown, N. Clarke, and P. Gardner, "Resonant Mie Scattering (RMieS) correction of infrared spectra from highly scattering biological samples," *Analyst*, vol. 135, pp. 268–277, 2010.

[165] J. E. Bertie, "Optical Constants," in *Handbook of Vibrational Spectroscopy*, 1st ed., J. M. Chalmers and P. R. Griffiths, Eds. John Wiley & Sons, Ltd, 2002.

[166] P. Bassan, A. Kohler, H. Martens, J. Lee, E. Jackson, N. Lockyer, P. Dumas, M. Brown, N. Clarke, and P. Gardner, "RMieS-EMSC correction for infrared spectra of biological cells: Extension using full Mie theory and GPU computing," *Journal of Biophotonics*, vol. 3, no. 8-9, pp. 609–620, 2010.

[167] M. J. Baker, R. Goodacre, C. Sammon, M. P. Marques, P. Gardner, W. Tipping, J. Sule-Suso, B. Wood, H. J. Byrne, M. Hermes, P. Matousek, C. J. Campbell, S. El-Mashtoly, J. Frost, C. Phillips, M. Diem, A. Kohler, K. Lau, S. Kazarian, W. Petrich, G. Lloyd, I. Delfino, G. Cinque, M. Isabelle, N. Stone, C. Kendall, L. Jamieson, D. Perez-Guaita, L. Clark, K. Gerwert, I. Notingher, L. Quaroni, R. Bhargava, A. Meade, and F. Lyng, "Single cell analysis/data handling: general discussion," *Faraday Discuss.*, vol. 187, pp. 299–327, 2016.

[168] P. Bassan, A. Sachdeva, A. Kohler, C. Hughes, A. Henderson, J. Boyle, J. H. Shanks, M. Brown, N. W. Clarke, and P. Gardner, "FTIR microscopy of biological cells and tissue: data analysis using resonant Mie scattering (RMieS) EMSC algorithm," *Analyst*, vol. 137, pp. 1370–1377, 2012.

[169] T. van Dijk, D. Mayerich, P. S. Carney, and R. Bhargava, "Recovery of Absorption Spectra from Fourier Transform Infrared (FT-IR) Microspectroscopic Measurements of Intact Spheres," *Applied Spectroscopy*, vol. 67, no. 5, pp. 546–552, 2013.

[170] R. Lukacs, R. Blumel, B. Zimmerman, M. Bagcioglu, and A. Kohler, "Recovery of absorbance spectra of micrometer-sized biological and inanimate particles," *Analyst*, vol. 140, pp. 3273–3284, 2015.

[171] T. Konevskikh, R. Lukacs, R. Blumel, A. Ponossov, and A. Kohler, "Mie scatter corrections in single cell infrared microspectroscopy," *Faraday Discuss.*, vol. 187, pp. 235–257, 2016.

[172] T. Konevskikh, R. Lukacs, and A. Kohler, "An improved algorithm for fast resonant Mie scatter correction of infrared spectra of cells and tissues," *Journal of Biophotonics*, pp. e201 600 307–n/a, 2017.

[173] A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[174] M. J. Adams, *Chemometrics in Analytical Spectroscopy*, 2nd ed., ser. RSC Analytical Spectroscopy Series, N. W. Barnett, Ed. The Royal Society of Chemistry, 2004.

[175] R. Bhargava, "Towards a practical Fourier transform infrared chemical imaging protocol for cancer histopathology," *Analytical and Bioanalytical Chemistry*, vol. 389, no. 4, pp. 1155–1169, 2007.

[176] E. J. (Swain) Marcsisin, C. M. Uttero, A. I. Mazur, M. Miljkovic, B. Bird, and M. Diem, "Noise Adjusted Principal Component reconstruction to optimize infrared microspectroscopy of individual live cells," *Analyst*, vol. 137, pp. 2958–2964, 2012.

[177] J. Trevisan, P. P. Angelov, P. L. Carmichael, A. D. Scott, and F. L. Martin, "Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives," *Analyst*, vol. 137, pp. 3202–3215, 2012.

[178] J. Coates, "Classical Methods of Quantitative Analysis," in *Handbook of Vibrational Spectroscopy*, 1st ed., J. M. Chalmers and P. R. Griffiths, Eds. John Wiley & Sons, Ltd, 2002.

[179] M. Zeaiter, J.-M. Roger, and V. Bellon-Maurel, "Robustness of models developed by multivariate calibration. Part II: The influence of pre-processing methods," *TrAC Trends in Analytical Chemistry*, vol. 24, no. 5, pp. 437 – 445, 2005.

[180] R. J. Barnes, M. S. Dhanoa, and S. J. Lister, "Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra," *Appl. Spectrosc.*, vol. 43, no. 5, pp. 772–777, 1989.

[181] R. G. Brereton, *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*. John Wiley & Sons, Ltd, 2003.

[182] K. Varmuza and P. Filzmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics*. Boca Raton, FL: CRC Press, 2009.

[183] R. Brereton, *Chemometrics for pattern recognition*. Wiley, 2009.

[184] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. Boston: Academic Press, 1990.

[185] C. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc, 2006.

[186] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer-Verlag New York, Inc, 2009.

[187] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," ser. The Morgan Kaufmann Series in Data Management Systems. Boston: Morgan Kaufmann, 2012.

[188] S. Wold, K. Esbensen, and P. Geladi, "Principal Component Analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37 – 52, 1987.

[189] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer-Verlag New York, Inc, 2002.

[190] R. Bro and A. K. Smilde, "Principal component analysis," *Analytical Methods*, vol. 6, no. 9, pp. 2812–2831, 2014.

[191] ——, "Centering and scaling in component analysis," *Journal of Chemometrics*, vol. 17, no. 1, pp. 16–33, 2003.

[192] P. C. Gillette and J. L. Koenig, "Noise Reduction via Factor Analysis in FT-IR Spectra," *Applied Spectroscopy*, vol. 36, no. 5, pp. 535–539, 1982.

[193] P. C. Mahalanobis, "On the generalised distance in statistics," *Proceedings of the National Institute of Sciences of India*, vol. 2, pp. 49–55, 1936.

[194] R. G. Brereton and G. R. Lloyd, "Re-evaluating the role of the mahalanobis distance measure," *Journal of Chemometrics*, vol. 30, no. 4, pp. 134–143, 2016.

[195] O. M. Kvalheim, "Introduction to Factor-based Approaches," in *Handbook of Vibrational Spectroscopy*, 1st ed., J. M. Chalmers and P. R. Griffiths, Eds. John Wiley & Sons, Ltd, 2002.

[196] S. Wold, M. Sjstrm, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109 – 130, 2001.

[197] M. Barker and W. Rayens, "Partial least squares for discrimination," *Journal of Chemometrics*, vol. 17, no. 3, pp. 166–173, 2003.

[198] J. Pijanka, G. D. Sockalingum, A. Kohler, Y. Yang, F. Draux, G. Parkes, K.-P. Lam, D. Collins, P. Dumas, C. Sandt, D. G. van Pittius, G. Douce, M. Manfait, V. Untereiner, and J. Sule-Suso, "Synchrotron-based FTIR spectra of stained single cells. Towards a clinical application in pathology," *Laboratory Investigation*, vol. 90, no. 5, pp. 797–807, May 2010.

[199] L. G. Brown, "A survey of image registration techniques," *ACM computing surveys*, vol. 24, no. 4, pp. 325–376, Dec. 1992.

[200] J. Maintz and M. A. Viergever, "A survey of medical image registration," *Medical Image Analysis*, vol. 2, no. 1, pp. 1–36, 1998.

[201] B. Zitová and J. Flusser, "Image registration methods: a survey," *Image and Vision Computing*, vol. 21, no. 11, pp. 977 – 1000, 2003.

[202] J. Modersitzki, *Numerical methods for image registration*. Oxford university press, 2004.

[203] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, pp. 1153–1190, July 2013.

[204] A. Sotiras, Y. Ou, N. Paragios, and C. Davatzikos, "Graph-based deformable image registration," in *Handbook of Biomedical Imaging: Methodologies and Clinical Research*, N. Paragios, J. Duncan, and N. Ayache, Eds. Boston, MA: Springer US, 2015, pp. 331–359.

[205] J. T. Kwak, S. M. Hewitt, S. Sinha, and R. Bhargava, "Multimodal microscopy for automated histologic analysis of prostate cancer," *BMC Cancer*, vol. 11, no. 1, pp. 1–16, 2011.

[206] C. Yang, D. Niedieker, F. Großerüschkamp, M. Horn, A. Tannapfel, A. Kallenbach-Thieltges, K. Gerwert, and A. Mosig, "Fully automated registration of vibrational microspectroscopic images in histologically stained tissue sections," *BMC Bioinf.*, vol. 16, no. 1, pp. 1–14, 2015.

[207] G. Lippolis *et al.*, "Automatic registration of multi-modal microscopy images for integrative analysis of prostate tissue sections," *BMC cancer*, vol. 13, no. 1, p. 408, 2013.

[208] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, ser. ICCV '99, vol. 2. Washington, DC, USA: IEEE Computer Society, 1999, pp. 1150–1157.

[209] ——, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[210] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.

[211] J. M. R. S. Tavares, "Analysis of biomedical images based on automated methods of image registration," in *Advances in Visual Computing: 10th International Symposium, ISVC 2014, Proceedings, Part I.* Springer International Publishing, 2014, pp. 21–30.

[212] A. P. Keszei, B. Berkels, and T. M. Deserno, "Survey of non-rigid registration tools in medicine," *Journal of Digital Imaging*, vol. 30, no. 1, pp. 102–116, 2017.

[213] R. M. Rangayyan, B. Acha, and C. Serrano, *Color Image Processing With Biomedical Applications (SPIE Press Monograph Vol. PM206)*. SPIE Press, 2011.

[214] J. Trevisan, P. P. Angelov, A. D. Scott, P. L. Carmichael, and F. L. Martin, "IRootLab: a free and open-source MATLAB toolbox for vibrational biospectroscopy data analysis," *Bioinformatics*, vol. 29, no. 8, pp. 1095–1097, 2013.

[215] D. Mayerich, M. Walsh, M. Schulmerich, and R. Bhargava, "Real-time interactive data mining for chemical imaging information: application to automated histopathology," *BMC Bioinformatics*, vol. 14, no. 1, pp. 1–11, 2013.

[216] W. R. Crum, T. Hartkens, and D. L. G. Hill, "Non-rigid image registration: theory and practice," *The British Journal of Radiology*, vol. 77, no. suppl_2, pp. S140–S153, dec 2004.

[217] A. Roche *et al.*, "The correlation ratio as a new similarity measure for multimodal image registration," in *Medical Image Computing and Computer-Assisted Interventation, MICCAI98*. Springer, 1998, pp. 1115–1124.

[218] A. Roche, G. Malandain, N. Ayache, and X. Pennec, "Multimodal image registration by maximization of the correlation ratio," INRIA, Tech. Rep. 3378, 1998.

[219] F. Maes, D. Vandermeulen, and P. Suetens, "Medical image registration using mutual information," *Proceedings of the IEEE*, vol. 91, no. 10, pp. 1699–1722, Oct 2003.

[220] J. Larrey-Ruiz, R. Verdú-Monedero, and J. Morales-Sánchez, "A Fourier domain framework for variational image registration," *Journal of Mathematical Imaging and Vision*, vol. 32, no. 1, pp. 57–72, 2008.

[221] R. Verdu-Monedero, J. Larrey-Ruiz, and J. Morales-Sanchez, "Frequency implementation of the Euler-Lagrange equations for variational image registration," *IEEE Signal Processing Letters*, vol. 15, pp. 321–324, 2008.

[222] A.-G. Legaz-Aparicio, R. Verdú-Monedero, J. Larrey-Ruiz, J. Morales-Sánchez, F. López-Mir, V. Naranjo, and Á. Bernabéu, "Efficient variational approach

to multimodal registration of anatomical and functional intra-patient tumorous brain data," *International Journal of Neural Systems*, p. 1750014, 2017.

[223] B. Fischer and J. Modersitzki, "A unified approach to fast image registration and a new curvature based registration technique," *Linear Algebra and its Applications*, vol. 380, pp. 107 – 124, 2004.

[224] P. H. Schönemann, "A generalized solution of the orthogonal Procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.

[225] D. L. G. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, "Medical image registration," *Physics in Medicine and Biology*, vol. 46, no. 3, p. R1, 2001.

[226] J. M. Fitzpatrick, J. B. West, and C. R. Maurer, "Predicting error in rigid-body point-based registration," *IEEE Transactions on Medical Imaging*, vol. 17, no. 5, pp. 694–702, Oct 1998.

[227] J. D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference, Fifth Edition (Statistics: Textbooks and Monographs)*. Chapman and Hall/CRC, 2010.

[228] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, "NIH image to ImageJ: 25 years of image analysis," *Nature Methods*, vol. 9, no. 7, pp. 671–675, jun 2012.

[229] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona, "Fiji: an open-source platform for biological-image analysis," *Nature Methods*, vol. 9, no. 7, pp. 676–682, jun 2012.

[230] D. Sage, D. Prodanov, J.-Y. Tinevez, and J. Schindelin, "MIJ: making interoperability between ImageJ and Matlab possible," in *ImageJ User & Developer Conference*, oct 2012, pp. 24–26.

[231] M. Krzywinski and N. Altman, "Points of significance: Visualizing samples with box plots," *Nature Methods*, vol. 11, no. 2, pp. 119–120, jan 2014.

[232] ——, "Points of significance: Comparing samples—part I," *Nature Methods*, vol. 11, no. 3, pp. 215–216, feb 2014.

[233] M. A. Cohenford and B. Rigas, "Cytologically normal cells from neoplastic cervical samples display extensive structural abnormalities on IR spectroscopy: Implications for tumor biology," *Proceedings of the National Academy of Sciences*, vol. 95, no. 26, pp. 15 327–15 332, 1998.

[234] J. M. Schubert, A. I. Mazur, B. Bird, M. Miljković, and M. Diem, "Single point vs. mapping approach for spectral cytopathology (SCP)," *Journal of Biophotonics*, vol. 3, no. 8-9, pp. 588–596, 2010.

[235] J. Doherty, G. Cinque, and P. Gardner, "Single-cell analysis using Fourier transform infrared microspectroscopy," *Applied Spectroscopy Reviews*, vol. 52, no. 6, pp. 560–587, 2017.

[236] M. Diem, M. Miljković, B. Bird, T. Chernenko, J. Schubert, E. Marcsisin, A. Mazur, E. Kingston, E. Zuser, K. Papamarkakis *et al.*, "Applications of infrared and Raman microspectroscopy of cells and tissue in medical diagnostics: Present status and future promises," *Journal of Spectroscopy*, vol. 27, no. 5-6, pp. 463–496, 2012.

[237] M. Miljković, B. Bird, K. Lenau, A. I. Mazur, and M. Diem, "Spectral cytopathology: new aspects of data collection, manipulation and confounding effects," *Analyst*, vol. 138, no. 14, pp. 3975–3982, 2013.

[238] M. Diem, "Infrared microspectroscopy of cells and tissue in medical diagnostics," in *Modern Vibrational Spectroscopy and Micro-Spectroscopy.* Wiley Online Library, 2015, ch. 13, pp. 283–338.

[239] B. Bird, M. J. Romeo, M. Diem, K. Bedrossian, N. Laver, and S. Naber, "Cytology by infrared micro-spectroscopy: Automatic distinction of cell types in urinary cytology," *Vibrational Spectroscopy*, vol. 48, no. 1, pp. 101–106, 2008.

[240] J. M. Schubert, B. Bird, K. Papamarkakis, M. Miljković, K. Bedrossian, N. Laver, and M. Diem, "Spectral cytopathology of cervical samples: detecting cellular abnormalities in cytologically normal cells," *Laboratory investigation*, vol. 90, no. 7, pp. 1068–1077, 2010.

[241] K. Papamarkakis, B. Bird, J. M. Schubert, M. Miljković, R. Wein, K. Bedrossian, N. Laver, and M. Diem, "Cytopathology by optical methods:

spectral cytopathology of the oral mucosa," *Laboratory Investigation*, vol. 90, no. 4, pp. 589–598, 2010.

[242] D. Townsend, M. Miljković, B. Bird, K. Lenau, O. Old, M. Almond, C. Kendall, G. Lloyd, N. Shepherd, H. Barr *et al.*, "Infrared micro-spectroscopy for cytopathological classification of esophageal cells," *Analyst*, vol. 140, no. 7, pp. 2215–2223, 2015.

[243] M. Diem, M. Miljković, B. Bird, A. I. Mazur, J. M. Schubert, D. Townsend, N. Laver, M. Almond, and O. Old, "Cancer screening via infrared spectral cytopathology (scp): results for the upper respiratory and digestive tracts," *Analyst*, vol. 141, no. 2, pp. 416–428, 2016.

[244] CLS Cell Lines Service GmbH, http://www.clsgmbh.de/, 2017, [Last access: 21/10/2017].

[245] DSMZ, "Leibniz Institute-German Collection of Microorganisms and Cell Cultures," https://www.dsmz.de/, 2017, [Last access: 21/10/2017].

[246] CRYSTAL GmbH, http://www.crystal-gmbh.com/, 2017, [Last access: 21/10/2017].

[247] L. Kastl, C. E. Rommel, B. Kemper, and J. Schnekenburger, "Standardized cell samples for midIR technology development," in *Proc. SPIE*, vol. 9315, 2015, pp. 931 507–1–931 507–6.

[248] L. Kastl, B. Kemper, G. R. Lloyd, J. Nallala, N. Stone, V. Naranjo, F. Peñaranda, and J. Schnekenburger, "Potential of mid IR spectroscopy in the rapid label free identification of skin malignancies," in *Proc. SPIE*, vol. 9703, 2016, pp. 970 307–1–970 307–8.

[249] ——, "Performance of mid infrared spectroscopy in skin cancer cell type identification," in *Proc. SPIE*, vol. 10060, 2017, pp. 1 006 006–1–1 006 006–10.

[250] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, Jan 1979.

[251] C. Hughes, M. D. Brown, F. J. Ball, G. Monjardez, N. W. Clarke, K. R. Flower, and P. Gardner, "Highlighting a need to distinguish cell cycle signatures from cellular responses to chemotherapeutics in SR-FTIR spectroscopy," *Analyst*, vol. 137, pp. 5736–5742, 2012.

[252] M. Jimenez-Hernandez, C. Hughes, P. Bassan, F. Ball, M. D. Brown, N. W. Clarke, and P. Gardner, "Exploring the spectroscopic differences of Caki-2 cells progressing through the cell cycle while proliferating in vitro," *Analyst*, vol. 138, pp. 3957–3966, 2013.

[253] D. R. Whelan, K. R. Bambery, L. Puskar, D. McNaughton, and B. R. Wood, "Synchrotron Fourier transform infrared (FTIR) analysis of single living cells progressing through the cell cycle," *Analyst*, vol. 138, pp. 3891–3899, 2013.

[254] T. Blaschke, M. Kelly, and H. Merschdorf, "Object-Based Image Analysis: Evolution, History, State of the Art, and Future Vision," in *Remotely Sensed Data Characterization, Classification, and Accuracies*, ser. Remote Sensing Handbook. CRC Press, 2015, pp. 277–294.

[255] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*. Pearson, 2007.

[256] J. C. Russ, *The Image processing handbook*, 7th ed. CRC Press LLC, 2016.

[257] J. Filik, A. V. Rutter, J. Sule-Suso, and G. Cinque, "Morphological analysis of vibrational hyperspectral imaging data," *Analyst*, vol. 137, pp. 5723–5729, 2012.

[258] F. Meyer and S. Beucher, "Morphological segmentation," *Journal of Visual Communication and Image Representation*, vol. 1, no. 1, pp. 21–46, 1990.

[259] P. Soille, *Morphological Image Analysis*. Springer Nature, 2004.

[260] E. Meijering, "Cell segmentation: 50 years down the road [life sciences]," *IEEE Signal Processing Magazine*, vol. 29, no. 5, pp. 140–145, Sept 2012.

[261] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu, "Methods for nuclei detection, segmentation, and classification in digital histopathology: A review-current status and future potential," *IEEE Reviews in Biomedical Engineering*, vol. 7, pp. 97–114, 2014.

[262] S. Beucher and C. Lantuejoul, "Use of Watersheds in Contour Detection," in *International Workshop on Image Processing: Real-time Edge and Motion Detection/Estimation, Rennes, France.*, Sep. 1979.

[263] F. Meyer, "Un algorithme optimal de ligne de partage des eaux," *Actes du*, vol. 2, pp. 847–859, 1991.

[264] J. Serra, "Courses on Mathematical Morphology," Center of Mathematical Morphology, http://cmm.ensmp.fr/~serra/cours/, 2000, [Last access: 25/07/2017].

[265] S. Beucher, "Image Segmentation and Mathematical Morphology," Center of Mathematical Morphology, http://cmm.ensmp.fr/~beucher/wtshed.html, 2010, [Last access: 25/07/2017].

[266] P. Mather and B. Tso, *Classification Methods for Remotely Sensed Data*, 2nd ed. CRC Press, 2009.

[267] J. Li and A. Plaza, "Hyperspectral Image Processing: Methods and Approaches," in *Remotely Sensed Data Characterization, Classification, and Accuracies*, ser. Remote Sensing Handbook. CRC Press, 2015, pp. 247–258.

[268] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 55–63, 1968.

[269] R. G. Brereton and G. R. Lloyd, "Partial least squares discriminant analysis: taking the magic away," *Journal of Chemometrics*, vol. 28, no. 4, pp. 213–225, 2014, cEM-13-0209.R1.

[270] R. G. Congalton and K. Green, *Assessing the Accuracy of Remotely Sensed Data*, 2nd ed. CRC Press, 2008.

[271] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[272] S. Wang and X. Yao, "Multiclass Imbalance Problems: Analysis and Potential Solutions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 1119–1130, 2012.

[273] T. Fearn, "Discriminant Analysis," in *Handbook of Vibrational Spectroscopy*, 1st ed., J. M. Chalmers and P. R. Griffiths, Eds. John Wiley & Sons, Ltd, 2002.