Novel chemometric approaches towards handling biospectroscopy datasets

by

Camilo de Lelis Medeiros de Morais

A thesis submitted in partial fulfilment for the requirements for the degree of Doctor of Philosophy at the University of Central Lancashire

June 2020

STUDENT DECLARATION FORM

Type of Award Doctor of Philosophy

School School of Pharmacy and Biomedical Sciences

1. Concurrent registration for two or more academic awards

I declare that while registered as a candidate for the research degree, I have not been a registered candidate or enrolled student for another award of the University or other academic or professional institution

2. Material submitted for another award

I declare that no material contained in the thesis has been used in any other submission for an academic award and is solely my own work

3. Collaboration

Where a candidate's research programme is part of a collaborative project, the thesis must indicate in addition clearly the candidate's individual contribution and the extent of the collaboration. Please state below:

The samples used in this project were provided by the Brain Tumour North West (BTNW) biobank. Experimental and clinical support were provided by the Neurosurgery and Neuropathology departments of the Lancashire Teaching Hospitals NHS Trust, Royal Preston Hospital. The PhD candidate was responsible for spectroscopy measurements, algorithm development, data analysis, interpretation of results and writing of scientific papers.

---

4. Use of a Proof-reader

No proof-reading service was used in the compilation of this thesis.

Signature of Candidate 

Print name: CAMILO DE LELIS MEDEIROS DE MORAIS

Abstract

Background

Chemometrics allows one to identify chemical patterns using spectrochemical information of biological materials, such as tissues and biofluids. This has fundamental importance to overcome limitations in traditional bioanalytical analysis, such as the need for laborious and extreme invasive procedures, high consumption of reagents, and expensive instrumentation. In biospectroscopy, a beam of light, usually in the infrared region, is projected onto the surface of a biological sample and, as a result, a chemical signature is generated containing the vibrational information of most of the molecules in that material. This can be performed in a single-spectra or hyperspectral imaging fashion, where a resultant spectrum is generated for each position (pixel) in the surface of a biological material segment, hence, allowing extraction of both spatial and spectrochemical information simultaneously. As an advantage, these methodologies are non-destructive, have a relatively low-cost, and require minimum sample preparation. However, in biospectroscopy, large datasets containing complex spectrochemical signatures are generated. These datasets are processed by computational tools in order to solve their signal complexity and then provide useful information that can be used for decision taking, such as the identification of clustering patterns distinguishing disease from healthy controls samples; differentiation of tumour grades; prediction of unknown samples categories; or identification of key molecular fragments (biomarkers) associated with the appearance of certain diseases, such as cancer. In this PhD thesis, new computational tools are developed in order to improve the processing of bio-spectrochemical data, providing better clinical outcomes for both spectral and hyperspectral datasets.

Materials and Methods

Sample splitting. A new sampling methodology, called the Morais-Lima-Martin (MLM) algorithm, was developed to improve data splitting for classification applications. The MLM algorithm was developed by modifying the Kennard-Stone (KS) sample selection method with the addition of a 10% random-mutation factor to the sample splitting methodology. This methodology was tested in one simulated and six real-world datasets (4 for infrared (IR) spectroscopy and 2 for Raman spectroscopy) aiming to maximize sample discrimination using principal component analysis linear discriminant analysis (PCA-LDA). The results were compared with other two data splitting approaches: the random-selection (RS) and the KS algorithm alone.

Hyperspectral imaging. Novel classification methods based on principal component analysis quadratic discriminant analysis (PCA-QDA), successive projections algorithm quadratic discriminant analysis (SPA-QDA), multivariate curve resolution alternating least squares (MCR-ALS), three-dimensional principal component analysis (3D-PCA) with linear discriminant analysis (3D-PCA-LDA) and quadratic discriminant analysis (3D-PCA-QDA) were developed to improve sample discrimination based on hyperspectral imaging datasets. PCA-QDA, SPA-QDA and MCR-ALS were applied to distinguish meningioma WHO Grade I (n = 66) and Grade II (n = 24) tissue samples based on Raman microspectroscopy imaging. 3D-PCA was applied to distinguish 5 ovarian cancers from 5 healthy controls samples using Raman hyperspectral images of blood plasma in an unsupervised approach, and 3D-PCA-LDA and 3D-PCA-QDA were

applied to discriminate a further dataset of 38 samples (20 benign controls and 18 ovarian cancer samples) based on Raman hyperspectral images of blood plasma.

Uncertainty estimation. A new method to calculate classification uncertainty for linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and support vector machines (SVM) was developed based on bootstrap. This methodology was tested for 4 datasets (one simulated and three real-world applications of IR spectroscopy) in order to estimate misclassification probability and evaluate classification robustness.

Data standardisation. A dataset of 10 blood plasma samples (5 from healthy controls, 5 from ovarian cancer patients) was measured under different experimental conditions: by three different attenuated total reflection Fourier-transform infrared (ATR-FTIR) instruments, operated by two different operators, using two different sets of co-additions scans and spectral resolution, and varying room temperature and air humidity in order to evaluate whether changes in instrumental and environmental conditions would alter the spectral response and modify sample discrimination. A novel protocol based on direct standardisation (DS) and piecewise direct standardisation (PDS) was developed in order to standardise spectral differences caused by instrumental and environmental changes.

Results

Sample splitting is an essential step for building classification models based on spectral data. The proposed MLM algorithm performed better than the KS and RS methods in terms of sensitivity and specificity. RS showed the poorest predictive response, followed by KS which showed good accuracy towards prediction, but relatively unbalanced sensitivities and specificities. The sensitivities and specificities obtained using MLM were more similar to each other, indicating a more reliable classification. MLM classification accuracies ranged from ~80–99% varying the dataset.

Meningioma Grade I and Grade II tissue samples were discriminated with 96.2% accuracy, 85.7% sensitivity and 100% specificity using PCA-QDA and SPA-QDA. The PCA loadings, SPA-QDA selected wavenumbers, and the recovered imaging profiles after MCR-ALS indicated the following wavenumbers responsible for class differentiation: 850 cm^{-1} (amino acids or polysaccharides), 1130 cm^{-1} (phospholipid structural changes), 1230–1360 cm^{-1} (Amide III and CH_2 deformation), 1450 cm^{-1} (CH_2 bending), and 1858 cm^{-1} (C=O stretching). For the ovarian cancer datasets, 3D-PCA was able to clearly distinguish healthy controls from ovarian cancer samples using the complete 3D hyperspectral images in the range from ~780–1858 cm^{-1}, and the new 3D-PCA-LDA and 3D-PCA-QDA algorithms were able to discriminate healthy controls from ovarian cancer samples with 100% accuracy in comparison with standard classification methodologies which provided 64% accuracy.

When applying the uncertainty estimation method to 4 different spectrochemical datasets, classification models with lower misclassification probabilities (m_p) were substantially more stable when the spectra were perturbed with white Gaussian noise. The m_p is a quantitative metric of uncertainty that ranges from 0 to 1, where m_p closer to 0 is an indicator of better classification robustness.

Experimental conditions greatly affected the spectral profiles in the standardisation study. DS and PDS were able to correct for instrumental-related changes improving the classification accuracy from 66.7% to 77.8% (DS) and 74.1% (PDS), and for operator-related changes improving the accuracy from 75.6% to 82.2% (DS) and 77.8% (PDS).

Discussion

Sample splitting is a process performed after spectral pre-processing and prior model construction and consists of dividing the experimental sample set into at least two subsets called training and test, where the training set is used for model construction and the test set for model validation. The MLM algorithm combines the good spectral representativeness in the test set provided by the KS algorithm with a small degree of randomness that may be found in biological applications. MLM generated a better predictive performance in comparison with standard methodologies, providing more well-equilibrated sensitivity and specificity results.

Classification of hyperspectral images is a process that requires large computational-cost, once the data are big in size and contain complex overlapping spectral features that require advanced chemometric tools for their analysis and interpretation. Herein, meningioma tissues were discriminated based on Raman microspectroscopy images where the classification results found by PCA-QDA and SPA-QDA are very promising, showing the potential of this methodology for aiding clinicians to delineate patient treatment. 3D-PCA was applied for ovarian cancer detection using Raman hyperspectral imaging, generating scores showing clear differences between the two classes on both principal components (PCs) 1 and 2; and the loadings profiles on these components indicate that the main biomarkers contributing for class differentiation are amino acids, lipids and DNA. 3D-PCA provided fast exploratory analysis for hyperspectral data, having potential for future applications in other types of spectrochemical imaging data. The new 3D discriminant analysis approaches (3D-PCA-LDA and 3D-PCA-QDA) provided fast class differentiation for multi-image hyperspectral datasets with superior discriminating performance compared to algorithms using unfolding procedures, which are often employed for this type of data.

Misclassification probability can be used as a new metric to assess classification quality since it contains information of the model uncertainty and is also associated with model robustness. This methodology was validated against propagation coefficients for SVMs and the results between our proposed methodology based on bootstrap and the established method were found at $R^2 = 0.971$, indicating agreement between the two methods.

Finally, we have constructed a protocol for model standardisation using DS and PDS transfer technologies described for FTIR spectrochemical applications. This is a critical step toward the construction of a practical spectrochemical analysis model for daily routine analysis, where uncertain and random variations are present in the data.

Conclusion

This thesis is focused on developing computational tools that will improve sample splitting; tools for exploratory analysis and classification of hyperspectral images; uncertainty estimation to evaluate the robustness of biospectral classification models; and development of a chemometrics and standardisation protocol for handling spectral data acquired under different experimental conditions. The requirement for such techniques is demonstrated by the fact that applications of deep-learning algorithms of complex datasets are being increasingly recognized as critical for extracting important information from biospectroscopy datasets and visualizing them in a readily interpretable form. Hereby, we have provided new chemometric approaches for biospectroscopy where successful applications of these techniques will then allow trial biospectroscopy in clinical settings, where fast, reliable, and highly accurate diagnosis would be obtained.

Table of Contents

Acknowledgements

This PhD was fully funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Brazil (grant 88881.128982/2016-01), Ministry of Education (MEC), Brazil. I greatly thank CAPES for the financial support provided for the realisation of this PhD project.

I also would like to thank Prof. Francis L. Martin at the University of Central Lancashire (UCLan) for supervising me during this PhD project. His insights and support greatly contributed to this thesis. I am also grateful to Dr. Andrew Shaw at UCLan who provided supervisory support when needed.

I would like to thank Dr. Taha Lilo, Prof. Pierre Martin-Hirsch, Prof. Timothy Dawson, Prof. Nihal Gurusinghe, Prof. Charles Davis and Mrs. Katherine Ashton at the Royal Preston Hospital for providing clinical support for this study. I am grateful to the Brain Tumour North West (BTNW) biobank for providing samples for this project, and to all patients who direct or indirectly contributed to this study.

Finally, I would like to thank my family for all the support provided; and my friends in the UK and abroad for the good moments during this PhD.

List of Tables

List of Figures

CHAPTER 1 | INTRODUCTION TO BIOSPECTROSCOPY **Page**

CHAPTER 2 | MULTIVARIATE CLASSIFICATION TECHNIQUES FOR VIBRATIONAL SPECTROSCOPY IN BIOLOGICAL SAMPLES

CHAPTER 3 | IMPROVING DATA SPLITTING FOR CLASSIFICATION APPLICATIONS IN SPECTROCHEMICAL ANALYSES EMPLOYING A RANDOM-MUTATION KENNARD-STONE ALGORITHM APPROACH

CHAPTER 4 | A COMPUTATIONAL PROTOCOL FOR SAMPLE SELECTION IN BIOLOGICAL-DERIVED INFRARED SPECTROSCOPY DATASETS USING MORAIS-LIMA-MARTIN (MLM) ALGORITHM

CHAPTER 9 | STANDARDIZATION OF COMPLEX BIOLOGICALLY-DERIVED SPECTROCHEMICAL DATASETS

Abbreviations

3D-PCA	Three-dimensional principal component analysis
3D-PCA-LDA	Three-dimensional principal component analysis linear discriminant analysis
3D-PCA-QDA	Three-dimensional discriminant analysis quadratic discriminant analysis
3-MCA	3-methylcholanthrene
AC	Accuracy
ALS	Asymmetric least squares
ANN	Artificial neural networks
ANOVA	Analysis of variance
Ant	Anthracene
ATR	Attenuated total reflection
ATR-FTIR	Attenuated total reflection Fourier-transform infrared
AUC	Area under the curve
AWLS	Automatic weighted least squares
B[a]P	Benzo[a]pyrene
BNN	Back-propagation neural network
BTNW	Brain Tumour North West
CA-125	Serum cancer antigen
CAT	Chemometric Agile Tool
CCD	Charge-coupled device
CD	Circular dichroism
CLT	Central limit theorem
CSF	Cerebrospinal fluid
CTCCA	Calibration transfer based on canonical correlation analysis
CTMMC	Calibration transfer based on the maximum margin criterion
CV	Cross-validation
DA	Discriminant analysis
DBM	Difference-between-mean
DF	Discriminant function
DoE	Design of experiments
DS	Direct standardization
EEM	Excitation-emission matrix
EMSC	Extended multiplicative signal correction
EV	Explained variance
FACT	Free Access Chemometrics Toolbox
FFPE	Formalin-fixed paraffin-embedded
FIR	Far infrared
FN	False negative
FP	False positive
FSIW-EFA	Fixed size image window-evolving factor analysis
FT-IR	Fourier-transform infrared
FT-NIR	Fourier-transform near-infrared
GA	Genetic algorithm
GLSW	Generalized least squares weighting
GMR	Guided model reoptimization
ICA	Independent component analysis

IDH	Isocitrate dehydrogenase
iPLS	Interval partial least squares
IR	Infrared
KNN	K nearest neighbour
KS	Kennard-Stone
LC-MS	Liquid chromatography–mass spectrometry
LDA	Linear discriminant analysis
LR-	Negative likelihood ratio
LR+	Positive likelihood ratio
LV	Latent variable
LVQ	Learning vector quantization
MCR	Multivariate curve resolution
MCR-ALS	Multivariate curve resolution alternating least squares
MCT	Mercury cadmium telluride
MIA	Multivariate image analysis
MIR	Mid-infrared
MLM	Morais-Lima-Martin
MLPCA	Maximum likelihood PCA
MRMR	Minimum redundancy maximum relevance
MS	Mass spectrometry
MSC	Multiplicative scatter correction
MU	Model updating
NCLS	Non-negatively constrained least squares
NIPALS	Non-linear iterative partial least squares
NIR	Near infrared
NMR	Nuclear magnetic resonance
NPV	Negative predictive value
OPLS	Orthogonal projections to latent structures
OSC	Orthogonal signal correction
PARAFAC	Parallel factor analysis
PARAFAC-LDA	Parallel factor analysis linear discriminant analysis
PARAFAC-QDA	Parallel factor analysis quadratic discriminant analysis
PC	Principal component
PCA	Principal component analysis
PCA-LDA	Principal component analysis linear discriminant analysis
PCA-QDA	Principal component analysis quadratic discriminant analysis
PCA-SVM	Principal component analysis support vector machines
PDS	Piecewise direct standardization
PLDA	Piecewise linear discriminant analysis
PLS	Partial least squares
PLS-DA	Partial least squares discriminant analysis
PMF	Positive matrix factorization
PPV	Positive predictive value
QDA	Quadratic discriminant analysis
RBF	Radial basis function
RDA	Regularized discriminant analysis
RMieS-EMSC	Resonant Mie scattering-extended multiplicative signal correction.
RMSECV	Root mean square error of cross-validation
ROC	Receiver operating characteristic
RS	Random selection

S/N	Signal-to-noise
SD	Standard-deviation
SEIRA	Surface-enhanced IR absorption
SENS	Sensitivity
SERS	Surface-enhanced Raman spectroscopy
SG	Savitzky-Golay
SHE	Syrian hamster embryo
SIMCA	Soft independent modelling by class analogy
SNR	Signal-to-noise ratio
SNV	Standard normal variate
SPA	Successive projections algorithm
SPA-QDA	Successive projections algorithm-quadratic discriminant analysis
SPEC	Specificity
SVD	Singular value decomposition
SVM	Support vector machines
SW	Shenk and Westerhaus method
TEAM	Transfer *via* extreme learning machine auto-encoder method
TN	True negative
TP	True positive
TTWD-DA	Trilinear Three-way Data – Discriminant Analysis
Tucker3-LDA	Tucker3 linear discriminant analysis
Tucker3-QDA	Tucker3 quadratic discriminant analysis
US	Ultrasound
UV	Ultraviolet
UV-Vis	Ultraviolet-Visible
VIP	Variable importance in projections
WHDS	Wavelet hybrid direct standardization
WHO	World Health Organization
YOU	Youden’s index

CHAPTER 1 | INTRODUCTION TO BIOSPECTROSCOPY

5.1 Vibrational Spectroscopy

The use of spectroscopy techniques for quantitative or qualitative analysis of biochemical materials is well known (Baker *et al.*, 2014; Butler *et al.*, 2016; Santos *et al.*, 2017). Spectroscopic methods provide a chemical signature of the material being analysed according to the interaction between an electromagnetic radiation wave at a specific frequency and the molecules present in the material. In this phenomenon, the electrons that constitute the atoms and molecules in the material absorb photons with a specific light frequency, hence, being excited to higher energy states. This process is called excitation. Then, the electrons in a higher energy state release the total or part of the absorbed radiation returning to a lower energy state. This process is called emission. The excitation and emission processes can occur at different light frequencies, which determines the type of spectroscopy technique. The electromagnetic spectrum comprises a series of frequencies, from long radio waves (3 Hz) until gamma rays (>30 Ehz). Figure 1.1 illustrates the different types of electromagnetic radiation signals according to their frequency.

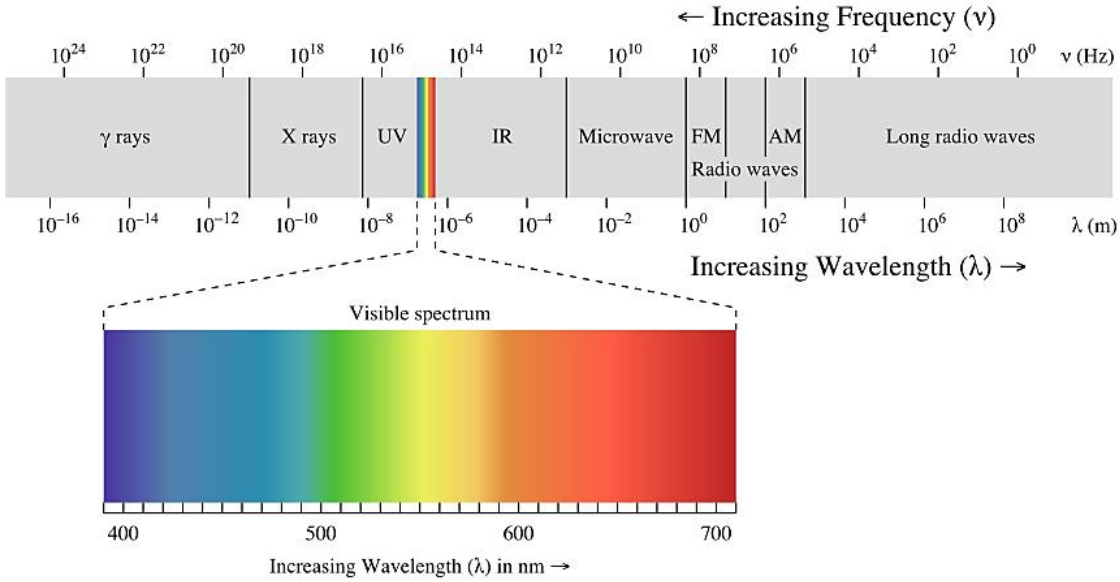


Figure 1.1. The electromagnetic spectrum labelled by frequency and wavelength regions. UV: ultra-violet; IR: infrared; FM: frequency modulation; AM: amplitude modulation. Inset: expanded visible spectrum. (CC BY-NC-SA, Chemistry Library < https://chem.libretexts.org/>).

The light frequency is proportional to the photon energy according to the Planck-Einstein relation (Logiurato, 2014):

$$E = h\nu \tag{1.1}$$

where E is the energy of the photon, h is the Planck's constant (6.626×10^{-34} J·s), and ν is the light frequency. ν can be replaced by the wavelength through the relationship between the wavelength (λ), frequency (ν) and the speed of light in vacuum (c, 2.998×10^{8} m/s):

$$c = \lambda\nu \tag{1.2}$$

$$E = \frac{hc}{\lambda} \tag{1.3}$$

Hence, the photon energy is inversely proportional to the wavelength. Vibrational spectroscopy encompasses techniques that excite molecules to absorb or emit energy in the infrared (IR) wavelength region between ~700 nm to 1 mm (frequency of 430 THz to 300 GHz, energy of 1.7 eV to 1.24 meV) (Skoog *et al.*, 2007). The spectrochemical signal is generated because all molecular bonds vibrate at any temperature above absolute zero (T > 0 K); and when a molecular bond is exposed to IR radiation with frequency equal to that fundamental vibration frequency, then it absorbs the radiation. IR radiation does not cause electronic transitions, only vibrational and rotational transitions. That is, the electron is not excited to a higher electronic energy levels but to a higher vibrational energy levels within the same electronic energy level (Figure 1.2).

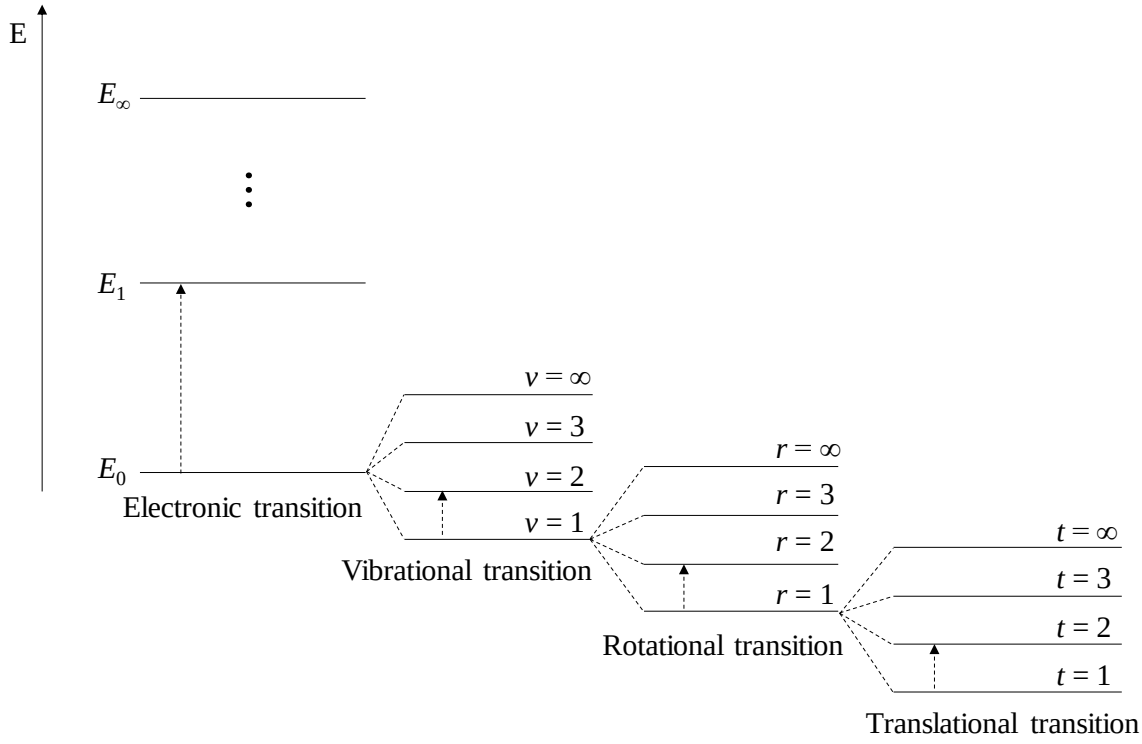


Figure 1.2. Energy diagram showing electronic, vibrational, rotational and translational transitions. E stands for total energy.

Each electronic energy level is composed of several vibrational energy levels which are composed of several rotational energy levels whose are composed of several translational energy levels. Electronic transitions are of higher energy frequency and are the basis of techniques such as ultraviolet-visible spectroscopy, while vibrational, rotational and translational transitions are of lower energy and associated with the techniques such as IR and Raman spectroscopy (vibrational transitions) and microwave spectroscopy (rotational transitions). Translational transitions are of a so small energy gap that any source of energy will cause this transition, hence, causing molecules to move.

There are two types of molecular vibration movements affected by IR radiation: stretching (axial deformation) and bending (angular deformation). Stretching vibrations can be either symmetric or asymmetric; while bending vibrations can be in-plane (symmetric or asymmetric) or out-of-plane (symmetric or asymmetric) (Nasdala *et al.*, 2004; Skoog *et al.*, 2007). These vibration motions are shown in Figure 1.3.

A molecule containing *N* atoms has 3*N* degrees of freedom, *i.e.*, every atom in the molecule has 3 coordinates (*x*-, *y*- and *z*- coordinate) in a tridimensional space. Also, every molecule has three types of movements: (1) translation: motion of the entire molecule

through the space; (2) rotation: rotational motion of the entire molecule around its center of gravity; and (3) vibration: the motion of each of its atoms relative to the other atoms. In vibrational spectroscopy, only vibration movements are measured, thus, translation and rotation movements are subtracted from the total number of degrees of freedom ($3N$). Therefore, the number of vibrational degrees of freedom for a polyatomic non-linear molecule is $3N - 6$, since there are 3 translation movements (along the x-, y- and z- axis) and 3 rotation movements (around the x-, y- and z- axis) for any non-linear molecule. For linear molecules, because the rotation along the bond axis is not valid, the number of vibrational degrees of freedom is $3N - 5$. Thus, by knowing the number of atoms in a molecule, it is possible to estimate how many vibrational signals will be shown ($3N - 6$ for non-linear molecules, and $3N - 5$ for linear molecules) (Skoog *et al.*, 2007). Further details about IR and Raman spectroscopy will be discussed hereafter.

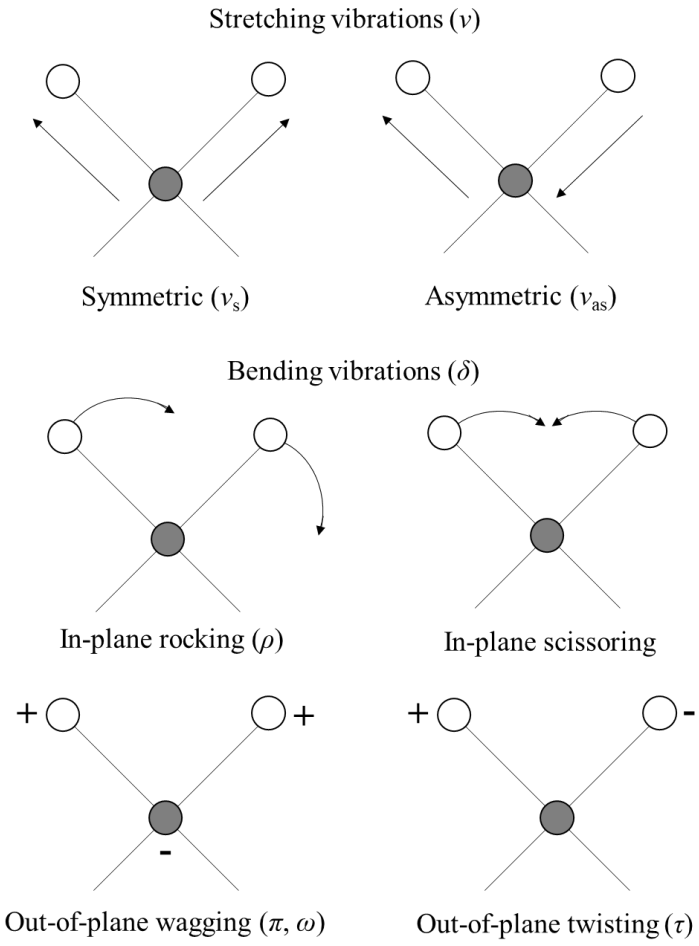


Figure 1.3. Type of molecular vibrations. + indicates motion from the page toward the reader and – indicates motion away from the reader.

5.1.2 Infrared Spectroscopy

The IR region is divided into three sub-regions: near infrared (NIR), 0.78 – 3 μm; mid-infrared (MIR) region, 3 – 50 μm; and, far-infrared (FIR) region, 50 – 1000 μm. To have a detectable IR signal, the molecule must have a variation of its resultant dipole moment different from zero ($\mu_R \neq 0$), therefore, homonuclear species such as O_2, N_2, Cl_2 and I_2 do not absorb IR radiation.

The interaction between the IR radiation and molecular bonds happens because any covalent bond while vibrating creates an electromagnetic field proportional to the variation of dipole moment. When a wave of IR radiation reaches this electromagnetic field generated with same oscillating frequency, the radiation is absorbed (constructive interfering). A bond vibration can be approximately explained by a mechanical model composed by two masses connected by a spring (harmonic oscillator model). Thus, a diatomic molecule can be compared to an ideal harmonic oscillator, defined by Hooke's law by (Skoog *et al.*, 2007):

$$F = -ky \tag{1.4}$$

where F is the force acting against the displacement y from its equilibrium position. The proportionality constant k is called the force constant. In an equilibrium position, the potential energy is zero; therefore, when the spring is compressed or stretched, the potential energy, E, will vary according to the work needed to move the mass:

$$\partial E = -F\partial y = ky\partial y$$

$$\int_0^E \partial E = k\int_0^y y\partial y$$

$$E = \frac{1}{2}ky^2 \tag{1.5}$$

The potential energy curve of a simple harmonic oscillator is a parabola (dashed line, Figure 4). However, real molecules do not follow an ideal model, therefore a more accurate potential energy curve for diatomic molecules resembles an anharmornic oscillator (continuous line, Figure 1.4). Each energy level in these oscillator curves represents a vibrational level of a molecule where the electrons are excited. These vibrational levels are within the electronic level (Figure 1.2). For NIR, the vibrational

transitions happen in regions of high energy, where the absorptions are called overtones or combination bands. This generates a very superposed signal of chemical features in comparison with MIR (fundamental vibration modes), though it can be used to asses relevant chemical information such as concentration. On the other hand, FIR works in low frequencies (< 650 cm^{-1}, > 15 μm), being particularly useful for studies involving bonds containing metallic atoms.

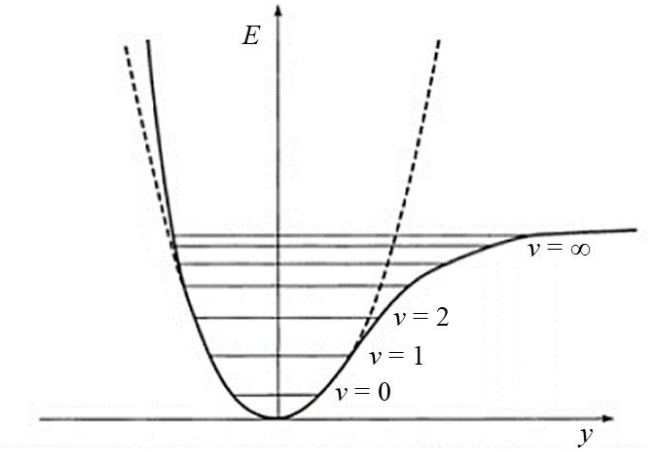


Figure 1.4. Harmonic (dashed line) and anharmonic (continuous line) oscillators model for infrared spectroscopy, where *E* represents the potential energy, *v* the vibrational energy level, and *y* the bond stretching distance.

MIR spectroscopy has been the main type of technique employed for analysing biological materials (Bajer *et al.*, 2014), since its signal contain fundamental vibration modes which are little convoluted in comparison with NIR and covers most of covalent bonds not detected by FIR. The main fundamental vibrational mode absorptions of functional groups within the MIR range are depicted in Figure 1.5 (Reusch, 1999). Most stretching vibrations occur in higher wavenumber region, since these require more energy; while bending vibrations tend to occur in lower wavenumber region. These signals shift towards higher or lower wavenumbers according to the molecular configuration and neighbouring functional groups.

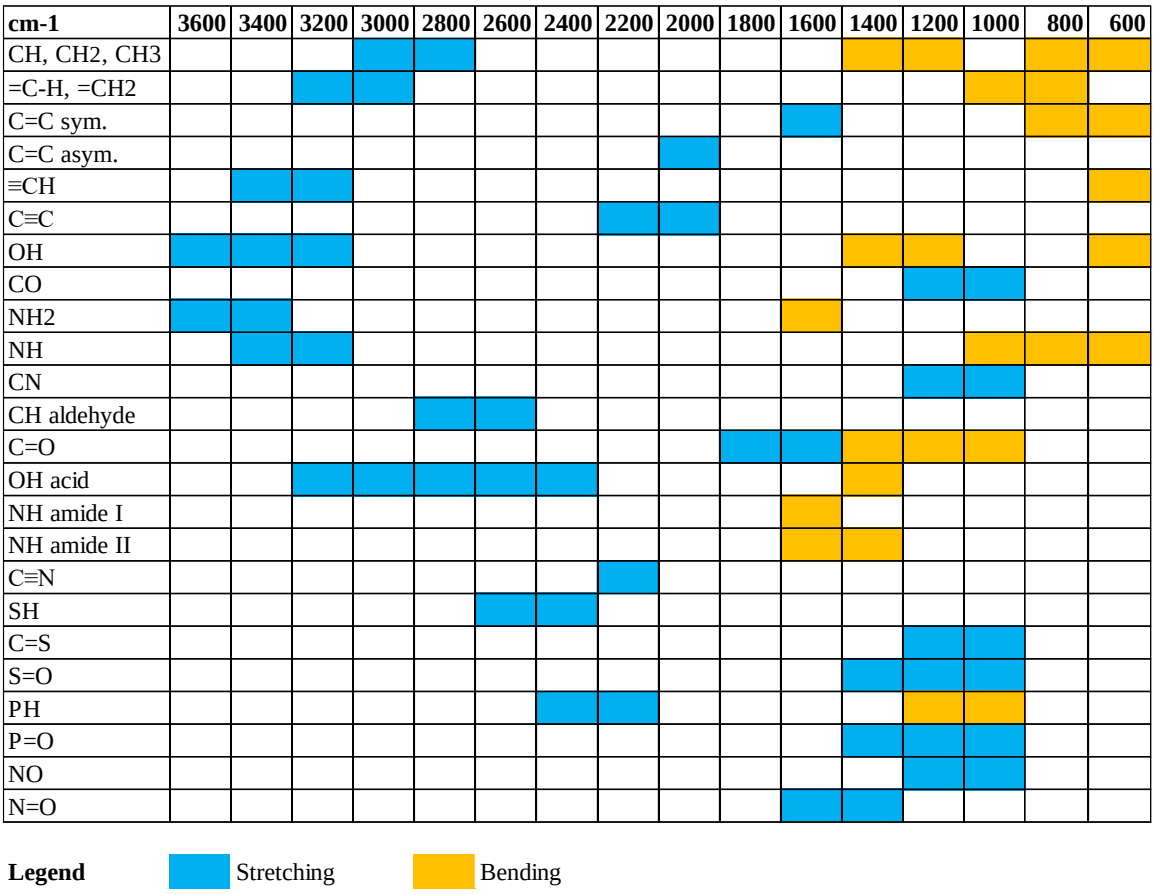


Figure 1.5. Approximation of the main vibration modes in IR spectroscopy (3600–600 cm^{-1}). Sym.: symmetric; asym.: asymmetric.

Experimental measurements using MIR spectroscopy are made using instruments composed basically of five parts: (1) light source, (2) spectrograph, (3) sampling area, (4) detector, and (5) computer module (Figure 1.6a). The light source is responsible for generating electromagnetic radiation at the infrared range; the spectrograph is an optical apparatus containing an interferometer, diffraction grating or prism for diffraction of the incident IR light; the sampling area contains the sample to be irradiated with IR light; the detector captures the diffracted IR light generating an electric potential response; and the computer module process the electric information transforming it into an interferogram and, by using a Fourier transform (FT), into a spectrum. The MIR spectrometer can work in two modes: transmission or reflection. In transmission mode (Figure 1.6b), the infrared light passes through the sample and reach the detector, so the final recorded IR signal is given by:

$$T = \frac{P}{P_0} \quad (1.6)$$

where T is the transmission (value between 0 and 1), P is the signal potential through the sample, and P_0 is the signal potential through the blank, *i.e.*, background signal (*e.g.*, reference material, air, sample matrix without substance of interest). Transmission does not have a linear relationship with concentration. For this reason, a mathematical transformation using the Beer-Lambert's law is applied:

$$A = -\log T \quad (1.7)$$

where A is the absorbance, which is linearly proportional to the chemical concentration, since:

$$A = \varepsilon bc \quad (1.8)$$

where ε is the molecular absorptivity coefficient, b is the length of the optical path, and c is the concentration.

In reflection mode, the IR light reaches the material surface and bounces in the material through a phenomenon called reflectance, where the incident light after a certain degree of depth penetration reflects back to the spectrometer in an angle close to 180º. The most popular technique for reflection is the attenuated total reflection (ATR) mode (Figure 1.6c), where a crystal, often diamond, is placed between the light source and sample to generate an evanescent wave that amplifies the signal intensity (Baker *et al.*, 2014). The main disadvantage of the transmission mode is that it is suitable only to analyse liquid in relative large volumes (> 2 mL) or thin materials; thus, reflectance by means of the ATR technique is more adequate to analyse solid, thick, or small volume of liquids. The penetration depth of ATR using a diamond crystal varies according to the incident radiation frequency and material properties, but it usually varies from 0.5 to 2 μm.

From an economical point of view, the use of MIR spectroscopy, in particular ATR-FTIR, has great advantages. This is because the instrumentation has a relative low-cost, requires a low-cost maintenance, does not require laborious or wet-chemistry sample preparation procedures, has a fast data acquisition, and is non-destructive, that is, the sample can be reused after analysis. Thus, applications of this technique in biochemical areas are of great importance as substitute or auxiliary methods to reference analysis usually employed using electrochemical, chromatographic, mass spectrometric, and thermogravimetric techniques, or combination of these.

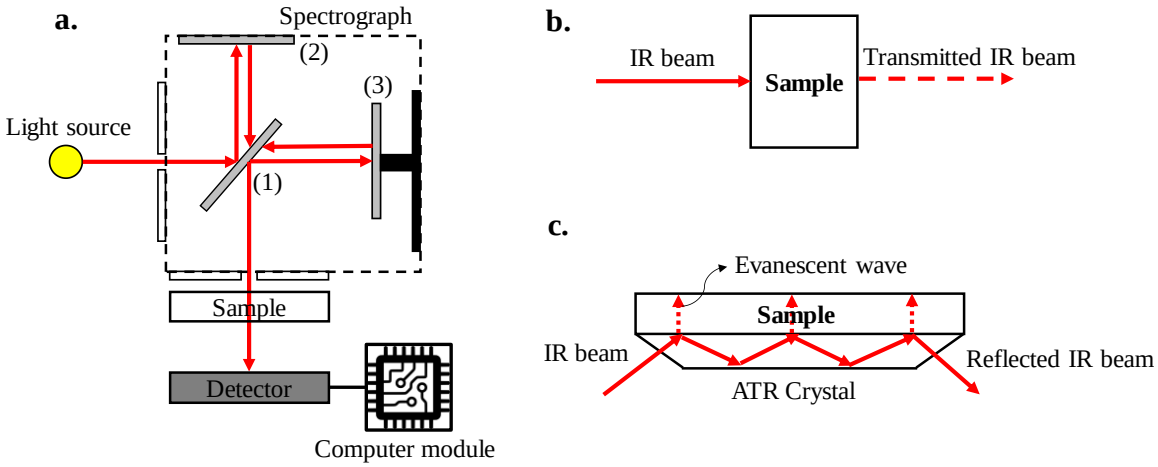


Figure 1.6. (a) FT-IR components spectrometer diagram, where the spectrograph is represented by a Michelson interferometer: (1) beam splitter, (2) fixed mirror, and (3) moving mirror; (b) transmission mode illustration; (c) ATR mode illustration.

1.1.2 Raman Spectroscopy

Raman spectroscopy is based on an anomalous scattering effect that happens with less than 1% of absorbed photons by a sample. More than 99% of the photons absorbed by a substance undergo elastic scattering (*e.g.*, Rayleigh scattering), which does not change the state of the material, since these photons are absorbed, increasing the vibrational energy state of the molecules, and then emitted from the excited vibrational energy state to their initial vibrational energy state. However, a few percent of absorbed photons undergo an inelastic process, where the molecule does not return to the initial vibrational energy state, emitting photons with a specific frequency of deviation to maintain the equilibrium of the system (Skoog *et al.*, 2007). The inelastic scattering can be the Stokes or anti-Stokes scattering, which occur in 1 out 10 million absorbed photons. The Stokes scattering occurs when a molecule absorbs part of the energy of the incoming monochromatic wavelength (*e.g.*, laser) and emits a wavelength of less energy than the wavelength absorbed; and the anti-Stokes scattering happens when the molecule emits a wavelength of greater energy than the absorbed wavelength (Figure 1.7). The latter happens under certain circumstances where the molecule is in a partially excited energy state before absorbing electromagnetic radiation.

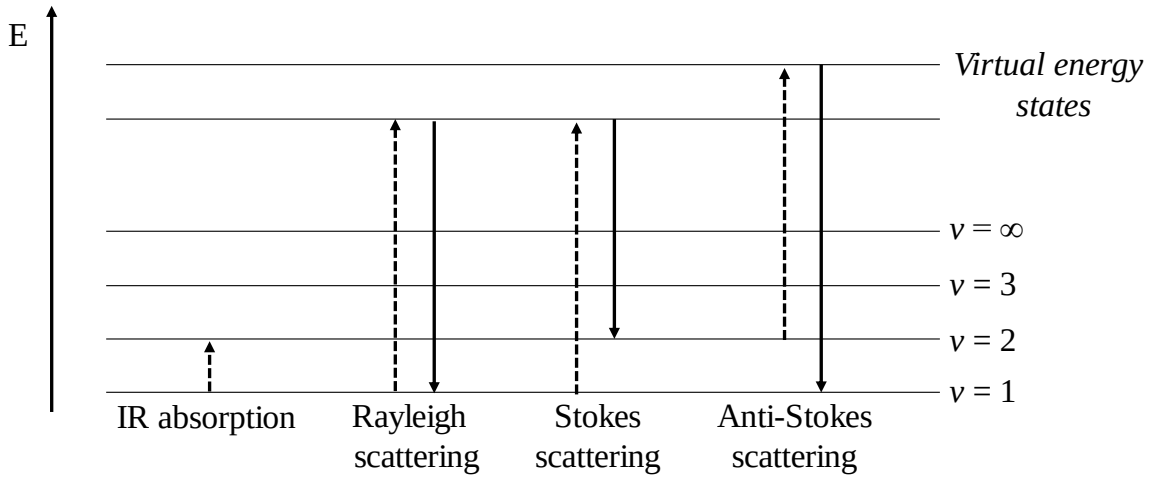


Figure 1.7. Energy diagram showing the IR absorption, Rayleigh scattering, Stokes and Anti-Stokes scattering in the vibrational energy states. E stands for total energy.

The elastic scattering is filtered by the instrument so that only the inelastic scattering is detected and used to produce the Raman spectrum. Both Stokes and anti-Stokes phenomena occur, but at room temperature, there is a lower population of molecules in an initial excited energy state to absorb radiation, therefore the anti-Stokes signal is weaker than the Stokes signal (Skoog *et al.*, 2007). For this reason, many spectrometers only work with the Stokes scattering. Differently from IR, Raman spectroscopy is based on the change of polarizability rather than the dipole moment, which makes the vibration bands in Raman quite different from IR. The main vibration modes for Raman are depicted in Figure 1.8 (HORIBA, 2020).

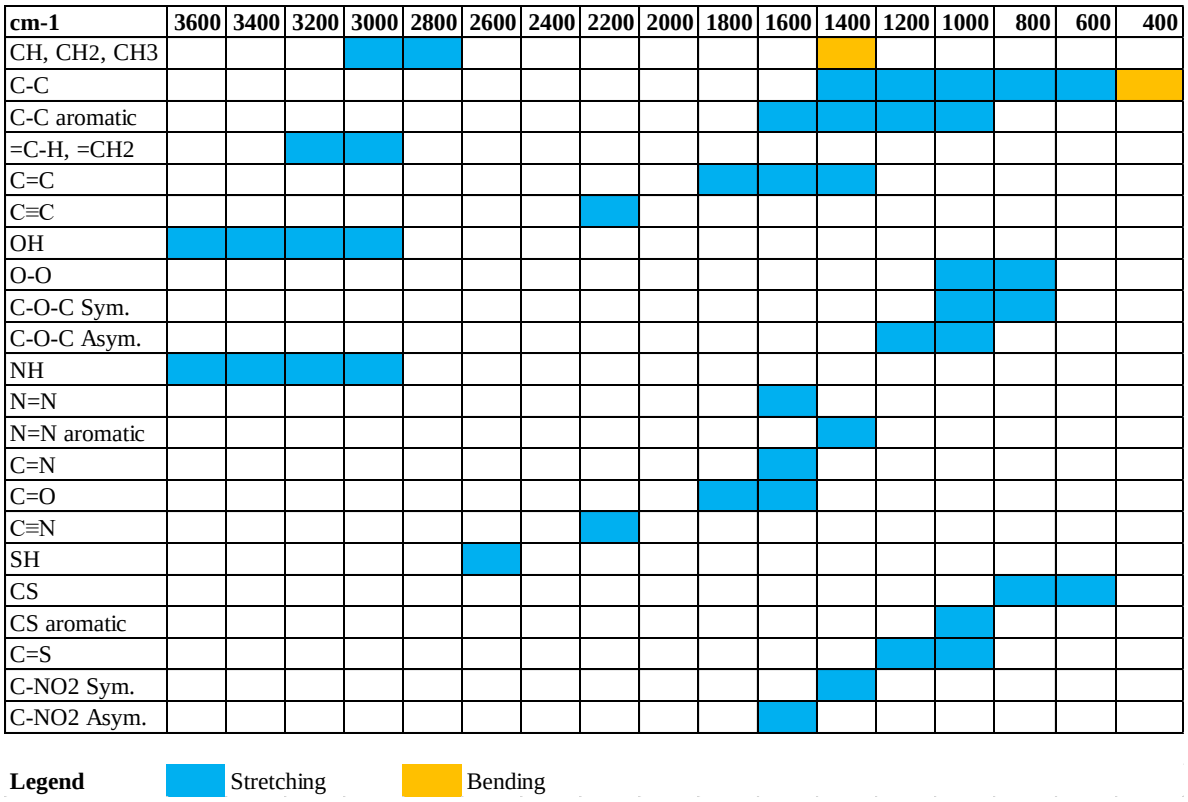


Figure 1.8. Approximation of the main vibration modes for Raman (Raman shift: 3600–400 cm^{-1}). Sym.: symmetric; asym.: asymmetric.

The main advantage of Raman in comparison with IR spectroscopy is that Raman is water transparent below 3000 cm^{-1}, since it has a weak signal after 3000 cm^{-1} (Skoog *et al.*, 2007). This is important for biological applications since liquid or fresh medium is composed mainly of water, which is a problem for IR. IR analysis is often performed on dry samples to minimise the water interferent signal. The Raman spectrometer is composed of 4 main parts: (1) a monochromatic laser source, (2) a spectrograph grating, (3) a charge-coupled device (CCD) detector, and (4) a computer module. The monochromatic laser light, often in the visible or NIR range, is used to excite the sample to the virtual energy states where the inelastic scattering phenomena occur. Most of Raman spectrographs use fixed diffraction grating instead of interferometers which have moving mirrors using in Fourier-transform infrared (FT-IR) spectrometers, hence, the detector is a CCD camera detector that records the whole spectrum at once, thus without needing a Fourier-transformation. The fix diffraction grating and CCD detector also makes the technique substantially faster than FT-IR for spectral acquisition in small wavenumber windows. A computer module is then responsible for converting the CCD detector signal into the spectrum. These parts are summarised in Figure 1.9.

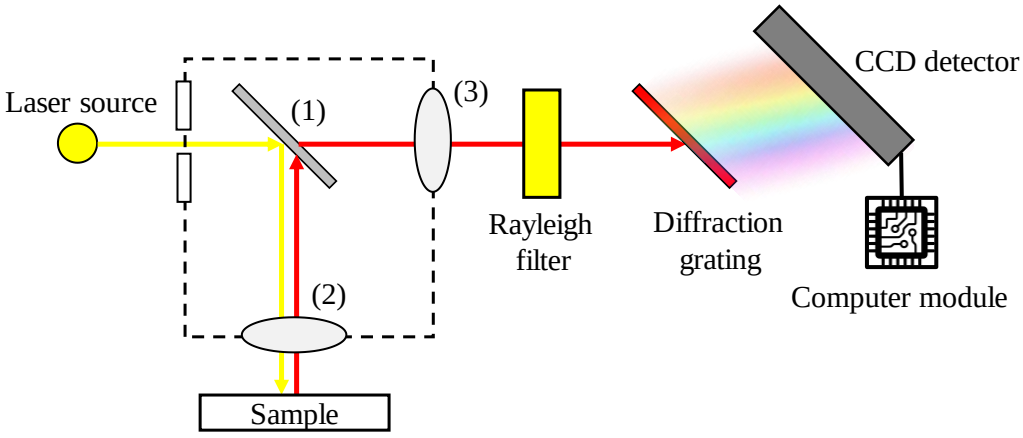


Figure 1.9. Diagram showing the Raman spectrometer components, where: (1) beam splitter, (2) objective lenses, and (3) lens.

1.2 Biospectroscopy

The identification and diagnosis of diseases such as cancer are a global scale problem. The detection of these complex diseases in hospital environments usually requires laborious and extreme invasive procedures, causing much discomfort to patients and even leading them to surgical procedures; in addition to costs associated with sample preparation and complex machinery to the facility.

Biospectroscopy is focused on using vibrational spectroscopy techniques to analyse biological materials (Trevisan *et al.*, 2012). The identification of chemical patterns within biological samples using their spectrochemical information has fundamental importance to overcome issues on disease diagnostic, and corresponds to a breakthrough area of technological innovation having great social impact. Biospectroscopy allows fast, non-destructive and low-cost analysis of biological materials, and has been used in a wide range of applications, including cancer detection based on liquid biopsies and tissue analysis (Kendall *et al.*, 2009; Trevisan *et al.*, 2012), toxicology assays (Heys *et al.*, 2017), environmental studies (Obinaju & Martin, 2013), taxonomic identification (Zimmermann *et al.*, 2015), and detection of cellular mechanisms (Miller & Dumas, 2010).

Biological samples have a specific spectrochemical signature within a region called the biofingerprint region, which is from 1800–900 cm^{-1} for IR and 2000–500 cm^{-1} for Raman (Kelly *et al.*, 2011). The biofingerprint region comprises key absorptions of

lipids (ν_s(C=O), δ_s(CH_2)), proteins (amide I, amide II, ν_s(COO^-)), nucleic acid (ν_{as}(PO_2^-), ν_s(PO_2^-)) and carbohydrates (ν_s(CO-O-C)) (Baker *et al.*, 2014). The main spectrochemical signatures of the spectral markers (or biomarkers) within the biofingerprint region for IR and Raman are shown in Figure 1.10.

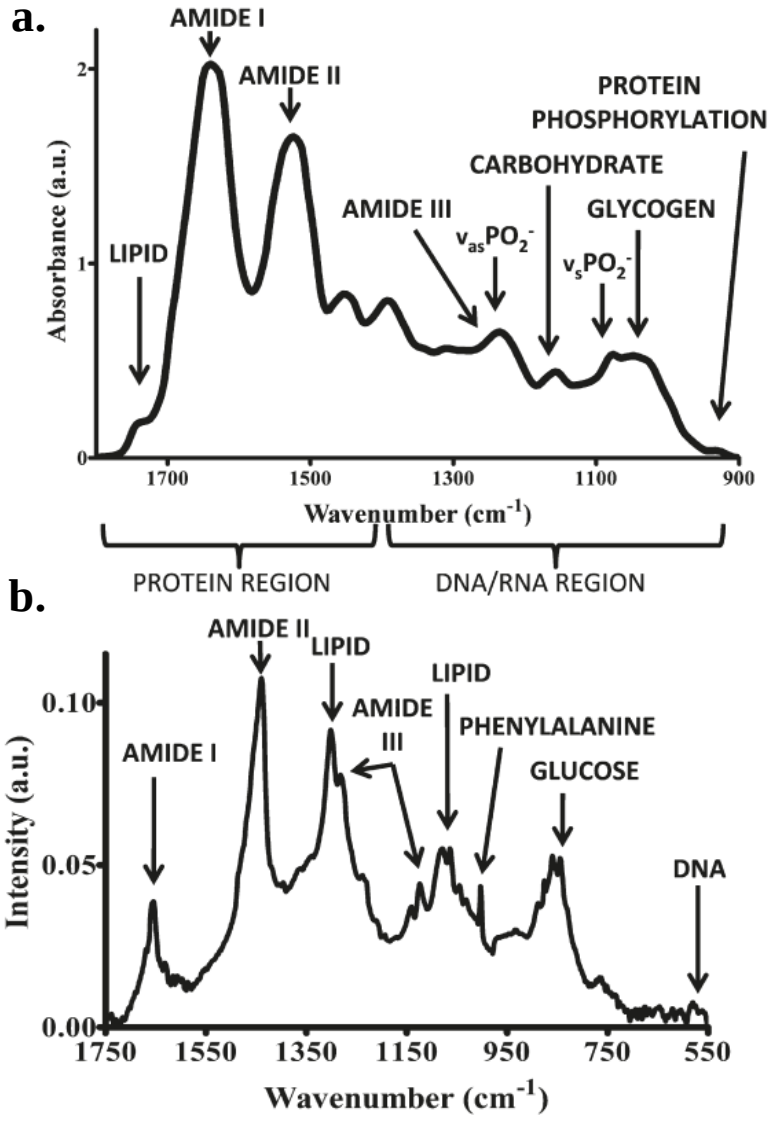


Figure 1.10. Biochemical-cell fingerprint of (A) IR and (B) Raman spectra with tentative peak assignments. (Reprinted (adapted) with permission from Kelly *et al.*, 2011. Copyright 2011 American Chemical Society).

Several types of clinical samples, including biofluids (cerebrospinal fluid (CSF), saliva, blood or urine) or tissue, can be analysed using both IR or Raman spectroscopy in order for disease screening and diagnosis using this biofingerprint region, since the spectral data experimentally acquired can be used as input information for category

classification models using computational tools, or chemometrics (Figure 1.11) (Mitchell *et al.*, 2014).

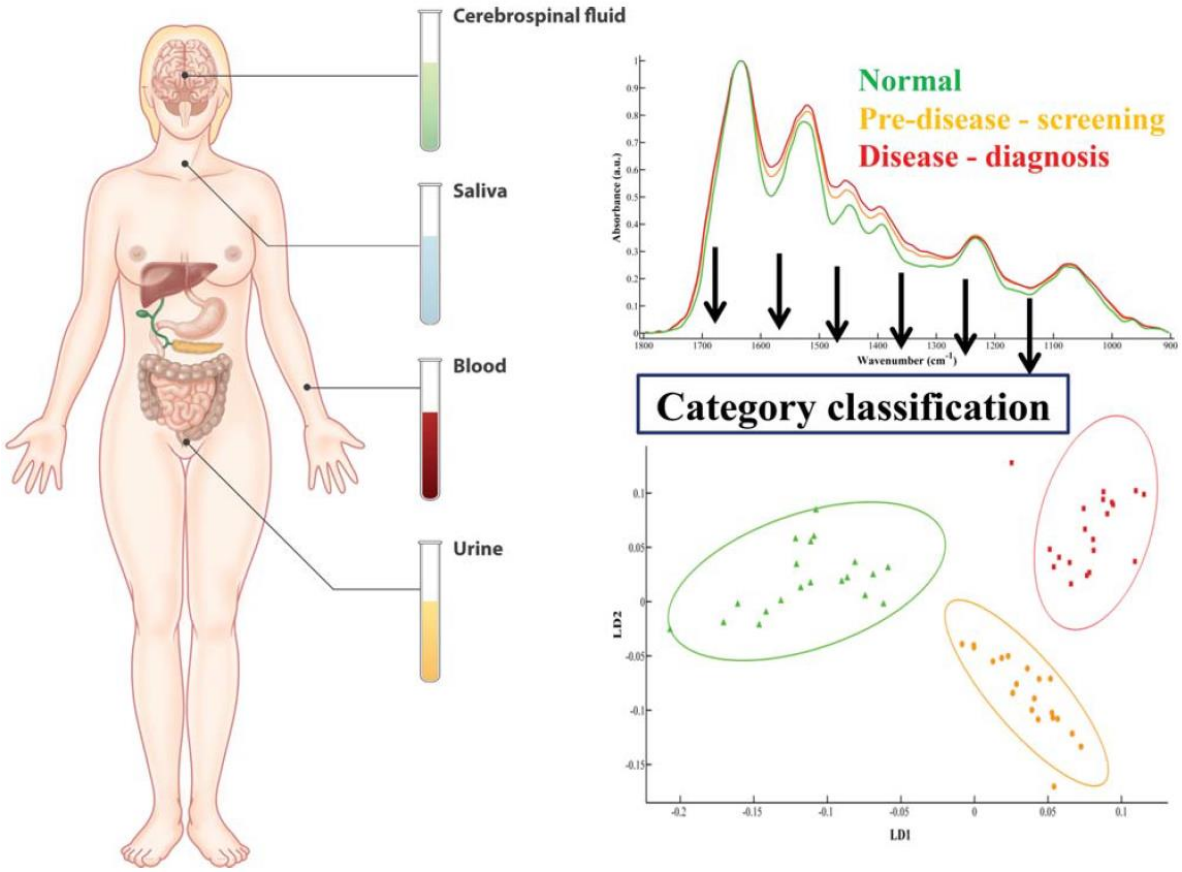


Figure 1.11. Diagram showing a variety of clinical samples (cerebrospinal fluid, saliva, blood, urine) that can be used for disease screening and diagnosis using their biofingerprint spectrochemical signature. (Reprinted with permission from Mitchell *et al.*, 2014. Copyright 2014 John Wiley & Sons, Inc.).

1.2.1 Sample Preparation

Sample preparation for biospectroscopy is minimal, since both IR and Raman are non-destructive techniques. Blood is often centrifuged for extraction of plasma or serum which are subsequently used for spectral measurement. Small aliquots of these biofluids (50 to 250 μL) are usually pipetted onto low-emission (low-E) or aluminium-covered glass slides and allowed to dry overnight at room temperature (Baker *et al.*, 2014; Paraskevaidi *et al.*, 2018c). The same procedure is performed for urine, where the centrifugation process is performed to precipitate solid material and the supernatant is further analysed (Maitra *et al.*, 2019). Saliva is measured as is after drying, while CSF

fluid is analysed diluted in an ethanol 70% (vol/vol) solution to avoid bacterial proliferation and contamination. Samples can be measured in liquid state using Raman since this technique is "water transparent", however they must be properly dried and stored in desiccators prior to FTIR measurements using ATR.

Tissues can be analysed fresh, snap-frozen, formalin-fixed paraffin-embedded (FFPE) or FFPE after dewaxing. Ideally fresh tissue provide the most authentic spectrochemical information, however, this type of sample decompose quickly, hence, most applications involve snap-frozen or FFPE tissue samples. Snap-frozen samples are preferably analysed by Raman spectroscopy due to water interference on FT-IR (Baker *et al.*, 2014; Butler *et al.*, 2016), while FFPE tissue can be analysed either using Raman or FT-IR spectrometers. Dewaxing can be performed prior to analysis in order to remove the paraffin signal interfering (Baker *et al.*, 2014), however there are computational ways to minimise the influence of the paraffin signal in the spectrum (Tfayli *et al.*, 2009), thus FFPE tissue can also be measured as is.

1.2.2 Chemometrics

Advancements on spectrometers instrumentation allow obtainment of complex spectral data of biological samples in a quick and almost automated fashion. However, the mathematical techniques to process the experimental data and extract relevant information are as important as the analytical instrumentation used to acquire them, where the information extraction process, and, as consequence, the clinical diagnosis is only possible to be achieved by using chemometric techniques. Chemometrics is defined as "the science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods" (Hibbert, 2016). Chemometrics is the tool used to convert complex spectrochemical data into meaningful information, thus enabling the analyst to draw relevant conclusions about the experimental data. Chemometric techniques must advance along the instrumentation and data complexity, and it is a keystone to obtain satisfactory and reliable diagnostic results.

Chemometrics include multivariate calibration and classification techniques, and should be used where the spectral signature has overlapping bands, unknown sources of variation or interference, poor signal-to-noise (S/N) ratio, non-linearity in absolute

wavelengths, or other adverse effects in the reactional medium (Brereton, 2003). Multivariate calibration techniques are employed when there are several discrete values as reference labels among the samples distributed in an order of importance (*e.g.*, 1, 2, 3…), hence, being mainly used to estimate chemical concentration values in quantitative analysis based on the spectral data. For qualitative analysis, where there are few category labels among the samples and the aim is to distinguish or classify them, multivariate classification techniques are employed.

Multivariate classification is divided into unsupervised and supervised methods (Beebe *et al.*, 1998). Unsupervised methods are used mainly for exploratory analysis in order to identify natural patterns or clustering that arise from the data, hence, no pre-defined labels are inputted in the model. The most common unsupervised method for exploratory analysis of spectral data is the principal component analysis (PCA) (Bro & Smilde, 2014). PCA decomposes the spectral data into a few principal components (PCs) responsible for most of the original data variance. Each PC is orthogonal to each other and they are distributed in a decreasing order of explained variance, so the 1st PC accounts the highest explained variance, followed by the 2nd PC and so on. Each PC is composed of scores (projections of the samples on the PC direction) and loadings (angle cosines of the wavenumbers projected on the PC direction) (Santos *et al.*, 2017). The PCA scores show the variance on sample direction, thus showing clustering and trending patterns and being used to assess similarities/dissimilarities among the samples; and the PCA loadings show the variance on wavenumber direction, thus being used to identify important wavenumbers or spectral markers (Bro & Smilde, 2014) (Figure 1.12).

In the example depicted in Figure 1.12, the features that distinguish the red from the blue class are characterized by an increase at the wavenumber 1590 cm^{-1} (+ loadings on PC1, + scores on PC1), and a decrease at the wavenumbers 1370 cm^{-1} and 1100 cm^{-1} (- loadings on PC1, - scores on PC1). However, although PCA is a very useful technique for exploratory analysis of the dataset, PCA is not a proper classification technique, since even though PCA gives an indication of the sample nature, it is not able to predict the category of a given blind sample by itself. For this reason, supervised classification techniques are employed in the PCA scores, selected wavenumber features, or in the whole original spectrum region in order to obtain diagnostic results.

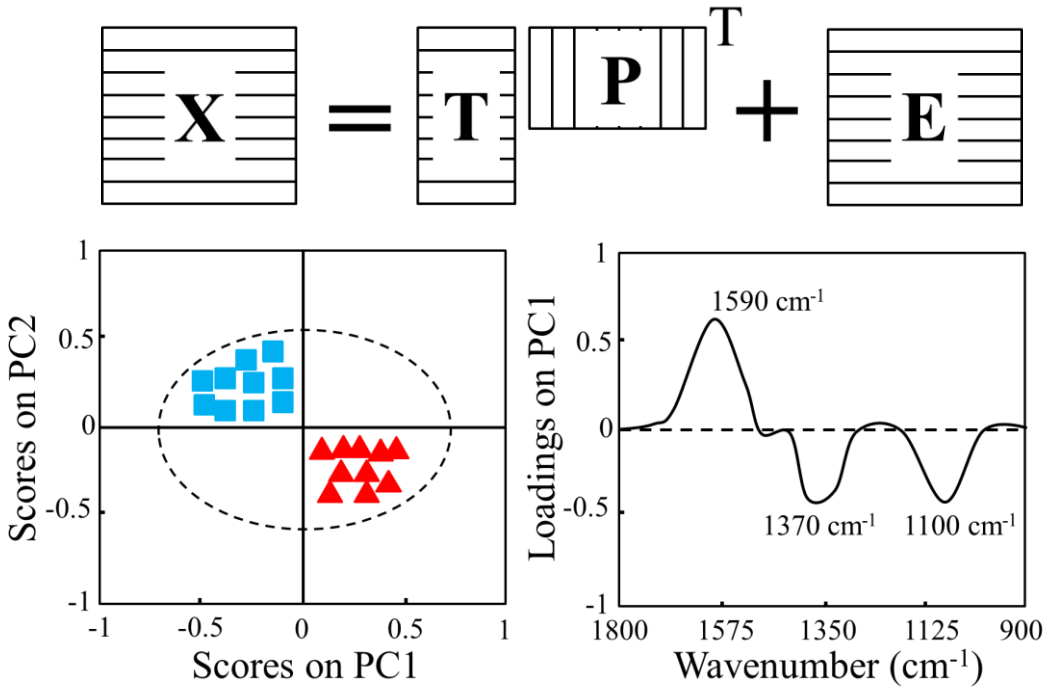


Figure 1.12. Example of PCA decomposition, where the scores show a segregation pattern between blue and red samples, and the loadings show the three wavenumbers (1590 cm^{-1}, 1370 cm^{-1} and 1100 cm^{-1}) responsible for class separation. **X** = spectra dataset, **T** = PCA scores, **P** = PCA loadings, **E =** residuals. Superscript T stands for the matrix transpose operation.

The most common supervised classification techniques are the K nearest neighbour (KNN), artificial neural networks (ANN), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and support vector machines (SVM). KNN is a method that classifiers the samples based on the distance from each of the samples in the training set using the *K* nearest samples, so the classification of an unknown sample is based on the concept of majority vote, that is, the sample is classified to the group that has the most members of training samples amongst its neighbours (Naes *et al.*, 2002). ANN is a more sophisticated method based on neuron interconnections, being inspired by how the human brain works. ANNs are based on a series of nodes that connects to each other in different depth layers. The contributions for all nodes are multiplied by constants and added before a non-linear transformation within the node, which are often a Gaussian transformation function that extract features from the dataset (Naes *et al.*, 2002).

LDA is one of the most common methods used in supervised classification of spectral data LDA calculate the discriminant function between two classes according to the Mahalanobis distance between the samples. QDA works similarly, however is less used than LDA probably due to the lack of QDA algorithms available. The main difference between LDA and QDA is that LDA uses a pooled covariance matrix for sample distance calculation, while QDA uses the individual variance-covariance matrix of each class for this calculation (Dixon & Brereton, 2009; Wu *et al.*, 1996). The LDA and QDA classification scores for sample *I* of class *k* are calculated as follows (Morais & Lima, 2018):

LDA,

$$L_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}} \mathbf{C}_{\mathrm{pooled}}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_k) - 2 \log_e \pi_k \quad (1.9)$$

QDA,

$$Q_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}} \mathbf{C}_k^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_k) + \log_e |\mathbf{C}_k| - 2 \log_e \pi_k \quad (1.10)$$

where $\mathbf{x}_i$ is the vector containing the input classification variables for sample *I*; $\bar{\mathbf{x}}_k$ is the mean vector for class *k*; $\mathbf{C}_{\mathrm{pooled}}$ is the pooled covariance matrix; $\mathbf{C}_k$ is the variance-covariance matrix of class *k*; and π_k is the prior probability of class *k*. These additional terms are calculated as follows:

$$\mathbf{C}_{\mathrm{pooled}} = \frac{1}{n} \sum_{k=1}^{K} n_k \mathbf{C}_k \quad (1.11)$$

$$\mathbf{C}_k = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}} \quad (1.12)$$

$$\pi_k = \frac{n_k}{n} \quad (1.13)$$

where *n* is the total number of samples in the training set, *K* is the total number of classes, and n_k is the number of samples of class *k* in the training set. These scores are used to calculate the discriminant function (DF) between two classes as follows:

LDA,

$$DF = L_{i1} - L_{i2} \quad (1.14)$$

QDA,

$$DF = Q_{i1} - Q_{i2} \quad (1.15)$$

SVM is defined as a method of supervised classification in which decision boundaries (hyperplanes) are determined that maximise the separation of data in different classes (Hibbert, 2016). In other words, SVMs are binary classifiers that work by finding a classification hyperplane in a high-dimensional space which separates two classes of samples providing the largest margin of separation (Harrington, 2015). SVM is a linear technique by nature, however it uses a non-linear step called the kernel transformation (Cortes & Vapnik, 1995). The SVM kernel function is responsible for transforming the data into a different feature space changing its classification ability (Dixon & Brereton, 2009). This provides an extra power to SVM compared to other discriminant algorithms such as LDA and QDA. The LDA and QDA discriminant functions, as well as the SVM discrimination, are illustrated in Figure 1.13.

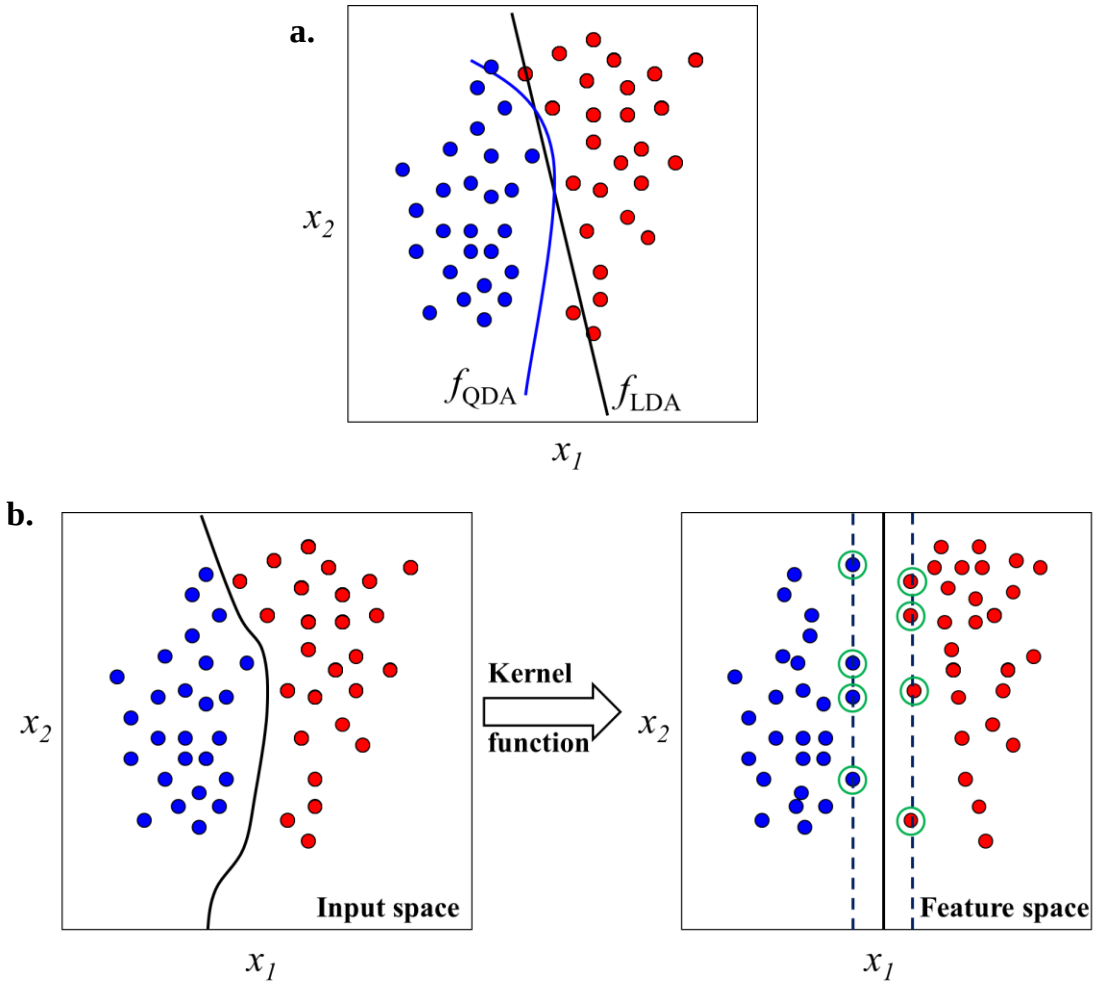


Figure 1.13. Illustration of discriminant functions for classification. (a) Discriminant functions for LDA (f_{LDA}) and QDA (f_{QDA}); (b) SVM kernel transformation and discrimination in the feature space. The circled samples are the closest samples to the class margins, denominated support vectors. x_1 and x_2 represent spectral variables.

LDA is less susceptible to overfitting and usually shows good classification ability when classes have similar variance structures, *i.e.*, when they have similar distributions, and also performs well even with small datasets; while QDA performs better than LDA when classes have different variance structures/distributions, but underperforms LDA in small datasets (Wu *et al.*, 1996). SVM generally performs better than LDA and QDA, however it is highly susceptible to overfitting and more time-consuming.

The input classification variables for LDA, QDA or SVM can be PCA scores in PCA-LDA, PCA-QDA and PCA-SVM algorithms, or specific wavenumbers selected by variable selection algorithms such as the successive projections algorithm (SPA) (Soares *et al.*, 2013) or genetic algorithm (GA) (McCall, 2005). SVM can also work with the full spectrum region, however LDA and QDA requires a number of variables equal or smaller than the number of samples in the training set. Further details about the chemometric techniques employed to analyse biospectroscopy data are discussed in Chapter 2.

1.3 Spectrochemical Imaging

Despite many advantages, conventional spectroscopy techniques lack one fundamental aspect for analysing complex heterogenous samples: spatial information. Spectroscopy techniques using single-spectrum acquisition rely exclusively on positional measures, and no information on spatial distribution is extracted. This narrows down its usability in potential applications involving structural investigations of components distributed over tissue segments. To overcome these limitations and obtain both spatial and spectral information combined, hyperspectral imaging techniques are used.

Hyperspectral techniques combine imaging and spectroscopy. For this, an image is generated for the segment being analysed but, for each image position (pixel), a spectrum is also generated, creating a three-dimensional (3D) object for each sample measured, also called a hyperspectral “data-cube”. This enriches the data obtained since one can access spatial information in the *x*- and *y*-axis, and chemical information in the *z*-axis. Figure 1.14illustrates an example of hyperspectral data.

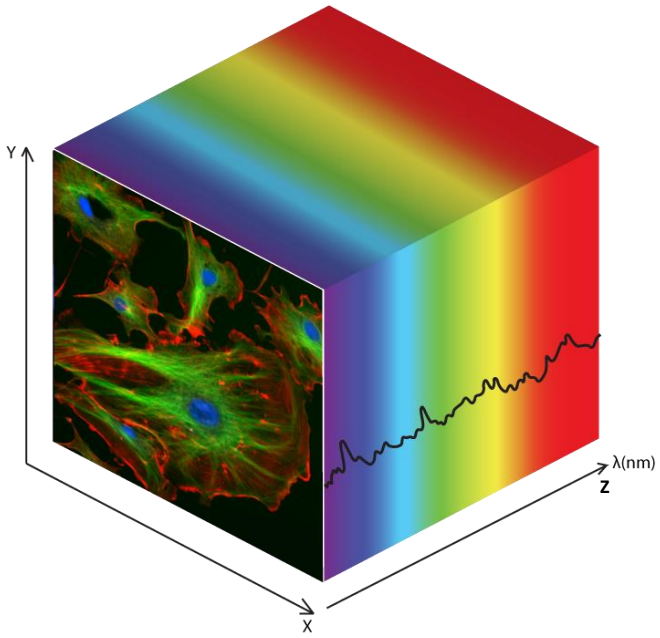


Figure 1.14. Hyperspectral "data-cube", where the spatial information is shown on the x- and y-axis, and the spectral information on the z-axis. (CC BY-NC 2.0, < https://imgbin.com/png/6s6tUKPJ/hyperspectral-imaging-data-cube-photon-etc-market-analysis-multispectral-png>).

1.3.1 Hyperspectral Imaging Techniques

Hyperspectral imaging techniques vary according to the frequency of the electromagnetic ration used for irradiating the sample, and most of classical spectroscopy techniques can be adapted for imaging. IR (MIR and NIR), UV or Vis (UV-Vis), fluorescence and Raman spectroscopy are the most used hyperspectral imaging techniques. MIR and Raman hyperspectral imaging can be used by coupling a microscope and automated sample stage to the conventional spectrometer, where the IR (for FT-IR) or laser (for Raman) light beam is focused by objective lenses onto a small area of the sample. Then, three data acquisition configurations are often used: point scan, line scan or area scan (Pu *et al.*, 2015). Although having a great level of detail, point scans require a long acquisition time, therefore being unsuitable for real-time applications. Area scans enable the acquisition of whole images at discrete wavelengths, but it is not capable of detecting objects in great detail. Therefore, the best acquisition mode is usually line scan, since it combines fast acquisition with detailed information obtained through a 2D array detector.

1.3.2 Data Acquisition

The instrumental setup vary according to the instrumental technique used. Parameters such as distance from the sample to the detector and light source power may affect the result. Hyperspectral images can be obtained in reflectance, transmittance, fluorescence or scattering modes. Before measurement, the instrument is usually calibrated using white or dark references to account for the uneven intensity of the light source, the spectral response of the device, and the dark radiation present in the measurement chamber that may affect the imaging detector. However, this calibration procedure does not take into consideration morphological information of the sample being measured. Flat samples are not affected by morphological changes, but non-flat samples (*e.g.*, round-shaped or cylindrical-shaped) are affected by their surface curvature that creates a gradient distance variation between the sample and detector, causing differences in the optical pathlength. This should be corrected by additional pre-treatments prior to analysis (Pu *et al.*, 2015).

1.3.3 Data Analysis

Often, feature extraction techniques are applied to reduce and compress hyperspectral images, since this type of data usually requires extensive storage space and processing power. To speed up computational analysis, spatial- and chemical-relevant features are extracted from images by employing multivariate-based techniques. This is possible because although the hyperspectral data contains millions of data points, only a few of them contain relevant information to the property being measured. This is the most critical part for processing hyperspectral images, since the overall sensitivity and robustness of the machine vision system strongly depends upon the responsiveness of the features containing the desired information for analysis. Firstly, one has to decide which type of features to extract: spectral features, textural features or a combination of both. The key issue is to determine whether the desired information is proportional to the number of pixels or spectral signatures, regardless of how they are distributed in the hyperspectral image or if the desired information is related to the spatial distribution of pixels (Duchesne *et al.*, 2012).

The simplest method of feature selection is to select wavelengths at peaks or troughs of the spectral curve according to the maximum or minimum intensity difference (Pu *et al.*, 2015). However, some relevant wavelengths might not be selected by this procedure due to the multivariate nature of the data, therefore, it might be necessary to use multivariate image analysis (MIA) procedures for this. MIA was introduced in the late 1980's to deal with images that represent more than one measurement per pixel. It can perform at either a global or pixel level. The first is focused on analysing each hyperspectral image as a 'sample', therefore covering all the information contained in the image to obtain an analytical parameter. The second is focused on analysing individual pixel spectra within each hyperspectral image (Prats-Montalbán *et al.*, 2011).

MIA is very efficient at extracting spectral features, where it starts unfolding the hyperspectral data cube into a 2D data matrix, so the spectral feature for each pixel refers to a row, and then some type of data reduction technique is employed. PCA and multivariate curve resolution alternating least squares (MCR-ALS) are the most used techniques. MCR-ALS decomposes the unfolded data **X** in a bilinear model as follows (Prats-Montalbán *et al.*, 2011):

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \tag{1.16}$$

The MCR results are the concentration distribution maps **C**, the pure spectra of image constituents **S,** and the residuals **E**. This decomposition separates the information of different constituents in an image to find its purest information or to detect pixels with selective information. It can be applied to one or more images together, where multi-image analysis is used when a multilayer image from a single sample or a series of images with related chemical composition (Prats-Montalbán *et al.*, 2011).

Similar methodologies focusing in finding 'purest' components is also explored by other algorithms, such as SIMPLISMA and fixed size image window-evolving factor analysis (FSIW-EFA). The purest pixels provide an approximation of the pure spectra sought, while the purest spectral channels, when unfolded, allow construction of approximate distribution maps of the purest constituents (Prats-Montalbán *et al.*, 2011). Other algorithms for feature selection in hyperspectral images are minimum redundancy maximum relevance (MRMR), GA, SPA and neural network approaches by extracting the weight of each input in the back-propagation (Pu *et al.*, 2015). Thereafter, the spectral

and/or spatial features extracted from the images are used as input variables for the classification algorithms detailed in section 1.2.2 and Chapter 2.

1.4 Current Challenges in Biospectroscopy

Biospectroscopy is still a science in development, where the actual main challenge is the final clinical implementation of the developed methodologies in hospital environments. This is a challenge still in the horizon nowadays since before proceeding to this final step there are a series of other issues that must be addressed and validated, and these are almost exclusively related to the analysis of the spectral data; since protocols for sample preparation and experimental measurements are already well established (Baker *et al.*, 2014; Butler *et al.*, 2016; Martin *et al.*, 2010).

1.4.1 Sample and Data Complexity

Biological samples are extremely complex by nature, since the number of chemical components, hence, spectral features are exceptionally overlapping. Therefore, only multivariate techniques covering the analysis of multiple wavenumbers at the same time are capable of working in such complex environment. Traditionally, a pre-defined spectral dataset is used for model construction, *i.e.*, to train the model, while an external test set is used to validate the model towards unknown samples. However, there is still a lack of understanding on how to select samples for training and test sets. Manual selection of samples spectra usually introduces bias, thus random selection is often used instead. But random selection adds a high-risk of extrapolation, that is, the classification model does not include enough source of variation that allow them to accurately predict unknown samples. Another strategy often used in analytical applications, is the Kennard-Stone (KS) algorithm (Kennard & Stone, 1969), that systematically select samples for the training set based on maximising the training space, but it does not take into account random or extreme behaviours that often happen in biological medium. Therefore, there is a need for algorithms to optimally select samples for training and test sets in biospectroscopy applications.

1.4.2 Processing Imaging Data

The major drawback of hyperspectral imaging is the time required to analyse the data experimentally acquired. Hyperspectral image datasets occupy gigabytes of space and depending on the number and size of the images, simple data analysis such as PCA may take days or weeks to process using standard computers. Supercomputing is often used in this situation, but this adds a cost that many institutions cannot afford. For this reason, faster and accurate chemometric methods to analysis hyperspectral data-cubes are required in biological applications, where often many samples must be analysed and compared quickly using lower-capacity computers.

1.4.3 Uncertainty in Diagnostic Accuracy

Overfitting is the major problem of predictive computational models. Many times the model is well trained and internally validated with a good accuracy, but the model is not robust enough to predict real external samples where subtle sources of variation or random noise is present. For this reason, mathematical tools to predict uncertainty in model diagnostic in order to foresee future performance are needed.

1.4.4 Environmental Variability

Measurements made in different centres, even using the same sample preparation, instrument manufacturer and measurement settings, are different. That is because there are underlying variations in the environment, such as humidity and CO_2 level, in the instrument, such as ageing of parts, and in sample handling that may affect the spectral response. For this reason, a protocol showing how to standardise spectral response across different centres using computational tools is needed as a final step before implementing biospectroscopy in real-routine applications, since in this scenario samples will need to be measured by different instruments, analysts and in different locations, and the diagnostic result will need to be consistent across these different environments.

1.5 Aims and Objectives

1.5.1 Research Aim

This PhD, entitled "*Novel chemometric approaches towards handling biospectroscopy datasets*", aims to provide a novel protocol for the analysis of biospectroscopy datasets, solving current issues present in four fronts of spectrochemical analysis: (1) sample selection, (2) processing of imaging data, (3) uncertainty estimation in diagnostic accuracy, and (4) standardisation methods to account for environmental variability.

1.5.2 Objectives

i. To produce a chemometric protocol for multivariate classification of biospectroscopy data.
ii. To develop a method for optimal sample selection in biospectroscopy application based on the combination of the Euclidian-distance-based Kennard-Stone (KS) selection with a random mutation factor.
iii. To speed exploratory analysis and classification of hyperspectral imaging data through a three-dimensional principal component analysis (3D-PCA) approach.
iv. To develop a methodology to estimate diagnostic uncertainty and model robustness in linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and support vector machines (SVM) classification.
v. To produce a protocol showing how to standardise biospectroscopy datasets collected across different centres.

1.6 Statement of Originality

All the chemometric techniques developed in Chapters 3 to 8 are new, and they were never applied in biological-derived spectrochemical datasets before. The protocols provided in Chapters 2 and 9, as well as the software interface provided in Chapter 10, are also new, and they were produced to assist the biospectroscopy community showing how to process complex spectrochemical data of biological materials. This thesis show how to obtain more reliable and accurate diagnostic performance and biomarkers identification for any biological or clinical application where at least two distinct categories are analysed. Successful development of such chemometric methods will allow to trial these tools in a clinical setting, where fast, reliable, and highly accurate diagnosis are required.

1.7 Thesis Structure

This thesis is structured in 11 chapters plus appendices. Chapter 2 is a protocol on how to analyse biospectroscopy data towards classification applications; Chapters 3 and 4 demonstrate a new algorithm for sample selection in biological applications; Chapters 5 to 7 are related to hyperspectral image analysis, where Chapter 5 demonstrate a clinical application of MCR-ALS and Raman hyperspectral imaging for meningioma WHO tumour grade detection, Chapter 6 demonstrate a new chemometric algorithm for three-dimensional exploratory analysis of hyperspectral images, and Chapter 7 is a further development if this algorithm coupled to LDA and QDA as discriminant analysis approaches. Chapter 8 demonstrates how to estimate uncertainty and model robustness for LDA, QDA and SVM applied to biospectroscopy datasets; and Chapter 9 is a protocol on how to standardise biospectroscopy datasets acquired across different centres. Chapter 10 demonstrates an user-friendly software interface developed to process trilinear 3D data; and Chapter 11 is the overall discussion of the thesis conclusions and future perspectives for the field. The Appendices contain details about supplementary materials for Chapters 2, 5 and 9; and the Ethics Approval of this project.

CHAPTER 2 | MULTIVARIATE CLASSIFICATION TECHNIQUES FOR VIBRATIONAL SPECTROSCOPY IN BIOLOGICAL SAMPLES

This chapter is a protocol on how to analyse biospectroscopy data towards classification applications using traditional chemometric techniques. This chapter is part of a paper published in Nature Protocols (IF 11.334):

- Morais CLM, Lima KMG, Singh M, Martin FL. Tutorial: multivariate classification for vibrational spectroscopy in biological samples. Nat. Protoc. **2020**. https://doi.org/10.1038/s41596-020-0322-8

Abstract: The use of vibrational spectroscopy techniques, such as Fourier-transform infrared (FTIR) and Raman spectroscopy, has been a successful method to study the interaction of light with biological materials and facilitate novel cell biology analysis. Disease screening and diagnosis, microbiological studies, forensic and environmental investigations make use of spectrochemical analysis very attractive due to its low cost, minimal sample preparation, non-destructive nature and substantially accurate results. However, there is now an urgent need for multivariate classification protocols allowing one to analyse biological-derived spectrochemical data in order to obtain accurate and reliable results. This is stimulated by the fact that applications of deep-learning algorithms of complex datasets are being increasingly recognized as critical towards extracting important information and visualizing it in a readily interpretable form. Hereby, we have constructed a protocol for multivariate classification analysis of vibrational spectroscopy data [FTIR, Raman and near-infrared (NIR)] highlighting a series of critical steps, such as pre-processing, data selection, feature extraction, classification and model validation. This is an essential aspect towards the construction of a practical spectrochemical analysis model for biological analysis in real-world applications, where fast, accurate and reliable classification models are fundamental.

Author contribution: C.L.M.M. performed the data analysis and wrote the manuscript.

Camilo L. M. Morais, PhD candidate

Prof. Francis L. Martin, Supervisor

2.1 Introduction

Vibrational spectroscopy comprises techniques related to electronic changes in the internal vibrational energy levels of molecules. Biomolecules that contain chemical bonds that vibrate generating a change in the dipole moment as a result of the transition are IR active (Martin *et al.*, 2010; Santos *et al.*, 2017). Infrared (IR) and Raman spectroscopy are the main spectroscopic techniques used to assess vibrational molecular modes, where the first is based on molecular dipole changes and the latter on molecular polarizability changes. IR spectroscopy is dived into near-IR (NIR), mid-IR (MIR) and far-IR (FIR) spectroscopy depending on the incident light frequency. MIR is the main technique used to analyse biological materials since it covers fundamental vibrational modes of important biomolecules. Vibrational spectroscopy can provide rapid, label-free, and objective analysis for the clinical domain. The fingerprint region, between 1800-900 cm^{-1}, include important absorptions of lipids (C=O symmetric stretching at ~1750 cm^{-1}, CH_2 bending at ~1470 cm^{-1}), proteins (Amide I at ~1650 cm^{-1}, Amide II at ~1550 cm^{-1}, Amide III at ~1260 cm^{-1}), carbohydrates (CO-O-C symmetric stretching at ~1155 cm^{-1}), nucleic acid (asymmetric phosphate stretching at ~1225 cm^{-1}, symmetric phosphate stretching at ~1080 cm^{-1}), glycogen (C-O stretching at ~1030 cm^{-1}), and protein phosphorylation (~970 cm^{-1}) (Baker *et al.*, 2014; Kelly *et al.*, 2011; Movasaghi *et al.*, 2008). The high-region, between 3700–2800 cm^{-1}, can also be used for analysis, where information of water (-OH stretching at ~3275 cm^{-1}), protein (symmetric -NH stretching at ~3132 cm^{-1}), fatty acids and lipids (=C-H asymmetric stretching at 3005 cm^{-1}, CH_3 asymmetric stretching at ~2970 cm^{-1}, CH_2 asymmetric stretching at ~2942 cm^{-1}, CH_2 symmetric stretching at ~2855 cm^{-1}) can be obtained (Paraskevaidi *et al.*, 2017b). NIR spectroscopy can also be applied as a biospectroscopy tool. This technique is mainly composed of MIR overtones, hence, the signal is very complex containing many overlapping features. Therefore, biomarkers identification using NIR is harder and more ambiguous, although this technique is a powerful tool for quantification and classification applications (Pasquini *et al.*, 2018). Raman spectroscopy is based on an inelastic scattering effect. Most of the photons absorbed by a molecule suffers elastic scattering; only a small portion of them (<1%) suffer inelastic scattering, where the released radiation has lower or higher energy than the initial incoming absorbed radiation (Santos *et al.*, 2017). Inelastic scattering can be Stokes (photons with lower energy are emitted) or anti-

Stokes (photons with higher energy are emitted), and both correspond to the Raman signal. Due to the small probability of molecules in an initial high energy state at room temperature, the anti-Stokes signal is not so strong, and the Stokes signal is usually recorded as the final Raman spectrum.

Since both IR and Raman are non-destructive and sensitive techniques with a relative low-cost, passive of automation, and translatable to portable devices, their use to investigate biological samples are of great interest (Baker *et al.*, 2014; Butler *et al.*, 2016). Biofluids offer an ideal diagnostic medium due to their ease and low cost of collection and daily use in clinical biology. Applications using IR and Raman spectroscopy to investigate biological samples for food (Jin *et al.*, 2015; Karoui *et al.*, 2010; Li-Chan *et al.*, 1996; Prieto *et al.*, 2017; Qu *et al.*, 2015; Scotter *et al.*, 1990), plant (Baranska *et al.*, 2013; Bittner *et al.*, 2013; Buitrago *et al.*, 2018; Cozzolino, 2014), microorganism (Jarvis & Goodacre, 2004; Lorenz *et al.*, 2017; Naumann *et al.*, 1991; Quintelas *et al.*, 2015; Rodriguez-Saona *et al.*, 2001; Schmitt & Flemming, 1998; Stöckel *et al.*, 2016; Strola *et al.*, 2014; Weiss *et al.*, 2019; Zarnowiec *et al.*, 2015) and clinical analysis (Baker *et al.*, 2018; Bunaciu *et al.*, 2014; De Bruyne *et al.*, 2018; Pence & Mahadevan-Jansen, 2016; Sakudo *et al.*, 2016) are many. These previously mentioned advantages associated with the application of multivariate statistical methods of data analysis make these techniques every day more attractive for routine application. Previous protocols for IR (Baker *et al.*, 2014) and Raman (Butler *et al.*, 2016) spectroscopy to analyse biological samples have been already published, but there is still a lack of good practical procedures on how to analyse the acquired data for classification applications where, for example, the spectral data can be used to determine if a given sample is healthy or disease, or if it belongs or not to a given group. This is critical since the results obtained by these studies are directly dependent on the data analysis methodology being used.

Bio-spectral data analysis is a science that requires multidisciplinary knowledge, where to obtain reliable and chemically-meaningful results, the application of chemometric techniques is fundamental. Chemometrics is defined as "the science of relating measurements made on a chemical system or process to the state of the system *via* application of mathematical or statistical methods" (Hibbert, 2016). The use of statistical methods to solve chemical problems trace back centuries, though in 1949, the first report of least squares regression, design of experiments and analysis of variance (ANOVA) appear in analytical chemistry, by Mandel (Mandel, 1949). In the early 1960's,

multivariate methods were first reported in a modern physical-chemistry context to determine the number of components in spectral mixtures as theoretical chemistry approaches (Wallace, 1960; Weber, 1961). Practical implementation into experimental analysis started with the influence of statistical approaches by Pearson and Fisher, whose published work in multivariate analysis in the 1920's and 1930's, acted as inspiration to apply ideas such as principal component analysis (PCA), factor analysis and discriminant modelling in a chemical context (Brereton *et al.*, 2017). During late 1960's and the 1970's, advancements in computer power and availability, the development of artificial intelligence algorithms, and the work done by Bruce Kowalski in the US and Svante Wold in Sweden enabled the introduction of multivariate methods in analytical chemistry in a modern fashion, and the word "chemometrics" was defined (Brereton *et al.*, 2017).

Multivariate methods in a chemical context can be seen as an expansion of the Lambert-Beer's law in a multi-component approach, where the absorbance (spectral response) is a linear combination of concentration time coefficients (Beebe *et al.*, 1998). For the notation, generally, bold uppercase characters (*e.g.*, **X**) represent matrices, bold lowercase characters (*e.g.*, **x**) represent vectors, and italic characters (*e.g.*, *n*) represent scalars. The concentrations are related to sample differences, thus being used to assess the real chemical concentration or to find similarities/dissimilarities between samples, while the coefficients represent the weight of each variable (*e.g.*, wavenumber) in the linear combination, hence, being used to find possible spectral markers. In classification applications, most of the algorithms employed to discriminate spectral data are a combination of feature extraction or feature selection methods followed by discriminant or class modelling techniques, which are mostly distance-based; a classical example is the partial least squares discriminant analysis (PLS-DA) algorithm (Brereton & Lloyd, 2014).

There are several steps to process biospectroscopic data towards classification applications. Firstly, before analysing the data, one must think if the experiment was performed correctly, if the number of sample is representative to solve their problem, and if the data are bilinear, that is, if the product of spectrum times concentration is a constant. If design of experiments (DoE) (Jacyna *et al.*, 2019) are required to acquire representative data, this should be performed. If the data are not bilinear, then non-linear data analysis approaches should be investigated. After data acquisition, the first step is to visualise the data. Anomalous spectral behaviours should be investigated and, depending on the

application of interest, removed from the dataset. Outlier detection is a powerful tool to systematically investigate anomalous spectral profiles, though one must always visualise the data during this procedure. Then, pre-processing, data selection, model construction and validation are the essential steps to obtain reliable results. These steps are summarized in Figure 2.1. We must stress that these steps are iterative and entwined and the user can go back and change them until convergence to a good model is achieved. For this reason, red arrows going backwards are shown in Figure 2.1; this means that in order to optimise the model the user must test different data selection (including outlier detection), pre-processing and model construction techniques in an interconnected way since there is no single route to validation.

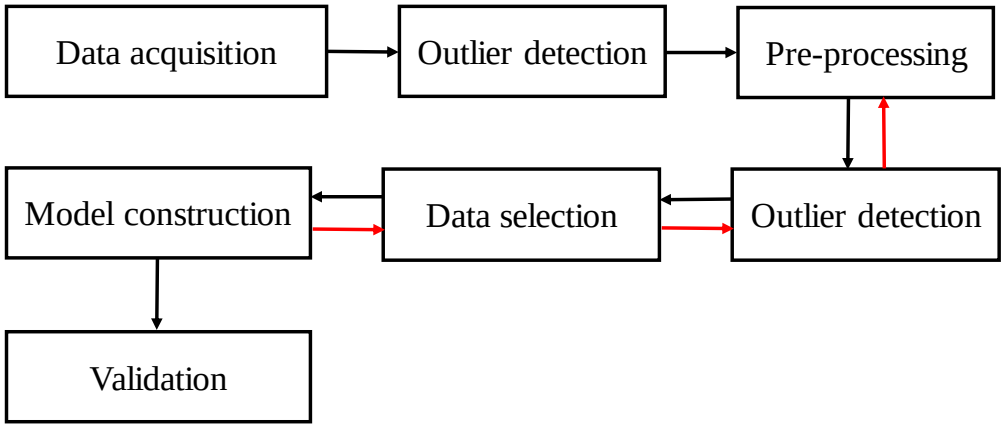


Figure 2.1. Spectral data analysis flowchart.

2.2 Experimental Design

More important than the data analysis itself is the experimental setup used to acquire the spectral data. Previous protocols demonstrate all the materials and steps needed for spectral data acquisition of biological-derived samples using both IR and Raman spectroscopy (Baker *et al.*, 2014; Butler *et al.*, 2016; Martin *et al.*, 2010; Morais *et al.*, 2019c). Therefore, in this protocol, we will focus solely on the spectral data analysis aspect.

2.2.1 Minimum Dataset Requirements

Before carrying on with the experiments, the number of samples must be defined. For pilot studies, power tests are recommend, where a power of 80% can be used as the minimum number of samples for the dataset (Jones *et al.*, 2003). Normally, 5 to 25 point spectra are collected per sample (Morais *et al.*, 2019c), and 10 point spectra have been suggested in a previous protocol for ATR-FTIR (Baker *et al.*, 2014). By increasing the number of spectra replicates, the standard-deviation between measurements is reduced, since the standard-deviation is proportional to $1/\sqrt{n}$, where n is the number of spectra replicates. Extra caution should be taken when analysing heterogeneously distributed samples (*e.g.*, tissues), where spectra replicates should be acquired in a way that covers the sample surface as uniformly as possible. Sample replicates are also recommended. For precision estimation, at least six replicates at three levels should be performed (Morais *et al.*, 2019c). When patient variability is being measured, *i.e.*, when the classification model is performed in a sample-basis, the spectral replicates per sample can be averaged so each resultant spectral response corresponds to a different patient. In this way, the chemometric model is modelled per patient rather than by spectral replicate. However, this requires a larger number of samples and might be difficult to be implemented in small pilot studies. In larger studies, especially before routine implementation, thousands of samples are necessary. This number is defined by the analyst experience and the classification rigour needed. The analyst while designing the experiment must think about confounding factors and the sources of variability that needs to be contemplated in the experiment. If needed, standardisation procedures should be performed to make sure that systematic variations due to environmental, instrumental or analyst changes do not affect the spectral response (Morais *et al.*, 2019c).

Also, the classes' sizes must be taken into consideration. Ideally, classes should have equal size; however, in real clinical scenarios it is unlikely this will occur. For example, for general screening applications, it is very common in clinical settings to have more healthy patients than disease; while, when investigating a specific type of disease, it is more likely that the patients being recruited contain the disease of interest whilst the control group is reduced. When both situations are present, the analyst must take extra care in the data analysis to avoid overfitting the model towards the biggest class size. Some solutions are the application of prior-probability terms based on the classes' size,

the use of non-parametric methods, or by increasing the number of samples for each class to ensure that the calibration model covers enough sources of variation for each classes. As well, according to the central limit theorem (CLT), by increasing the number of samples the data will tend to a normal distribution, which will make parametric classification methods more efficient.

Before pre-processing, the data can be evaluated visually and through some statistical methods in order to identify anomalous behaviours or biased patterns. This is first performed by visual inspection (*e.g.*, plotting the data to identify anomalous spectral features), followed by Hotelling's T^2 *versus* Q residuals charts using only the mean-centred raw spectra. Principal component analysis (PCA) residuals can be explored to identify experimental bias, in which heteroscedastic distributions indicate biased experimental measurements, whereas homoscedastic distributions are associated with good sampling (Beebe *et al.*, 1998). The signal-to-noise ratio (SNR) can be estimated by dividing the signal power (P_{signal}) by the power of the noise (P_{noise}), that is, $SNR = P_{signal}/P_{noise} = \left(A_{signal}/A_{noise}\right)^2$, where A is the amplitude; or by the inverse of the coefficient of variation, when only non-negative variables are measured (Morais *et al.*, 2019c). Collinearity can be evaluated by calculating the condition number, which shows how sensitive the result is to perturbations in the spectral data and to roundoff errors made during the solution process (this value is naturally elevated for spectral data, which indicates high collinearity) (Morais *et al.*, 2019c). Visualising the data by plotting them throughout all the data analysis steps summarized in Figure 2.1 is essential. Prior scientific knowledge of the problem being addressed is also important. The data must be meaningful and the analyst has to make decisions based on how the spectral data look. All data analysis algorithms generate numbers based on the input data, thus if the data are not meaningful (*i.e.*, the signal of interest is absent) the model will generate untrue values. Thus, adequate instrumental techniques allied with good chemometric practices is fundamental.

2.2.2 Pre-processing

Pre-processing is applied to the spectral data in order to remove or reduce the contribution of signals which are not related to the analyte or target property, or to the sample discrimination (which depends on the chemical composition). Pre-processing of the raw data reduces chemically irrelevant variations with the goal of improving accuracy and precision of qualitative and quantitative analyses. The primary role of pre-processing is to transform the spectrum to the best fit condition and to ensure that optimum performance can be achieved in later steps. This process is essential to correct for physical interferences, such as light scattering due to different particle sizes, different sample thickness or different optical paths; and random instrumental noise. However, pre-processing techniques also carry the risk of generating correlations in the noise structure, which would impact negatively on the quality of the multivariate model; thus one should use pre-processing techniques with caution and not overuse them.

In biological applications, the first pre-processing usually consist of truncating the biofingerprint region: 1800-900 cm^{-1} for IR data (Kelly *et al.*, 2011), 2000 to 500 cm^{-1} for Raman (Kelly *et al.*, 2011), and 900 to 2600 nm for NIR (Paraskevaidi *et al.*, 2018b). This removes spectral artefacts such as water and CO_2 absorptions present in other parts of the IR spectrum, and additional baseline distortions that may be present in the spectrum (Morais *et al.*, 2019c). The high-region associated mainly to lipids (3700 to 2800 cm^{-1}) can also been used for IR (Paraskevaidi *et al.*, 2017b) and Raman (Pence & Mahadevan-Jansen, 2016), however this region is highly affected by water absorption in IR (free ν(O-H) at 3600-3650 cm^{-1}; hydrogen-bonded ν(O-H) at 3300-3400 cm^{-1}) (Pavia *et al.*, 2008) and Raman (fully hydrogen-bonded ν(O-H) at 3250 cm^{-1}; partly hydrogen-bonded ν(O-H) at 3300-3630 cm^{-1}) (Hu *et al.*, 2013). Usually, the model performance in the fingerprint region is better than in the high-region due to less water interference and the presence of more complex chemical features (Callery *et al.*, 2019; Paraskevaidi *et al.*, 2017b).

Figure 2.2 depicts the effect of each pre-processing for a given spectral dataset; and Figure 2.3 shows a flowchart to define which pre-processing technique to use after removal of substrate contributions. Pre-processing techniques should be used in the most parsimonious way (Seasholtz & Kowalski, 1993). The order in which the pre-processing steps are performed is fundamental; they must be performed in a logical order so that the next pre-processing step does not mask the signal of interest highlighted with the previous

pre-processing (Morais *et al.*, 2019c). Each pre-processing has its advantage, disadvantage and optimization step, which will be discussed hereafter.

Digital removal of substrate contributions. Sometimes substrate contributions originating from components such as glass or wax are present in the spectral data. These effects can be mitigated or eliminated through digital filters, such as digital de-waxing (de Lima *et al.*, 2017; Tfayli *et al.*, 2009). For example, the extended multiplicative signal correction (EMSC) algorithm has been reported to neutralise variability caused by paraffin signal and allow selection of unique spectral features related to the sample composition in vibrational spectroscopy (de Lima *et al.*, 2017; Tfayli *et al.*, 2009); independent component analysis (ICA) and non-negatively constrained least squares (NCLS) are also common methods of digital de-waxing for vibrational spectroscopy (Ibrahim *et al.*, 2017; Meksiarun *et al.*, 2017; Tfayli *et al.*, 2009). Glass contributions are reduced in the high-wavenumber region of the mid-IR spectrum, thus allowing spectral data analysis within the region between ~2500–3800 cm^{-1} (Bassan *et al.*, 2014); and can be reduced in NIR spectroscopy by subtracting the glass spectrum from the sample spectrum and by working in the wavelength range of 1850–2150 nm (Paraskevaidi *et al.*, 2018b).

Smoothing. Smoothing is made by spectral filters that remove random noise while preserving useful spectral information. The most used smoothing technique is the Savitzky-Golay (SG) algorithm (Savitzky & Golay, 1964). It is based on a polynomial equation fitted in a least squares sense within a pre-defined interval of spectral points, where the central point from the interval is removed and used as a fitting criteria. This interval is then displaced to the next point of the spectrum and the fitting procedure is repeated. SG smoothing is excellent to remove large instrumental noise and can be applied to any type of vibrational spectroscopy technique. Its major disadvantage is that the polynomial order and the window size used in the polynomial fitting affect the result, so one must use an polynomial order similar to the spectral shape features (*e.g.*, 2^{nd} order polynomial for vibrational spectroscopy data), and the window size must be an odd number not too small (which keeps the noise) nor too large (changing the spectral shape) (Morais *et al.*, 2019c). In addition, moving-window mathematical pre-processing techniques introduce correlations in the noise structure, and this may complicate the use of chemometric models assuming that the noise is identically and independently distributed (iid) (Geladi *et al.*, 1985).

Light scattering correction. Light scattering is present when particles with different sizes, especially smaller than the spectral electromagnetic wavelength, is present on the material being analysed. This shifts the absorbance or spectral intensity in a systematic fashion, affecting the *y*-axis. Light scattering effects are also generated by different probe pressures when portable spectrometers are used to analyse solid samples, or by different lengths of optical path. Light scattering is very common in NIR spectroscopy, and some techniques such as multiplicative scatter correction (MSC) (Geladi *et al.*, 1985) and standard normal variate (SNV) (Barnes *et al.*, 1989) can be used to correct this problem. MSC corrects light scattering (Mie scattering) maintaining the original spectral shape and the same spectral scale. As main disadvantage, it needs a reference spectrum representative of all measurements. Usually, this reference spectrum is not available, and then it is substituted by the average spectrum across all training samples (Morais *et al.*, 2019c). SNV also corrects light scattering (Mie scattering) maintaining the original spectral shape with no need of a reference spectrum, but it creates an artificial absorbance scale with negative values since the data are centralized to zero in the *y*-scale (Morais *et al.*, 2019c). Resonant Mie scattering is also a frequent issue in IR spectroscopy of biological materials (Bassan *et al.*, 2009), where a dispersion artefact occurs through a light scattering when there is simultaneous absorption. This can be often observed by a severe baseline distortion followed by a systematic shift in the *y*-axis (Bassan *et al.*, 2009; Bassan *et al.*, 2010). Bassan *et al.* (2009, 2010) have proposed a modified version of the EMSC algorithm to correct for resonant Mie scattering, named the RmieS-EMSC algorithm. Additionally, Mie scattering (elastic scattering) is also present in Raman spectroscopy (Kiefer *et al.*, 1997), contributing to baseline distortions which are often mistakenly assigned to a fluorescence background. This can be also corrected by applying a modified version of the EMSC algorithm through the addition of polynomial extensions to the basic EMSC algorithm in order to correct for fluctuating baseline features (Liland *et al.*, 2016).

Baseline correction. Baseline correction techniques remove background absorptions interferences. Baseline distortions are commonly present in all types of vibrational spectroscopy techniques. For NIR, the baseline distortions are mainly a result of light scattering, which can be corrected by MSC or SNV; however, for IR and Raman spectroscopy, this effect is more apparent, especially in the latter due to fluorescence interferences. There are several techniques of baseline correction, most of them already

included in spectrometers' software, in which the main ones are the rubber-band baseline correction, polynomial baseline correction, asymmetric least squares (ALS), automatic weighted least squares (AWLS), and Whittaker filter. The baseline correction technique being chosen affect the final result, therefore, they must be kept consistent throughout all the data analysis, especially if new samples are added after the model is developed.

Spectral differentiation. First and second derivatives can be applied to the spectral data in order to correct both light scattering and baseline distortions. Also, these techniques highlight smaller spectral differences between the samples' spectra, which can be critical to find distinctive spectral features amongst complex samples. It can also be coupled to SG smoothing in a single routine, making these procedures computationally easier. However, derivatives have great disadvantages. Spectral differentiation is not indicated to correct for resonant Mie scattering since it does not correctly deal with these spectral distortions. The order of the derivative function must be chosen carefully to avoid increasing the noise level too much. In addition, derivatives using moving-window procedures also carry the same risk of introducing correlations in the noise structure discussed to smoothing techniques, which affects the use of chemometric models assuming that the noise is iid. Also, derivatives change the spectral scale (*y*-axis scale) to mathematical coefficients instead of absorbance, thus the spectral intensity of derivative bands cannot be used for direct correlation with chemical concentrations; and spectral markers (biomarkers) identification needs to be performed carefully, since derivatives shift the spectral band positions in $i \times d$ wavenumbers, where i is the derivative order and d is the data spacing resolution. Some software automatically correct this spectral shift by deleting the first i wavenumber position(s), thus matching the size of the spectral response (derivative result) with the reference wavenumber vector; but some software do not offer this correction.

Normalisation. Spectral normalisation techniques are commonly employed in IR and Raman spectroscopy to correct for different sample thickness and concentration, hence, avoiding the influence of non-desired spectral signatures among the samples. However, this procedure must be performed only when needed and with care, since the normalisation might hide important spectral bands that could be discriminant features among the samples, such as amide I and amide II absorptions; and it also may introduce non-linearities to the data (Morais *et al.*, 2019c). Amide I and vector normalisation are the commonest type of normalisation for IR data; the first can be used when the amide I

band is not a distinguishing feature between the classes; and the latter when this information is unknown but different sample thickness or concentration correction is needed.

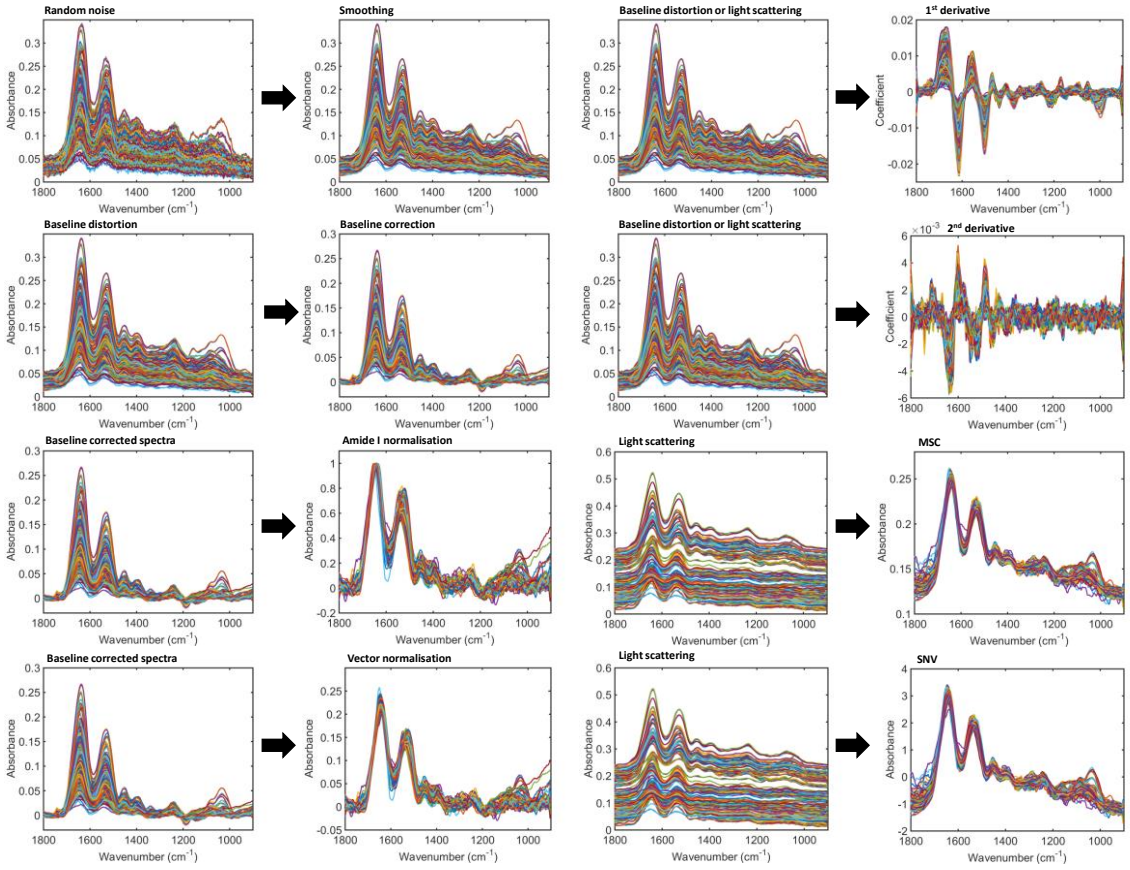


Figure 2.2. Effect of different pre-processing applied to an IR dataset. MSC: multiplicative scatter correction; SNV: standard normal variate.

Raman spectrometers using CCD detectors also surfers from cosmic rays interferences, which create spikes in the spectral data compressing important Raman spectral signatures. Spikes removal is an essential step when analysing Raman data and must be performed before any data pre-processing. Most Raman spectrometers' software have spikes removal routines. Finally, scaling methods (also referred as "standardization" by Hastie *et al.* (2009)) are fundamental when dealing with multivariate methods, in particular PCA and partial least squares (PLS). Mean-centring is a very reasonable approach to use with spectral data before modelling, after which all variables in the dataset will have zero mean. When data contain information represented by different scales (*e.g.*, after data fusion using both IR and Raman spectra), block-scaling should be used. In this case, each block of data (*i.e.*, data for each instrumental technique) would

have the same sum of squares (normally after mean-centring) (Morais *et al.*, 2019c). For discrete data with different scales, autoscaling is recommend.

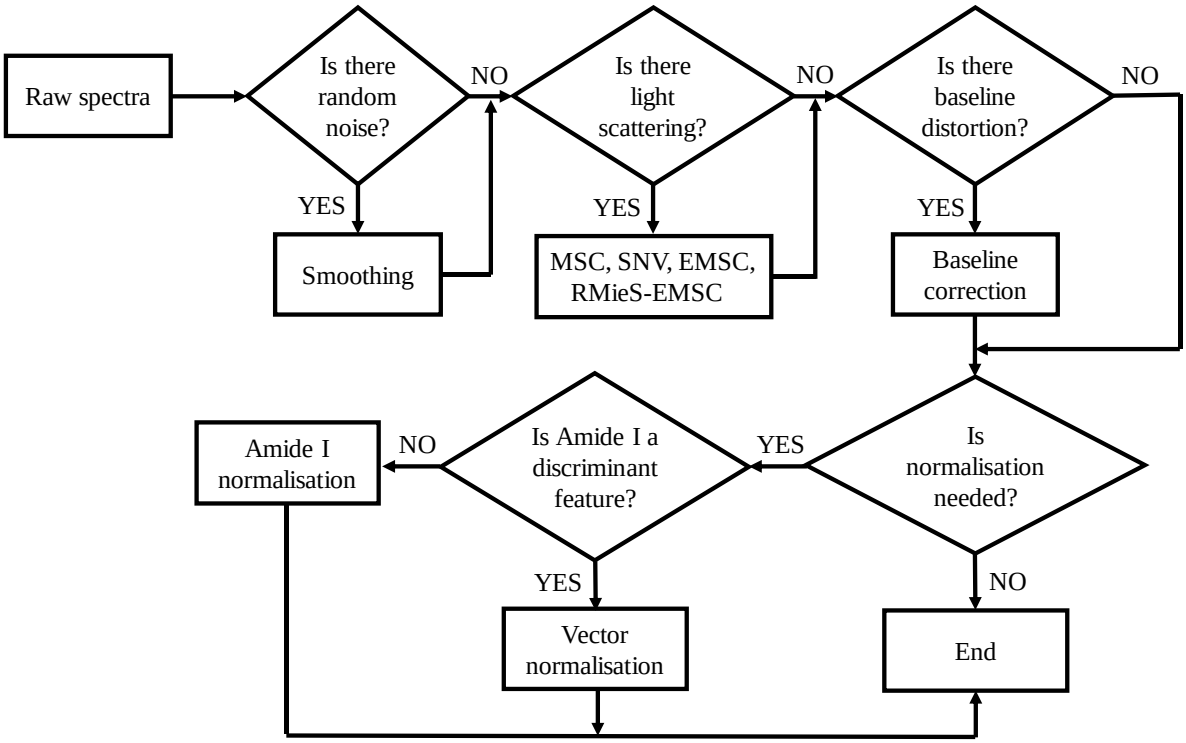


Figure 2.3. Decision tree to define the pre-processing technique for a spectral dataset. MSC: multiplicative scatter correction; SNV: standard normal variate; EMSC: extended multiplicative signal correction; RmieS-EMSC: resonant Mie scattering – extended multiplicative signal correction.

2.2.3 Outlier Detection

When analysing real data, it often occurs that some observations are different from the majority. Such observations are called outliers. The spectral signal for some samples might differ from the spectral signal for the majority of the samples being measured. This can happen either by substantial differences in chemical structure or concentration for these specific samples, or by a measurement error. In the first case, we usually refer to an extreme sample, that is, a sample that belongs to the measurement set but with an extreme property value. This sample is characterized by a high Hotelling's T^2, and usually does not skew the model in a high degree; although it is recommended exclusion of this sample

from the dataset before modelling. In the latter case, when the spectral abnormality is caused by a measurement error, this sample is a true outlier, being characterized by a high Q residuals. This sample should be removed from the dataset before analysis. Both extreme samples and outliers should be investigated in order to find possible sources of abnormalities.

There are several techniques for outlier detection, such as the Jack-knife (Martens & Martens, 2000), Z-score (Rousseuw & Hubert, 2011) and K-mode clustering (Jiang *et al.*, 2016). However, one of the most popular and visually intuitive technique for outlier detection is the Hotelling's T^2 *versus* Q residuals test (Bakeev, 2010). In this test, a chart is created using the Hotelling's T^2 values (sum of the normalised squared scores, which is the distance from the multivariate mean to the sample projection onto the PCA principal components (PCs) space) in the *x*-axis and the Q residuals (sum of squares of each sample in the PCA error matrix, representing the residuals between a sample and its projection onto the PCs space) in the *y*-axis, generating a scatter plot (Morais *et al.*, 2019c). All samples far from the origin of this chart are considered candidates to outliers and should be investigated and removed. Samples should be removed one at a time, since PCA is highly influenced by the samples that are included in the model. Samples with high values in both Hotelling's T^2 and Q residuals are the outliers with the greatest effect in PCA, while the samples with high values in only one of these parameters are the outliers with the second-greatest effect on the PCA model (Morais *et al.*, 2019c). Figure 2.4 illustrates 4 outliers detected amongst a set of 700 IR spectra by a Hotelling's T^2 *versus* Q residuals chart applied to the pre-processed data (AWLS baseline correction and vector normalisation). The outliers were spectra corresponding to background noise measured within the experimental set, most likely by a mistake made by the analyst when placing the samples in the attenuated total reflection (ATR) apparatus.

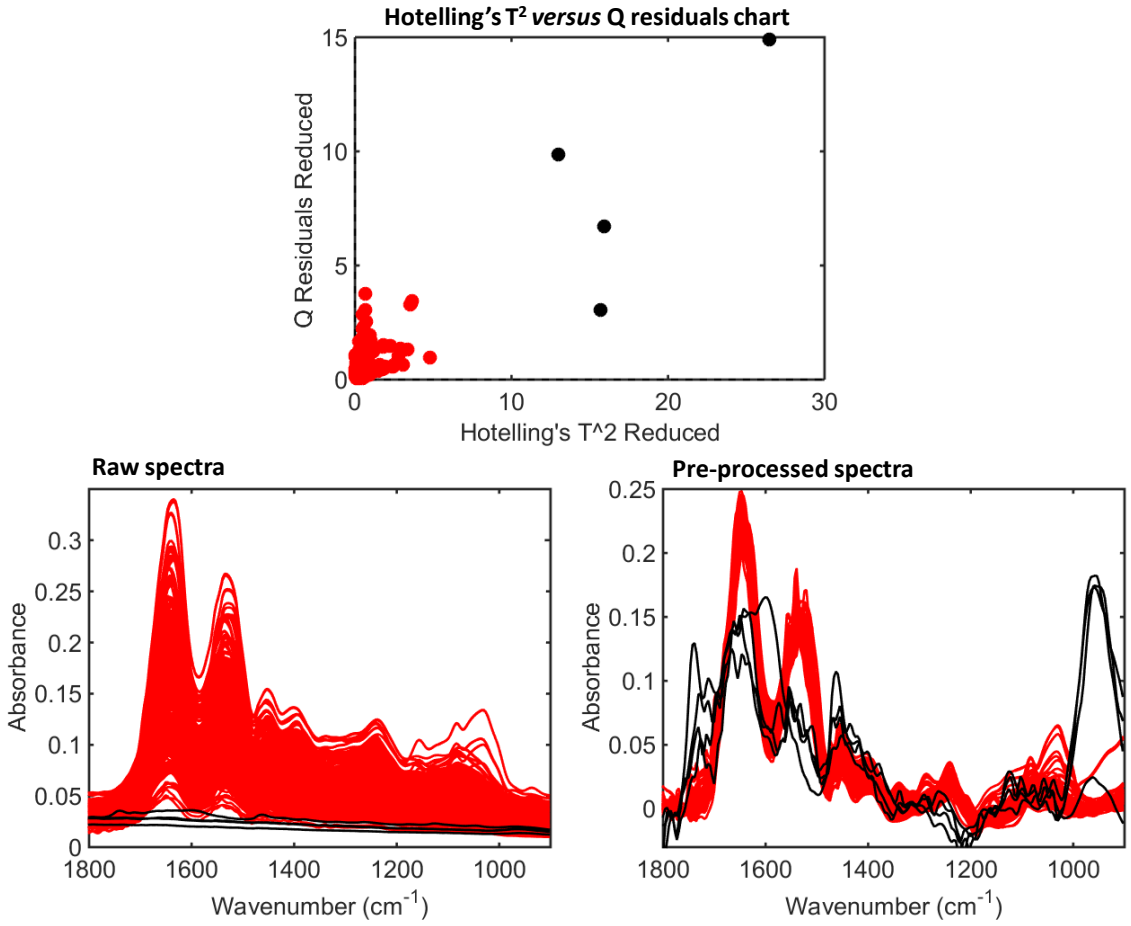


Figure 2.4. Outlier detection test by a Hotelling's T^2 *versus* Q residuals chart. Pre-processing: AWLS baseline correction and vector normalisation. PCA model built with 8 PCs (94.3% cumulative explained variance). Spectra in black: outliers.

2.2.4 Data Selection

A fundamental step towards building predictive chemometric models is data selection, that is, splitting an initial experimental dataset into at least two subsets: training and test. The training set contemplates the major fraction of the samples and is used to build the classifier, whereas the test set includes the remaining fraction of samples and is used to evaluate the model classification performance, since, although they are measured during the same experiment, the test set is considered external to the model (blind), thus reflecting the expected model behaviour toward new observations (Morais *et al.*, 2019d). When two subsets are used, cross-validation is recommended to optimize the model parameters. Cross-validation uses samples from the training set to optimize model parameters, such as the number of PCs in PCA-based models or latent variables (LVs) in PLS-based models, in an iterative internal validation process. This is made by first

removing a certain number of samples from the training set and then building the model with the remaining samples, where the removed samples are predicted as a temporary validation set (Morais *et al.*, 2019d). This is performed for a certain number of repetitions usually until all training samples are excluded once from the training set and predicted as an external validation set. One of the most popular cross-validation methods is the leave-one-out cross-validation (also called leave-one-spectrum-out cross-validation). In this case, only one sample spectrum is removed from the training set per each interaction. Although much used, leave-one-out cross-validation is only indicated for small size datasets, usually with no more than 20 samples in the training set (Morais *et al.*, 2019c). When this number is larger, other cross-validation approaches are recommended, such as venetian blinds or random subset selection. When there are replicate spectra, leave-one-spectrum-out cross-validation should not be used at all, but rather a continuous-block cross-validation (also called leave-one-patient-out cross-validation when the number of replicate spectra is equal for each sample and organised in a sequential way within the spectral matrix **X**), otherwise during the cross-validation procedure the training and temporary validation sets will have spectra from the same sample, hence, giving overoptimistic cross-validation results. In continuous-block or leave-one-patient-out cross-validation, the whole set of replicas for a same sample is transferred to the temporary validation set during cross-validation, thus the training and temporary validation sets will not have spectra for the same sample, hence, giving more realistic results.

When a large number of samples is measured, generally more than 100, it is recommended to split the experimental dataset into three groups: training, validation and test. In this case, an extra validation set that does not contain training samples is used to optimize the model. It is important to stress that the training, validation and/or test sets cannot contain spectra of the same sample distributed among them, *i.e.*, the samples in each set must be independent. Extra caution must be taken when multiple spectral replicates are used to feed the model, to ensure that they do not overlap in different sets.

There are several ways to split the samples into training, validation and/or test sets. Manual splitting is not recommend, since it can introduce bias to the model. Thus, computational-based methodologies are recommended instead. Random-selection and the Kennard-Stone (KS) algorithm (Kennard & Stone, 1969) are some well-known approaches. Random-selection is a simpler approach where samples are assigned to the

training, validation and/or test sets randomly. The KS algorithm works based on a Euclidian distance calculation by first assigning the sample with the maximum distance from all other samples to the training set, and then by selecting the samples that are as far away as possible from the selected sample to this set, until the designed number of selected samples is reached. This ensures that the training model will contain samples that uniformly cover the complete sample space, for which no or minimal extrapolation of the remaining samples is necessary (Morais *et al.*, 2019c). KS has been proved to be superior than random-selection alone (Morais *et al.*, 2019d), but for biological-derived samples, we have recently proposed a modification to the KS algorithm by adding a small degree of randomness to it, where the model predictive performance increased; this is called the MLM algorithm (Morais *et al.*, 2019d).

2.2.5 Modelling

Exploratory analysis is the first step towards analysing complex spectral data, where the analyst can initially assess the data in order to identify clustering patterns and trends, thus helping them to draw conclusions about the nature of samples, outliers and experimental errors (Morais *et al.*, 2019c). PCA is the most common method of exploratory analysis, in which the pre-processed spectral data are decomposed into a few number of PCs responsible for the majority of the variance within the original dataset. The PCs are orthogonal to each other and are generated in a decreasing order of explained variance, so that the first PC explain most of the data variance, followed by the second PC and so on (Bro & Smilde, 2014). PCA decomposition takes the form:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \qquad (2.1)$$

where **X** represents the pre-processed spectral data, **T** is the PCA scores, **P** the loadings, and **E** the residuals.

PCA is then often the first step of the data analysis, followed by classification, cluster analysis, or other multivariate techniques. The PCA scores represent the variance in the sample direction, being used to detect clustering patterns related to chemical similarities/dissimilarities between the samples. The PCA loadings represent the variance in the wavenumber direction, being used to identify spectral variables with high degree of importance for the pattern observed in the scores distribution (Morais *et al.*, 2019c).

The PCA loadings are commonly used for searching spectral markers that distinguish samples from different biological classes. This can be performed by identifying the spectral bands with the highest absolute loadings coefficients (positive or negative) on the discriminant PCs directions (Morais *et al.*, 2019c). Another strategy proposed by Martin *et al.* (2007) is the cluster vector approach, where a median score is calculated for each of the 3 PCs that represent the best samples' clustering in a three-dimensional space and, thereafter, the three loading vectors for these PCs weighted by the median score are summed. As a result, a new loading vector is generated representing the effective loadings profile for the clustering (Martin *et al.*, 2007; Santos *et al.*, 2017). The PCA residuals represent the difference between the decomposed and original pre-processed data, being used to identify experimental errors. Ideally, the PCA residuals should be random and close to zero (homoscedastic distribution); otherwise they indicate experimental bias (Morais *et al.*, 2019c).

PCA is a fast, intuitive and reliable method to identify differences between spectral data, however, it is important to stress that PCA is not a classification technique. PCA is a data reduction and exploratory analysis method, but PCA solely cannot be used to systematically classify samples. For this, supervised classification techniques are needed. Supervised classification methods build computer-based classifiers that can predict future samples based on their training spectral profiles. Therefore, data selection as mentioned previously is fundamental before building supervised classification models.

Classification methods are separated into two groups: one-class modelling (also called class-modelling) and discriminant models. In one-class modelling, the classification model output does not solely depend on the training classes, thus it can assume values such as "unknown" or that the test sample does not belong to any of the training classes. On the other hand, in discriminant models the model outputs are always referent to one of the training classes. The one-class modelling approach is very useful when only one class is modelled and the model output is whether the sample belongs or not to the reference class. However, one-class approaches requires a large number of samples, since the class boundaries must include all the sample space as much as possible, and usually provides worst classification results than discriminant models, since in one-class modelling slightly extreme samples to the reference class could be interpreted as not belonging to the class. These problems are not present in discriminant models, since the model output is always one of the training classes, and the class space is much larger.

Discriminant approaches cannot predict samples that do not belong to the training classes, thus building meaningful training sets is fundamental in this type of analysis.

The main algorithm of one-class modelling is the soft independent modelling by class analogy (SIMCA) (Wold & Sjöström, 1977). In SIMCA, each class is modelled by an independent PCA model of opportune dimensionality. Then, the class space is defined according to some statistically defined outlier detection criterion, which is often the distance-to-the-model criterion (Marini, 2010). This is made by calculating the probability distributions for the T^2 statistics and Q statistics for the PCA model of each class, where a threshold corresponding to a determined confidence level (usually 95%) is chosen for both statistics to define the class space (Marini, 2010). Other ways to define the class space are possible, such as the method proposed by Pomerantsev (2008), although the T^2 and Q statistics is the most common approach.

There are several discriminant analysis algorithms, most of them based on distance calculations on the real or transformed sample space. The main discriminant analysis algorithms employed in biological-derived spectrochemical applications will be discussed below.

Linear discriminant analysis (LDA). LDA is a discriminant analysis algorithm based on a Mahalanobis distance calculation between the samples for each class (Dixon & Brereton, 2009). This calculation can be performed with or without Bayesian probability terms, which can be applied when classes have different sizes (Dixon & Brereton, 2009). LDA uses the pooled variance-covariance matrix in the distance calculation, hence, the distance between a test sample and a given class centroid is weighted according to the overall variance of each spectral variable (Dixon & Brereton, 2009). This is particularly useful when the classes have similar variance structures or when the sample size is small (Wu *et al.*, 1996). However, LDA is highly affected when classes have different variance structures, which often happens in complex biological medium. In addition, LDA is a parametric method that assumes the samples follow a normal distribution and cannot be applied to ill-conditioned data, *e.g.*, when the number of spectral variables is larger than the number of samples (Morais *et al.*, 2019c). Although spectral data usually do not perfectly follow a normal distribution, LDA is robust enough to handle spectroscopy data and, according to the CLT, this effect can be reduced by increasing the sample size. The issue related to ill-conditioned data can be solved by the application of PCA or variable

selection techniques to the pre-processed spectral data prior LDA, such as the principal component analysis linear discriminant analysis (PCA-LDA) algorithm, where LDA is applied to the PCA scores (Morais & Lima, 2018).

Quadratic discriminant analysis (QDA). Similarly to LDA, QDA is a discriminant analysis algorithm based on a Mahalanobis distance calculation between the samples for each class, which can use Bayesian probability terms to correct for classes having different sizes (Dixon & Brereton, 2009). However, differently from LDA, QDA forms a separate variance model for each class, thus using a different variance-covariance matrix for each class (Dixon & Brereton, 2009). For this reason, QDA outperforms LDA when classes exhibiting different within-category variances are being analysed (Morais *et al.*, 2019a). Like LDA, QDA is also a parametric method that is highly affected by ill-conditioned data; however, these issues can be solved in the same manner as described for LDA. Often, QDA is applied to the PCA scores in the principal component analysis quadratic discriminant analysis (PCA-QDA) algorithm (Morais & Lima, 2018). Its main disadvantages are that QDA underperforms LDA for small size datasets and it has a higher risk of overfitting than LDA (Morais *et al.*, 2019a; Morais *et al.*, 2019b; Wu *et al.*, 1996).

Partial least squares discriminant analysis (PLS-DA). PLS-DA (Brereton & Lloyd, 2014) is a feature extraction and classification algorithm that usually performs better than PCA followed by LDA, as the scores from PCA do not necessarily describe the difference between the samples, but the variance in the spectral data (Morais *et al.*, 2019b). In PLS-DA, a PLS model (Geladi & Kowalski, 1986) is applied to the pre-processed spectral data reducing the original spectral variables to a small number of latent variables (LVs), where then a linear discriminant classifier is used for classifying the groups (Hibbert, 2016). It is important to stress that there are different ways of performing the PLS model, such as using the SIMPLS (de Jong, 1993) or the non-linear iterative partial least squares (NIPALS) algorithm (Wold *et al.*, 2001), and that the classification rule of PLS-DA vary according to the application or software being used. Linear classifiers based on Euclidian distance to centroids (Brereton & Lloyd, 2014), LDA (Pomerantsev & Rodionova, 2018) and Bayesian decision rule (Pérez *et al.*, 2009) are some examples that can be used in PLS-DA. In addition, PLS-DA can be adapted to one-class modelling, as described by Pomerantsev and Rodionova (2018). The main disadvantage of PLS-DA is that this algorithm is greatly affected by classes having different sizes and it requires optimization

of the number of LVs, which is often performed by cross-validation (Morais *et al.*, 2019b). Also, it is important to highlight that PLS-DA is a binary classifier, that is, when more than two classes are analysed a PLS2 model is built where the classes are codded in a matrix with size *m* (rows, samples) x *n* (columns, classes) containing zeros when the sample does not belong to the target class and ones when the sample belong to the target class. In PLS-DA, class-coding cannot be made in a sequential manner (*e.g.*, 1, 2, 3, …) since this imply a distance relationship between the samples (*e.g.*, samples from class 1 are farther from class 3 than the samples in class 2). Some softwares allow the input of sequential class-coding, but this information is internally convert into a zeros and ones matrix before model construction.

K-nearest neighbours (KNN). KNN (Cover & Hart, 1967) is a local non-parametric classification method where samples are classified based on the "majority vote" approach, that is, a given test sample spectrum is projected in a feature space and based on the calculation of a distance or dissimilarity metric (Manhattan, Euclidian, Minkowski or Mahalanobis distance; or by correlation), depending on the number of nearest surrounding neighbour training samples to this test sample, the sample is classified towards the majority observed class. The main advantage of KNN is that it can be applied to almost all type of data independent of its probability distribution or condition number, and does not require a particular ratio between the number of samples and the number of spectral wavenumbers (Marini, 2010). KNN main disadvantages are that the model tends to overfit by skewing towards the bigger class size when unequal classes sizes are analysed, and that the model is highly sensitivity towards random spectral noise and to the "curse of dimensionality" (Marini, 2010; Morais *et al.*, 2019c). In addition, KNN requires the optimization of the distance calculation method and the *k* value (number of neighbours), which can be performed through cross-validation (Morais *et al.*, 2019c).

Support vector machines (SVM). SVM is a binary linear classifier with a non-linear step called the kernel transformation (Cortes & Vapnik, 1995). A kernel function transforms the input spectral space into a feature space by applying a mathematical transformation which is often non-linear. Then, a linear decision boundary is fit between the closest samples to the border of each class (called support vectors), thus defining the classification rule. Although being highly accurate to classify spectral data, SVM requires many parameters optimization, such as the type of kernel function and its parameters, and it is highly susceptible to overfitting; besides being a highly time-consuming algorithm

(Morais *et al.*, 2019c). The radial basis function (RBF) kernel is often the best kernel to use in SVM, since it can adapt to different data distribution. To avoid overfitting, cross-validation should be always performed to estimate the best kernel parameters (Morais *et al.*, 2019c). Multiclass SVMs are possible through the implementation of approaches such as one *vs.* all, one *vs.* one, fuzzy rules and directed acyclic graph trees (Brereton & Lloyd, 2010), although the first approach is the most common.

When data complexity increases, for instance, when the spectra data are not following a bilinear rule or when the components complexity are too excessive to be analysed by the previous methods, "black box" algorithms, *i.e.*, machine learning techniques where the classification rules are hard to interpret, can be applied. Most of these algorithms were developed for applications such as face recognition, where a high degree of non-linearity is observed between the measurements, but they have found their way into spectrochemical applications. Artificial neural networks (ANN) (Marini *et al.*, 2008), random forests (Fawagreh *et al.*, 2014) and deep-learning approaches (LeCun *et al.*, 2015) are common classification methods applied in such situations. All these techniques have a non-linear classification nature and higher accuracy in comparison with more simpler methods, however in order for these algorithms work properly with a low-risk of overfitting many parameters need to be optimised, which depends on the analyst skills and usually demands high computation cost (Morais *et al.*, 2019c). Classification techniques should be used in a parsimonious order (Seasholtz & Kowalski, 1993), in which the simplest algorithms should be performed first before testing more complex algorithms. A suggested order for running these classification algorithms is: LDA > PLS-DA > QDA > KNN > SVM > ANN > Random forests > deep-learning approaches (Morais *et al.*, 2019c).

Apart from these discriminant methods, there are many other discriminant analysis algorithms that are known but not much applied to biological-derived spectral data, such as learning vector quantization (LVQ) (Dixon & Brereton, 2009) and regularized discriminant analysis (RDA) (Wu *et al.*, 1996). These algorithms are not much used probably due to the lack of available software containing these routines. The main chemometric softwares for classification applications are shown in Table 2.1. Apart from these main softwares, there are many open source freely-available options with specific algorithms for classification of spectral data, such as the MultiDA (Yang *et al.*, 2012) toolbox for MATLAB that contains some classification routines; the Biodata (De

Gussem *et al.*, 2009) toolbox for MALTAB that contain PCA-LDA routines; the SAISIR (Cordella & Bertrand, 2014) toolbox that contains PCA-LDA, PLS-DA and QDA routines; the ParLeS (Rossel, 2008) software that contain routines including PCA and PLS; the Raman Processing Program (Reisner *et al.*, 2011) that contains LDA, ANN and SVM routines; the PML (Jing *et al.*, 2014) toolbox for machine learning; the DD-SIMCA (Zontov *et al.*, 2017) toolbox for MATLAB that contain SIMCA routines; the libPLS (Li *et al.*, 2018) library for MATLAB that contain PLS-DA routines; and the LIBSVM (Chang & Lin, 2011) library for SVM. Other classification routines can be found in spectrometer softwares or by specific libraries or toolbox available online for MATLAB, Octave, Scilab, R and Python.

Octave and Scilab are open source freely-available platforms with syntax very similar to MATLAB and may interchangeable routines (Alsberg & Hagen, 2006), *i.e.*, routines made for MATLAB often work in Octave or Scilab. Freely-available chemometric toolboxes for Octave and Scilab include the SAISIR toolbox (Cordella & Bertrand, 2014) (https://www.chimiometrie.fr/saisir_webpage.html) and the FACT (Free Access Chemometrics Toolbox) (https://www.scilab.org/fact-free-access-chemometrics-toolbox). R is another powerful open source statistical platform often used for chemometric applications (Wehrens, 2011). Apart from the CAT (Chemometric Agile Tool) toolbox showed in Table 2.1, there are other freely-available chemometrics packages for R, such as the Chemometrics package (https://www.rdocumentation.org/packages/chemometrics/versions/1.4.2) and the CRAN package Chemometrics (Varmuza & Filzmoser, 2009) (https://cran.r-project.org/web/packages/chemometrics/index.html). Python is a high-level computer programming language which is also becoming popular for chemometric applications. Jarvis *et al.* (2006) developed an open source chemometric toolbox named PYCHEM for multivariate analysis of spectral data using Python. PYCHEM is freely-available at http://pychem.sourceforge.net/.

Table 2.1. Main chemometric softwares for multivariate classification.

Software	Website	Classification algorithms	Availability
IrootLab (Trevisan *et al.*, 2013)	http://trevisanj.github.io/irootlab/	LDA, PCA-LDA, ANN, fuzzy classification, KNN, logistic regression, SVM, PCA-SVM, binary decision trees.	Free
Classification Toolbox for MATLAB (Ballabio & Consonni, 2013)	http://www.michem.unimib.it/	LDA, QDA, PCA-LDA, PCA-QDA, classification trees (CART), PLS-DA, SIMCA, unequal class models (UNEQ), potential functions, SVM, KNN, backpropagation neural networks.	Free
CAT (Chemometric Agile Tool)	http://gruppochemiometria.it/index.php/software	LDA, QDA, KNN.	Free
PLS_Toolbox	http://www.eigenvector.com/	PLS-DA, SVM, SIMCA, KNN.	Commercial
Statistics and Machine Learning Toolbox for MATLAB	https://mathworks.com/	Binary decision trees, LDA, QDA, naïve Bayes classifier, KNN, SVM, random forest.	Commercial
Unscrambler X	https://www.camo.com/unscrambler/	SIMCA, LDA, PLS-DA, SVM.	Commercial
Pirouette	https://infometrix.com/	KNN, SIMCA, PLS-DA.	Commercial
SIMCA Umetrics	https://umetrics.com/	SIMCA, PLS-DA, orthogonal partial least squares discriminant analysis (OPLS-DA).	Commercial

2.2.6 Feature Extraction and Selection

The feature extraction stage is responsible for producing a smaller number of variables that are more informative than the original whole set of wavenumber/variables. Feature selection is commonly applied as a stage prior to classification as a means to prevent overfitting and to circumvent the "curse of dimensionality". Feature extraction

and selection can be used to reduce data complexity, to reduce redundant information, to speed computation-time, and to aid biomarkers identification. Feature extraction techniques allow one to extract spectral features related to important chemical components within the spectral dataset, and feature selection techniques significantly reduces the pre-processed spectral dataset to a small set of variables responsible for class differentiation. The most straightforward way to identify important spectral features is by plotting the colour-coded mean-centred data. As mentioned before, mean-centring is a key scaling step applied to the data before multivariate analysis. By plotting the mean-centred data, the analyst can see spectral regions where the general data trend between the classes diverge. Spectral regions where one can clearly see a spectral difference are often the most important spectral regions within the dataset. These regions can be selected for model construction; though it is a trial and error procedure. *I.e.*, the analyst needs to evaluate the model validation performance using the whole spectrum, some selected spectral regions, and spectral variables selected by other methods. Also, knowing the nature of the phenomena being measured can aid and guide the analyst to select important spectrochemical features.

Feature extraction techniques can also be used directly or indirectly to identify important spectral variables. PCA and PLS-DA are two very common feature extraction techniques. PCA loadings and the PLS-DA regression coefficients can be used to identify important spectral features. Some approaches based on PLS such as variable importance in projections (VIP) scores (Ferrés *et al.*, 2015), selectivity ratio (Ferrés *et al.*, 2015) and interval partial least squares (iPLS) (Nørgaard *et al.*, 2000) are useful tools to find important spectral variables. VIP scores, selection ratio and iPLS are very different techniques where the first two are methods employing the PLS loadings or the regression coefficients, which carry some risk of misinterpretation, since regression vectors of inverse calibration methods, such as PLS, can exhibit extremely complex behaviour in even the most simplistic circumstances (Brown & Green, 2009). On the other hand, methods employing the minimum error as guidance, such as iPLS, do not carry this risk and are in some degree more reliable for qualitative spectral interpretation.

Another feature extraction technique particularly useful for hyperspectral imaging data is the multivariate curve resolution alternating least squares (MCR-ALS) algorithm (de Juan & Tauler, 2006; Jaumot *et al.*, 2015). MCR-ALS can be applied to reduce the spectral dataset into important spectral features and aid biomarker identification through

the interpretation of the concentration and spectral profiles containing the purest components in the experimental spectral matrix. MCR-ALS allows the introduction of chemical constraints in the factor analysis model, improving the resolution of spectral mixtures through adjusting chemically-meaningful parameters; and allow the construction of concentration distribution maps based on the recovered MCR-ALS concentration profiles for hyperspectral images datasets (de Juan *et al.*, 2004). For first-order spectral data, MCR-ALS can also be applied to solve mixtures and to recover concentration and pure spectral profiles towards kinetic, quantitative or qualitative applications. However, MCR-ALS is limited by the fact that its bilinear decomposition may show a large degree of rotational ambiguity, precluding the obtainment of reliable results.

Feature selection techniques allows the identification of specific spectral wavenumbers within the original spectral space responsible for maximizing the class differences. Some examples of feature selection algorithms include the minimum redundancy maximum relevance (MRMR) algorithm (Radovic *et al.*, 2017), in which the variable selection process is based on the maximization of the relevance of extracted features and simultaneous minimization of redundancy between them (Morais *et al.*, 2019c); the successive projections algorithm (SPA) (Soares *et al.*, 2013), which is an iterative forward feature selection method operating by solving co-linearity problems within the spectral dataset, thus selecting wavenumbers whose information content is minimally redundant (Theophilou *et al.*, 2018); and the genetic algorithm (GA) (McCall, 2005), which is a iterative combinational algorithm inspired by Mendelian genetics where a set of initial variables undergo selection, cross-over combinations and mutations until the fittest selected variables are found (Theophilou *et al.*, 2018). The variables selected by these techniques can be used as input for the classification methods described previously, which is important since these techniques reduce the data size and collinearity, hence, improving the model accuracy and analysis time. Adaptations of PCA, PLS, SPA and GA (as feature extraction/selection techniques) to LDA, QDA, SVM, KNN and ANN (as classifiers) are well known (Siqueira & Lima, 2016a; Siqueira & Lima, 2016b; Siqueira *et al.*, 2018).

2.2.7 Model Validation

The performance of any classification model must be validated by calculating some quality metrics, or figures of merit, for a validation or test set. The training or cross-validation performance reflects the model fitting but it does not reflect the model predictive performance towards unknown samples. For this reason, figures of merit such as accuracy (AC), sensitivity (SENS), specificity (SPEC), positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (LR+), negative likelihood ratio (LR-), F-Score and G-Score are often calculated for external validation or test sets (Morais & Lima, 2017; Siqueira *et al.*, 2017). The equations to calculate these parameters are depicted in Table 2.2.

For binary models, *i.e.*, models containing two classes, these parameters are calculated only once, where the positive class is the class of interest (*e.g.*, disease) and the negative class is the control class (*e.g.*, healthy controls). When more than two classes are modelled, then these parameters must be calculated individually per class. Often, receiver operating characteristic (ROC) curves, including the area under the curve (AUC) value; and confusion matrices containing the predicted number of samples or predicted classification rate per class are reported in a form of a table or graphically to aid the reader evaluating the model classification performance. Also, based on the confusion matrix, the Cohen's kappa coefficient (κ) (Warrens, 2011) can be calculated, which is a weighted average of the model performance. Other parameters, such as model uncertainty, can be calculated. Uncertainty is related to the probability of misclassification and model robustness (Morais *et al.*, 2019b), and can be calculated for LDA, QDA and SVM models (Morais *et al.*, 2019b); PLS-DA (de Almeida *et al.*, 2013; Rocha & Sheen, 2016); and ANN (Allegrini & Olivieri, 2016). The number of quality metrics to report depends on the application and rigor of the study. We recommend reporting at least the accuracy, sensitivity and specificity for small studies, while all the metrics in Table 2.2 can be reported for bigger studies.

Table 2.2. Quality parameters to evaluate the model classification performance. TP stands for true positives, FP for false positives, TN for true negatives, and FN for false negatives.

Parameter	Equation	Meaning
Accuracy (AC) / %	$\frac{TP + TN}{TP + FP + TN + FN} \times 100$	Number of samples correctly classified considering true and false negatives. Optimal value: 100%.
Sensitivity (SENS) / %	$\frac{TP}{TP + FN} \times 100$	Proportion of positive samples (*e.g.*, disease) that are correctly classified. Optimal value: 100%.
Specificity (SPEC) / %	$\frac{TN}{TN + FP} \times 100$	Proportion of negative samples (*e.g.*, healthy controls) that are correctly classified. Optimal value: 100%.
Positive predictive value (PPV) / %	$\frac{TP}{TP + FP} \times 100$	Number of test positives that are true positives. Optimal value: 100%.
Negative predictive value (NPV) / %	$\frac{TN}{TN + FN} \times 100$	Number of test negatives that are true negatives. Optimal value: 100%
Positive likelihood ratio (LR+)	$\frac{SENS}{1 - SPEC}$	Ratio between the probability of predicting a sample as positive when it is truly positive and the probability of predicting a sample as positive when it is actually negative. SENS and SPEC are not in percentage. Optimal value: infinite.
Negative likelihood ratio (LR-)	$\frac{SPEC}{1 - SENS}$	Ratio between the probability of predicting a sample as negative when it is actually positive and the probability of predicting a sample as negative when it is truly negative. SENS and SPEC are not in percentage. Optimal value: 0.
F-Score	$\frac{2 \times SENS \times SPEC}{SENS + SPEC}$	Model performance considering imbalanced classes. Optimal value: 100%.
G-Score	$\sqrt{SENS \times SPEC}$	Model performance not accounting for the classes size. Optimal value: 100%.

2.3 Procedure

Loading the data ● Timing 5 min – 2 d, depending on the size of the dataset

1. The spectra data must be loaded into the software for data analysis. Usually, the spectral data need to be converted to suitable .txt or .csv files within the spectrometer software, or saved in an extension format readable by the software used for data analysis. MATLAB commands such as 'csvread' and 'importdata', or the option right click > 'Import Data…' over the file in the "Current Folder" window of MATLAB, can be used to load standard .csv or .txt files. IrootLab86 toolbox for MATLAB contain an interface called "mergertool" to load different spectral formats: .DAT files, .csv files, OPUS binary files from Fourier transform infrared (FTIR) Bruker® spectrometers, and .txt and Wire files from Renishaw® Raman spectrometers. We strongly suggest saving the spectral files in .csv or .txt file formats since these are universal formats most chemometric software read regardless the instrument manufacturer brand.

▲ **CRITICAL** Experimental procedures for sample preparation and Raman and FTIR spectral acquisition are demonstrated in other protocols (Baker *et al.*, 2014; Butler *et al.*, 2016; Martin *et al.*, 2010; Morais *et al.*, 2019c).

▲ **CRITICAL** The routine to load the spectral data depends on the file format, spectrometer manufacturer, and software being used to analyse the data.

? TROUBLESHOOTING

In case of fail to load directly the spectral data into the software for data analysis, export the spectral data into readable .txt, .csv or .xls formats within the spectrometer software and load them into Microsoft® Excel or any spreadsheet software. Then, copy and paste the spectral data from the spreadsheet to the data analysis software.

■ **PAUSE POINT** Save the spectral dataset in the data analysis software format (*e.g.*, .mat for MATLAB) into a known folder for further analysis.

Data quality evaluation ● Timing 40 min – 4 h, depending on the size of the dataset

2. Evaluate the raw spectral data by plotting them and by performing quality tests to identify anomalous spectra or biased patterns before applying processing. This

can be done by visual inspection of the spectral profiles, followed by plotting Hotelling's T^2 *versus* Q residuals charts using only the mean-centred data, and the analysis of the PCA residuals. Samples far from the origin of the Hotelling's T^2 *versus* Q residuals chart should be investigated and removed. Outliers must be removed one at the time from the PCA model. PCA residuals should be random and close to zero. Further instructions about data quality evaluation can be found in 'Minimum dataset requirements' and 'Outlier detection' in the 'Experimental design' section. Hotelling's T^2 *versus* Q residuals charts can be built using the automatic Outlier Detection algorithm (Morais *et al.*, 2019c) for MATLAB at https://figshare.com/articles/Outlier_Detection/7066613/2.

Data pre-processing ● Timing 15 min – 4 h, depending on the size of the dataset

▲ **CRITICAL** Steps 3 – 8 below can be modified depending on the nature of the dataset. Pre-processing effects are depicted in Fig. 2.2 and a pre-processing decision flowchart is shown in Fig. 2.3. Further details about pre-processing techniques can be found in 'Pre-processing' in the 'Experimental design' section.

3. *Selecting the biofingerprint region*. Truncate the spectra dataset to the biofingerprint region to eliminate atmospheric interference present in other regions of the spectra. FTIR: 1800 – 900 cm^{-1}; Raman: 2000 – 500 cm^{-1}; NIR: 900 – 2600 nm.
4. *Savitzky-Golay (SG) smoothing for removing spectral noise*. When random noise is present, SG smoothing should be applied. Window size varies according to the spectral dataset resolution and size. The window size must be an odd number, since a central data point is required for the smoothing process. Try different window sizes from 3 to 21 and observe how the spectra change (in shape) and how the noise is reduced. Use the smallest window size that removes a considerable amount of the noise while maintaining the original spectral shape. *E.g.*, using a spectral resolution of 4 cm^{-1}, the IR biofingerprint region (900-1800 cm^{-1}) usually contains 235 wavenumbers; in this case, a window size of 5 points should be used. The polynomial order of the SG fitting should be second order for IR, Raman and NIR datasets due to the quadratic band shape of the spectrum.
5. *Light scattering correction using either MSC, SNV or second derivative.* First, try using MSC or SNV, as MSC maintains the spectral scale and both methods

maintain the original spectral shape. If the results are not satisfactory (*e.g.*, classification accuracy < 75% in the validation set), try using the second-derivative spectra.

▲ **CRITICAL** If resonant Mie scattering (Bassan *et al.*, 2009) is present in the spectra (often detected by a severe baseline distortion followed by a systematic shift in the *y*-axis), then the RmieS-EMSC algorithm (Bassan *et al.*, 2010) should be used for spectral correction instead of MSC, SNV, second derivative or other baseline correction technique.

6. *Perform baseline correction using AWLS or rubber-band baseline correction*. If EMSC or spectral differentiation is applied as the light scattering correction method, baseline correction is not necessary.
7. *Normalisation*. Normalise the spectrum to the Amide I or Amide II peak, or perform a vector normalisation (2-norm, length = 1) to correct different scales across spectra (*e.g.*, due to different sample thickness when using FTIR in transmission mode).
8. *Scaling*. Mean-centre the spectral dataset. In case of data fusion, block-scaling should be used.

▲ **CRITICAL** Plot the spectral data throughout all the pre-processing steps to identify anomalous behaviours. For parsimonious reason, only use the pre-processing methods that are needed for the dataset (see Fig. 3 and 'Pre-processing' in the 'Experimental design' section).

■ **PAUSE POINT** Save the pre-processed spectral dataset in the data analysis software format (*e.g.*, .mat for MATLAB) into a known folder for further analysis.

Exploratory analysis ● Timing 1 h – 4 d, depending on the size of the dataset

9. Perform a PCA model with the mean-centred pre-processed spectral data to identify clustering patterns, trends and outliers within the dataset. Determine the number of PCs by plotting the number of PCs *versus* the model explained variance, where the selected number of PCs should be the one that contains the majority of the cumulative explained variance before a constant trend is observed in the next following PCs. Usually the number of PCs should not exceed 10 PCs, since this can add random noise to the model.

10. Plot the PCA scores on PC1 *versus* PC2, and investigate other combinations of PCA scores plot on different PCs according to the number of selected PCs to identify possible clustering patterns or trends. Colour-code the samples to facilitate visualisation. If a clear segregation pattern between the classes is observed on the PCA scores space, this is an indication that PCA-based discriminant models, such as PCA-LDA and PCA-QDA, might work well with the dataset.
11. Plot the Hotelling's T^2 *versus* Q residuals chart for the PCA model built in order to identify possible outliers still within the spectral dataset. The outliers should be removed from the dataset before proceeding to the next steps.

▲ **CRITICAL** The pre-processed spectral data must be mean-centred before PCA.

▲ **CRITICAL** PCA is not a classification method, thus the PCA scores plot is not the final classification model.

? TROUBLESHOOTING

If no segregation trend is observed in the PCA scores plot, this is an indicative of the dataset complexity. The visualisation of the PCA scores is limited to 3-demenions (3D) plots, hence, no apparent segregation trend does not mean that the dataset cannot be discriminated in the PCA scores space. Therefore, PCA-based discriminant models can still be built by using 4 to 10 PCs, or more.

Data selection ● Timing 10 min – 4 h, depending on the size of the dataset

12. Separate the samples that will be used for the training and test sets. Data selection should be performed before model construction. The samples can be split into training (70%) and test (30%) sets, using a cross-validated model; or they can be split into training (70%), validation (15%) and test (15%) sets without using cross-validation. To maintain consistency and account for well-balanced training models, the KS or MLM algorithms are suggested to separate the samples into the sub-sets. The KS algorithm is freely available at https://doi.org/10.6084/m9.figshare.7607420.v1; and the MLM algorithm is freely available at https://doi.org/10.6084/m9.figshare.7393517.v2.

▲ **CRITICAL** Spectrum replicas for a same sample cannot be present in more than one sub-set; that is, spectral replicates cannot be distributed amongst the training, validation and/or test sets.

? TROUBLESHOOTING

When spectral replicates are present in the dataset, the data selection algorithm can be applied in a way to keep the spectrum replicas together, or by averaging the spectral replicates before applying the data selection algorithm.

? TROUBLESHOOTING

If the percentages of samples (70%, 30% or 15%) for each sub-set generate numbers with decimal places, round them to the closest integer values.

■ **PAUSE POINT** Save the training, validation and/or test sets in the data analysis software format (*e.g.*, .mat for MATLAB) into a known folder for further analysis.

Model construction ● Timing 1 h – 4 d, depending on the size of the dataset

▲ **CRITICAL** Feature extraction (*e.g.*, by means of PCA) or feature selection (*e.g.*, by means of SPA or GA) should be used to reduce data collinearity and speed up data processing and analysis time. PLS-DA is already a feature extraction method; thus performing a feature extraction technique prior PLS-DA is not necessary. KNN, SVM and ANN algorithms can be applied either without or after feature extraction/selection techniques. The classification technique being tested must follow a parsimony order: LDA > PLS-DA > QDA > KNN > SVM > ANN > random forests > deep-learning approaches.

13. Apply the feature extraction or selection technique. The optimization of the number of PCs during PCA-based methods or LVs during PLS-DA can be performed using an external validation set (15% of the original dataset) or using cross-validation (leave-one-out for small datasets (≤20 samples); venetian blinds or random subsets with 10 data splits can be used for large datasets (>20 samples); or continuous-block (*i.e.*, leave-one-patient-out cross-validation) when replicate spectra are present). GA should be performed three times, starting from different initial populations, and the best result using an external validation set (15% of the original dataset) should be used. Cross-over probability should be set to 40% and mutation probability should be set to 1–10%, according to the dataset size.

14. The classification method should be used optimized by an external validation set or by using cross-validation, especially for selecting the number of LVs of PLS-DA, and the kernel parameters for SVM. The kernel function for SVM should be the radial basis function (RBF) kernel, due to its adaptation to different data distribution. To avoid overfitting, cross-validation should be always performed during model construction to estimate the best RBF parameters in SVM.

▲ **CRITICAL** The final classification model must be built with the optimum classifier parameters.

■ **PAUSE POINT** Save the training model parameters for further analysis.

Model validation ● Timing 1 h – 8 h, depending on the size of the dataset

15. After model construction using the training set, the model must be blindly validated by an external test set. The samples in the test set cannot be present in the training set; and the model output for the test set must be statistically compared with reference known values.
16. Based on the model output for the training and test sets, calculate the accuracy, sensitivity and specificity for each set. The metrics for the training set are used to assess the fitting of the model, but they do not reflect the true model behaviour towards unknown samples. The metrics for the test set are the true expected results representing the predictive classification ability of the model.

2.4 Troubleshooting

5.1.2 Loading the Data

The file format in which the spectral data is saved must be readable by the data analysis software. Check the importing data routines in the data analysis software beforehand to save the experimental files in a suitable file format.

2.4.2 Data Pre-processing

Data pre-processing techniques should be used in a parsimonious way, and they cannot mask the signal of interest. Testing different pre-processing techniques is

recommend to find the best solution in terms of cross-validation or validation performance.

2.4.3 Model Construction

In case of unsatisfactory classification results, the complexity of the model being tested should increase in the following order: LDA > PLS-DA > QDA > KNN > SVM > ANN > random forests > deep-learning approaches. Changing the type of classifier, the feature extraction/selection technique and the type of pre-processing are ways to narrow down the classification results and find the best classification model. The performance testing of candidates models can be made by cross-validation or by using an external validation test. The final performance of the classification model must be calculated using an external test set containing independent samples (samples not present in the training set).

5.1 Timing

2.5.1 Loading the Data

Step 1, importing the data to the data analysis software: 5 min – 2 d, depending on the size of the dataset.

2.5.2 Data Quality Evaluation

Step 2(A), plotting and inspecting the spectral profiles: 15 min – 1 h, depending on the size of the dataset.

Step 2(B), inspecting Hotelling's T^2 *versus* Q residuals charts for the mean-centred raw data: 15 min – 2 h, depending on the size of the dataset.

Step 2I, analysis of PCA residuals: 10 min.

2.5.3 Data Pre-processing

Step 3, selecting the biofingerprint region: 10 min.

Step 4, Savitzky-Golay (SG) smoothing for removing spectral noise: 2 min – 20 min, depending on the size of the dataset.

Step 5, light scattering correction using either MSC, SNV, second derivative or RmieS-EMSC: 2 min – 20 min, depending on the size of the dataset.

Step 6, performing baseline correction using AWLS or rubber-band baseline correction: 2 min – 40 min, depending on the size of the dataset.

Step 7, normalisation: 1 – 10 min, depending on the size of the dataset.

Step 8, scaling: 1 – 10 min, depending on the size of the dataset.

2.5.4 Exploratory Analysis

Step 9, building a PCA model with the mean-centred pre-processed spectral data to identify clustering patterns: 10 min – 4 h, depending on the size of the dataset.

Step 10, plot the PCA scores on PC1 *versus* PC2, and investigate other combinations of PCA scores plot on different PCs to identify possible clustering patterns or trends: 30 min – 12 h, depending on the number of selected PCs.

Step 11, checking the Hotelling's T^2 *versus* Q residuals chart to identify outliers: 10 min – 2 h, depending on the size of the dataset.

2.5.5 Data Selection

Step 12, sample splitting: 10 min – 4 h, depending on the size of the dataset.

2.5.6 Model Construction

Step 13, application of feature extraction or selection techniques: 15 min – 8 h, depending on the size of the dataset and the feature selection technique being used.

Step 14, construction of the classification model: 15 min – 4 d, depending on the size of the dataset, type of cross-validation and type of classifier.

2.5.7 Model Validation

Step 15, obtaining the model outputs for the test set: 15 min – 1 h, depending on the size of the test set.

Step 16, calculation of model performance metrics: 30 min – 8 h, depending on the number of metrics being calculated and the number of classes in the dataset.

5.1 Anticipated Results

To illustrate how this protocol can be used to analyse spectral data, we will conduct some classification models for 3 real datasets: (1) Syrian hamster embryo (SHE) cells dataset composed of FTIR spectra, free available as part of IrootLab toolbox (http://trevisanj.github.io/irootlab/); (2) Raman spectra of blood plasma to detect ovarian cancer, free available at https://doi.org/10.6084/m9.figshare.6744206.v1; and (3) NIR spectra of corn samples, free available at http://www.eigenvector.com/data/Corn/index.html.

Dataset 1 (SHE cells) was originally published by Trevisan *et al.* (2010) and is composed of originally 10 classes, containing attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectra of cells exposed to 5 contaminants in two levels (non-transformed and transformed). In this example, only two classes are being used for analysis: class 1 (cells exposed to benzo[a]pyrene (B[a]P) non-transformed, n = 59) and class 2 (cells exposed to B[a]P transformed, n = 62). The spectra at the fingerprint region (1800-900 cm^{-1}) was pre-processed by 2^{nd} differentiation. The step-by-step analysis of this dataset using PCA-LDA/QDA is demonstrated in the Appendix A. Dataset 2, originally published by Paraskevaidi *et al.* (2018a), is also composed of two classes: class 1 containing 182 Raman spectra of blood plasma from healthy individuals, and class 2 containing 189 Raman spectra of blood plasma from ovarian cancer patients. The raw spectra were pre-processed by cutting the Raman fingerprint region (2000-500 cm^{-1}), followed by Savitzky-Golay 2^{nd} differentiation (window of 21 points, 2^{nd} order polynomial fit) and vector normalisation. Dataset 3 consists of 80 corn samples measured on 3 different NIR spectrometers. The spectra were acquired in the range between 1100 – 2498 nm at 2 nm intervals. The classification models for this dataset were based on the spectra collected by instrument 5 ('m5spec') where class 1 was defined as the samples

with protein content ≤ 8.5 (n = 36), and class 2 the samples with protein content > 8.5 (n = 44). The spectra for this dataset were pre-processed by SNV. The raw spectra for datasets 1–3 are show in the Appendix A, Figure A1.1.

A PCA was applied for exploratory analysis of the pre-processed spectral data followed by an outlier detection algorithm. The PCA scores plot and Hotelling's T^2 *versus* Q residuals charts for datasets 1–3 are show in the Appendix A, Figure A1.2. For dataset 1, it is possible to identify a segregation trend between the samples from class 1 and 2 along PC1, where the samples from class 1 are mostly distributed on the left-side, and the samples from class 2 on the right-side of the PCA scores plot (Figure S2a). The Hotelling's T^2 *versus* Q residuals chart (Figure S2b) does not show any sample significantly far from the origin, thus no outlier is present in this dataset. In dataset 2, the PCA scores plot on PC1 *versus* PC2 do not show any clear segregation between the classes (Figure S2c); and the Hotelling's T^2 *versus* Q residuals chart (Figure S2d) indicates the presence of 4 outliers in class 2. These 4 spectra were removed from the dataset before further analysis. The PCA scores plot for dataset 3 (Figure S2e) show a separation trend along PC2, where the samples from class 1 are mostly on the positive side of the scores on PC2, and the samples from class 2 are mostly in the negative side of the scores on PC2. The Hotelling's T^2 *versus* Q residuals chart for this dataset (Figure S2f) does not indicate the presence of outliers.

After data selection using the MLM algorithm (70% of samples for training and 30% of the samples for test), the mean-centred pre-processed spectra were used to build PCA-LDA, PCA-QDA and PLS-DA classification models. The training and cross-validation accuracies for datasets 1–3 using PCA-LDA and PLS-DA are show in Table 2.3.

The optimal number of factors was selected by cross-validation (Figures S3–S5). Overall, the PLS-DA models are superior to the PCA-LDA models, which often happens since the PLS decomposition takes into consideration the reference classes labels for the training set in a way that the latent variables maximize the covariance between the samples, which emphasize the differences between the classes; while PCA decomposition only describe the variance in the data, which might not be totally related to class differences (Morais *et al.*, 2019c).

Table 2.3. Training accuracies for PCA-LDA and PLS-DA algorithms applied to datasets 1–3. Cross-validation using venetian-blinds with 10 data splits. PCs stands for principal components; LVs stands for latent variables; and EV stands for cumulative explained variance.

Dataset	Algorithm	Number of factors	Training accuracy	Cross-validation accuracy
1	PCA-LDA	10 PCs (97% EV)	93%	90%
	PLS-DA	5 LVs (86% EV)	95%	92%
2	PCA-LDA	9 PCs (27% EV)	61%	61%
	PLS-DA	2 LVs (6% EV)	89%	72%
3	PCA-LDA	8 PCs (100% EV)	91%	84%
	PLS-DA	6 LVs (98% EV)	93%	88%

The performance of the discriminant analysis models in Table 2.3 applied to an external test set is depicted in Table 2.4. PLS-DA models show superior predictive performance, where higher accuracies, sensitivities and specificities are observed in the test set in comparison to PCA-LDA. The mean pre-processed spectrum and discriminant function plot for PLS-DA applied in datasets 1–3 are show in Figure 2.5. The PLS-DA regression coefficients and ROC curves are show in the Appendix A, Figures A1.6–9.

Table 2.4. Test performance of PCA-LDA and PLS-DA models applied to datasets 1–3.

Dataset	Algorithm	Accuracy	Sensitivity	Specificity
1	PCA-LDA	86%	95%	78%
	PLS-DA	89%	95%	83%
2	PCA-LDA	67%	59%	75%
	PLS-DA	80%	75%	85%
3	PCA-LDA	83%	69%	100%
	PLS-DA	88%	77%	100%

The classification performances of the PLS-DA models in Table 2.4 are satisfactory. Usually, in clinical applications, the minimum threshold for accuracy, sensitivity or specificity is 75%, the level often found in routine clinical procedures. The AUC values for the PLS-DA models in Table 2.4 were 0.99 (dataset 1), 0.96 (dataset 2) and 0.99 (dataset 3), indicating excellent predictive performance. Nevertheless, the

classification performance of these models might improve by changing the type of pre-processing or by increasing the degree of complexity of the classification technique, such as by using feature selection techniques (*e.g.*, SPA and GA) or non-linear classifiers (*e.g.*, SVM and ANN). For sake of simplicity, herein only the results obtained by PCA-LDA and PLS-DA, which are the most common classification algorithms, are reported.

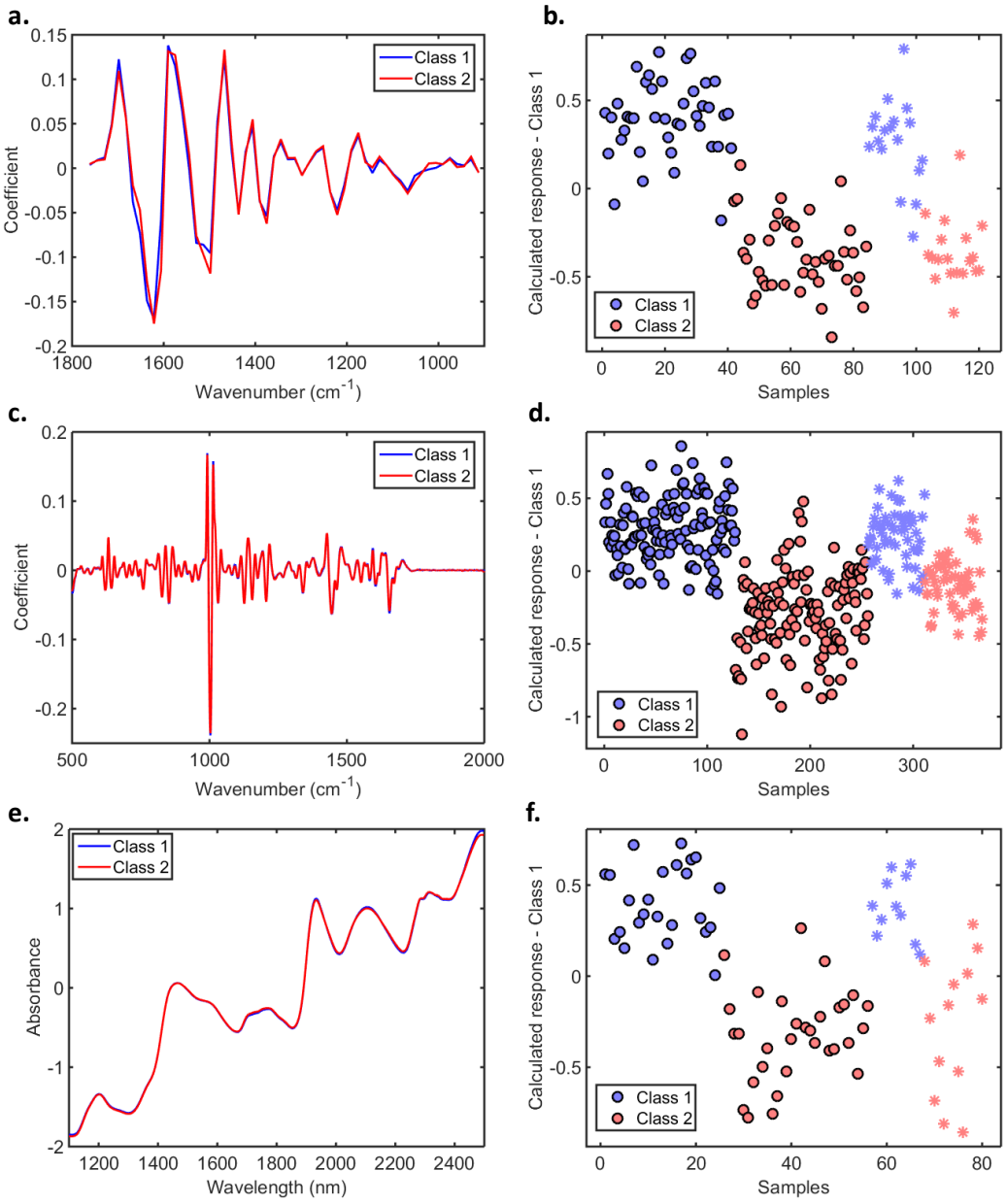


Figure 2.5. Results for PLS-DA models in datasets 1–3. (a) Mean pre-processed FTIR spectra (2nd derivative) for dataset 1; (b) calculated PLS-DA response for dataset 1, where o = training samples and * = test samples; (c) mean pre-processed Raman spectra (2nd Savitzky-Golay derivative (window of 21 points, 2nd order polynomial function) and vector normalisation) for dataset 2; (d) calculated PLS-DA response for dataset 2, where o = training samples and * = test samples; I mean pre-processed NIR spectra (SNV) for dataset 3; (f) calculated PLS-DA response for dataset 3, where o = training samples and * = test samples.

CHAPTER 3 | IMPROVING DATA SPLITTING FOR CLASSIFICATION APPLICATIONS IN SPECTROCHEMICAL ANALYSES EMPLOYING A RANDOM-MUTATION KENNARD-STONE ALGORITHM APPROACH

This chapter is published in Bioinformatics (IF 4.531). It demonstrates a new algorithm called MLM for sample selection in biospectroscopy datasets:

- Morais CLM, Santos MCD, Lima KMG, Martin FL. Improving data splitting for classification applications in spectrochemical analyses employing a random-mutation Kennard-Stone algorithm approach. Bioinformatics **2019**; 35(24): 5257–5263. https://doi.org/10.1093/bioinformatics/btz421

Abstract: Motivation: Data splitting is a fundamental step for building classification models with spectral data, especially in biomedical applications. This approach is performed following pre-processing and prior to model construction, and consists of dividing the samples into at least training and test sets; herein, the training set is used for model construction and the test set for model validation. Some of the most-used methodologies for data splitting are the random selection (RS) and the Kennard-Stone (KS) algorithms; here, the former works based on a random splitting process and the latter is based on the calculation of the Euclidian distance between the samples. We propose an algorithm called the Morais-Lima-Martin (MLM) algorithm, as an alternative method to improve data splitting in classification models. MLM is a modification of KS algorithm by adding a random-mutation factor. **Results:** RS, KS and MLM performance are compared in simulated and six real-world biospectroscopic applications using principal component analysis linear discriminant analysis (PCA-LDA). MLM generated a better predictive performance in comparison with RS and KS algorithms, in particular regarding sensitivity and specificity values. Classification is found to be more well-equilibrated using MLM. RS showed the poorest predictive response, followed by KS which showed good accuracy towards prediction, but relatively unbalanced sensitivities and specificities. These findings demonstrate the potential of this new MLM algorithm as a sample selection method for classification applications in comparison with other regular methods often applied in this type of data. **Availability:** MLM algorithm is freely available for MATLAB at https://doi.org/10.6084/m9.figshare.7393517.v1.

Author contribution: C.L.M.M. developed the algorithm, performed the data analysis and wrote the manuscript.

Camilo L. M. Morais, PhD candidate — Prof. Francis L. Martin, Supervisor

3.1 Introduction

Data splitting is a process used to separate a given dataset into at least two subsets called ‘training’ (or ‘calibration’) and ‘test’ (or ‘prediction’). This step is usually implemented after pre-processing, when the samples’ spectra have been corrected for noise or undesired variability. These subsets are used towards constructing chemometric models for quantification or classification applications. In quantification, calibration models are built to assign a concentration or discrete value to a sample based on its spectral signature, whilst in classification applications, samples or experimental observations are assigned to ‘classes’ based on their spectrochemical signature. This is made by using chemometric methods such as principal component analysis linear discriminant analysis (PCA-LDA) (Morais & Lima, 2018), partial least squares discriminant analysis (PLS-DA) (Brereton & Lloyd, 2014), or support vector machines (SVM) (Cortes & Vapnik, 1995). Sometimes, especially for large datasets, an extra subset called ‘validation’ is also obtained, containing measurements observations used for optimising factors in the chemometric model, such as the number of principal component (PCs) in PCA-LDA, latent variables in PLS-DA, and kernel parameters in SVM. When the validation set is not present, cross-validation is applied. In this case, samples from the training set are used in an iterative validation process for optimising these models parameters. This is made by firstly removing a certain number of samples from the training set and then building the classification model with the remaining samples, where the removed samples are predicted as a temporary validation set. This is performed for a certain number of repetitions until all training samples are excluded once from the training set and predicted as a temporary validation set. One of the most popular cross-validation methods is the leave-one-out cross-validation, where only one sample is removed from the training set per each iteration. A misclassification error is then calculated for this temporary validation set, where different models parameters, such as different number of factors or principal components, are tested. The training model with the lowest cross-validation error is then chosen as final, where the classification parameters that led to the lowest cross-validation error value are selected. The samples primarily excluded from modelling (test set) are used for final model evaluation, since they are considered as being external to the model (blind). In this case, one simulates how the model would behave in the presence of new observations, though they are often measured in the same experiment with the training samples.

To avoid the presence of bias introduced by manual data splitting, there are a number of computational methods that can be used for sample selection, such as based on leverage (Wang *et al.*, 1991), random selection (RS) or Kennard-Stone (KS) algorithm (Kennard & Stone, 1969). RS and KS are the most used methods for sample selection; the former due to its simplicity and the latter due to its adaptation to analytical chemistry applications, since it allows a training model covering most sources of variations within the dataset, ensuring the training model is more representative of the whole dataset. Currently, the original KS paper (Kennard & Stone, 1969) has >1,000 citations, being the method of choice in many classification applications.

Although including as much variability as possible within the training model provides a good predictive performance, sometimes random phenomena might occur with new samples in a test set, in particular when samples come from complex matrices. An example of this is biological-derived samples. Biological samples can be affected by a series of factors that are difficult to include in relatively small datasets. For example, in clinical applications the spectrochemical response of a 'healthy' and 'disease' sample may vary according to changes in diet and lifestyle (Lindon *et al.*, 2017). The same applies for bacteria or viruses extracted from certain media, since environmental variations may also change their spectral signature. Additionally, random factors such as genetic mutations might affect the predictive performance of a classification model for biological samples in the future. These phenomena add a degree of 'randomness' in the predictive behaviour of a classifier, since more extrapolations might be needed to address all of these issues. Thus, having in mind the inclusion of as much representativeness as possible in the training model but with a small degree of randomness, we propose a new algorithm based on a random-mutation Kennard-Stone approach; we call this the Morais-Lima-Martin (MLM) algorithm.

Towards comparison of the predictive response of MLM with RS and KS, we tested classification models on six real-world spectrochemical datasets using PCA-LDA, where the predictive performance in terms of accuracy, sensitivity and specificity were evaluated. In addition, simulations with normally distributed random data were performed to evidence the performance of the MLM algorithm in comparison with the RS and KS method.

3.2 System and Methods

3.2.1 Datasets

Six real-world datasets were used towards comparing the classification performance of RS, KS and MLM algorithms. Dataset 1 contains 280 infrared (IR) spectra of two *Cryptococcus* fungi specimens acquired *via* attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectroscopy. This dataset is publically available at https://doi.org/10.6084/m9.figshare.7427927.v1. Class 1 is composed of 140 spectra of *Cryptococcus neoformans* samples and class 2 of 140 spectra of *Cryptococcus gattii* samples. Spectra were acquired in the 400–4000 cm^{-1} spectral range with a resolution of 4 cm^{-1} and 16 co-added scans using a Bruker VERTEX 70 FTIR spectrometer (Bruker Optics, Ltd., UK). The spectral data were pre-processed by excising the biofingerprint region (900–1800 cm^{-1}), which was followed by automatic weighted least squares (AWLS) baseline correction and normalisation to the Amide I peak (1650 cm^{-1}). More details regarding this dataset can be found in literature (Costa *et al.*, 2016; Morais *et al.*, 2017).

Dataset 2 contains 240 IR spectra derived from formalin-fixed paraffin-embedded brain tissues separated into two classes. Class 1 contains 140 spectra from normal brain tissue, and class 2 contains 100 spectra from glioblastoma brain tissue. Spectra were collected *via* ATR-FTIR spectroscopy using a Bruker VECTOR 27 FTIR spectrometer with a Helios ATR attachment (Bruker Optics, Ltd., UK). The raw spectra, acquired in the 400–4000 cm^{-1} spectral range with a resolution of 8 cm^{-1} and 32 co-added scans, were pre-processed by excising the biofingerprint region (900-1800 cm^{-1}), which was followed by rubberband baseline correction and normalisation to the Amide I peak (1650 cm^{-1}). This dataset is publicly available as part of the IrootLab toolbox (http://trevisanj.github.io/irootlab/) (Trevisan *et al.*, 2013), and more information about it can be found in Gajjar (2013).

Dataset 3 contains 183 IR spectra distributed into 3 classes. Class 1 contains 59 spectra of Syrian hamster embryo (SHE) cells treated with benzo[*a*]pyrene (B[*a*]P), class 2 contains 62 spectra of SHE cells treated with 3-methylcholanthrene (3-MCA) and class 3 contains 62 spectra of SHE cells treated with anthracene (Ant). Spectra were acquired in the 400–4000 cm^{-1} spectral range with a resolution of 8 cm^{-1} by using a Bruker TENSOR 27 spectrometer with a Helios ATR attachment (Bruker Optics, Ltd., UK). Pre-

processing was performed by excising the biofingerprint region (900–1800 cm^{-1}), which was followed by rubberband baseline correction and normalisation to the Amide I peak (1650 cm^{-1}). This dataset is publicly available as part of the IrootLab toolbox (http://trevisanj.github.io/irootlab/) (Trevisan *et al.*, 2013), and further information can be found in Trevisan *et al.* (2010).

Dataset 4 contains 270 IR spectra from blood samples divided into four classes. Class 1 is composed of 90 IR spectra of control samples, class 2 contains 88 spectra from patients with Dengue, class 3 contains 66 spectra from patients with Zika, and class 4 contains 26 spectra from patients with Chikungunya. This dataset is publically available at https://doi.org/10.6084/m9.figshare.7427933.v1. Spectra were collected in ATR mode by using a Bruker VERTEX 70 FTIR spectrometer (Bruker Optics, Ltd., UK). Acquisition was performed in the 400–4000 cm^{-1} spectral range with a resolution of 4 cm^{-1} and 16 co-added scans. Pre-processing was performed by excising the biofingerprint region (900–1800 cm^{-1}), which was followed by Savitzky-Golay smoothing (window of 7 points) (Savitzky and Golay, 1964), AWLS baseline correction, and normalisation to the Amide I peak (1650 cm^{-1}). Further details about this dataset can be found in Santos *et al*. (2018).

Dataset 5 contains 351 Raman spectra of blood plasma divided into two classes: 162 spectra of healthy individuals (class 1), and 189 spectra of ovarian cancer patients (class 2). This dataset is publicly available at https://doi.org/10.6084/m9.figshare.6744206.v1. Raman spectra were collected using an InVia Renishaw Raman spectrometer coupled with a charge-coupled device (CCD) detector and Leica microscope, with 5% laser power (785 nm), 5x objective magnification, 10s exposure time and 2 accumulations in the spectral range of 400-2000 cm^{-1}. The spectral data were pre-processed by Savitzky-Golay smoothing (window of 15 points), AWLS baseline correction, and vector normalisation. Further details about this dataset can be found in Paraskevaidi *et al.* (2018).

Dataset 6 contains 322 surface-enhanced Raman spectroscopy (SERS) spectra of blood plasma also divided into two classes: 133 spectra of healthy individuals (class 1), and 189 spectra of ovarian cancer patients (class 2). This dataset is publicly available at https://doi.org/10.6084/m9.figshare.6744206.v1. SERS spectra were collected using the same settings for dataset 5 but, in this case, silver nanoparticles were mixed with the biofluid before spectral acquisition. The spectral pre-processing was performed using

Savitzky-Golay smoothing (window of 15 points), AWLS baseline correction, and vector normalisation. Further details about this dataset can be found in Paraskevaidi *et al.* (2018).

Simulations were also performed with simulated data. This data were generated for each simulation (1000 simulations) based on a normally distributed random matrix with size of 100 × 1000 for class 1, and 100 × 1000 for class 2 (100 observations, 1000 variables per observation). The matrix values ranged randomly from -10 to 10 units. A shift of 5 units was randomly added to class 2 to create a difference between the classes. The codes to produce class 1 and class 2 in MATLAB are 'class_1 = randn(100,1000).*randn(100,1000);' and 'class_2 = (randn(100,1000)+5).*randn(100,1000);'. Class 1 and class 2 were generated for each simulation (1000 times), where all algorithms (RS, KS, and MLM) were independently applied per each simulation.

3.2.2 Software

Data analysis was performed within the MATLAB R2014b (MathWorks, Inc., USA) environment. Pre-processing was performed using PLS Toolbox 7.9.3. (Eigenvector Research, Inc., USA) and classification was performed using the Classification Toolbox for MATLAB (http://www.michem.unimib.it/) (Ballabio & Consonni, 2013). RS, KS and MLM algorithms were performed using laboratory-generated routines. MLM algorithm is public available at https://doi.org/10.6084/m9.figshare.7393517.v1.

3.2.3 Sample Selection

Samples were divided into training (70%) and test (30%) sets using, independently, the RS, KS or MLM algorithms. RS is based on a random sample selection where spectra from the original dataset are randomly assigned to training or test. KS algorithm is based on an Euclidian distance calculation, where the sample with maximum distance to all other samples are selected, then the samples which are as far away as possible from the selected samples are selected, until the selected number of samples is reached. This means that the samples are selected in such a way that they will uniformly cover the complete sample space, reducing the need for extrapolation of the remaining samples. MLM algorithm, based on a KS-based approach, applies a KS method to the

data, as described before; then, a random-mutation factor is used in the KS results, where some samples from the training set are transferred to the test set, and some samples from the test set are transferred to training. Herein, the mutation factor was set at 10%. This value is inspired in the mutation probability of genetic algorithms (Morais *et al.*, 2019c), where 10% is a common threshold employed to keep a balance between the degree of randomness and model convergence. MLM algorithm is visually illustrated in Figure 3.1.

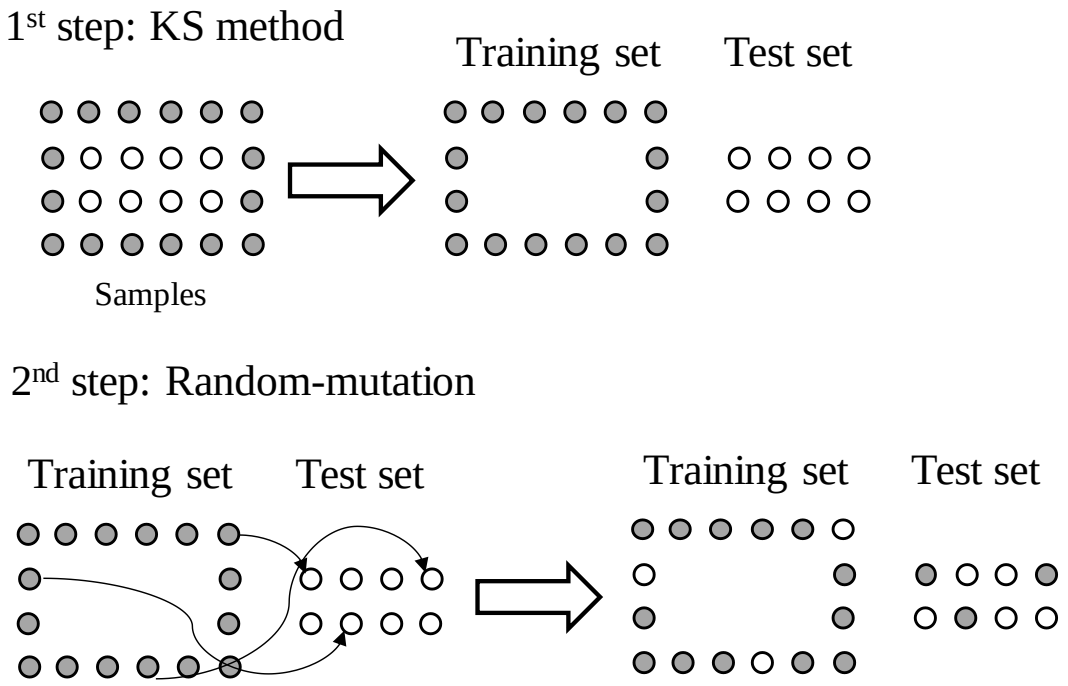


Figure 3.1. Illustration of the MLM algorithm based on a random-mutation of the Kennard-Stone (KS) method. Adapted from Morais *et al.* (2018a).

3.2.4 Classification

Classification was performed based on a PCA-LDA algorithm. For this, initially a principal component analysis (PCA) model is applied to the pre-processed data, decomposing the spectral space into a small number of PCs representing most of the original data-explained variance (Bro & Smilde, 2014). Each PC is composed of scores and loadings, the former representing the variance on samples direction, and the latter the variance on variables (*e.g.*, wavenumber) direction. Then, the PCA scores are used as input for a linear discriminant analysis (LDA) classifier. LDA performs a Mahalanobis distance calculation to linearly classify the input space (PCA scores) into at least two

classes (Morais & Lima, 2018; Dixon & Brereton, 2009). The LDA classification scores (L_{ik}) can be calculated in a non-Bayesian form as (Morais & Lima, 2018; Dixon & Brereton, 2009):

$$L_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}} \mathbf{C}_{\mathrm{pooled}}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_k) \quad (3.1)$$

where $\mathbf{x}_i$ is a vector containing the input variables for sample i; $\bar{\mathbf{x}}_k$ is the mean vector of class k; $\mathbf{C}_{\mathrm{pooled}}$ is the pooled covariance matrix between the classes; and, T represents the matrix transpose operation. Model optimization was performed using cross-validation venetian blinds with 10 splits.

The PCA-LDA classification performance was evaluated by means of accuracy, sensitivity and specificity calculations. Accuracy represents the total number of samples correctly classified considering true and false negatives; sensitivity measures the proportion of positives that are correctly identified; and, specificity measures the proportion of negatives that are correctly identified (Morais & Lima, 2017). These parameters are calculated as follows:

$$\text{Accuracy (\%)} = \left(\frac{\mathrm{TP+TN}}{\mathrm{TP+FP+TN+FN}}\right) \times 100 \quad (3.2)$$

$$\text{Sensitivity (\%)} = \left(\frac{\mathrm{TP}}{\mathrm{TP+FN}}\right) \times 100 \quad (3.3)$$

$$\text{Specificity (\%)} = \left(\frac{\mathrm{TN}}{\mathrm{TN+FP}}\right) \times 100 \quad (3.4)$$

where TP stands for true positives; TN for true negatives; FP for false positives; and, FN for false negatives.

3.3 Results and Discussion

Six real-world datasets were evaluated using different data splitting techniques: RS, KS, and our new MLM algorithm. These datasets are composed of IR and Raman spectra from biological-derived applications involving: IR spectra of fungi (dataset 1); IR spectra of cancer brain tissue (dataset 2); IR spectra for toxicological study (dataset 3); IR spectra of viruses (dataset 4); Raman spectra of plasma for ovarian cancer detection (dataset 5); and, SERS spectra of plasma for ovarian cancer detection (dataset 6). Figure 3.2 shows the pre-processed mean spectrum with standard deviation for each class in datasets 1–6. The pre-processed spectra from these datasets were used as input for the sample selection techniques, where their classification performances were evaluated *via* the PCA-LDA algorithm.

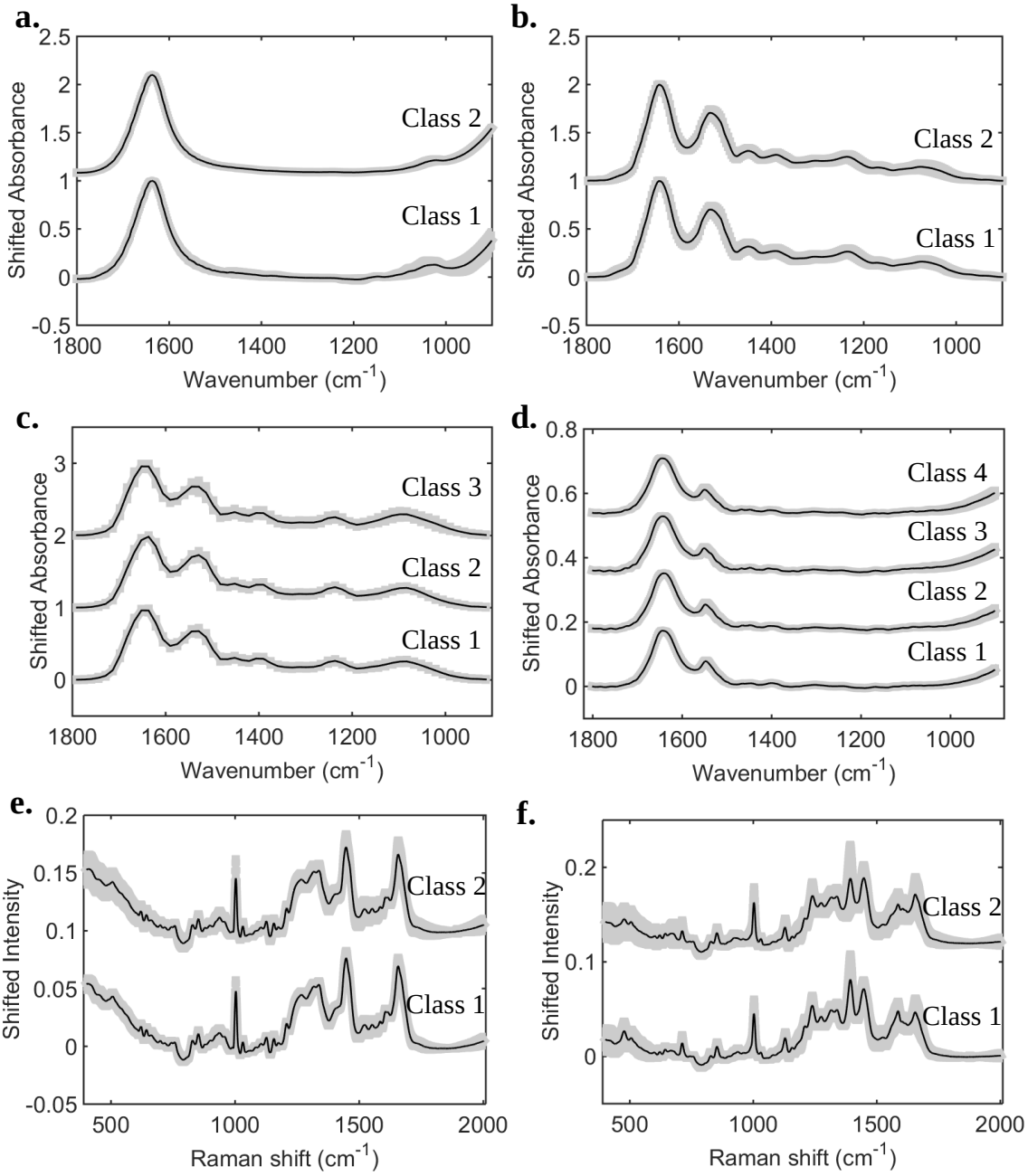


Figure 3.2. Mean pre-processed spectrum with standard deviation (shaded) for each class in dataset 1 (a), 2 (b), 3 (c), 4 (d), 5 I, and 6 (f).

Dataset 1 is composed of 280 IR spectra for two fungi specimens groups (*Cryptococcus neoformans* [class 1]; *Cryptococcus gattii* [class 2]), each class having 170 spectra each. Both fungi classes are pathogenic agents responsible for causing Cryptococcosis in humans, differing in their epidemiology, host range, virulence, antifungal susceptibility and geographic distribution (Morais *et al.*, 2017). From a clinical point of view, *Cryptococcus neoformans* is a pathogen with a tendency to attack the central nervous system and its effects are mainly noted in immunosuppressed patients,

whereas *Cryptococcus gattii* targets the lungs of immunocompetent, healthy individuals (Morais *et al.*, 2017). RS, KS and MLM were independently applied to the pre-processed spectra separating 70% of them for training and 30% for testing. Cross-validated PCA-LDA was applied for model construction using three PCs (99% cumulative explained variance) selected according to the minimum cross-validation error rate within the minimum number of PCs (Figure 3.3). The model fitting performance is shown in Table 3.1, where the best training (84%) and cross-validation (83%) accuracy are observed using RS algorithm. KS generates the worst fitting performance with 80% accuracy in both training and cross-validation. The MLM algorithm shows an intermediary performance with 83% and 82% accuracy in training and cross-validation, respectively.

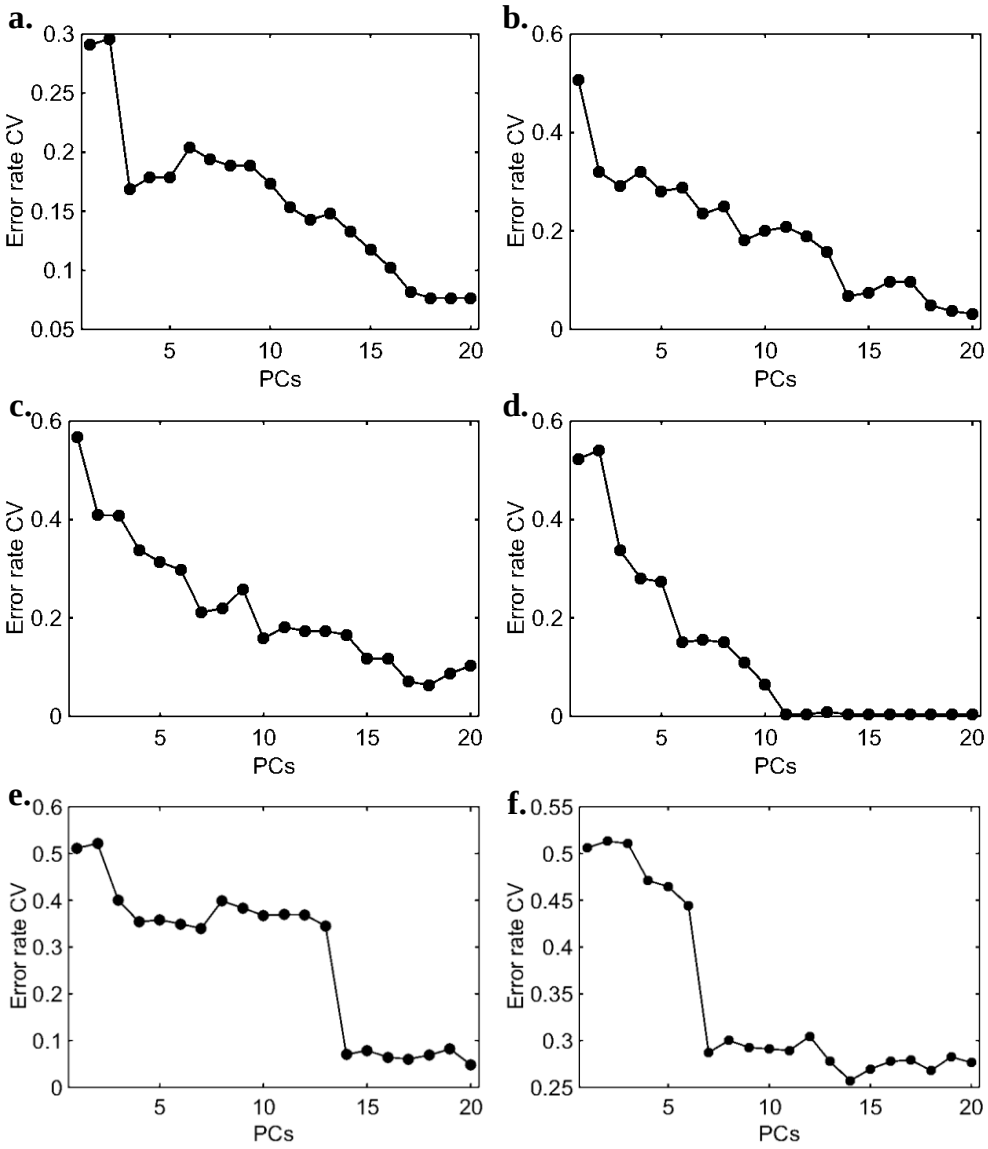


Figure 3.3. PCA-LDA cross-validation error rate for datasets 1 (a), 2 (b), 3 (c), 4 (d), 5 I, and 6 (f). CV: cross-validation; PCs: principal components.

Table 3.1. PCA-LDA fitting accuracy for training and cross-validation (CV) varying with the sample selection method (RS: random selection; KS: Kennard-Stone; MLM: Morais-Lima-Martin) applied in datasets 1–6.

Dataset	Sample selection method	Training Accuracy (%)	CV Accuracy (%)
1	RS	84	83
	KS	80	80
	MLM	83	82
2	RS	85	83
	KS	81	80
	MLM	82	77
3	RS	86	84
	KS	83	82
	MLM	84	80
4	RS	92	91
	KS	90	90
	MLM	93	90
5	RS	93	93
	KS	89	88
	MLM	91	88
6	RS	74	72
	KS	75	72
	MLM	76	75

Although the best fitting accuracy, the RS-based model exhibits a very poor sensitivity, at 69%, in the test set (Table 3.2). The specificity is high (88%), but the model seems to have a poor balance in terms of sensitivity and specificity, indicating that one class is much better classified than the other. The KS-based model with the worst fitting gives the best specificity (98%), but the sensitivity remains the same. On the other hand, the MLM-based model shows the best well-balanced performance, where the specificity falls to 78%, but the sensitivity increases to 74%, indicating that both classes are well-classified, and the model is not skewed towards a good classification of just one of the classes. Overall accuracy varying with the sample selection method is depicted in Figure 3.4, where the accuracy for dataset 1 using MLM (81%) is close to the KS algorithm (83%), which achieves the best accuracy due to the great specificity of this model. RS has the worst accuracy (79%), indicating that the performance of this method in the test set is inferior to the other algorithms that had worst fitting; thus, confirming that good fitting is not necessarily associated with good predictions.

Table 3.2. Sensitivity and specificity for the test set obtained by PCA-LDA varying with the sample selection method (RS: random selection; KS: Kennard-Stone; MLM: Morais-Lima-Martin) applied in datasets 1–6.

Dataset	Sample selection method	Sensitivity (%)	Specificity (%)
1	**RS**	69	88
	KS	69	98
	MLM	74	78
2	**RS**	79	63
	KS	79	80
	MLM	81	80
3	**RS**		
	Class 1	83	87
	Class 2	79	89
	Class 3	89	100
	KS		
	Class 1	94	97
	Class 2	100	92
	Class 3	84	100
	MLM		
	Class 1	94	92
	Class 2	95	95
	Class 3	84	100
4	**RS**		
	Class 1	96	100
	Class 2	100	100
	Class 3	85	98
	Class 4	88	95
	KS		
	Class 1	100	100
	Class 2	100	100
	Class 3	90	98
	Class 4	88	97
	MLM		
	Class 1	100	100
	Class 2	100	100
	Class 3	95	98
	Class 4	88	99
5	**RS**	94	88
	KS	94	95
	MLM	94	91
6	**RS**	70	70
	KS	72	84
	MLM	72	89

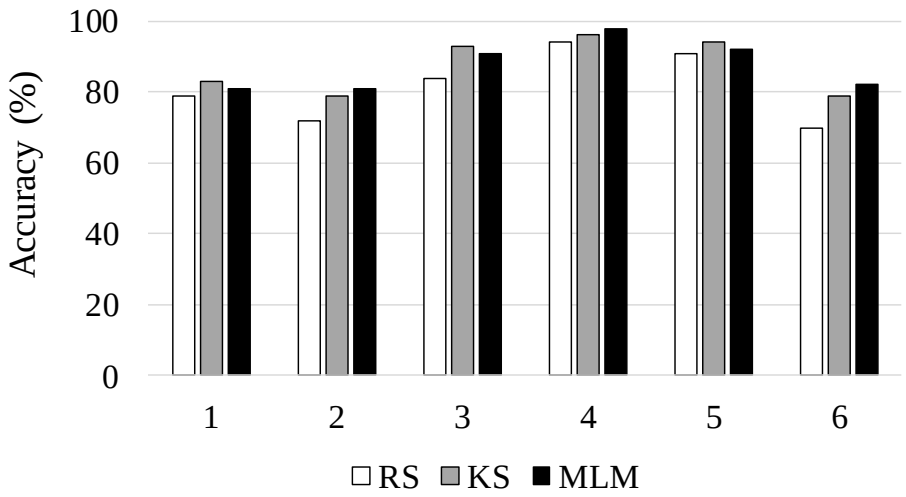


Figure 3.4. Accuracy in the test set obtained by PCA-LDA varying with the sample selection method (RS: random selection; KS: Kennard-Stone; MLM: Morais-Lima-Martin) applied in datasets 1–6.

Dataset 2 is composed of 140 spectra of normal (class 1) and 100 spectra of glioblastoma (class 2) brain tissue samples. Glioblastoma is the brain cancer type with the poorest survival rate (Gajjar *et al.*, 2013). Reference methods for detecting these types of cancer, such as immunohistochemical detection of isocitrate dehydrogenase (IDH), suffers from some limitations, especially their subjective nature (Gajjar *et al.*, 2013). The use of IR spectroscopy has the potential to aid tumour differentiation based on a non-analyst dependent, fast and non-destructive methodology. In this dataset, both tumour types are differentiated based on their IR spectrochemical signature. The pre-processed IR spectra for dataset 2 are shown in Figure 3.2b. As before, RS, KS and MLM algorithms were applied to this dataset separating the data into training and test sets. PCA-LDA was applied as a classification method using 9 PCs (Figure 3.3b), accounting to 99% of cumulative explained variance. The training performance of this model in dataset 2 is shown in Table 3.1, where the RS algorithm presents the best fitting (training and cross-validation accuracy of 85% and 83%, respectively). The other algorithms (KS and MLM) have the lowest fitting performance with accuracies around 80%. Nevertheless, as before, the situation is reversed in the test set, where the RS algorithm has the worst sensitivity and specificity values (Table 3.2). In the test set, the best sensitivity and specificity values are obtained using MLM, with a slightly superior performance than KS algorithm. The overall model accuracy also is better for MLM (Figure 3.4), where the accuracy in the

test set is observed at 81% using MLM, at 79% using KS, and at 72% using RS. This confirms MLM to be the method of choice for this dataset.

Dataset 3 consists of spectra derived from SHE cells treated with one of three agents: B[*a*]P, class 1; 3-MCA, class 2; or, Anthracene, class 3. Class 1 is composed of 59 IR spectra, and both class 2 and 3 of 62 spectra. Pre-processed spectra for this dataset are shown in Figure 3.2c. PCA-LDA model was built using 10 PCs (99% cumulative explained variance) (Figure 3.3c). The best training performance was found using RS algorithm, followed by MLM and KS, which had similar fitting (Table 3.1). KS and MLM algorithms exhibit similar performance in the test set, with sensitivities and specificities for class 1 and 2 >90%. For class 3, both algorithms show 100% specificity and 84% sensitivity. On the other hand, the RS algorithm presents a slightly better sensitivity for class 3 (89%), but lower sensitivity and specificities for the other classes (<90%). Accuracy in the test set was found to be superior for KS (93%), followed by MLM (91%) and RS (84%) (Figure 3.4). Similarly to dataset 1, KS has a slightly better performance than MLM; however, the figures of merit for MLM are more well-balanced, where extreme situations in KS (100% sensitivity or specificity) are not found, but more coherent values between these two metrics (*i.e.*, sensitivity and specificity values closer to each other).

Dataset 4 is composed of control and typical virus-infected blood samples. Class 1 contains 90 IR spectra of control samples; class 2 contains 88 spectra of blood from patients with Dengue; class 3 contains 66 spectra of blood from patients with the Zika virus; and, class 4 contains 26 spectra of blood from patients with Chikungunya. These viruses are transmitted by mosquitos of genus *Aedes*, having many chemical similarities (*e.g.*, Dengue and Zika are from the same family, Flaviviridae), in particular in their surface proteins (Santos *et al.*, 2018). Fast clinical diagnosis using reference methodologies is difficult; however, IR spectroscopy can be used as an alternative tool for viral infection differentiation (Santos *et al.*, 2018). Pre-processed spectra for dataset 4 are shown in Figure 3.2d. PCA-LDA model was built using 6 PCs (Figure 3.3d), accounting for 97% of cumulative explained variance using RS and MLM sample selection methods, and 96% using KS sample selection method. RS and MLM exhibit similar fitting performance, with accuracies >90% in the training set. KS shows a slightly lower training performance with an accuracy of 90% in the training set (Table 3.1). In the test set, MLM algorithm shows the best sensitivity and specificity values (Table 3.2), followed by KS and RS. The overall accuracy in the test set also follows this trend, where

the MLM algorithm has an accuracy of 98%, followed by KS (96%) and RS (94%) (Figure 3.4).

Both datasets 5 and 6 are for diagnosis of ovarian cancer based respectively on the Raman and SERS spectra of blood plasma. These techniques have great potential towards liquid biopsy diagnosis of ovarian cancer in a minimally-invasive, rapid and objective fashion (Paraskevaidi *et al.*, 2018). Both datasets contain 2 classes, where dataset 5 is divided into 162 Raman spectra for class 1 (healthy controls) and 189 Raman spectra for class 2 (ovarian cancer); and dataset 6 is divided into 133 SERS spectra for class 1 (healthy controls) and 189 SERS spectra for class 2 (ovarian cancer). These spectra are shown in Figures 3.2e and 3.2f, respectively. Model construction was performed with PCA-LDA using 14 PCs (Figures 3.3e and 3.3f, respectively), which accounted to 98% of cumulative variance in dataset 5 and 94% of cumulative variance in dataset 6. Training performance was superior using RS in dataset 5 and MLM in dataset 6 (Table 3.1), while for prediction of the external test set, the MLM algorithm showed similar classification performance in comparison with KS for dataset 5 and the best performance amongst all three algorithms in dataset 6 (Table 3.2 and Figure 3.4), where the test accuracy for the MLM algorithm was equal to 92% in dataset 5 and 82% in dataset 6, in comparison with 94% (dataset 5) and 79% (dataset 6) using the KS algorithm, and 91% (dataset 5) and 70% (dataset 6) using the RS algorithm.

Finally, 1000 simulations using a normally distributed randomly data were performed in order to compare the performance of the RS, KS and MLM algorithms in a more robust way. As depicted in Figure 3.5, the MLM algorithm achieved the best classification performance in terms of accuracy among all algorithms tested, with an average accuracy of 67% in the range between 53-82%. RS algorithm achieved the worst accuracy values, with an average of 66% and range 50-80%, while KS achieved an accuracy value similar to MLM (67%), but with a poorer lower-limit, where accuracies ranged between 50-82%. In addition, the histogram profiles in Figure 3.5 show that amongst all 1000 simulations, MLM algorithm achieved the highest frequency peak (>150 times) above the average accuracy of 67%, while for RS and KS algorithms the highest frequency peak is below the average accuracy of 67%.

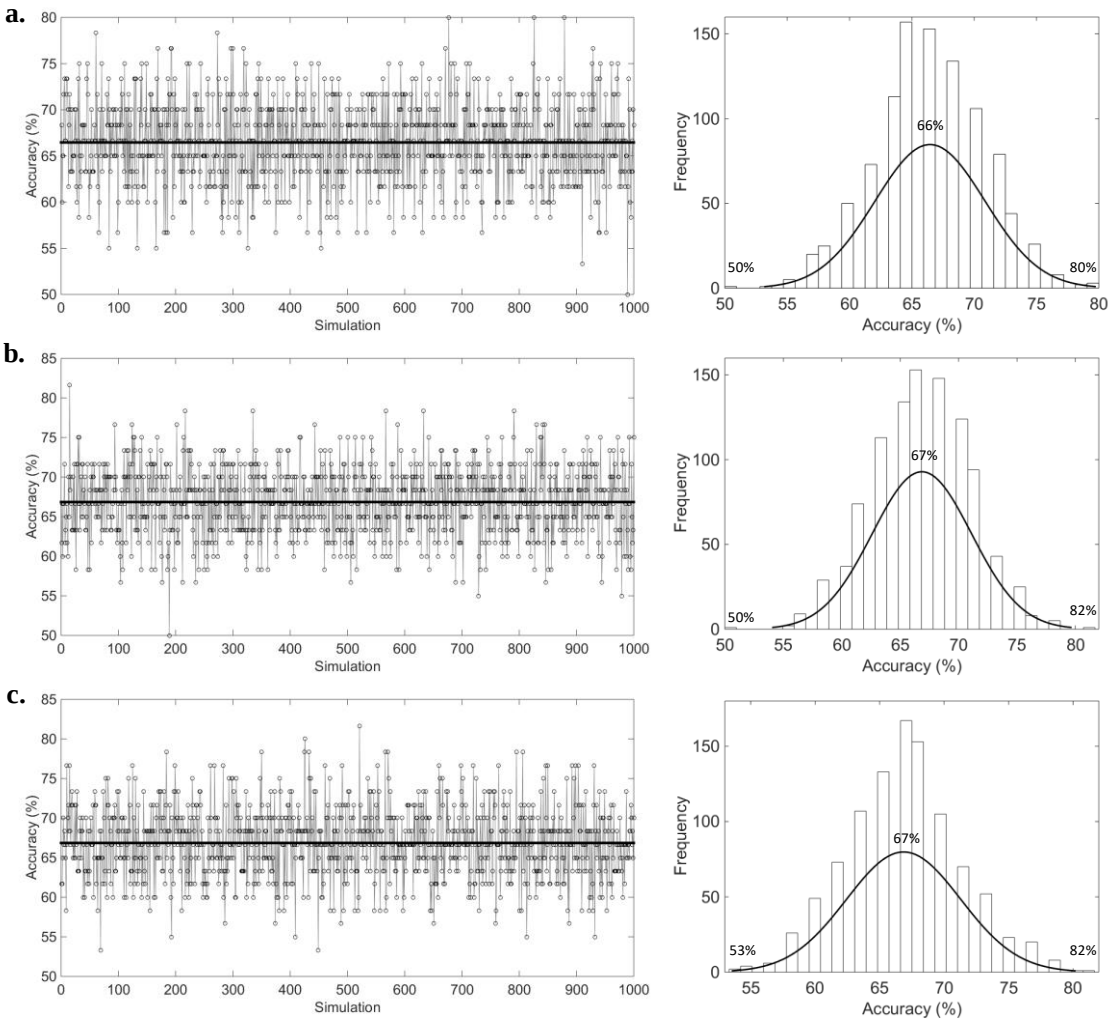


Figure 3.5. PCA-LDA accuracy distribution and histogram for 1000 simulations using normally distributed randomly data, where (a) RS, (b) KS, and (c) MLM algorithm.

These findings confirm the hypothesis that our new MLM algorithm based on a random-mutation KS algorithm approach presents a better overall performance than using RS or KS algorithms independently, especially due to the well-balanced sensitivity and specificity values in the prediction set for real-world samples. The fact that RS individually achieved good fitting but a lower predictive performance indicates that this algorithm might not include a representative variance in the training model. This reinforces the hypothesis that not necessarily an algorithm with good fitting, as demonstrated using RS, will generate good predictive results towards external samples.

3.4 Conclusion

Herein, a new data splitting algorithm, called MLM, is proposed. RS, KS and MLM algorithms were compared for sample selection using six real-world datasets from spectrochemical applications (IR spectroscopy for fungi differentiation, IR spectroscopy for brain cancer tissue analysis, IR spectroscopy to investigate agent-treated cells, IR to identify viral infection, Raman spectroscopy to detect ovarian cancer, and SERS to detect ovarian cancer), where their classification performance are evaluated by means of PCA-LDA models. The RS algorithm showed the best training performance, having the best fitting accuracies. However, when testing external samples, the RS performance was the worst amongst the algorithms tested, indicating that RS might not include all sources of variation within the training set, creating a non-representative model. The KS algorithm, on the other hand, provided much better classification results in comparison with the RS algorithm, where the accuracy in the test set is much superior. However, unbalanced sensitivities and specificities were often found in the test set using the KS algorithm, indicating that one class is far better classified than the other. The MLM algorithm combined the good predictive performance of the KS algorithm in terms of accuracy in the prediction samples, as also demonstrated with normally distributed randomly-based simulations, and has a well-balanced performance of sensitivities and specificities, since when using this algorithm these values are closer to each other for real-world datasets. The MLM algorithm might be the best algorithm for sample selection in spectrochemical applications, since it combines the good spectral representativeness in the test set provided by the KS algorithm, with a small degree of randomness that may be found in biological applications.

CHAPTER 4 | A COMPUTATIONAL PROTOCOL FOR SAMPLE SELECTION IN BIOLOGICAL-DERIVED INFRARED SPECTROSCOPY DATASETS USING MORAIS-LIMA-MARTIN (MLM) ALGORITHM

This chapter is published in Nature Protocol Exchange. It contains a protocol showing the use of the MLM algorithm for sample selection in biospectroscopy datasets:

- Morais CLM, Martin FL, Lima KMG. A computational protocol for sample selection in biological-derived infrared spectroscopy datasets using Morais-Lima-Martin (MLM) algorithm. Protocol Exchange, **2018**. https://doi.org/10.1038/protex.2018.141

Abstract: Infrared (IR) spectroscopy is a powerful analytical technique that can be applied to investigate a wide range of biological materials (e.g., biofluids, cells, tissues), where a specific biochemical signature is obtained representing the 'fingerprint' signal of the sample being analysed. This chemical information can be used as an input data for classification models in order to distinguish or predict samples groups based on computational algorithms. One fundamental step towards building such computational models is sample selection, where a fraction of the samples measured during an experiment are used for building the classifier, whereas the remaining ones are used for evaluating the model classification performance. This protocol shows how sample selection can be performed in a computational environment (MATLAB) by using a combination of Euclidian-distance calculation and random selection, named Morais-Lima-Martin (MLM) algorithm, as a previous step before building classification models in biological-derived IR datasets.

Author contribution: C.L.M.M. developed the algorithm, performed the data analysis and wrote the manuscript.

Camilo L. M. Morais, PhD candidate — Prof. Francis L. Martin, Supervisor

4.1 Introduction

Infrared (IR) spectroscopy is a vibrational spectroscopy technique that generates a unique chemical signature representing most of the molecules present in a material. It is much used to analyse biological materials (Baker *et al.*, 2014), since it allows building protocols for analysing tissues, cells and biofluids in a non-destructive, fast and low-cost fashion (Baker *et al.*, 2014; Martin *et al.*, 2010). Computational methods are used to maximize processing time and extract relevant information. Chemometric methods are often applied to build predictive models where the complex spectral data are transformed to chemically-relevant and easy-to-interpret information by means of multivariate analysis techniques. In classification applications, samples are assigned to groups based on their IR spectrochemical signature. This includes, for example, differentiation of brain tumour types (Bury *et al.*, 2019b), identification of neurodegenerative diseases (Paraskevaidi *et al.*, 2017b), cervical cancer screening (Neves *et al.*, 2016), endometrial and ovarian cancer identification (Paraskevaidi *et al.*, 2018d), identification of prostate cancer tissue samples (Siqueira *et al.*, 2017), differentiation of endometrial tissue regions (Theophilou *et al.*, 2018), toxicology screening (Duan *et al.*, 2019; Morais *et al.*, 2018b), and microbiologic studies involving fungi and virus identification (Costa *et al.*, 2016; Morais *et al.*, 2017; Santos *et al.*, 2017).

However, before model construction, a fundamental step is to split the spectral dataset into at least two subsets: training and test. The training set is used for model construction and the test set for final model evaluation. Model optimization is often performed using cross-validation, where samples from the training set are used in an interactive process of model validation. Figure 4.1a contains a flowchart illustrating the fundamental steps for model construction. Usually, sample splitting is performed by random-selection or Euclidian-distance using the Kennard-Stone (KS) algorithm (Kennard & Stone, 1969). This protocol provides a computational methodology for sample splitting based on a combination of the Euclidian-distance methodology of KS with a random-mutation factor to optimize sample selection, maximizing classification rates. This algorithm, named Morais-Lima-Martin (MLM), is illustrated in Figure 4.1b.

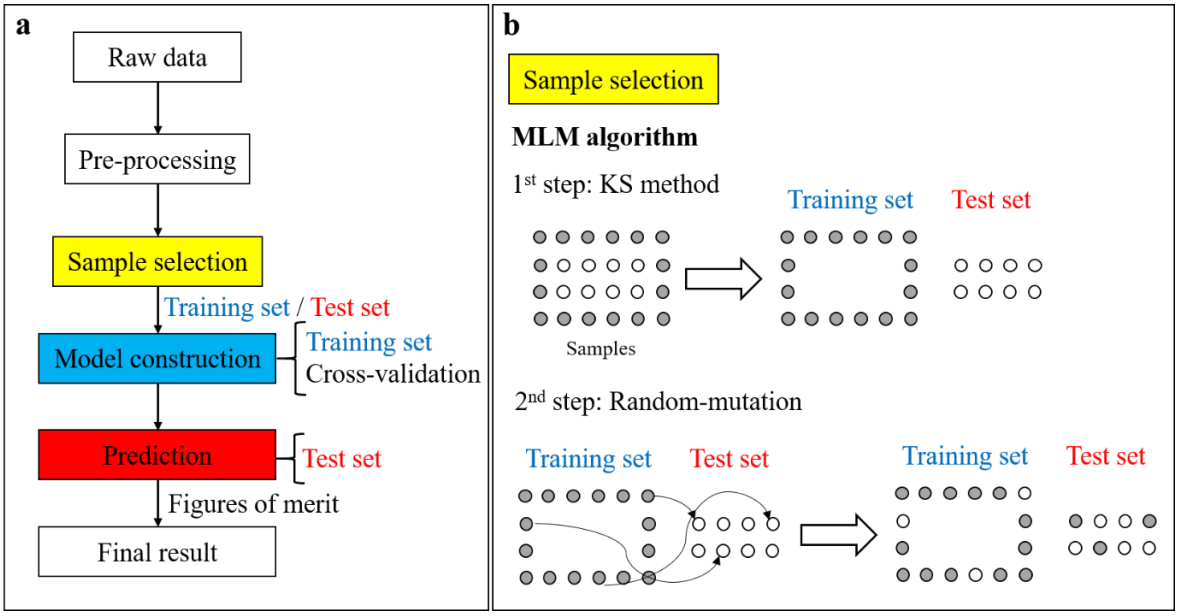


Figure 4.1. A computational methodology for sample splitting based on a combination of the Euclidian-distance methodology of KS with a random-mutation factor to optimize sample selection. (a) Flowchart for IR data processing in classification applications; (b) illustration of sample selection using MLM algorithm.

4.2 Equipment

4.2.1 Requirements for Running this Protocol

- MATLAB R2014b (version 8.4) or above (https://www.mathworks.com). The algorithm, however, might work in older versions of MATLAB;
- MLM algorithm, available for download at https://doi.org/10.6084/m9.figshare.7393517.v1;
- A classed spectroscopy dataset (a sample dataset is provided together with the algorithm).

4.2.2 Preparing Data Files

MLM algorithm only works within MATLAB environment. Data should be loaded and saved in .mat format. Spectral data must be organized into matrices, where each spectrum corresponds to a row, and spectral variables are distributed among the

columns. Figure 4.2a illustrates an example of dataset with 2 classes within MATLAB environment.

CAUTION. IR spectra must be pre-processed before sample selection. Pre-processing methodologies for IR spectral data of biological materials can be found elsewhere (Baker *et al.*, 2014).

4.3 Procedure

Algorithm installation

(1) Download and extract the "MLM.zip" file to a folder of choice;

(2) start MATLAB;

(3) navigate within MATLAB to the folder where the "MLM.zip" file was extracted;

(4) within MATLAB, right click on the folder "MLM" and select "Add to Path > Selected Folders and Subfolders".

Selecting the dataset

To execute the example dataset, go to the folder "MLM > DATASET" within MATLAB, and double-click on the file 'DATASET.mat'. For running the algorithm with another dataset, navigate within MATLAB to the "work" folder (i.e., the folder containing the dataset of interest), and double-click on it.

Using MLM algorithm

MLM algorithm was built to divide the spectral cohort into training and test sets. The training set should contain 70% of the samples, and the test set 30% of the samples. For this, firstly it is necessary to calculate how many samples must be assigned to the training and test set. For example, in the example dataset depicted in Figure 4.2a, class 1 is divided into 98 samples for training (70%, 0.7 140 = 98) and 42 samples for test (30%, 0.3 140 = 42); and class 2 is divided into 70 samples for training (70%, 0.7 100 = 70) and 30 samples for test (30%, 0.3 100 = 30).

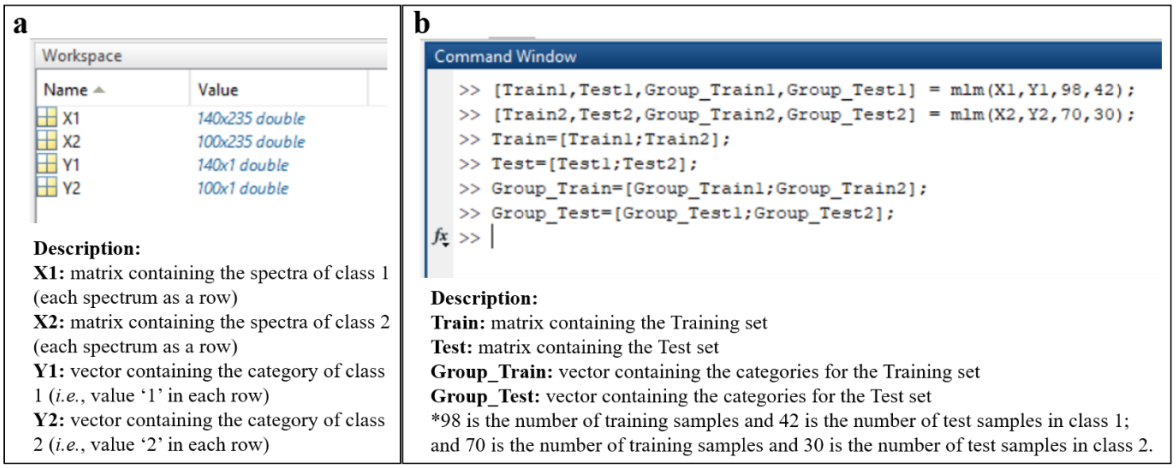


Figure 4.2. Using the MLM algorithm (a) Example dataset within MATLAB, containing 140 spectra for class 1 and 100 spectra for class 2; (b) commands for running MLM algorithm.

After the number of training and test sample for each class is calculated, the algorithm should be applied by typing the commands depicted in Figure 4.2b in the MATLAB Command Window. In this figure, the following steps are performed:

Sample splitting for class 1, where 98 is the number of training samples and 42 is the number of test samples:

[Train1,Test1,Group_Train1,Group_Test1] = mlm(X1,Y1,98,42);

(2) Sample splitting for class 2, where 70 is the number of training samples and 30 is the number of test samples:

[Train2,Test2,Group_Train2,Group_Test2] = mlm(X2,Y2,70,30);

(3) Building the Training set by combining the training samples of class 1 and 2:

Train=[Train1;Train2];

(4) Building the Test set by combining the test samples of class 1 and 2:

Test=[Test1;Test2];

(5) Building the group category representing the training samples:

Group_Train=[Group_Train1;Group_Train2];

(6) Building the group category representing the test samples:

Group_Test=[Group_Test1;Group_Test2];

For more than two classes, the procedure is the same, where the sample splitting is performed for each class separately. The random-mutation factor is set as 10% (default).

CAUTION. The number of training and test samples for each class must be an integer value. In the case of 70% and 30% generate numbers with decimal places, they must be rounded to the closest integer value (e.g., 25.7 to 26; 14.2 to 14; 70.9 to 71; etc).

4.4 Timing

Time is dependent on the computer setup, number of spectra, and number of variables (wavenumbers) in the dataset. Time of analysis of each dataset was practically instantaneous (<1 second) using the follow computational settings: Intel® CoreTM i7 (2.80 GHz) processor with 16.0 GB of RAM memory.

4.5 Troubleshooting

If MLM algorithm does not work: verify that the MLM folder containing the MATLAB routines was added to the MATLAB path. Also, verify if the input numbers of samples (i.e., number of training samples + number of test samples) are equal to the total number of samples.

If you cannot load the sample dataset: verify that your current working directory within MATLAB is the folder containing the dataset (folder named 'DATASET').

4.6 Anticipated Results

The sample dataset used in this protocol is composed of 140 spectra representing control brain tissue samples (class 1) and 100 spectra representing cancer (glioblastoma) brain tissue samples (class 2) (Figure 4.3a). Further details about this dataset can be found in Gajjar *et al.* (2013). Samples were divided into training (70%) and test (30%) sets as

depicted in Figure 4.2b. Two classification algorithms were applied: principal component analysis linear discriminant analysis (PCA-LDA) (Morais & Lima, 2018) and partial least squares discriminant analysis (PLS-DA) (Brereton & Lloyd, 2014). PCA-LDA was applied using 9 principal components (99% cumulative explained variance) with cross-validation venetian blinds (10 data splits). Similarly, PLS-DA was performed using 9 latent variables (98% cumulative explained variance) with cross-validation venetian blinds (10 data splits). Models were built using the Classification Toolbox for MATLAB (http://www.michem.unimib.it/) (Ballabio & Consonni, 2013) and the PLS Toolbox version 7.9.3 (Eigenvector Research, Inc., US). Data were mean-cantered before analysis. The classification performance of these algorithms in the training and test sets are shown in Table 4.1. In both PCA-LDA and PLS-DA, the accuracy values of the training and test sets are similar, indicating absence of overfitting. Also, MLM algorithm provided well-balanced sensitivities and specificities, indicating that the classification methods have similar predictive performance in both classes (control and cancer).

PLS-DA model achieved the best classification performance, with an accuracy of 94% in the test set. Figure 4.3b shows the discriminant function (DF) graph of PCA-LDA, where some superposition between control and cancer samples are observed. On the other hand, the DF graph for PLS-DA (Figure 4.3c), shows a clear separation between the two group of samples, with only a few cancer samples misclassified as control. The receiver operating characteristic (ROC) curve for PLS-DA shows the great performance of this algorithm towards differentiation of control and cancer brain tissue, where an area under the curve (AUC) value of 0.971 is obtained (Figure 4.3d). Glioblastoma is the type of brain cancer with the poorest survival rate, particularly due to its poor prognosis, and its clinical diagnosis is much dependent on subjective and time-consuming analysis (Gajjar *et al.*, 2013). New clinical methodologies for tumour detection are needed in order to overcome these limitations; and IR spectroscopy, due to its non-destructive nature, fast data acquisition and processing, and relative low-cost might aid this type of diagnosis in the future. This protocol demonstrates the usage of sample selection, by means of MLM algorithm, for building classification models with good predictive performance in IR spectral datasets of biological-derived applications.

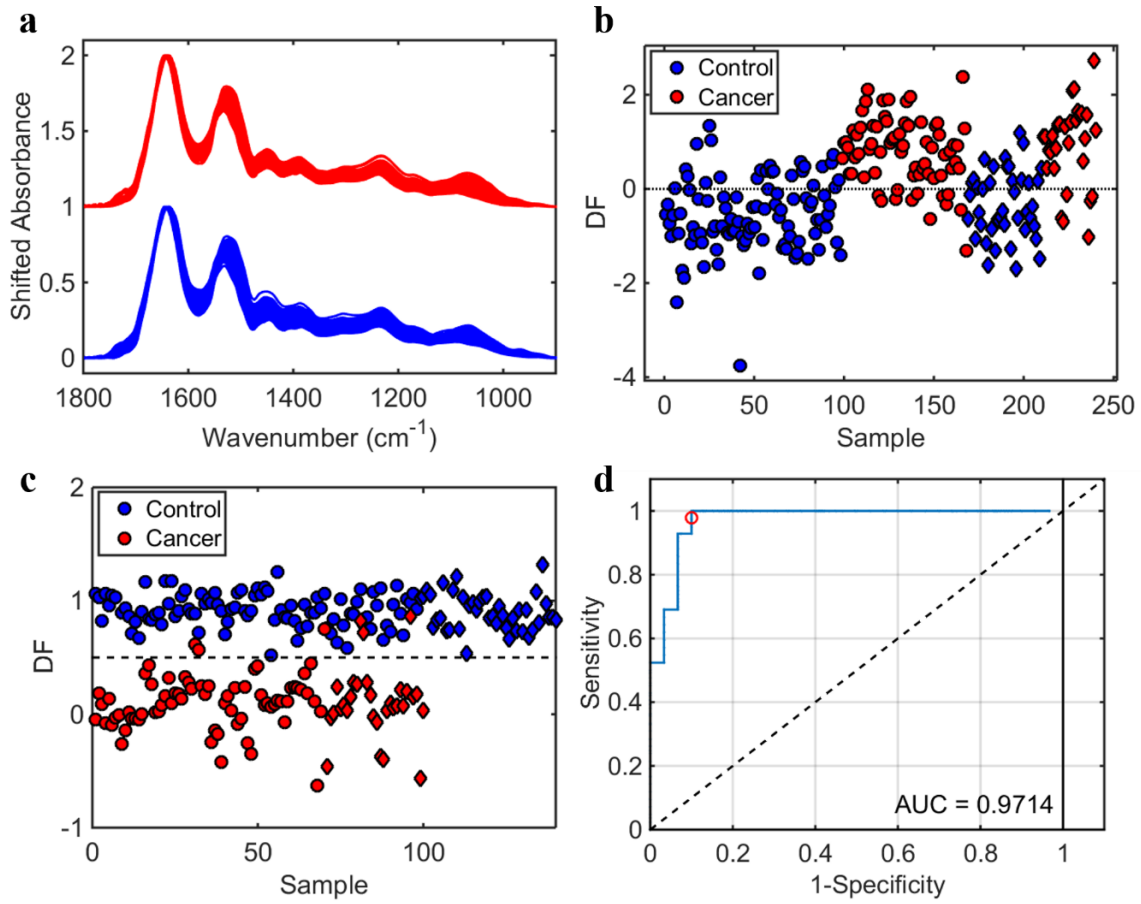


Figure 4.3. The sample dataset used in this protocol. (a) Pre-processed spectra (in blue: control samples; in red: cancer samples); (b) discriminant function (DF) graph representing the canonical variables of PCA-LDA (circles: training samples; diamonds: test samples); (c) discriminant function (DF) graph showing the predicted values of PLS-DA (circles: training samples; diamonds: test samples); (d) Receiver operating characteristic (ROC) curve for PLS-DA, where AUC stands for area under the curve.

Table 4.1. Classification performance of PCA-LDA and PLS-DA algorithms applied to the sample dataset.

Algorithm	Subset	Accuracy (%)	Sensitivity (%)	Specificity (%)
PCA-LDA	Training	82	83	80
	Cross-validation	77	82	71
	Test	81	81	80
PLS-DA	Training	98	99	97
	Cross-validation	96	97	96
	Test	94	98	90

CHAPTER 5 | DETERMINATION OF MENINGIOMA BRAIN TUMOUR GRADES USING RAMAN MICROSPECTROSCOPY IMAGING

This chapter is published in Analyst (IF 4.019). It demonstrates the use of Raman microspectroscopy imaging to discriminate meningioma brain tumour grades based on chemometric techniques:

- Morais CLM, Lilo T, Ashton KM, Davis C, Dawson TP, Gurusinghe N, Martin FL. Determination of meningioma brain tumour grades using Raman microspectroscopy imaging. Analyst **2019**; 144: 7024–7031. https://doi.org/10.1039/C9AN01551E

Abstract: Raman spectroscopy is a powerful technique used to analyse biological materials, where spectral markers such as proteins (1500–1700 cm^{-1}), carbohydrates (470–1200 cm^{-1}) and phosphate groups of DNA (980, 1080–1240 cm^{-1}) can be detected in a complex biological medium. Herein, Raman microspectroscopy imaging was used to investigate 90 brain tissue samples in order to differentiate meningioma Grade I and Grade II samples, which are the commonest types of brain tumour. Several classification algorithms using feature extraction and selection methods were tested, in which the best classification performances were achieved by principal component analysis-quadratic discriminant analysis (PCA-QDA) and successive projections algorithm-quadratic discriminant analysis (SPA-QDA), resulting in accuracies of 96.2%, sensitivities of 85.7% and specificities of 100% using both methods. A biochemical profiling in terms of spectral markers was investigated using the difference-between-mean (DBM) spectrum, PCA loadings, SPA-QDA selected wavenumbers, and the recovered imaging profiles after multivariate curve resolution alternating least squares (MCR-ALS), where the following wavenumbers were found to be associated with class differentiation: 850 cm^{-1} (amino acids or polysaccharides), 1130 cm^{-1} (phospholipid structural changes), the region between 1230–1360 cm^{-1} (Amide III and CH_2 deformation), 1450 cm^{-1} (CH_2 bending), and 1858 cm^{-1} (C=O stretching). These findings highlight the potential of Raman microspectroscopy imaging for determination of meningioma tumour grades.

Author contribution: C.L.M.M. performed the experiments, data analysis and wrote the manuscript.

Camilo L. M. Morais, PhD candidate

Prof. Francis L. Martin, Supervisor

5.1 Introduction

Raman spectroscopy provides sensitive spectrochemical signatures of materials based on their molecular polarisability changes (Kelly *et al.*, 2011). Raman is based on an inelastic scattering phenomenon that occurs in less than 1% of the absorbed photons by a molecule. This inelastic scattering is composed of Stokes and anti-Stokes scattering: the former occurs when the molecule emits a photon with less energy than the absorbed incoming radiation, and the latter happens when the molecule emits a photon with higher energy than the absorbed incoming radiation (Santos *et al.*, 2017). At room temperature, the Stokes scattering is more frequent, thus most instruments filter the elastic and anti-Stokes scattering and record the Stokes scattering signal as the final Raman spectrum.

Microspectroscopy Raman imaging allows one to obtain microscopically spatially distributed spectral data, where each position in the image is composed of a Raman spectrum in a specific wavenumber range. The hyperspectral image data are represented by three-dimensional (3D) arrays, where the spatial coordinates are present in the *x*- and *y*-axis while the spectral information is in the *z*-axis. A major advantage of Raman imaging is that it can be non-destructive depending on the incident laser frequency, has minimum water interference, and has a relatively low cost in comparison with other analytical techniques.

Raman imaging has been used in a wide range of applications, including pharmaceutical analysis (Kandpal *et al.*, 2018), forensic investigations (Almeida *et al.*, 2017), food quality control (Yaseen *et al.*, 2017), and to analyse biological materials (Butler *et al.*, 2016). In the latter, cancer detection plays an important role, where Raman imaging has been successfully applied to investigate breast (Abramczyk & Brozek-Pluska, 2013), cervical (Diem *et al.*, 2013), lung (Diem *et al.*, 2013), skin (Lui *et al.*, 2012), cancer (Kirsch *et al.*, 2010), and ovarian cancer (Morais *et al.*, 2019e).

Most of brain cancers are gliomas or meningioma tumours (Gajjar *et al.*, 2013). Gliomas are more aggressive types of tumours and have been widely investigated using Raman spectroscopy (Bury *et al.*, 2019b; Desroches *et al.*, 2018; Gajjar *et al.*, 2013; Livermore *et al.*, 2019), while meningiomas remain to be intensively investigated using vibrational spectroscopy. Meningiomas represent 20% to 35% of all primary intracranial tumours (Takahashi *et al.*, 2019). The majority of them occur in a supratentorial location; however, a few of them can arise in the posterior cranial fossa and, more rarely, as

extracranial meningiomas (Takahashi *et al.*, 2019). It usually manifests as single or sporadic lesions, causing symptoms such as sensory and motor deficits and gait disturbance; while multiple meningiomas are often associated with neurofibromatosis type II (Yeo *et al.*, 2019). Meningiomas can be divided into WHO Grade I, Grade II and Grade III. Grade I meningiomas are the commonest type of tumours, with slower growth and lower likelihood of recurrence; Grade II meningiomas also have a slower growth but higher likelihood of recurrence; and Grade III meningiomas are a very rare type of tumour with fast growing rate and much higher likelihood of recurrence. Surgical outcomes and treatment are dependent on the meningioma grade and histological subtypes (Yeo *et al.*, 2019).

In this thesis, Raman microspectroscopy imaging is applied to distinguish Grade I and Grade II meningiomas via the application of several chemometric approaches, including combination of feature extraction and selection methods with discriminant analysis techniques, and multivariate curve resolution alternating least squares (MCR-ALS) for profiling and differentiation of Grade I and Grade II tumour tissues.

5.2 Materials and Methods

5.2.1 Samples

Ninety brain tissue samples (66 meningiomas WHO Grade I, 24 meningiomas WHO Grade II) were analysed by a Renishaw InVia Basis Raman spectrometer coupled to a confocal microscope (Renishaw plc, UK). All samples were sourced from the Brain Tumour North West (BTNW) biobank (NRES14/EE/1270). All experiments were performed in accordance with the STEMH (Science, Technology, Engineering, Medicine and Health) Guidelines at the University of Central Lancashire, and approved by the ethics committee at the University of Central Lancashire (STEMH 917). Informed consents were obtained from human participants of this study. Formalin-fixed paraffin-embedded (FFPE) tissue specimens (10-μm-thick) were placed onto aluminium-covered glass slides for spectroscopy measurement. Microspectroscopy imaging was performed with an acquisition area of approx. 100 x 50 μm (50× magnification, 785 nm laser, 50% laser power (150 mW), 0.1 s exposure time, 780–1858 cm^{-1} spectral range) using the StreamHR^{TM} imaging technique (high-confocality mode) with a grid area of 42 x 28

pixels, resulting in 1176 spectra for each image (1 cm^{-1} data spacing). The laser power was set relatively high to ensure a good signal-to-noise ratio. To minimize any potential photodamage to the sample, the laser exposure time was set to only 0.1 s. Moreover, no damage was visually observed in the samples after measurement. The imaging acquisition time was approx. 8 min for each sample.

5.2.2 Computational Analysis

The Raman images were converted into suitable .txt files using the Renishaw WiRE software, and processed using MATLAB R2014b (MathWorks, Inc., USA) with lab-made routines. All the samples' images were pre-processed by cosmic rays (spikes) removal, Savitzky-Golay smoothing (window of 15 points, 2^{nd} order polynomial fitting), and asymmetric least squares baseline correction. The window size in the Savitzky-Golay smoothing was determined visually by testing different window sizes, where the smallest window size that removed random noise and kept the same spectral shape and intensity without smoothing-out relevant spectral peaks was chosen. MCR-ALS was applied to the image data using the HYPER-Tools toolbox in MATLAB (Mobaraki & Amigo, 2018).

First-order classification. Each pre-processed image with size 42 x 28 x 1015 was averaged into a single spectrum (1 x 1015) as the classification was performed on a sample basis. Initially, an outlier detection test was performed by a Hotelling T^2 *versus* Q residuals test (Morais *et al.*, 2019c). The remaining samples after outlier removal were split into training (60%), validation (20%) and test (20%) sets using the MLM sample selection algorithm (Morais *et al.*, 2018a; Morais *et al.*, 2019d). All data were mean-centred before further analysis.

For feature extraction and classification, principal component analysis combined with linear discriminant analysis (PCA-LDA), quadratic discriminant analysis (PCA-QDA) and support vector machines (PCA-SVM) were applied to the pre-processed data. PCA reduces the pre-processed spectral variables to a small number of principal components (PCs) responsible for the majority of the original data-explained variance. Each PC is orthogonal to each other and is generated in a decreasing order of explained variance, where the first PC explains most of the data variance, followed by the second PC, and so on. The PCs are composed of scores and loadings, the scores representing the variance on the sample direction, thus being used to identify similarities and

dissimilarities between the samples; and, the loadings represent the variance on the wavenumber direction, being used to identify possible spectral markers associated with class differentiation (Bro & Smilde, 2014). PCA decomposition takes the form (Bro & Smilde, 2014):

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \tag{5.1}$$

where **X** is a matrix containing the mean-centred pre-processed spectral data; **T** is a matrix containing the PCA scores for a determined number of PCs; **P** is a matrix containing the PCA loadings for a determined number of PCs; **E** is a residual matrix; and the superscript **T** represents the matrix transpose operation.

In PCA-LDA, PCA-QDA and PCA-SVM, a PCA model is applied to the pre-processed data and then a LDA, QDA or SVM classifier is applied to the PCA scores, respectively. LDA and QDA are discriminant analysis methods based on a Mahalanobis distance calculation. LDA assumes classes having similar variance structures, therefore using a pooled covariance matrix to calculate the classification score for each class, while QDA assumes classes having different variance structures, therefore using the variance-covariance matrix for each class individually when calculating the classification score (Dixon & Brereton, 2009; Morais & Lima, 2018). The LDA ($\boldsymbol{L_{ik}}$) and QDA ($\boldsymbol{Q_{ik}}$) classification scores can be calculated in a non-Bayesian form by (Dixon & Brereton, 2009; Morais & Lima, 2018):

$$\boldsymbol{L_{ik}} = (\mathbf{x_i} - \bar{\mathbf{x}}_{\boldsymbol{k}})^{\mathbf{T}}\mathbf{C}_{\mathbf{pooled}}^{-\mathbf{1}}(\mathbf{x_i} - \bar{\mathbf{x}}_{\boldsymbol{k}}) \tag{5.2}$$

$$\boldsymbol{Q_{ik}} = (\mathbf{x_i} - \bar{\mathbf{x}}_{\boldsymbol{k}})^{\mathbf{T}}\mathbf{C}_{\boldsymbol{k}}^{-\mathbf{1}}(\mathbf{x_i} - \bar{\mathbf{x}}_{\boldsymbol{k}}) \tag{5.3}$$

where $\mathbf{x_i}$ is a vector containing the input classification variables (*e.g.*, PCA scores) for sample $\boldsymbol{i}$; $\bar{\mathbf{x}}_{\boldsymbol{k}}$ is the mean vector of input classification variables for class $\boldsymbol{k}$; $\mathbf{C_{pooled}}$ is the pooled covariance matrix; and $\mathbf{C}_{\boldsymbol{k}}$ is the variance-covariance matrix of class $\boldsymbol{k}$. $\mathbf{C_{pooled}}$ and $\mathbf{C}_{\boldsymbol{k}}$ are calculated as follows (Morais & Lima, 2018):

$$\mathbf{C_{pooled}} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{n_k}\mathbf{C}_{\boldsymbol{k}} \tag{5.4}$$

$$\mathbf{C}_{\boldsymbol{k}} = \frac{\mathbf{1}}{\boldsymbol{n_k - 1}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n_k}}(\mathbf{x_i} - \bar{\mathbf{x}}_{\boldsymbol{k}})(\mathbf{x_i} - \bar{\mathbf{x}}_{\boldsymbol{k}})^{\mathbf{T}} \tag{5.5}$$

in which $\boldsymbol{n}$ is the total number of samples in the training set; $\boldsymbol{K}$ is the total number of classes; and $\boldsymbol{n_k}$ is the number of samples in class $\boldsymbol{k}$.

SVM is a binary linear classifier with a non-linear step called the kernel transformation (Cortes & Vapnik, 1995). A kernel function transforms the input data

space into a feature space by a applying a mathematical transformation which is often non-linear. Then, a linear decision boundary is fit between the closest samples to the border of each class (called support vectors), where each class is defined. SVM classification is performed as follows (Cortes & Vapnik, 1995; Morais *et al.*, 2017):

$$\boldsymbol{f(x) = \mathbf{sign}\left(\sum_{i=1}^{N_{SV}} \alpha_i y_i \phi\left(\mathbf{x}_i, \mathbf{z}_j\right) + b\right)} \quad (5.6)$$

where $\mathbf{x}_i$ and $\mathbf{z}_j$ are vectors containing sample measurement vectors (*e.g.*, PCA scores); N_{SV} is the number of support vectors; α_i is the Lagrange multiplier for sample i; y_i is the class membership of sample i (±1); $\phi\left(\mathbf{x}_i, \mathbf{z}_j\right)$ is the kernel function; and b is the bias parameter.

SVM was performed using a radial basis function (RBF) kernel, which is defined by (Morais *et al.*, 2017):

$$\phi\left(\mathbf{x}_i, \mathbf{z}_j\right) = \exp\left(-\gamma \left\|\mathbf{x}_i - \mathbf{z}_j\right\|^2\right) \quad (5.7)$$

where γ is the kernel parameter that determines the RBF width. Cross-validation venetian blinds with 10 data splits was performed to optimise the bias and kernel parameter.

Some feature selection techniques were used to analyse the image spectral data. Successive projections algorithm (SPA) (Soares *et al.*, 2013) and genetic algorithm (GA) (McCall, 2005) were used coupled with LDA, QDA and SVM. SPA is a forward feature selection method which operates by minimising the co-linearity of original pre-processed spectra; thus, selecting wavenumbers whose information content is minimally redundant (Theophilou *et al.*, 2018). GA is an iterative algorithm inspired by Mendelian genetics, where the pre-processed spectral data is reduced to a set of selected wavenumbers based on an evolutionary process (McCall, 2005). For this, a set of variables is randomly chosen to go through combinations, cross-overs and mutations until the best set of variables reaches the minimum of a pre-defined cost function (McCall, 2005; Santos *et al.*, 2017). The optimum number of variables for SPA and GA is obtained by minimizing the average risk G of misclassification in the validation set (Theophilou *et al.*, 2018; Siqueira *et al.*, 2017):

$$G = \frac{1}{N_V} \sum_{n=1}^{N_V} g_n \quad (5.8)$$

where N_V is the number of samples in the validation set and g_n is defined by:

$$g_n = \frac{r^2\left(\mathbf{x}_n, m_{I(n)}\right)}{\min_{I(m) \neq I(n)} r^2\left(\mathbf{x}_n, m_{I(m)}\right)} \quad (5.9)$$

where $r^2(\mathbf{x}_n, \boldsymbol{m}_{I(n)})$ is the squared Mahalanobis distance between sample $\mathbf{x}_n$ of class $I(n)$ and the centre of its true class ($\boldsymbol{m}_{I(n)}$); and $r^2(\mathbf{x}_n, \boldsymbol{m}_{I(m)})$ is the squared Mahalanobis distance between object $\mathbf{x}_n$ and the centre of the closest incorrect class ($\boldsymbol{m}_{I(m)}$). The GA routine was carried out using 100 generations containing 200 chromosomes each. Cross-over and mutation probabilities were set to 60% and 1%, respectively. The algorithm was repeated three times, starting from different random initial populations, and the best solution in terms of fitness value was employed.

MCR-ALS. Multivariate curve resolution alternating least squares (MCR-ALS) assumes a bilinear model that is the multi-wavelength extension of the Beer-Lambert's law. It decomposes an experimental matrix **D** into concentration and spectral profiles as follows (Jaumot *et al.*, 2015):

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \tag{5.10}$$

where **C** is a matrix containing the concentration profiles for a determined number of pure components in **D**; **S** is a matrix containing the spectral profiles for the pure components in **D**; and **E** is a residual matrix.

MCR-ALS can remove noise and physical/chemical interferences from the spectral matrix **D**, and allow one to recover the pure concentration and spectral profiles of the components that make the spectral matrix **D**. MCR-ALS is very useful to handle image data since it allows the reconstruction of image maps based on the recovered concentration profiles, where one can identify spatial and chemical differences between the samples being imaged (Prats-Montalbán *et al.*, 2011).

Model validation. The models were validated by calculating some quality parameters such as accuracy, sensitivity, specificity, and F-score. Accuracy represents the total number of samples correctly classified considering true and false negatives; sensitivity represents the proportion of positives that are correctly classified; specificity represents the proportion of negatives that are correctly classified; and, F-score measures the model performance considering imbalanced data (Morais & Lima, 2017). The equations to calculate these parameters are depicted in Table 5.1. In addition, the area under the curve (AUC) of the receiver operating characteristic (ROC) curve was evaluated to assess model quality. AUC values between 0.7 and 0.8 are considered acceptable, between 0.8 and 0.9 are considered excellent, and above 0.9 are considered outstanding (Mandrekar, 2010).

Table 5.1. Quality parameters for model validation. Where: TP stands for true positive, TN for true negative, FP for false positive, and FN for false negative.

Parameter	Equation
Accuracy (%)	$\left(\frac{\mathbf{TP + TN}}{\mathbf{TP + FP + TN + FN}}\right) \times \mathbf{100}$
Sensitivity (%)	$\left(\frac{\mathbf{TP}}{\mathbf{TP + FN}}\right) \times \mathbf{100}$
Specificity (%)	$\left(\frac{\mathbf{TN}}{\mathbf{TN + FP}}\right) \times \mathbf{100}$
F-score	$\frac{\mathbf{2 \times Sensitivity \times Specificity}}{\mathbf{Sensitivity + Specificity}}$

5.3 Results and Discussion

Ninety brain tissue samples (66 meningiomas Grade I, 24 meningiomas Grade II) were analysed by Raman microspectroscopy imaging. The median microscopic and Raman image for meningiomas Grade I and Grade II are depicted in Figures 5.1a–1d (the colour figures represent the mean response (average Raman intensity between 780–1858 cm^{-1}) of the median image for each group). Notably, each image presents different visual features due to the different distributions of chemicals on the sample surface, but their spectrochemical profile are very similar as shown in Figure 5.1e and 5.1f, indicating that chemical differences between meningiomas Grade I and Grade II are not visually clear.

The pre-processed spectra from the images acquired in the spectral range between 780–1858 cm^{-1} (Figure 5.1f) were used for further analysis. This spectral region includes the Raman fingerprint region, hence, encompassing spectrochemical signals of the main biomolecules present in the tissue samples (Kelly *et al.*, 2011). The assignment of the main peaks of the pre-processed Raman spectrum is depicted in Figure 5.1f. These include C-C stretching [$\nu(C\text{-}C)_1$] in amino acids or polysaccharides at 850 cm^{-1}, C-C stretching [$\nu(C\text{-}C)_2$] in proteins at 890 cm^{-1}, C-C stretching [$\nu(C\text{-}C)_3$] in amino acids at 930 cm^{-1}, C-C stretching [$\nu(C\text{-}C)_4$] in phenylalanine at 1003 cm^{-1}, phospholipid structural changes at 1130 cm^{-1}, Amide III peak at 1265 cm^{-1}, CH_2 bending [$\delta(CH_2)_1$] in lipids at 1296 cm^{-1}, CH_3/CH_2 deformation modes in DNA/RNA at 1336 cm^{-1}, CH_2 bending [$\delta(CH_2)_2$] in malignant tissues at 1450 cm^{-1}, NH_2 bending [$\delta(NH_2)$] in cytosine at 1610 cm^{-1}, and Amide I absorption at 1665 cm^{-1} (Movasaghi *et al.*, 2007). Some of these peaks are discriminant features between the samples and some of them are common amongst the

tumour types. The identification of relevant distinguishing spectral features between Grade I and Grade II samples are achieved by chemometric techniques.

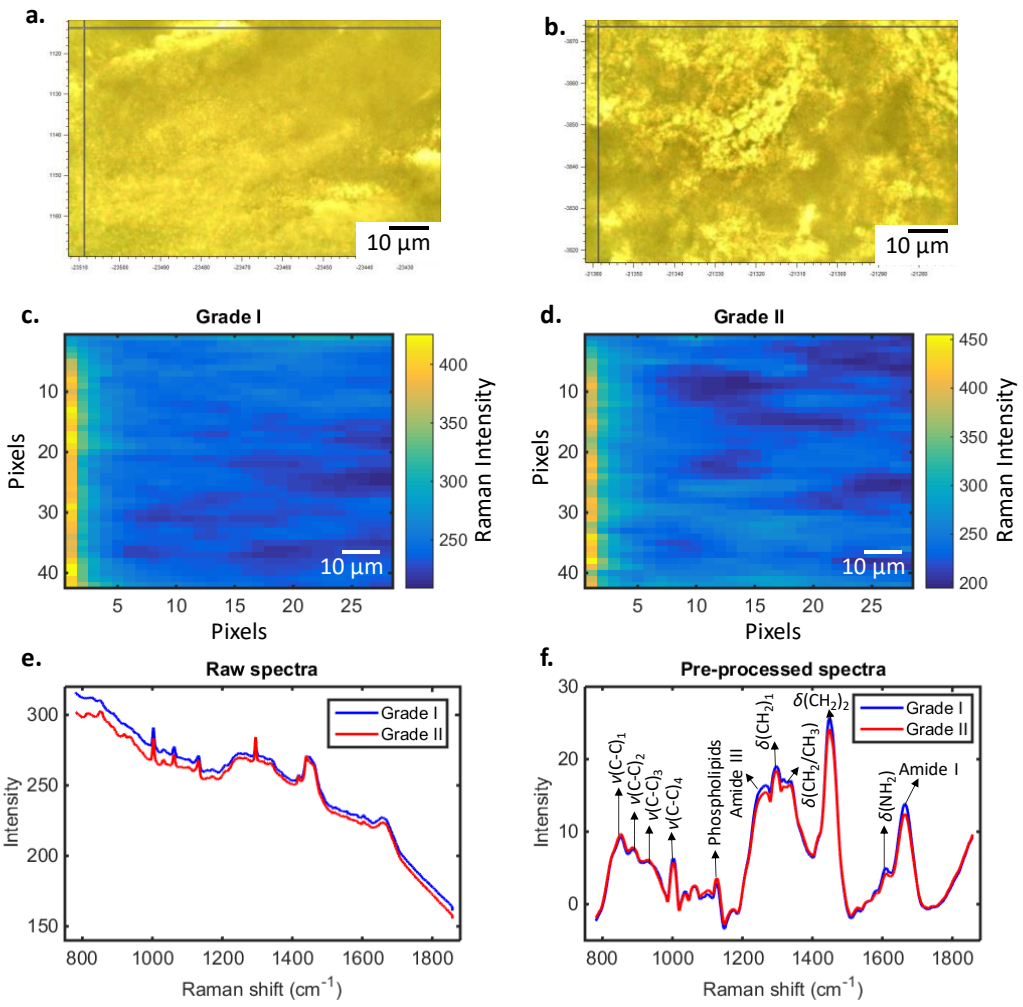


Figure 5.1. Median Raman miscrospectroscopy images. (a) Microscopic image of Grade I meningioma tissue; (b) microscopic image of Grade II meningioma tissue; (c) median raw image for meningioma Grade I samples; (d) median raw image for meningioma Grade II samples; (e) median raw spectra for meningiomas Grade I and Grade II; (f) median pre-processed spectra (Savitzky-Golay smoothing and asymmetric least squares baseline correction) for meningiomas with a tentative assignment of the main Raman peaks. Grade I and Grade II. Colour bar: Raman intensity. ν: stretching vibration, δ: bending.

Initially, outlier detection was performed by a Hotelling T^2 versus Q residuals test, where 4 samples (2 meningiomas Grade I, 2 meningiomas Grade II) were removed (see Appendix B Figure B1). First-order algorithms were used to analyse the pre-processed spectral data after outlier removal.

Feature extraction and classification by means of PCA-LDA, PCA-QDA and PCA-SVM; and feature selection and classification by means of SPA-LDA, SPA-QDA, SPA-SVM, GA-LDA, GA-QDA and GA-SVM, were applied to distinguish meningiomas Grades I and II on sample basis. Amongst the PCA-based algorithms (using 8 PCs, 98.94% explained variance, see ESI Figure S2 Appendix‡), the best performance was obtained with PCA-QDA (96.2% accuracy, 85.7% sensitivity, 100% specificity, and F-score = 92.3%). Also, SPA-QDA was the best algorithm amongst SPA-based methods, with the same performance of PCA-QDA. GA-based methods showed overall poorer performance, where the best algorithm (GA-QDA) achieved 73.1% accuracy but 0% sensitivity, indicating that GA-based models are most likely overfitted. More details about the predictive performance of each of these algorithms are provided in Table 5.2.

The ROC curve for PCA-QDA and SPA-QDA models are shown in Figure 5.2, where the AUC value was found at 0.929 indicating an outstanding classification performance for both algorithms.

Table 5.2. Quality parameter for distinguishing Grade I and Grade II meningiomas in the test set.

Algorithm	Accuracy	Sensitivity	Specificity	F-score
PCA-LDA	46.2%	85.7%	31.6%	46.2%
PCA-QDA	96.2%	85.7%	100%	92.3%
PCA-SVM	61.6%	28.6%	73.7%	41.2%
SPA-LDA	57.7%	100%	42.1%	49.3%
SPA-QDA	96.2%	85.7%	100%	92.3%
SPA-SVM	34.6%	71.4%	21.1%	32.5%
GA-LDA	61.5%	57.1%	63.2%	60.0%
GA-QDA	73.1%	0%	100%	0%
GA-SVM	42.3%	42.9%	42.1%	42.5%

QDA-based algorithms exhibit superior performance in comparison with LDA- and SVM-based methods. Usually, for complex biological data, QDA outperforms LDA since QDA-based algorithms model each class variance individually, while LDA assumes classes having similar variance structures (Morais & Lima, 2018). This occurs because the performance of QDA ultimately depends on the variance structure of the data. QDA is expected to work better than LDA for most biological applications, since quite commonly biological samples are composed of complex chemical matrices with different variances structures for each class. For example, diseases' samples can have a smaller variance distribution than healthy control samples, since the latter can be composed of individuals with different life habits, while patients with a same specific disease usually

have a similar life-style. The same can occur with different tumour grades, where one class can assume a different variance distribution in comparison with the other. The only situation where QDA underperforms LDA is when the number of samples in the dataset is small (Wu *et al.*, 1996), since the variance of each group might not be totally covered by QDA hence increasing the degree of extrapolation needed and commonly leading the model to overfitting.

SVM-based models seem to be highly overfitted, since the training performance for these algorithms are excellent (see ESI Table S1, Appendix‡), with near 100% correct classification rates; however, test performance is highly affected as demonstrated in Table 5.2. SVM classification performance would probably improve by adding more samples to the training set, thus creating a most representative training model. Nevertheless, PCA-QDA and SPA-QDA performance are both excellent in the test set, indicating that these algorithms are robust to provide a satisfactory prediction towards external samples.

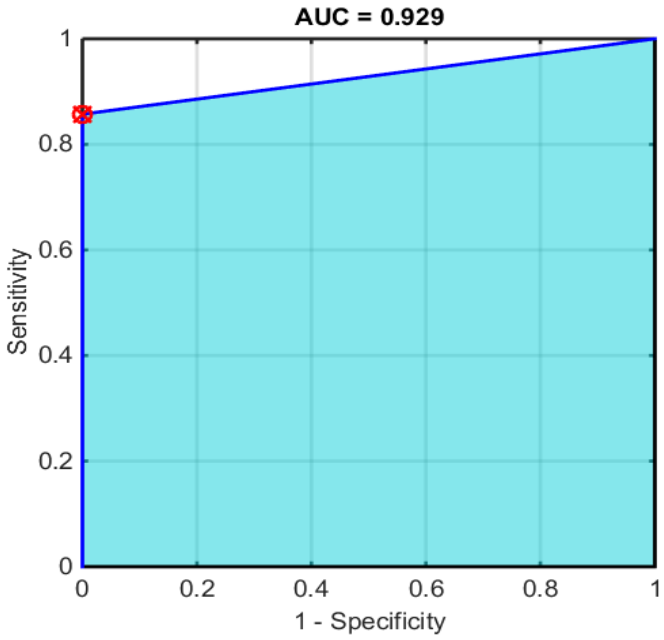


Figure 5.2. Receiver operating characteristic (ROC) curve for PCA-QDA and SPA-QDA. AUC: area under the curve.

The difference-between-mean (DBM) spectrum, PCA loadings on PC1 (56.64% explained variance), and SPA-QDA selected variables are shown in Figure 5.3. The PCA loadings indicate higher coefficients at ~850 cm^{-1}, ~1003 cm^{-1}, ~1130 cm^{-1}, ~1337 cm^{-1}, ~1450 cm^{-1}, ~1665 cm^{-1}, and ~1858 cm^{-1}; and the SPA-QDA selected variables are: ~850

cm^{-1}, ~1130 cm^{-1}, ~1245 cm^{-1}, ~1337 cm^{-1}, ~1450 cm^{-1}, and ~1858 cm^{-1}. Only the variable at 1245 cm^{-1} selected by SPA-QDA does not have a high PCA loadings, while the other variables selected by SPA-QDA are very close or are a perfect match with the ones observed in PCA-QDA. The list of PCA and SPA-QDA selected variables and tentative assignment according to Movasaghi *et al.* (2007) are shown in Table 5.3. The Raman shift at 1858 cm^{-1} is unknown based on this reference, but this wavenumber has been associated to C=O stretching in other literature (Mayo *et al.*, 2003). The peak at around 850 cm^{-1} has been previously detected in meningioma samples as belonging to tyrosine (Mehta *et al.*, 2018), an α-amino acid that constitute important structures in proteins responsible for signal transduction processes (Kato *et al.*, 1993); and the peaks at 1003 cm^{-1} (phenylalanine) and 1450 cm^{-1} (CH_2 bending in DNA) have also been reported as biomarkers of meningioma tumours (Mehta *et al.*, 2018; Zhou *et al.*, 2012). Phospholipids (1130 cm^{-1}), Amide III (1245 cm^{-1}) and Amide I (1665 cm^{-1}) have been reported for brain tumours in general (Gajjar *et al.*, 2013; Zhou *et al.*, 2012).

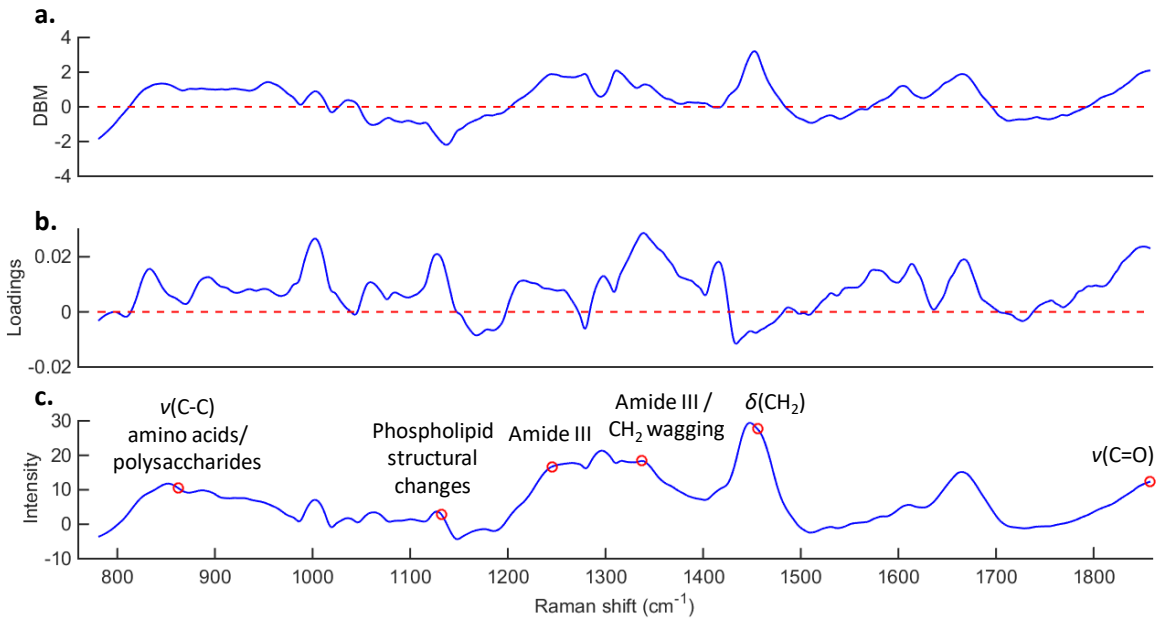


Figure 5.3. PCA loadings and SPA-QDA selected variables. (a) Difference-between-mean (DBM) spectrum (+ values: higher intensity in meningioma Grade I samples; - values: higher intensity in meningioma Grade II samples); (b) PCA loadings on PC1; (c) average training set spectrum and SPA-QDA selected variables (red circles) with their tentative assignment. ν: stretching vibration, δ: bending.

Table 5.3. Tentative assignment of PCA and SPA-QDA selected variables to distinguish meningiomas Grade I and Grade II. DBM: difference-between-mean spectrum, where ↑ represents higher intensity in meningioma Grade I samples, and ↓ represents higher intensity in meningioma Grade II samples.

Peak	Algorithm	Assignment	DBM
850 cm^{-1}	PCA/SPA-QDA	Amino acids or polysaccharides	↑
1003 cm^{-1}	PCA	C-C in phenylalanine	↑
1130 cm^{-1}	PCA/SPA-QDA	Phospholipid structural changes	↓
1245 cm^{-1}	SPA-QDA	Amide III	↑
1337 cm^{-1}	PCA/SPA-QDA	Amide III and CH_2 wagging vibrations	↑
1450 cm^{-1}	PCA/SPA-QDA	CH_2 bending	↑
1665 cm^{-1}	PCA	Amide I	↑
1858 cm^{-1}	PCA/SPA-QDA	C=O stretching	↑

MCR-ALS was employed to resolve the median Grade I and Grade II meningioma images in order to identify spectrochemical changes associated with tumour aggressiveness. MCR-ALS was performed with 4 components selected by singular value decomposition (99.99% explained variance, 0.21 lack of fit, non-negativity in concentration mode). The recovered concentration and spectral profiles of the 4 components are depicted in the ESI Figure S3 (Appendix‡). The 1st component of MCR-ALS was found to be associated with Grade II appearance (Figure 5.4a), once it is clearly present in the Grade II tissue sample. The spectral profile of the 1st component (Sopt 1) indicates distinguishing features at the region between 1230 cm^{-1} and 1360 cm^{-1} in comparison with the spectral profiles for other components (see ESI Figure S3 Appendix‡), where three peaks (1265 cm^{-1}, 1296 cm^{-1} and 1336 cm^{-1}) are presents. These peaks are associated with Amide III, CH_2 deformation in lipids, and CH_2/CH_3 twisting in polynucleotide chains, respectively (Movasaghi *et al.*, 2007). This region encompasses the wavenumber at 1337 cm^{-1} (amide III and CH_2 wagging vibrations) in Table 5.3. Similarly to Figure 5.1f, the peaks at 850 cm^{-1} [$\nu(C\text{-}C)_1$, amino acids or polysaccharides], 890 cm^{-1} [$\nu(C\text{-}C)_2$, proteins], 930 cm^{-1} [$\nu(C\text{-}C)_3$, amino acids], 1003 cm^{-1} [$\nu(C\text{-}C)_4$, phenylalanine], 1130 cm^{-1} (phospholipids), 1265 cm^{-1} (Amide III), 1296 cm^{-1} [$\delta(CH_2)_1$, lipids], 1336 cm^{-1} [$\delta(CH_3/CH_2)$, DNA/RNA], 1450 cm^{-1} [$\delta(CH_2)_2$, malignant tissue], and 1665 cm^{-1} (Amide I) are also present. In addition, other peaks at 1060 cm^{-1} [$\nu(PO_2^-)$, DNA/RNA], 1100 cm^{-1} [$\nu(C\text{-}C)_5$, lipids], and a small arm at 1459 cm^{-1} [$\delta(CH_2)_3$, deoxyribose] ((Movasaghi *et al.*, 2007) are observed as distinguishing features in the MCR-ALS Sopt 1 profile.

Bury *et al.* (2019c) have recently used Raman spectroscopy to discriminate meningioma Grade I brain tissue among different brain pathologies (low-grade glioma,

high-grade glioma, metastasis, lymphoma, and no-tumour) with 94.8% accuracy, 63.9% sensitivity and 97.1% specificity using PCA-LDA with smear-based samples; and, meningioma Grade I brain tissue among low-grade glioma, high-grade glioma, metastasis and lymphoma with 90.8% accuracy, 91.7% sensitivity and 90.8% specificity using PCA-LDA with FFPE samples. Mehta *et al.* (2018) have recently used Raman spectroscopy to discriminate healthy controls and meningioma patients based on serum using PCA-LDA. Seventy patients (35 controls, 35 meningiomas) were analysed, resulting in 70% accuracy to distinguish meningiomas versus controls in an independent test set; 72% accuracy to distinguish meningiomas Grade I versus controls; and 80% accuracy to distinguish meningiomas Grade II versus controls. The results reported herein (96.2% accuracy, 85.7% sensitivity, 100% specificity) are very promising to distinguish meningioma tissue grades, which is critical to delineate patient treatment; and also evidences the potential of Raman spectroscopy to investigate brain tumour tissues.

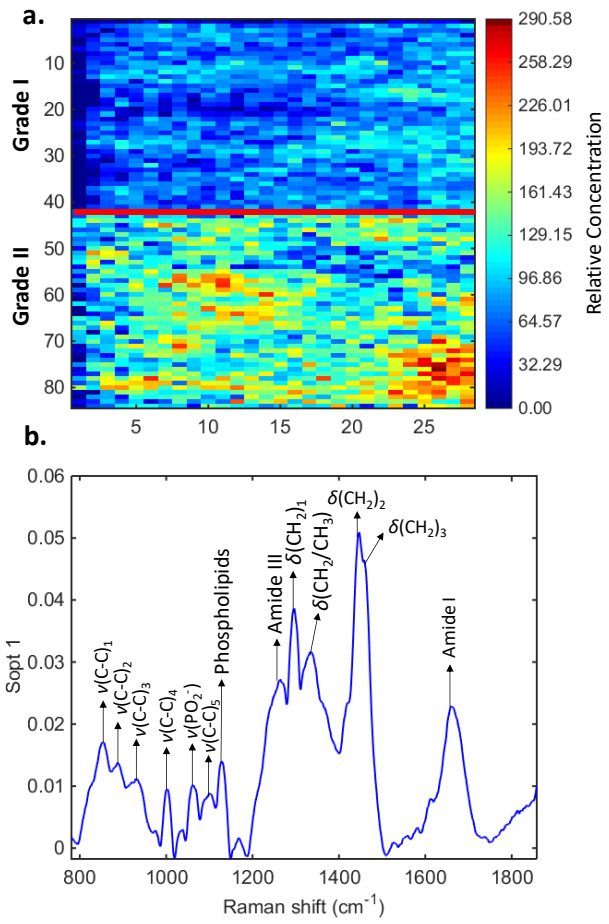


Figure 5.4. MCR-ALS results. (a) Recovered image using the MCR-ALS concentration profile for the 1st component; (b) MCR-ALS spectral profile for the 1st component with its tentative spectral markers assignment. Colour bar: relative concentration.

5.4 Conclusion

Ninety meningioma brain tissue samples (66 meningiomas Grade I, 24 meningiomas Grade II) were investigated using Raman microscpectroscopy imaging. Several chemometric algorithms were applied to distinguish the samples according to the tumour grade, where PCA-QDA and SPA-QDA were found to have to best classification performance at 96.2% accuracy, 85.7% sensitivity and 100% specificity (AUC = 0.929). Spectral bio-markers at 850 cm^{-1}, 1130 cm^{-1}, 1337 cm^{-1}, 1450 cm^{-1} and 1858 cm^{-1} were found in common using both PCA-QDA and SPA-QDA, and a further analysis using MCR-ALS indicated distinguishing features at the region between 1230–1360 cm^{-1} associated with increases in the WHO meningioma tumour grade. The classification results found by PCA-QDA and SPA-QDA are very promising, and show the potential of Raman microspectroscopy to distinguish meningioma tissue grades, thus aiding clinicians to delineate patient treatment.

CHAPTER 6 | A THREE-DIMENSIONAL PRINCIPAL COMPONENT ANALYSIS APPROACH FOR EXPLORATORY ANALYSIS OF HYPERSPECTRAL DATA: IDENTIFICATION OF OVARIAN CANCER SAMPLES BASED ON RAMAN MICROSPECTROSCOPY IMAGING OF BLOOD PLASMA

This chapter is published in Analyst (IF 4.019). It demonstrates a new method (3D-PCA) for exploratory analysis of hyperspectral images:

- Morais CLM, Martin-Hirsch PL, Martin FL. A three-dimensional principal component analysis approach for exploratory analysis of hyperspectral data: identification of ovarian cancer samples based on Raman microspectroscopy imaging of blood plasma. Analyst **2019**; 144: 2312–2319. https://doi.org/10.1039/C8AN02031K

Abstract: Hyperspectral imaging is a powerful tool to obtain both chemical and spatial information of biological systems. However, few algorithms are capable of working with full three-dimensional images, in which reshaping or averaging procedures are often performed to reduce the data complexity. Herein, we propose a new algorithm of three-dimensional principal component analysis (3D-PCA) for exploratory analysis of complete 3D spectrochemical images obtained through Raman microspectroscopy. Blood plasma samples of ten patients (5 healthy controls, 5 diagnosed with ovarian cancer) were analysed by acquiring hyperspectral imaging in the fingerprint region (~780–1858 cm^{-1}). Results show that 3D-PCA can clearly differentiate both groups based on its scores plot, where higher loadings coefficients were observed in amino acids, lipids and DNA regions. 3D-PCA is a new methodology for exploratory analysis of hyperspectral imaging, providing fast information for class differentiation.

Author contribution: C.L.M.M. developed the algorithm, performed the experiments, data analysis and wrote the manuscript.

Camilo L. M. Morais, PhD candidate

Prof. Francis L. Martin, Supervisor

6.1 Introduction

In spectrochemical imaging, a spectrum is generated for each pixel in the original image, where both spatial and chemical information are considered. The data are represented by three-dimensional (3D) arrays for each sample measured, where the spatial coordinates are present in the *x*- and *y*-coordinates and the wavenumbers in the *z*-coordinate. Thus, each wavenumber response (a 2D image) is stacked up one above the other in a manner similar to paper sheets in a book in order to form a 3D object (Porro-Muñoz *et al.*, 2011), informally called a "data cube".

There are many types of instrumental techniques that generate 3D spectrochemical imaging (*i.e.*, multispectral or hyperspectral imaging), *e.g.*, near-infrared (NIR), infrared (IR), Raman and mass spectrometry (MS) (Abramczyk & Brozek-Pluska, 2013; Aguayo *et al.*, 1986; Buchberger 2018; Türker-Kaya & Huck, 2017). Several matrices have been analysed by using spectrochemical imaging, *e.g.*, food (Amigo *et al.*, 2013; Pierna JAF *et al.*, 2012), soil (Eylenbosch *et al.*, 2017), atmospheric particulate matter (Ofner *et al.*, 2015), and tissues (Olmos *et al.*, 2017). Many chemometric techniques can be used for analysing this type of data, such as principal component analysis (PCA), partial least squares (PLS), multivariate curve resolution (MCR), among others (Amigo *et al.*, 2015; Mobaraki & Amigo, 2018); however, in many cases, reshaping, averaging procedures, and data compression are performed in order to reduce dimensionally (Amigo *et al.*, 2015; Morais & Lima, 2017). Recently, some adaptations of first-order algorithms used for classical spectroscopy data, such as linear discriminant analysis (LDA) and PCA, were produced for 2D data obtained *via* excitation-emission matrix (EEM) fluorescence spectroscopy (da Silva *et al.*, 2016; Morais & Lima, 2017). These algorithms, named 2D-LDA and 2D-PCA, are found to have excellent performance using 2D data without using previous dimensional reduction techniques (da Silva *et al.*, 2016; Morais & Lima, 2017), hence its usage could be extended for chemical imaging.

One of the imaging techniques that has found increasingly applications is Raman microspectroscopy (Butler *et al.*, 2016; Zhang *et al.*, 2010). Raman imaging has been used in a wide range of applications, including investigation of drug delivery systems (Smith *et al.*, 2015), pharmaceutical analysis (Tian *et al.*, 2011), food quality control (Yaseen *et al.*, 2017), and analysis of biological materials (Butler *et al.*, 2016; Lohumi *et al.*, 2017). For instance, in cancer detection, Raman imaging has been applied to diagnose

breast (Abramczyk & Brozek-Pluska, 2013), skin (Lui *et al.*, 2012), cervical (Diem *et al.*, 2013), lung (Diem *et al.*, 2013), and brain cancers (Kirsch *et al.*, 2010). A major advantage is that the use of Raman imaging provides both chemical and structural information of the sample being analysed with minimum water interference.

Ovarian cancer affects some 7,300 women in the UK alone per year and results in around 4,100 deaths per year (Paraskevaidi *et al.*, 2018d). For standard ovarian cancer diagnosis, women with symptoms undergo a pelvic examination followed by measurement of serum cancer antigen (CA-125). If symptoms persist in the absence of raised CA-125 levels, an abdominal and transvaginal ultrasound is performed (Jayson *et al.*, 2014; Paraskevaidi *et al.*, 2018d). However, ovarian cancer often presents late symptoms in which the cancer has already metastasized within the abdomen, resulting in late-stage and poor prognoses (Jayson *et al.*, 2014; Paraskevaidi *et al.*, 2018d). Besides these limitations, the diagnosis tends to be extremely invasive, expensive and time-consuming. Therefore, alternative methodologies to detect ovarian cancer that can reduce these drawbacks are of major importance, especially towards early-stage diagnosis. Herein, we propose a new algorithm of 3D principal component analysis (3D-PCA) for hyperspectral image analysis, exemplified in the exploratory analysis of plasma samples of healthy controls and ovarian cancer patients analysed by Raman microspectroscopy imaging.

6.2 Methods

6.2.1 Sample

Ten plasma samples of five healthy controls and five patients diagnosed with ovarian cancer were analysed by a Renishaw InVia Basis Raman spectrometer coupled to a confocal microscope (Renishaw plc, UK). All experiments were performed in accordance with Royal Preston Hospital Guidelines, and approved by the ethics committee at Royal Preston Hospital UK (16/EE/0010). Informed consents were obtained from all human participants of this study. For analysis, 50 μL of plasma were deposited on aluminium covered glass slides and left to air-dry overnight. Samples were analysed with an acquisition area of 50 μm × 50 μm using 50× magnification and a laser power of 100% at 785 nm with 0.1 ms exposure time. Hyperspectral images were acquired via

StreamHR^{TM} imaging technique (high confocality mode) with a grid area of 57 × 57 pixels, resulting in 3,249 spectra in the range of ~780–1858 cm^{-1} generated for each image (1 cm^{-1} data spacing, 1,016 wavenumbers per spectrum). Thus, each sample's image was composed by a data array with dimension 57 × 57 × 1016.

6.2.2 Software

The Raman images were converted into suitable .txt files using Renishaw WiRE software; and processed using MATLAB R2014b (MathWorks, Inc., USA) with lab-made routines. All the samples' images were pre-processed by cosmic rays (spikes) removal and Savtizky-Golay smoothing (window of 9 points, 2nd order polynomial fitting). All data were mean-centred before further data analysis. A personal computer (16 GB of RAM memory, Intel® CoreTM i7 processor 2.81 GHz) was used for data processing.

6.2.3 3D-PCA

PCA is an exploratory analysis technique characterized by the decomposition of a given spectral data matrix **X** into a few number of principal component (PCs) responsible for the majority of the original data variance. Each PC is orthogonal to each other, being composed of scores (projections of the samples on the PC direction) and loadings (angle cosines of the variables projected on the PC direction) (Bro & Smilde, 2014; Geladi & Kowalski, 1986; Santos *et al.*, 2017). The PCA decomposition of a spectral matrix **X** into scores (**T**), loadings (**P**) and residuals (**E**) takes the form:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{T} + \mathbf{E} \quad (6.1)$$

The scores **T** represent the variability on sample direction; the loadings **P** the variability on variables (*e.g.*, wavenumbers) direction; and the residuals **E** the unexplained data after decomposition. **T** is used for assessing similarities/dissimilarities among the samples in an exploratory analysis context, whereas **P** contains the weights for each variable in the decomposition.

In 3D-PCA, a regular PCA decomposition (eqn. 6.1) using nonlinear iterative partial least squares (NIPALS) algorithm is applied to each point (*i*,*j*) on the surface of the hyperspectral image data set. However, before PCA, each point in the image is transformed into a temporary 2D structure $\mathbf{X}_{ij}^{*}$ having *s* rows (samples) and *k* columns (variables) in order to keep the scores and loadings with their original meanings:

$$\mathbf{X}_{ij}^{*} = \mathbf{T}_{ij}\mathbf{P}_{ij}^{\mathrm{T}} + \mathbf{E}_{ij} \tag{6.2}$$

The number of PCs is selected based on the singular values obtained by singular value decomposition (SVD) (Bro & Smilde, 2014) of the hyperspectral imaging, in a similar manner as described by Morais and Lima for florescence data (Morais & Lima, 2017). After the number of PCs is selected, the scores $\mathbf{T}_{ij}$ and loadings $\mathbf{P}_{ij}$ are combined for all points (*i*,*j*) and separated for each PC. Hence, new three-dimensional arrays $\underline{\mathbf{T}}_{c}$ (*s* × *n* × *m*) and $\underline{\mathbf{P}}_{c}$ (*k* × *n* × *m*) are created for each PC, *c*. Figure 6.1 illustrate the 3D-PCA graphically.

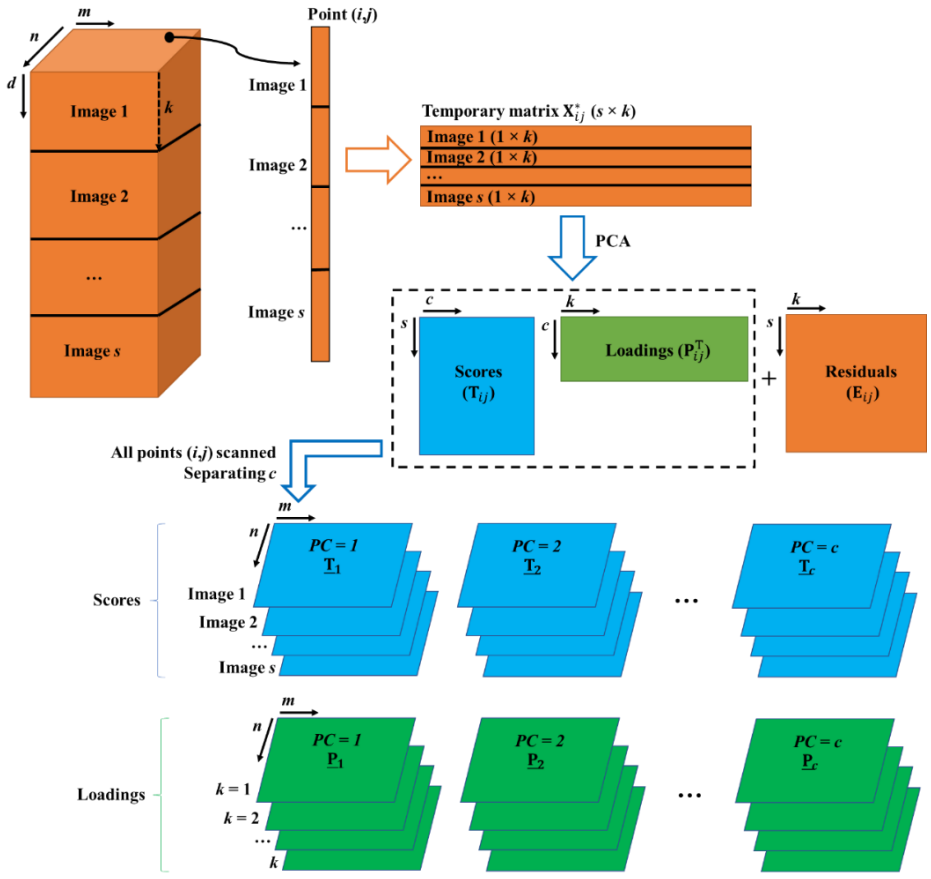


Figure 6.1. Illustration of data processing using 3D-PCA. *d* represents the z-axis coordinate dimension with size of *k* (number of wavenumbers) × *s* (number of images); *n* the number of pixels in the x-axis coordinate; *m* the number of pixels in the y-axis coordinate; and *c* the number of principal components (PCs).

6.3 Results and Discussion

Ten plasma samples (5 health controls and 5 of patients diagnosed with ovarian cancer) were analysed by Raman microspectroscopy imaging. Their hyperspectral images were generated with dimension of 57 × 57 × 1016, accounting 3,300,984 data points for each sample. The hyperspectral images for healthy controls and ovarian cancer samples are depicted in Figure 6.2 and Figure 6.3, respectively. Notably, each image presents distinct visual features, characterized by physical differences, such as dents and surface anomalies, of the samples analysed. However, chemically they should be grouped into at least two clusters (healthy vs. cancer).

The images were acquired in the spectral range of ~780–1858 cm^{-1}, which includes the fingerprint region; therefore, encompassing Raman signals of the major biochemical molecules present in the samples (Kelly *et al.*, 2011). 3D-PCA was applied to the pre-processed images using only 2 PCs (34.23% cumulative explained variance) (Table 6.1). The 3D-PCA took approximately 1 min to run the entire data set, which accounted to more than 33 million of data points (10 images × 3,300,984 data points/image), using a standard personal computer. The 3D-PCA scores on PC1 and PC2 are shown in Figure 6.4.

The scores on PC1 and PC2 across the x-axis (Figure 6.4A and 6.4B, respectively) show a separation tendency between healthy controls and ovarian cancer patients. However, a across the y-axis, the scores on both PC1 and PC2 are very noisy (Figure 6.4C and 6.4D, respectively); although, a separation pattern is observed on the scores on PC2 (Figure 6.4D). Combining the average scores on PC1 and PC2, the PC1 vs. PC2 scores plot (Figure 6.4E) shows a clear formation of two clusters separated along both PC1 and PC2. Healthy control patients are located in the bottom-right side of the graph, while ovarian cancer patients on the upper-left side. Only one ovarian cancer sample is within the healthy control cluster. Figure 6.5 shows the boxplots for comparing the 3D-PCA scores individually along the axis and averaged. In all cases, statistical difference between healthy controls and ovarian cancer patients were observed at a 95% confidence level ($p < 0.05$): $p \approx 10^{-25}$ for scores on PC1 across *x*-axis (Figure 6.5A); $p \approx 10^{-27}$ for scores on PC2 across *x*-axis (Figure 6.5B); $p \approx 10^{-46}$ for scores on PC1 across *y*-axis (Figure 5C); $p \approx 10^{-100}$ for scores on PC2 across y-axis (Figure 6.5D); $p \approx 0.004$ for

average scores on PC1 (Figure 6.5E); and $p \approx 0.002$ for average scores on PC2 (Figure 6.5F).

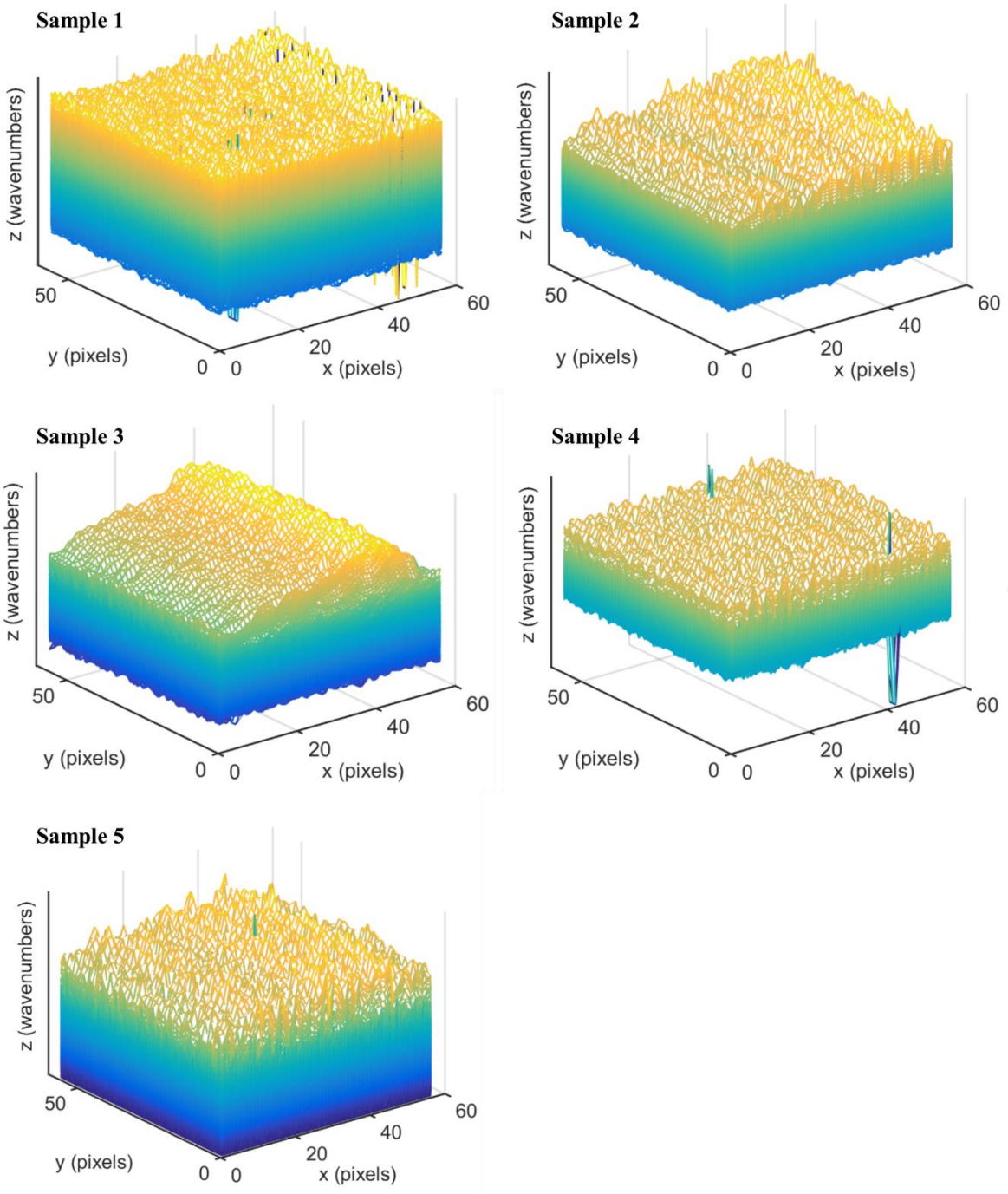


Figure 6.2. Raman hyperspectral images of healthy control samples.

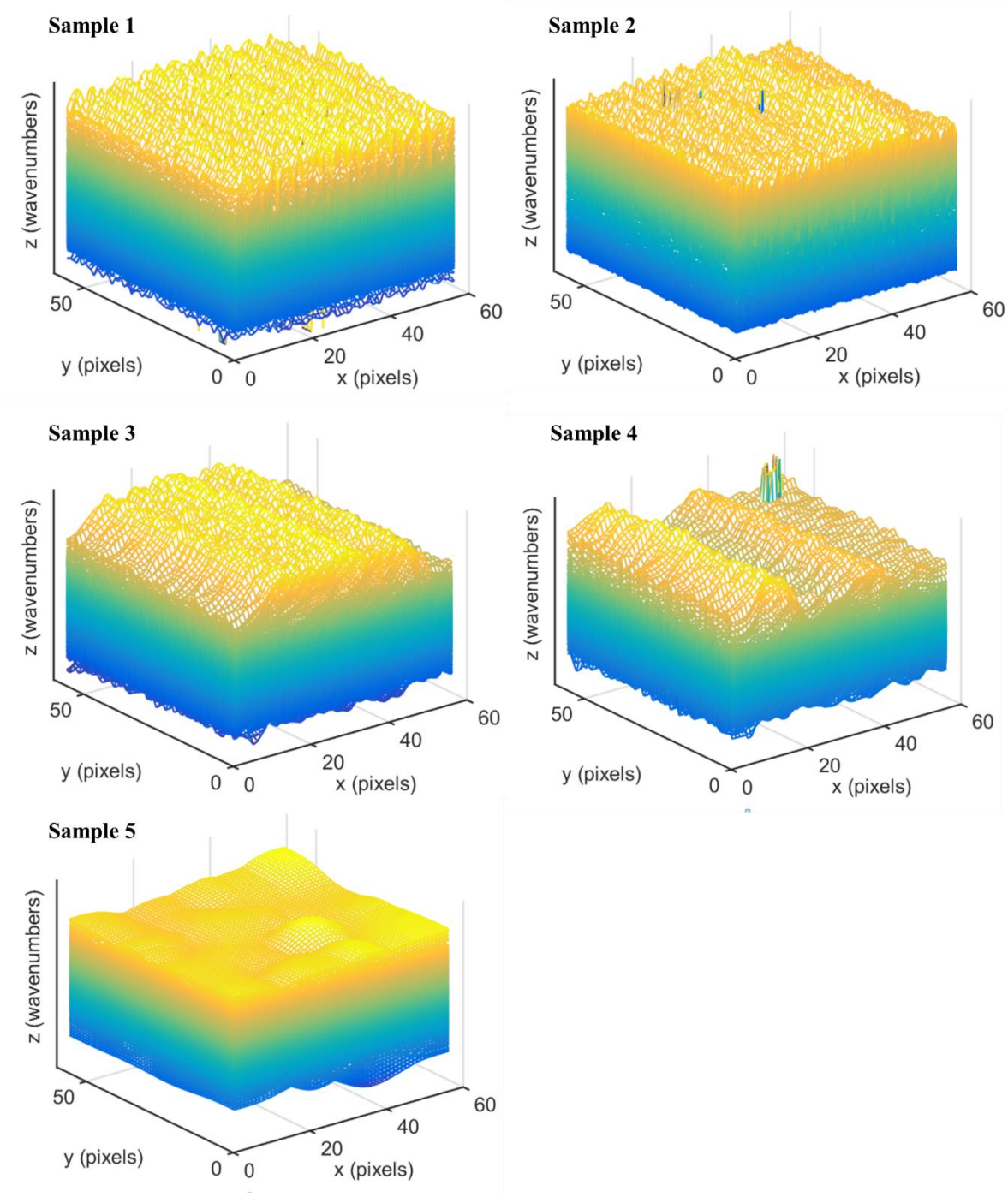


Figure 6.3. Raman hyperspectral images of ovarian cancer samples.

Table 6.1. Explained variance for 3D-PCA.

PC	Explained variance (%)	Cumulative explained variance (%)
1	20.78	20.78
2	13.45	34.23
3	11.34	45.57
4	10.32	55.88
5	9.61	65.50
6	9.12	74.62
7	8.73	83.34
8	8.46	91.80
9	8.20	100

The loadings profiles show larger coefficients around the Raman shift at 1400 cm^{-1} for PC1 (Figure 6.6A), a region containing N-H in-plane deformation and (C=O)-O-stretching in amino acids; and at ~1800 cm^{-1} and ~825 cm^{-1} representing C=O stretching in lipids and O-P-O stretching vibration in DNA, respectively (Movasaghi *et al.*, 2007). Vibrations around 820 cm^{-1} and 1400 cm^{-1} have been reported as protein biomarkers for cervical tumours (Movasaghi *et al.*, 2007; Utzinger *et al.*, 2001).

The fast data processing and clear scores segregation between healthy controls and ovarian cancer patients depicts the power of 3D-PCA as an exploratory analysis method for assessing between-samples differences in hyperspectral images. Even being an unsupervised method, statistical differences were found at a 95% confidence level between the 3D-PCA scores of the two different classes, indicating its potential usage towards classification applications. However, to build proper classification models in this case, a large cohort should be analysed by means of supervised classification techniques, which can be easily adapted to 3D-PCA by employing discriminant analysis techniques (Siqueira *et al.*, 2017) or support vector machines (Cortes & Vapnik, 1995) to the 3D-PCA scores.

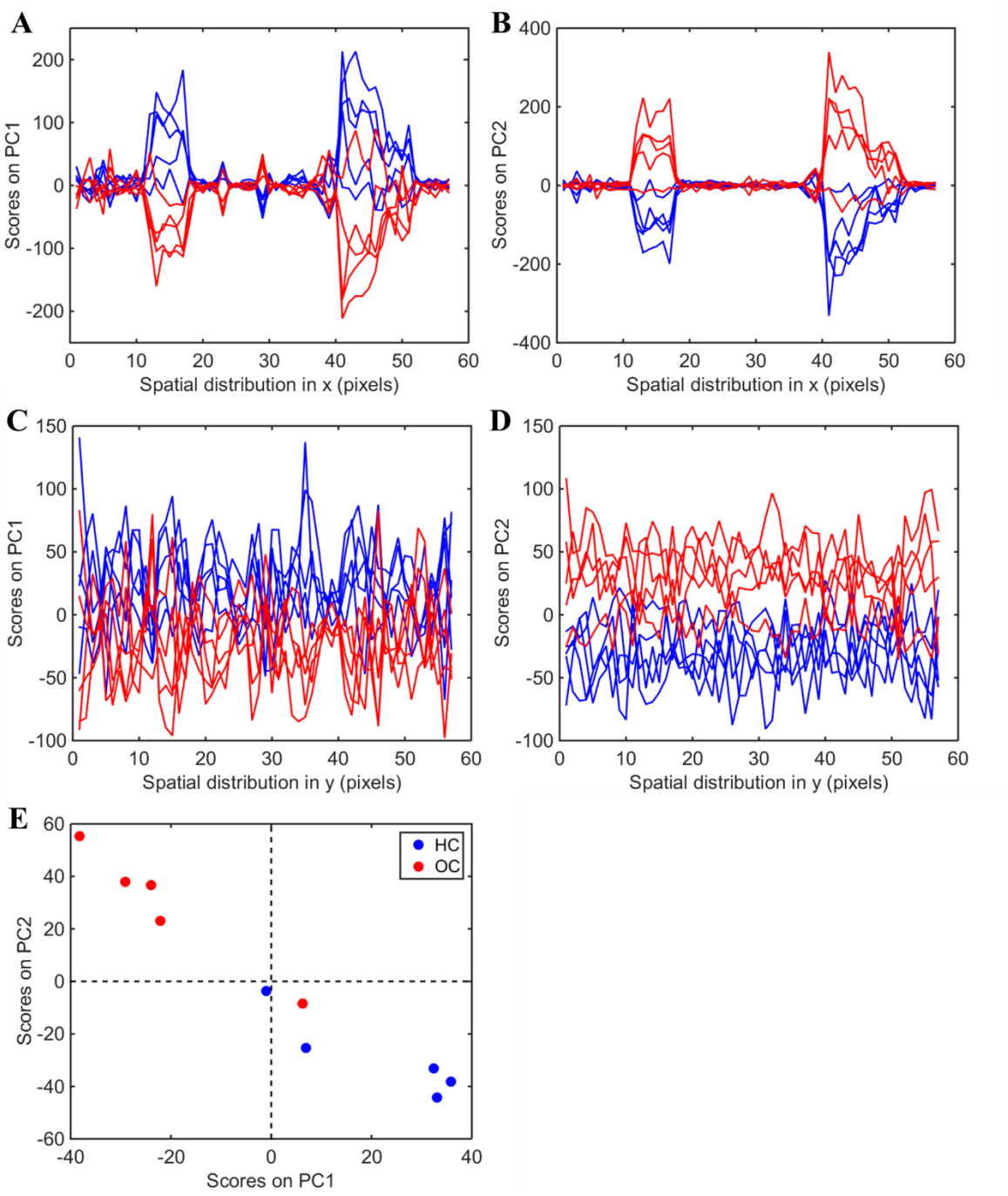


Figure 6.4. 3D-PCA scores plot. (A) Scores on PC1 and (B) PC2 across *x*-axis; (C) scores on PC1 and (D) PC2 across *y*-axis; (E) average scores on PC1 *versus* PC2. HC: healthy controls (in blue); OC: ovarian cancer (in red).

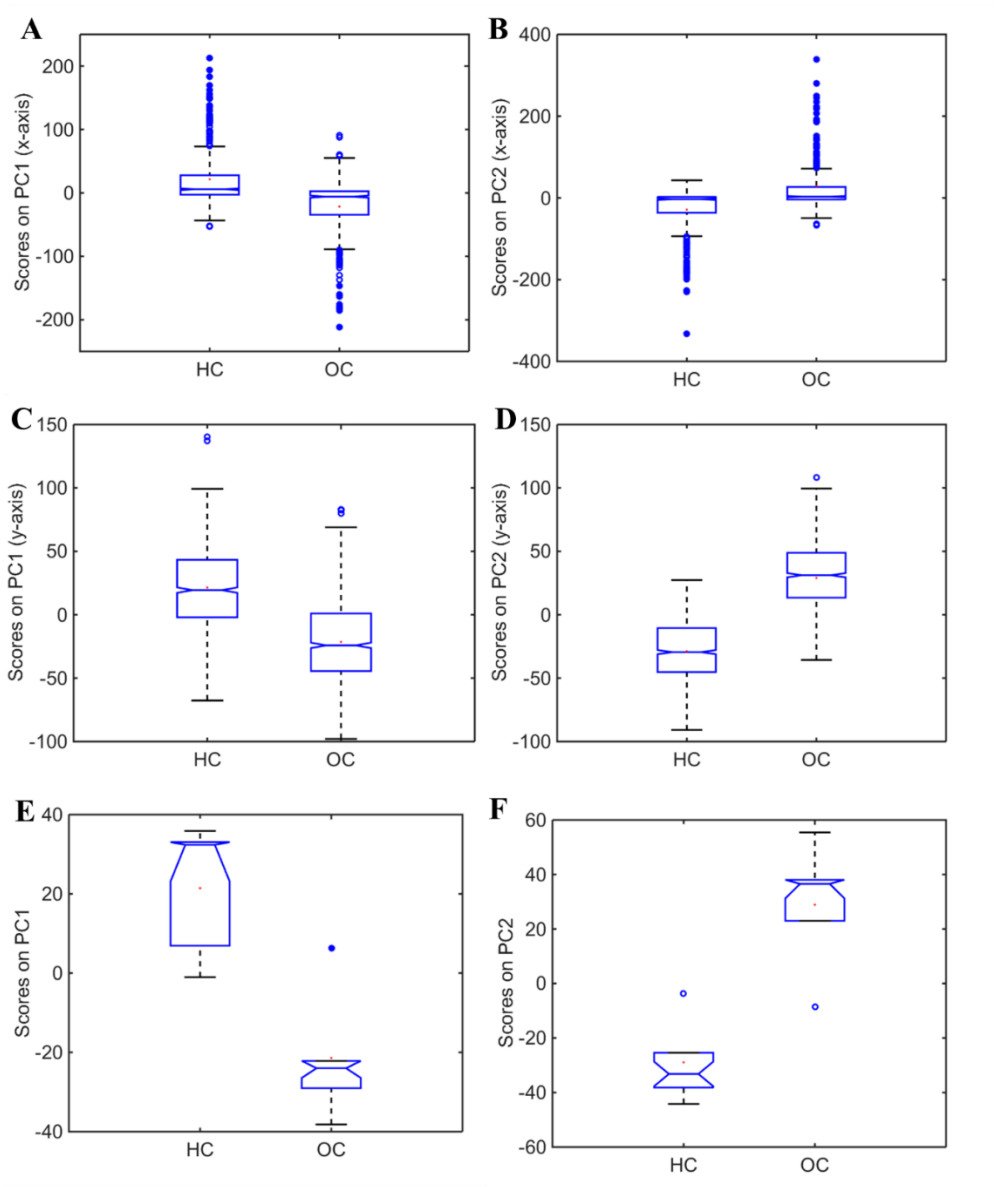


Figure 6.5. Boxplots for 3D-PCA scores. (A) Scores on PC1 across *x*-axis ($p = 1.903\times10^{-25}$); B) scores on PC2 across *x*-axis ($p = 4.884\times10^{-27}$); C) scores on PC1 across *y*-axis (6.118×10^{-46}); (D) scores on PC2 across *y*-axis (6.239×10^{-100}); (E) average scores on PC1 ($p = 0.004$); (F) average scores on PC2 ($p = 0.002$). HC: healthy controls; OC: ovarian cancer.

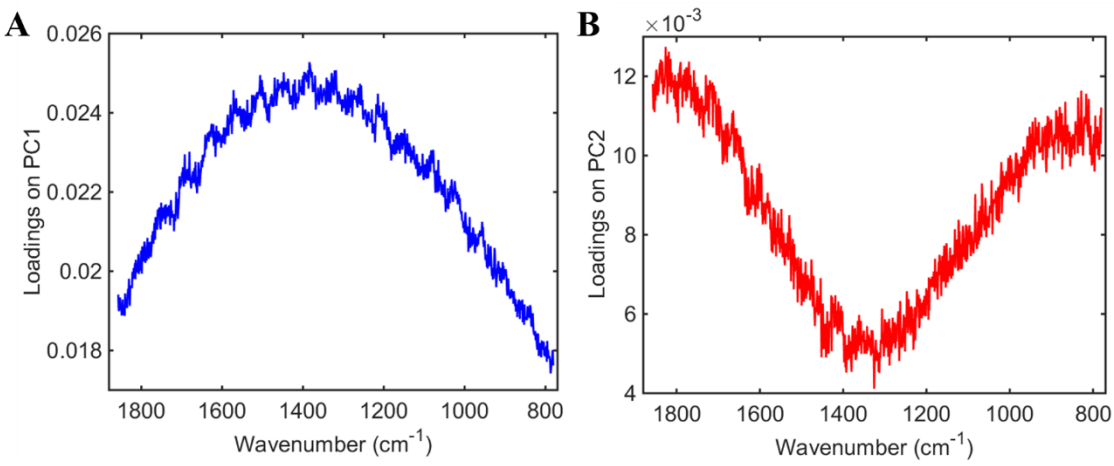


Figure 6.6. 3D-PCA loadings. (A) Loadings on PC1; (B) loadings on PC2.

6.4 Conclusion

This chapter reports a new 3D-PCA algorithm applied for exploratory analysis of plasma samples of healthy controls and ovarian cancer patients. Ten samples (5 healthy controls and 5 ovarian cancer) were analysed by Raman miscrospectrocopy imaging in the region of ~780–1858 cm^{-1}, generating data tensors with size of 57 × 57 × 1016 data points. 3D-PCA was applied to the whole dataset, generating scores showing clear differences between the two classes on both PC1 and PC2; and the loadings profiles on these components indicate that the main biomarker contributing for class differentiation are amino acids, lipids and DNA. 3D-PCA provided fast exploratory analysis for hyperspectral data, having potential for future applications in other types of spectrochemical imaging techniques.

CHAPTER 7 | A THREE-DIMENSIONAL DISCRIMINANT ANALYSIS APPROACH FOR HYPERSPECTRAL IMAGES

This chapter is *under review* in Analyst (IF 4.019). It demonstrates new chemometric techniques for discriminant analysis (3D-PCA-LDA and 3D-PCA-QDA) of hyperspectral images.

Abstract: Raman hyperspectral imaging is a powerful technique that provides both chemical and spatial information of a sample matrix being studied. The generated data are composed of three-dimensional (3D) arrays containing the spatial information across the *x*- and *y*-axis, and the spectral information in the *z*-axis. Unfolding procedures are commonly employed to analyze this type of data in a multivariate fashion, where the spatial dimension is reshaped and the spectral data fits into a two-dimensional (2D) structure and, thereafter, common first-order chemometric algorithms are applied to process the data. There are only a few algorithms capable of working with the full 3D array. Herein, we propose new algorithms for 3D discriminant analysis of Raman hyperspectral images based on a three-dimensional principal component analysis linear discriminant analysis (3D-PCA-LDA) and a three-dimensional discriminant analysis quadratic discriminant analysis (3D-PCA-QDA) approach. The analysis was performed in order to discriminant benign controls and ovarian cancer samples based on Raman hyperspectral imaging, in which 3D-PCA-LDA and 3D-PCA-QDA achieved much superior performance than their respective algorithms using unfolding procedures (PCA-LDA and PCA-QDA), where the classification accuracies improved from 64% to 100% after employing the 3D techniques. 3D-PCA-LDA and 3D-PCA-QDA are new approaches for discriminant analysis of hyperspectral images multisets to provide faster and superior classification performance than traditional techniques.

Author contribution: C.L.M.M. developed the algorithms, performed the data analysis and wrote the manuscript.

Camilo L. M. Morais, PhD candidate — Prof. Francis L. Martin, Supervisor

7.1 Introduction

Hyperspectral imaging techniques allows one to obtain specially distributed spectral data, where each image position (pixel) is composed of a spectrum in a specific wavelength range. These data are represented by three-dimensional (3D) arrays where the spatial coordinates are present in the *x*- and *y*-axis and the spectral information in the *z*-axis. Looking at another point of view, each wavelength response represents a two-dimensional (2D) image being stacked up one above the other to form a 3D object, informally called a "data cube" (Morais *et al.*, 2019e).

There are a number of hyperspectral imaging techniques depending on the electromagnetic radiation frequency of the light source or the spectrometric technique used to obtain the spectral response. For instance, several studies have been performed using visible, near-infrared, mid-infrared, and mass spectrometry hyperspectral imaging (Buchberger *et al.*, 2018; Pilling & Gardner, 2016; Türker-Kaya & Huck, 2017; Zuzak *et al.*, 2002). One of these imaging techniques that has found increasingly applications is Raman hyperspectral imaging (Lohumi *et al.*, 2017), which is a generally non-destructive technique where a spectral response is obtained based on molecular polarizability changes (Santos *et al.*, 2017). Raman hyperspectral imaging has been used in a wide range of applications, such as, pharmaceutical analysis (Kandpal *et al.*, 2018), food quality control (Yaseen *et al.*, 2017), forensic studies (Almeida *et al.*, 2017), and to investigate biological materials (Butler *et al.*, 2016). Some advantages of using Raman hyperspectral imaging to analyze biological samples include its relative low-cost, minimal or no sample preparation, high sensitivity to chemically-relevant information, and minimum water interference. For example, in cancer detection, Raman imaging has been successful applied to identify brain (Abramczyk & Brozek-Pluska, 2013), breast (Diem *et al.*, 2013), cervical (Diem *et al.*, 2013 Diem *et al.*, 2013), lung (Diem *et al.*, 2013), and skin cancer (Lui *et al.*, 2012).

Hyperspectral imaging data are analyzed by means of multivariate image analysis (MIA) techniques, where two approaches can be used: MIA at a pixel level (*e.g.*, "within-image" analysis), where chemical features are analyzed within a single image based on the spatial distribution of their spectral signatures, or MIA at a global image level (*e.g.*, "between-image" analysis), where the chemical features of each image are compared to a set of different images (Prats-Montalbán *et al.*, 2011). An imaging processing workflow

usually contemplates the following steps: pre-processing, feature extraction, feature selection and analysis, acquisition of desired information, and incorporation into prediction, monitoring or control schemes (Duchesne *et al.*, 2012); in which as series of algorithms are employed to perform these tasks, *e.g.*, principal component analysis (PCA) for feature extraction and exploratory analysis (Bro & Smilde, 2014), partial least squares (PLS) for feature extraction and quantification (Wold *et al.*, 2001), partial least squares discriminant analysis (PLS-DA) for feature extraction and classification (Brereton & Lloyd, 2014), and multivariate curve resolution alternating least squares (MCR-ALS) for feature extraction, exploratory analysis, calibration and construction of concentration distribution maps (Jaumot *et al.*, 2015; Prats-Montalbán *et al.*, 2011).

Since most algorithms used to process hyperspectral images are first-order-based, *i.e.*, applied to an one-dimensional vectoral data, unfolding strategies are often performed to handle hyperspectral 3D arrays. In this process, a 3D array with size m × n × k is unfolded to a 2D matrix with size m * n × k. This process is very useful when doing "within-image" analysis, once the spatial information of the image is distributed on the row-wise direction. However, for "between-image" analysis, when multiple images/samples are compared, the unfolding process might affect the variance structure of the data, once the relationship between neighboring pixels is lost. Some strategies to deal with 2D and 3D arrays without unfolding have been reported, *e.g.*, da Silva *et al.* (2016) reported a 2D linear discriminant analysis (2D-LDA) algorithm to classify three-way chemical data; Morais and Lima (2017) reported a 2D principal component analysis with linear discriminant analysis (2D-PCA-LDA), quadratic discriminant analysis (2D-PCA-QDA), and support vector machines (2D-PCA-SVM) to classify excitation-emission matrix (EEM) fluorescence data; and, Morais *et al.* (2019e) have reported a 3D-PCA approach to perform exploratory analysis in hyperspectral images.

In this thesis, we propose new 3D discriminant analysis approaches to classify hyperspectral images, named three-dimensional principal component analysis linear discriminant analysis (3D-PCA-LDA) and three-dimensional principal component analysis quadratic discriminant analysis (3D-PCA-QDA). Results are reported to discriminate benign controls and ovarian cancer patients based on the Raman hyperspectral imaging.

7.2 Methods

7.2.1 Samples

Thirty-eight samples (20 benign control individuals, 18 ovarian cancer patients) were analyzed by a Renishaw InVia Basis Raman spectrometer coupled to a confocal microscope (Renishaw plc, UK). The samples were collected with ethics approval by the East of England – Cambridge Central Research Ethics Committee (REC reference number 16/EE/0010, IRAS project ID 195311). Informed consents were obtained from all human participants of this study. For spectroscopic analysis, 30 μL of blood plasma were deposited on an aluminum-covered glass slide and left to air-dry overnight. Samples were measured with an acquisition area of 100 × 50 μm using 20× magnification and laser power of 50% at 785 nm with 0.1 s exposure time. Hyperspectral images were acquired *via* StreamHR^{TM} imaging technique (high-confocality mode) with a grid area of 22 × 13 pixels. Each image was composed of a 3D array with dimensions 22 × 13 × 1015, where 1015 wavenumbers were recorded per pixel (1 cm^{-1} data spacing, 725–1813 cm^{-1}).

7.2.2 Software

The Raman images were imported and processed in MATLAB R2014b (MathWorks, Inc., USA). All the samples' images were firstly pre-processed by cosmic rays (spikes) removal using a lab-made routine, followed by Savitzky-Golay (SG) smoothing (window of 15 points, 2nd order polynomial fitting) and automatic weighted least squares (AWLS) baseline correction using PLS Toolbox 7.9.3 (Eigenvector Research, Inc., USA). First-order discriminant analysis (PCA-LDA, PCA-QDA) were performed using the Classification Toolbox for MATLAB (Ballabio & Consonni, 2013), and 3D discriminant analysis (3D-PCA-LDA, 3D-PCA-QDA) were performed using lab-made algorithms. The pre-processed hyperspectral images were split into training (70%) and test (30%) sets using the Kennard-Stone uniform sample selection algorithm (Kennard & Stone, 1969).

7.2.3 Computational Analysis

Unfolded PCA-LDA and PCA-QDA were compared with 3D discriminant algorithms (3D-PCA-LDA and 3D-PCA-QDA). PCA is an exploratory analysis technique where a spectral data matrix **X** is decomposed into a few number of principal components (PCs) responsible for the majority of the original data variance. The first PC explains the biggest proportion of the data variance, followed by the second PC, and so on. Each PC is orthogonal to each other, being composed of scores (projections of the samples on the PC direction) and loadings (angle cosines of the variables projected on the PC direction). The scores represent the variability on sample direction, thus being used to assess similarities/dissimilarities among the samples based on their distribution pattern, and the loadings contain the weights for each variable in the decomposition, being used to find potential spectral markers (Bro & Smilde, 2014; Geladi & Kowalski, 1986; Morais *et al.*, 2019e).

In this 3D-PCA approach (Morais *et al.*, 2019e), a local bilinear PCA model is performed for each pixel position across the hyperspectral image dataset as follows:

$$\mathbf{X}_{ij}^{*} = \mathbf{T}_{ij}\mathbf{P}_{ij}^{\mathrm{T}} + \mathbf{E}_{ij} \tag{7.1}$$

where $\mathbf{X}_{ij}^{*}$ is a temporary matrix at the position (*i*,*j*) where rows represent samples, and columns represent wavenumbers; $\mathbf{T}_{ij}$ are the PCA scores at position (*i*,*j*); $\mathbf{P}_{ij}$ are the PCA loadings at position (*i*,*j*); $\mathbf{E}_{ij}$ are the residuals at position (*i*,*j*); and the superscript T represents the matrix transpose operation. At the end, 3D-PCA generates three 3D arrays representing the scores (**T**), loadings (**P**), and residuals (**E**). This is different of 3D-PCA for trilinear data based on Tucker3 ("true 3D-PCA"), which decomposes a trilinear three-dimensional array into three loadings and a core matrix (Tucker, 1966; Morais *et al.*, 2019a); and also different of Tucker3 model with orthogonal factors, known as "three-way PCA" (Gemperline *et al.*, 1992; Kroonenberg *et al.*, 2004).

In these 3D-PCA-LDA and 3D-PCA-QDA approaches, a linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) classifier are employed to the mean scores of 3D-PCA, respectively. The 3D-PCA-LDA (L_{sk}) and 3D-PCA-QDA (Q_{sk}) classification scores are thus calculated as follows (Morais & Lima, 2018):

$$L_{sk} = (\mathbf{x}_s - \overline{\mathbf{x}}_k)^{\mathrm{T}}\mathbf{C}_{\mathrm{pooled}}^{-1}(\mathbf{x}_s - \overline{\mathbf{x}}_k) - 2\log_e \pi_k \tag{7.2}$$

$$Q_{sk} = (\mathbf{x}_s - \overline{\mathbf{x}}_k)^{\mathrm{T}}\mathbf{C}_k^{-1}(\mathbf{x}_s - \overline{\mathbf{x}}_k) + \log_e|\mathbf{C}_k| - 2\log_e \pi_k \tag{7.3}$$

where $\mathbf{x}_s$ is a row-vector 1 × N representing the mean scores of $\underline{\mathbf{T}}$ for sample s for each principal component N; $\overline{\mathbf{x}}_k$ is a row-vector 1 × N representing the mean scores of class k for each principal component N; $\mathbf{C}_{\text{pooled}}$ is the pooled covariance matrix; $\mathbf{C}_k$ is the variance-covariance matrix of class k; and π_k is the prior probability of class k. $\mathbf{C}_{\text{pooled}}$, $\mathbf{C}_k$ and π_k are calculated as follows:

$$\mathbf{C}_{\text{pooled}} = \frac{1}{n}\sum_{k=1}^{K} n_k \mathbf{C}_k \tag{7.4}$$

$$\mathbf{C}_k = \frac{1}{n_k - 1}\sum_{s=1}^{n_k} (\mathbf{x}_s - \overline{\mathbf{x}}_k)(\mathbf{x}_s - \overline{\mathbf{x}}_k)^{\mathrm{T}} \tag{7.5}$$

$$\pi_k = \frac{n_k}{n} \tag{7.6}$$

where n is the total number of samples in the training set; K is the total number of classes; and n_k is the number of samples of class k.

The calculation procedure is the same in the unfolded PCA-LDA and PCA-QDA, in which equations 7.2–7.6 are performed with the PCA scores of the unfolded 3D array. Both unfolded and 3D models were built using cross-validation leave-one-out, and evaluated using an external test set.

7.2.4 Model Evaluation

The unfolded and 3D models were evaluated by means of the following figures of merit calculated in the test set: accuracy (total number of samples correctly classified considering true and false negatives), sensitivity (proportion of positives correctly classified), and specificity (proportion of negatives correctly classified) (Morais & Lima, 2017). These parameters are calculated as follows:

$$\text{Accuracy (\%)} = \left(\frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}\right) \times 100 \tag{7.7}$$

$$\text{Sensitivity (\%)} = \left(\frac{\text{TP}}{\text{TP+FN}}\right) \times 100 \tag{7.8}$$

$$\text{Specificity (\%)} = \left(\frac{\text{TN}}{\text{TN+FP}}\right) \times 100 \tag{7.9}$$

where TP stands for true positives; TN stands for true negatives; FP stands for false positives; and FN stands for false negatives. Additionally, confusion matrices containing the correct classification rates in the training, cross-validation and test sets were produced.

7.3 Results and Discussion

The mean Raman hyperspectral images for the 20 samples of the benign control group and 18 samples of the ovarian cancer group are depicted in Figures 7.1a and 7.1b, respectively. Distinct visual features characterized by surface abnormalities are observed on the images. This adds a degree of variance in the image data across the spatial domain for each sample. The hyperspectral images were acquired in the region between 725–1813 cm^{-1}, which includes the fingerprint region that contains Raman signatures of the main biochemical molecules present in the sample (Kelly *et al.*, 2011). The raw and pre-processed (SG smoothing and AWLS baseline correction) mean spectra for both groups of samples are depicted in Figure 7.2.

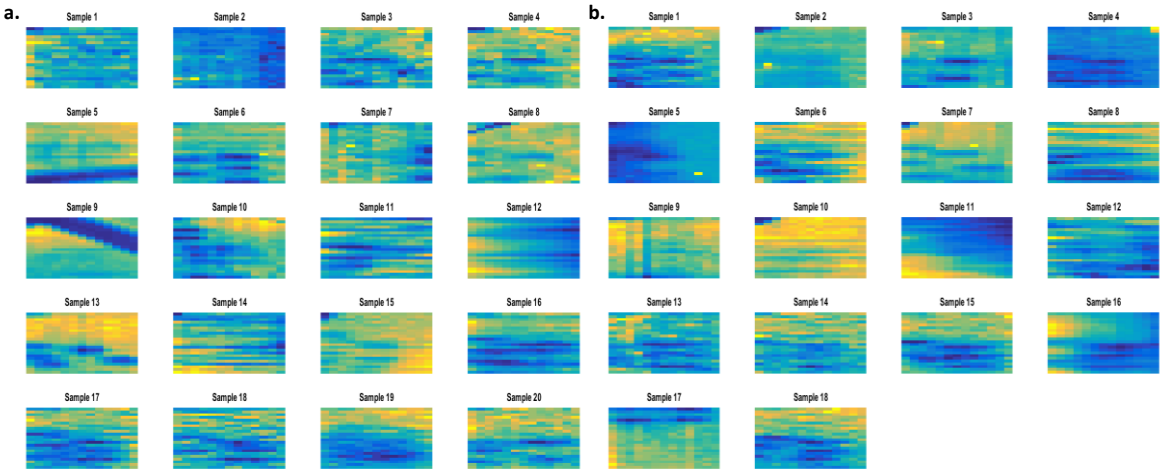


Figure 7.1. Raw Raman hyperspectral images. (a) Benign controls; (b) ovarian cancer patients. False-color images represented by the mean of the spectral dimension (725–1813 cm^{-1}).

Unfolded PCA-LDA and PCA-QDA were applied to the pre-processed data using 2 PCs (90.51% explained variance). Figures 7.3a and 7.3c show the unfolded PCA-LDA and PCA-QDA calculated classification boundaries between benign controls and ovarian cancer samples. The PCA scores response were average per sample, so each point in Figure 7.3 represents a sample (image). The superposition pattern observed in Figures 7.3a and 7.3c reflects the poor classification of unfolded PCA-LDA and PCA-QDA as demonstrated in Table 7.1, where ovarian cancer samples are being highly misclassified in the test set (80% misclassification), and in Table 7.2, where accuracies (64%) and sensitivities (20%) for PCA-LDA and PCA-QDA are substantially low.

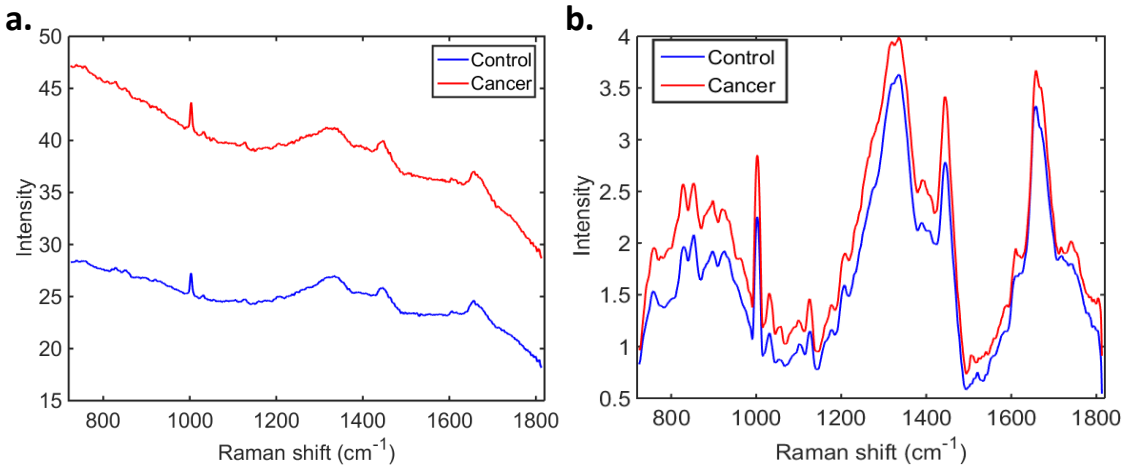


Figure 7.2. Mean Raman spectra for benign controls and ovarian cancer samples. **(**a) Raw; and (b) pre-processed Raman spectra. Pre-processing: Savitzky-Golay (SG) smoothing (window of 15 points, 2nd order polynomial fitting) and automatic weighted least squares (AWLS) baseline correction.

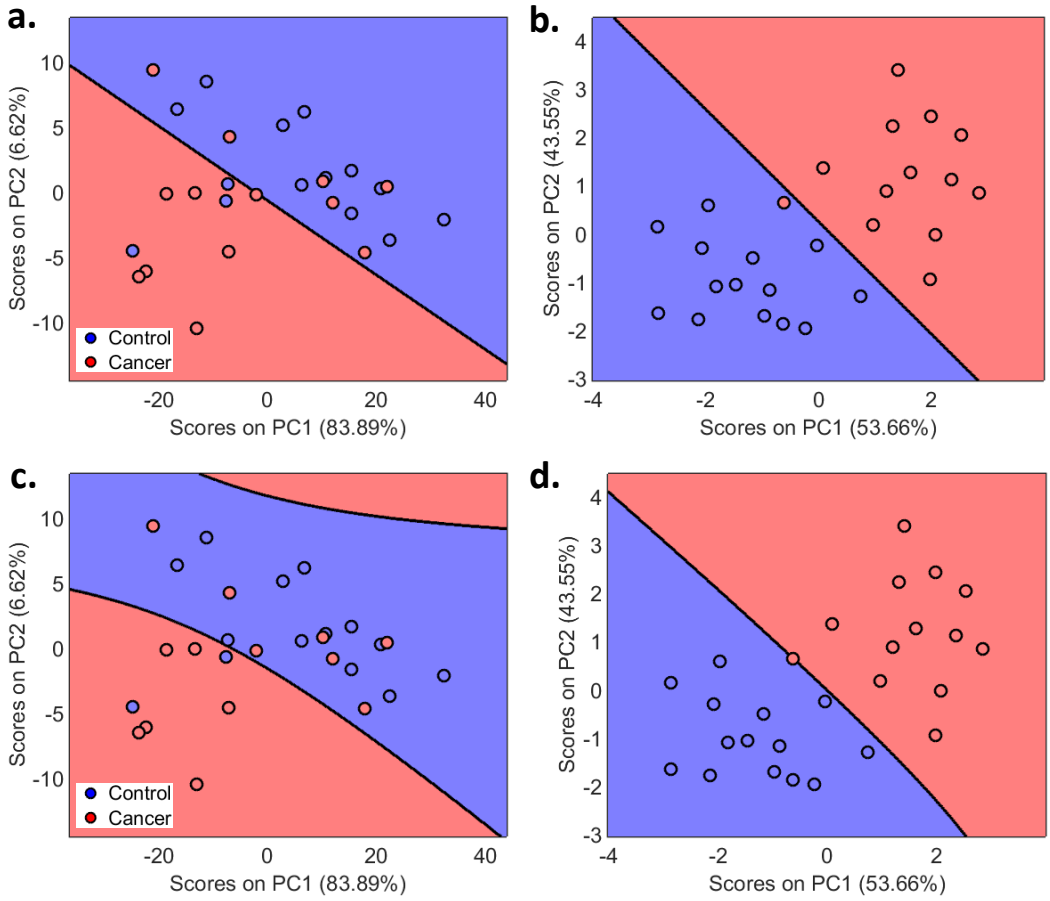


Figure 7.3. Calculated class boundaries on the PCA scores. (a) Unfolded PCA-LDA; (b) 3D-PCA-LDA; (c) Unfolded PCA-QDA; and (d) 3D-PCA-QDA. Numbers inside parenthesis on the *x*- and *y*-labels represent the percentage of explained variance in each principal component (PC).

Table 7.1. Confusion matrices for the training, cross-validation and test sets using the unfolded and 3D hypespectral images. PCA-LDA: principal component analysis linear discriminant analysis; PCA-QDA: principal component analysis quadratic discriminant analysis; 3D-PCA-LDA: three-dimensional principal component analysis linear discriminant analysis; 3D-PCA-QDA: three-dimensional principal component analysis quadratic discriminant analysis; Control: benign control group; Cancer: ovarian cancer patients; CV: cross-validation.

	PCA-LDA		PCA-QDA		3D-PCA-LDA		3D-PCA-QDA	
Training	Control	Cancer	Control	Cancer	Control	Cancer	Cancer	Cancer
Control	79%	21%	86%	14%	100%	0%	100%	0%
Cancer	46%	54%	54%	46%	8%	92%	8%	92%
CV								
Control	70%	30%	72%	28%	99%	1%	93%	7%
Cancer	51%	49%	57%	43%	7%	93%	8%	92%
Test								
Control	100%	0%	100%	0%	100%	0%	100%	0%
Cancer	80%	20%	80%	20%	0%	100%	0%	100%

Table 7.2. Quality parameters for the models using the unfolded hyperspectral images (PCA-LDA, PCA-QDA) and the full three-dimensional arrays (3D-PCA-LDA, 3D-PCA-QDA) for discriminating benign controls from ovarian cancer patients. PCA-LDA: principal component analysis linear discriminant analysis; PCA-QDA: principal component analysis quadratic discriminant analysis; 3D-PCA-LDA: three-dimensional principal component analysis linear discriminant analysis; 3D-PCA-QDA: three-dimensional principal component analysis quadratic discriminant analysis.

Data	Model	Accuracy	Sensitivity	Specificity
Unfolded	PCA-LDA	64%	20%	100%
	PCA-QDA	64%	20%	100%
3D	3D-PCA-LDA	100%	100%	100%
	3D-PCA-QDA	100%	100%	100%

On the other hand, by using the 3D-based algorithms (3D-PCA-LDA and 3D-PCA-QDA), the classification performance improved substantially. These algorithms were applied to the whole hyperspectral dataset without unfolding with a computation time of approximately 2 min per model using a standard laptop computer. Figures 7.3b and 7.3d show the 3D-PCA-LDA and 3D-PCA-QDA calculated classification boundaries between benign controls and ovarian cancer samples. There is a clear separation between the classes in both cases. For 3D-PCA-LDA (Figure 7.3b), one ovarian cancer sample of the training set is within the benign controls space; while in 3D-PCA-QDA this sample is

projected over the class boundary. This sample reduced the training and cross-validation fitting for these models, in which a correct classification rate of 92% was observed for the ovarian cancer group in the training set using both 3D-PCA-LDA and 3D-PCA-QDA; and 93% and 92% in cross-validation for 3D-PCA-LDA and 3D-PCA-QDA, respectively (Table 7.1). In the test set, the classification performance was perfect using the 3D algorithms (accuracy, sensitivity and specificity equal to 100%) (Table 7.2). These findings indicate the potential of 3D discriminant analysis compared to the unfolding procedure.

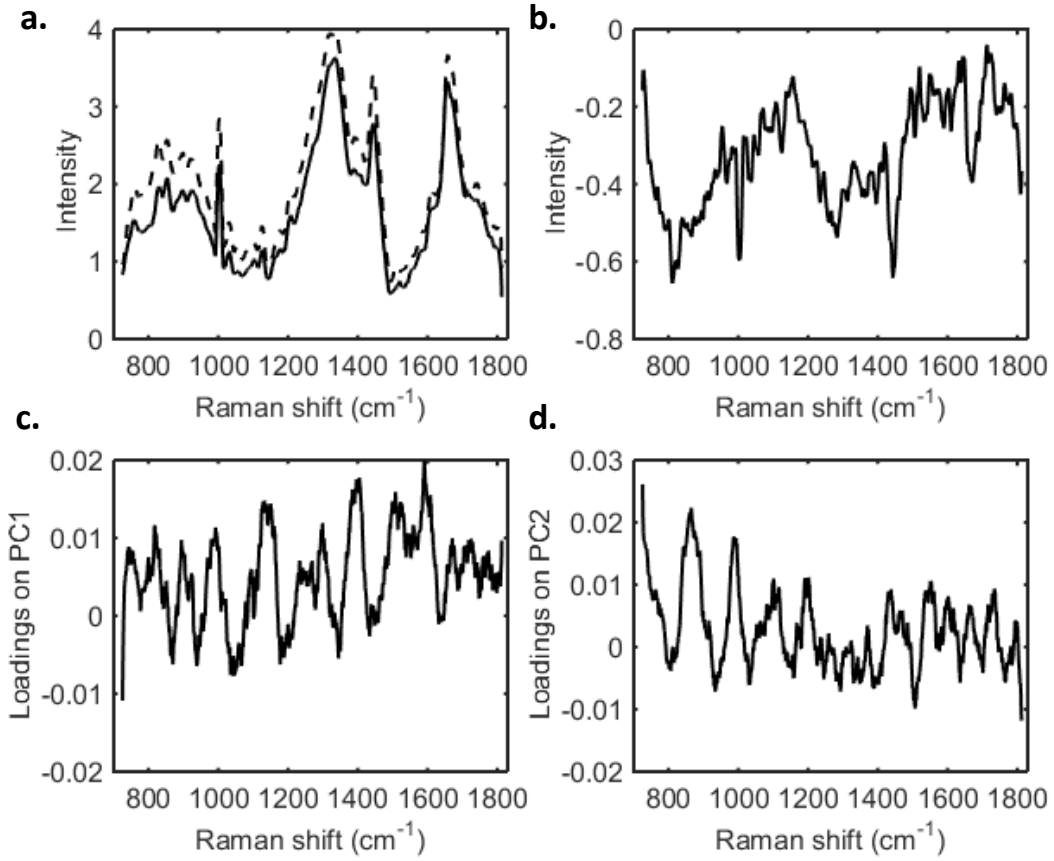


Figure 7.4. 3D-PCA loadings. (a) Average pre-processed spectra for benign controls (continuous line) and ovarian cancer (dashed line) samples; (b) difference-between-mean spectrum for benign controls and ovarian cancer samples (negative signal indicates higher intensity in ovarian cancer samples); (c) 3D-PCA loadings on PC1; (d) 3D-PCA loadings on PC2.

The difference-between-mean (DBM) spectrum and 3D-PCA loadings are shown in Figure 7.4. The ovarian cancer samples spectra appear to have overall higher intensity values than benign controls, as demonstrated in Figure 7.4a and 7.4b, where the negative

values in the latter indicate higher intensity influence in the ovarian cancer group. The 3D-PCA loadings on PC1 contain higher coefficients at: 820 cm^{-1} (C-C stretching in protein), 990 cm^{-1} (C-C stretching in glucose/collagen), 1140 cm^{-1} (fatty acids), 1400 cm^{-1} (NH in-plane deformation), 1510 cm^{-1} (ring breathing modes in DNA bases), 1592 cm^{-1} (C=C stretching) (Figure 7.4c) (29). The 3D-PCA loadings on PC2 contain higher coefficients at: 727 cm^{-1} (C-C stretching in collagen), 860 cm^{-1} (phosphate group) and 986 cm^{-1} (C-C stretching β-sheet in proteins) (Figure 7.4d) (Movasaghi *et al.*, 2007). PC1 seems to be related to wavenumbers of higher energy, encompassing mainly fatty acids, lipids and protein vibrations; while PC2 contain higher weights toward wavenumbers of lower energy, including collagen, phosphate groups of RNA, and C-C in proteins (Movasaghi *et al.*, 2007). Vibrations around 820 cm^{-1} and 1400 cm^{-1} (PC1) have been previously reported as protein markers for cervical tumors (Utzinger *et al.*, 2001) and ovarian cancer (Morais *et al.*, 2019e).

7.4 Conclusion

This paper reports new 3D discriminant analysis approaches named three-dimensional principal component analysis linear discriminant analysis (3D-PCA-LDA) and three-dimensional discriminant analysis quadratic discriminant analysis (3D-PCA-QDA) for classification of hyperspectral images datasets. These algorithms were compared with their unfolded versions (PCA-LDA and PCA-QDA), where a much superior performance was obtained with the 3D-based techniques to discriminate benign controls and ovarian cancer patients based on the Raman hyperspectral imaging. An improvement in the accuracy (64% to 100%) and sensitivity (20 to 100%) in the test set was observed when the 3D discriminant algorithms were applied. 3D-PCA loadings indicated spectral markers associated with proteins, lipids and DNA along PC1 and PC2 for class differentiation. These new 3D discriminant analysis approaches provide fast class differentiation for multi-image hyperspectral datasets with a superior discriminating performance compared to algorithms using unfolding procedures, which are often employed for this type of data.

CHAPTER 8 | UNCERTAINTY ESTIMATION AND MISCLASSIFICATION PROBABILITY FOR CLASSIFICATION MODELS BASED ON DISCRIMINANT ANALYSIS AND SUPPORT VECTOR MACHINES

This chapter is published in Analytica Chimica Acta (IF 5.256). It demonstrates a new method to estimate model uncertainty and misclassification probabilities for classification models in spectral data using discriminant analysis and support vector machines:

- Morais CLM, Lima KMG, Martin FL. Uncertainty estimation and misclassification probability for classification models based on discriminant analysis and support vector machines. Anal. Chim. Acta **2019**; 1063: 40–46. https://doi.org/10.1016/j.aca.2018.09.022

Abstract: Uncertainty estimation provides a quantitative value of the predictive performance of a classification model based on its misclassification probability. Low misclassification probabilities are associated with a low degree of uncertainty, indicating high trustworthiness; while high misclassification probabilities are associated with a high degree of uncertainty, indicating a high susceptibility to generate incorrect classification. Herein, misclassification probability estimations based on uncertainty estimation by bootstrap were developed for classification models using discriminant analysis [linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA)] and support vector machines (SVM). Principal component analysis (PCA) was used as variable reduction technique prior classification. Four spectral datasets were tested (1 simulated and 3 real applications) for binary and ternary classifications. Models with lower misclassification probabilities were more stable when the spectra were perturbed with white Gaussian noise, indicating better robustness. Thus, misclassification probability can be used as an additional figure of merit to assess model robustness, providing a reliable metric to evaluate the predictive performance of a classifier.

Author contribution: C.L.M.M. developed the algorithms, performed the data analysis and wrote the manuscript.

Camilo L. M. Morais, PhD candidate

Prof. Francis L. Martin, Supervisor

8.1 Introduction

Multivariate classification models are commonly employed to segregate clusters based on a supervised learning approach. Commonly, the data are initially divided into training and external validation sets, where the first is used for model construction and the latter to assess the model performance. The predictive capacity of classification models is assessed by quality parameters also called "figures of merit". The most used ones are the accuracy (total number of samples correctly classified considering true and false negatives), sensitivity (proportion of positives correctly identified) and specificity (proportion of negatives correctly identified) (Morais & Lima, 2017). Additional figures of merit can also be estimated to confirm the predictive performance of a classification model, such as precision (classifier ability to avoid wrong predictions), F-score (overall performance of the model considering imbalanced data), G-score (overall performance of the model not accounting for class sizes), area under the curve (AUC) of receiver operating characteristic curves, positive and negative prediction values, positive and negative likelihood ratios, and Youden's index (Ballabio *et al.*, 2018; Morais & Lima, 2017; Neves *et al.*, 2018; Parikh *et al.*, 2016; Siqueira *et al.*, 2017). The latter three are more commonly used for biomedical applications, where the ratio of true and false positives and negatives are an important factor towards making clinical decisions.

However, none of these figures of merit brings information of the degree of uncertainty in the classification model. Uncertainty is always present in any analytical measurement where a prior univariate or multivariate model is used to provide information of the property being analysed. For being non-specific, vibrational spectroscopy techniques generate thousands of data points for all chemical components that are susceptible to the radiation source incident on the sample, creating a very complex array of data for each sample analysed. To elucidate and extract information of the chemical components present in the spectrum, chemometric techniques are often employed. Multivariate calibration techniques, such as principal component regression (PCR) and partial least squares (PLS) regression, are used for quantification applications; and classification techniques, such as discriminant analysis (DA) and support vector machines (SVM), for qualitative applications (Naes *et al.*, 2002).

In spectroscopy applications, due to problems of collinearity and ill-conditioned data, variable reduction or selection techniques are often employed prior to classification

analysis. Principal component analysis (PCA) is one of the most popular methods of variable reduction, since it reduces all the spectral variables into a small number of principal components accounting for the majority of the original variance in the data (Bro & Smilde, 2014). Since the principal components are orthogonal to each other, the computation of inverse matrix operations used in discriminant analysis are achieved with high accuracy.

Uncertainty estimation for calibration models is well known (Cacuci & Ionescu-Bujor, 2010; Caja *et al.*, 2015). However, for classification techniques, uncertainty estimation is still a new topic, so far mainly explored for partial least squares discriminant analysis (PLS-DA) (de Almeida *et al.*, 2013; Rocha & Sheen, 2016). Herein, we propose an uncertainty estimation method based on bootstrap for calculation of misclassification probabilities in linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and SVM models applied to four different datasets, where the classification stability is also evaluated by adding white Gaussian noise to the spectral data.

8.2 Experimental

8.2.1 Datasets

Four datasets were used for testing. Dataset 1 is composed of simulated spectra generated using a normal distribution function. Class 1 contains 30 spectra with 301 variables each, with mean ranging from 0.15 to 0.42 intensity units and standard deviation ranging from 0.41 to 1.14 intensity units between samples. Class 2 contains also 30 spectra with 301 variables each, with mean ranging from 0.19 to 0.35 intensity units and standard deviation ranging from 0.35 to 0.86 intensity units between samples.

Dataset 2 is composed of 280 infrared (IR) spectra of two *Cryptococcus* fungi specimens acquired *via* attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectroscopy. Class 1 contains 140 spectra of *Cryptococcus neoformans* samples, and class 2 contains 140 spectra of *Cryptococcus gattii* samples. Spectra were acquired in the range of 400-4000 cm^{-1} with resolution of 4 cm^{-1} and 16 co-added scans using a Bruker VEXTER 70 FTIR spectrometer (Bruker Optics Ltd., UK). The spectra were pre-processed by cut in the biofingerprint region (900-1800 cm^{-1}), followed by automatic weighted least squares baseline correction and normalisation to the Amide I peak (1650

cm^{-1}). More information about this dataset can be found in literature (Costa *et al.*, 2016; Morais *et al.*, 2017).

Dataset 3 is composed of 240 IR spectra for two classes of formalin-fixed paraffin-embedded brain tissues measured using ATR-FTIR spectroscopy. Class 1 contains 140 spectra for normal brain tissue samples, and class 2 contains 100 spectra for glioblastoma brain tissue samples. Spectra were acquired in the range of 400-4000 cm^{-1} with resolution of 8 cm^{-1} and 32 co-added scans using a Bruker Vector 27 FTIR spectrometer with a Helios ATR attachement (Bruker Optics Ltd., UK). The spectra were pre-processed by cut in the biofingerprint region (900-1800 cm^{-1}), followed by ruberband baseline correction and normalisation to the Amide I peak (1650 cm^{-1}). This dataset is public available as part of IRootLab toolbox (http://trevisanj.github.io/irootlab/) (Martin *et al.*, 2010; Trevisan *et al.*, 2013) and more information about it can be found in Gajjar *et al.* (2013).

Dataset 4 is composed of 183 IR spectra separated into 3 classes. Class 1 is composed of 59 spectra of Syrian hamster embryo (SHE) cells contaminated with benzo[*a*]pyrene (B[*a*]P); class 2 is composed of 62 spectra of SHE cells contaminated with 3-methylcholanthrene (3-MCA); and class 3 is composed of 62 spectra of SHE cells contaminated with anthracene (Ant). Spectra were acquired by using a Bruker TENSOR 27 spectrometer with a Helios ATR attachement (Bruker Optics Ltd., UK). Spectra were recorded in the range of 400-4000 cm^{-1} with a resolution of 8 cm^{-1}. Pre-processing was performed by cut in the biofingerprint region (900-1800 cm^{-1}), rubberband baseline correction and normalisation to the Amide I peak (1650 cm^{-1}). This dataset is public available as part of IRootLab toolbox (http://trevisanj.github.io/irootlab/) (Martin *et al.*, 2010; Trevisan *et al.*, 2013); further information can be found in Trevisan *et al.* (2010).

8.2.2 Software

Data analysis was performed within MATLAB R2014b environment (The MathWorks, Inc., USA) using lab-made routines. Pre-processing was performed using PLS Toolbox 7.9.3 (Eigenvector Research, Inc., USA). Samples were divided into training (70%) and external validation (30%) sets using Kennard-Stone sample selection algorithm (Kennard & Stone, 1969).

8.2.3 Classification Techniques

Data were initially processed by PCA in order to reduce the number of variables and solve ill-condition problems. PCA decomposes the original spectral matrix **X** into scores (**T**), loadings (**P**) and residuals (**E**) as follows (Bro & Smilde, 2014):

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{8.1}$$

The PCA scores were used as input variables for the classification models (LDA, QDA and SVM) with the number of principal components selected by singular value decomposition (SVD) (Bro & Smilde, 2014; Morais & Lima, 2017) and root mean square error of cross-validation (RMSECV) values obtained with cross-validated PCA (Brereton, 2003). The cumulated explained variance was calculated based on SVD as follows (Morais & Lima, 2017):

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{-1} \tag{8.2}$$

$$\mathrm{v}(\%) = \left[\frac{\mathrm{diag}(\mathbf{S})}{\sum \mathrm{diag}(\mathbf{S})}\right] \times 100 \tag{8.3}$$

where v(%) is the explained variance; **U** and **V** are orthogonal matrices; and **S** is a matrix containing nonzero singular values on its diagonal.

The LDA (L_{ik}) and QDA (Q_{ik}) classification scores were calculated in a non-Bayesian form as follows (Dixon & Brereton, 2009; Morais & Lima, 2018):

$$L_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}}\mathbf{C}_{\mathrm{pooled}}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_k) \tag{8.4}$$

$$Q_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}}\mathbf{C}_k^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_k) \tag{8.5}$$

where $\mathbf{x}_i$ are the input variables for sample i; $\bar{\mathbf{x}}_k$ is the mean vector of class k; $\mathbf{C}_{\mathrm{pooled}}$ is the pooled covariance matrix; and $\mathbf{C}_k$ is the variance-covariance matrix of class k. $\mathbf{C}_k$ and $\mathbf{C}_{\mathrm{pooled}}$ are estimated as follows:

$$\mathbf{C}_k = \frac{1}{n_k - 1}\sum_{i=1}^{n_k}(\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}} \tag{8.6}$$

$$\mathbf{C}_{\mathrm{pooled}} = \frac{1}{n}\sum_{k=1}^{K} n_k \mathbf{C}_k \tag{8.7}$$

where n_k is the number of samples of class k; n is the total number of samples in the training set; and K is the number of classes.

SVM was applied to the PCA scores using a radial basis function (RBF) kernel (Cortes & Vapnik, 1995). The SVM classifier takes the form of (Morais *et al.*, 2017):

$$r_i = \text{sign}\left(\sum_{i=1}^{N_{SV}} \alpha_i y_i K\left(\mathbf{x}_i, \mathbf{z}_j\right) + b\right) \tag{8.8}$$

where r_i is the classification response for sample i; N_{SV} is the number of support vectors; α_i is the Lagrange multiplier; y_i is the class membership (±1) of sample i; $K\left(\mathbf{x}_i, \mathbf{z}_j\right)$ is the kernel function; and b is the bias parameter.

8.2.4 Misclassification Probability Estimation

The uncertainty estimation was based on Bootstrap (Wehrens *et al.*, 2000), a random sampling method with replacement that allows confidence intervals to be placed on the model predictions based on uncertainties of the original data (Rocha & Sheen, 2016). The procedure for calculating uncertainties based on residual bootstrap was originally presented by de Almeida *et al.* (2013) and adapted herein for LDA, QDA and SVM-based models. For comparison, uncertainty propagation estimate for SVM was calculated by differentiation of Eq. 8.8 based on a previous uncertainty estimation for RBF kernel in artificial neural networks (ANN), assuming that noise only affects the test sample (Allegrini & Oliveiri, 2016):

$$\mathrm{d}r = \sum_{i=1}^{N_{SV}} \alpha_i y_i \frac{\mathrm{d}K\left(\mathrm{x}_i, \mathrm{z}_j\right)}{\mathrm{d}x_i} \mathrm{d}x_i = \mathbf{b}_{\mathrm{SVM}}^{\mathrm{T}} \mathrm{d}\mathbf{x} \tag{8.9}$$

where $\mathbf{b}_{\mathrm{SVM}}^{\mathrm{T}}$ represents the uncertainty propagation of SVM using RBF kernel.

For bootstrap uncertainty estimation, initially, the residuals for LDA, QDA or SVM models are calculated using:

$$\mathbf{f}^* = \frac{\mathbf{f}}{\sqrt{1 - {}^{D_f}/_n}} \tag{8.10}$$

where $\mathbf{f}^*$ is the weighted model residual; $\mathbf{f}$ is the model residual; and D_f is the pseudo-degrees of freedom (van der Voet, 1999). $\mathbf{f}$ is estimated for LDA, QDA or SVM models as:

$$\mathbf{f} = \mathbf{y} - \hat{\mathbf{y}} \tag{8.11}$$

where $\mathbf{y}$ is the reference class category for all samples; and $\hat{\mathbf{y}}$ is the model response for LDA [$\hat{\mathbf{y}} = (L_1, \cdots, L_n)$]; QDA [$\hat{\mathbf{y}} = (Q_1, \cdots, Q_n)$]; or SVM [$\hat{\mathbf{y}} = (r_1, \cdots, r_n)$].

Then, bootstrapping is applied by removing sample i whose uncertainty is being estimated by the model. A new response matrix $\mathbf{y}^*$ is generated by replacing the remaining values in $\mathbf{y}$ with the model predicted $\hat{\mathbf{y}}$. Then, a new random residual vector $\mathbf{f}^*_{\text{boot}}$ is generated by bootstrapping. The bootstrapping residual $\mathbf{f}^*_{\text{boot}}$ is added to the $\hat{\mathbf{y}}$ predicted, generating a new response vector $\mathbf{y}^{**}$:

$$\mathbf{y}^{**} = \hat{\mathbf{y}} + \mathbf{f}^*_{\text{boot}} \tag{8.12}$$

A new classification model is then created using $\mathbf{y}^{**}$ as reference categories. Finally, a new residual vector $\hat{\mathbf{f}}^*$ is created by subtracting the bootstrapping predicted values $\hat{\mathbf{y}}^{**}$ from the model predicted $\hat{\mathbf{y}}$:

$$\hat{\mathbf{f}}^* = \hat{\mathbf{y}} - \hat{\mathbf{y}}^{**} \tag{8.13}$$

The confidence intervals are calculated for sample i based on the residual vector $\hat{\mathbf{f}}^*$. For a 95% confidence interval, the lower bound ($\mathbf{c}_{\text{low}}$) and the upper bond ($\mathbf{c}_{\text{up}}$) are given by:

$$\mathbf{c}_{\text{low}} = 0.25\hat{\mathbf{f}}^* \tag{8.14}$$

$$\mathbf{c}_{\text{up}} = 0.975\hat{\mathbf{f}}^* \tag{8.15}$$

For misclassification probability calculation, the classification categories $\mathbf{y}$ are treated as being normally distributed with mean equal to $\hat{\mathbf{y}}$ and standard deviation $\sigma = {}^1/_4\left(\mathbf{c}_{\text{low}} - \mathbf{c}_{\text{up}}\right)$. The probability that sample i is class k=1, denoted $P_{1,i}$, is equivalent to the probability that $\hat{y}_i$ is lower than the threshold value that separates the classes, y_{bound}. $P_{1,i}$ is given by the cumulative distribution function for the normal distribution (Rocha & Sheen, 2016):

$$P_{1,i} = P(\hat{y}_i \leq y_{\text{bound}}) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{y_{\text{bound}} - \hat{y}_i}{\sqrt{2}\sigma_i}\right)\right] \tag{8.16}$$

Similarly, the probability that sample i is class k=2, denoted $P_{2,i}$, is equal to $1 - P_{1,i}$. The misclassification probability of sample i, $m_{p,i}$, is therefore determined based on the classification of sample i as:

$$m_{p,i} = P_{1-y_i} \tag{8.17}$$

The m_p values range from 0 (no misclassification probability) to 1 (maximum misclassification probability). Values above 0.5 indicate higher probability of misclassification. A graphical flowchart illustrating the processing steps for misclassification probability calculation for PCA-LDA, PCA-QDA and PCA-SVM models is depicted in Figure 8.1.

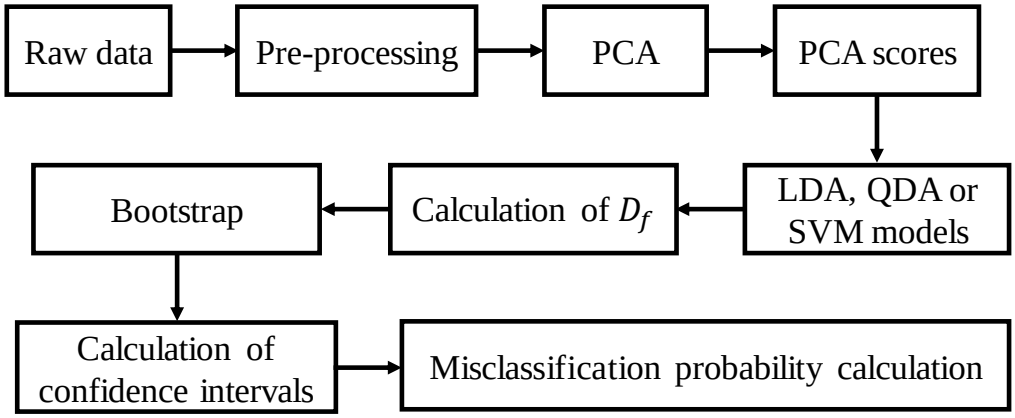


Figure 8.1. Flowchart illustrating data processing steps for misclassification probability calculation. D_f stands for pseudo-degrees of freedom.

8.3 Results and Discussion

Datasets 1-4 were analysed in order to estimate the misclassification probability associated with the trustworthiness and robustness of three classification algorithms: PCA-LDA, PCA-QDA and PCA-SVM. Pre-processed spectra with mean and standard-deviation for these datasets are depicted in Figure 8.2.

Dataset 1 is composed of simulated spectra (Figure 8.2a). Although this dataset has no chemical meaning, simulated data are commonly used as a primary source to evaluate discriminatory performance of classification algorithms (Morais & Lima, 2017). PCA was applied to the data and 10 PCs were selected according to SVD and RMSECV values (Figure 8.3a and b) (cumulative variance of 97.2%). PCA-LDA did not show a good classification, with an accuracy of 44.4%. The average misclassification rate for the test set was equal to 0.520. This high misclassification probability indicates a large degree of uncertainty for the PCA-LDA model, which is confirmed by the high misclassification probability (m_p > 0.5). On the other hand, by applying a QDA classifier, the classification performance improved substantially. The accuracy in the external validation set was found at 88.9% with average misclassification probability of 0.113. QDA performance

was superior than the one found by LDA due to the difference variance structures of class 1 and 2, as evidenced in the standard-deviation in Figure 8.2a. LDA assumes classes having similar variance structures, using a pooled covariance model. In contrast, QDA assumes classes having different variance structures, which improves considerably its performance over LDA when this condition happens (Dixon & Brereton, 2009; Morais & Lima, 2018). Additional figures of merit are depicted in Table 8.1.

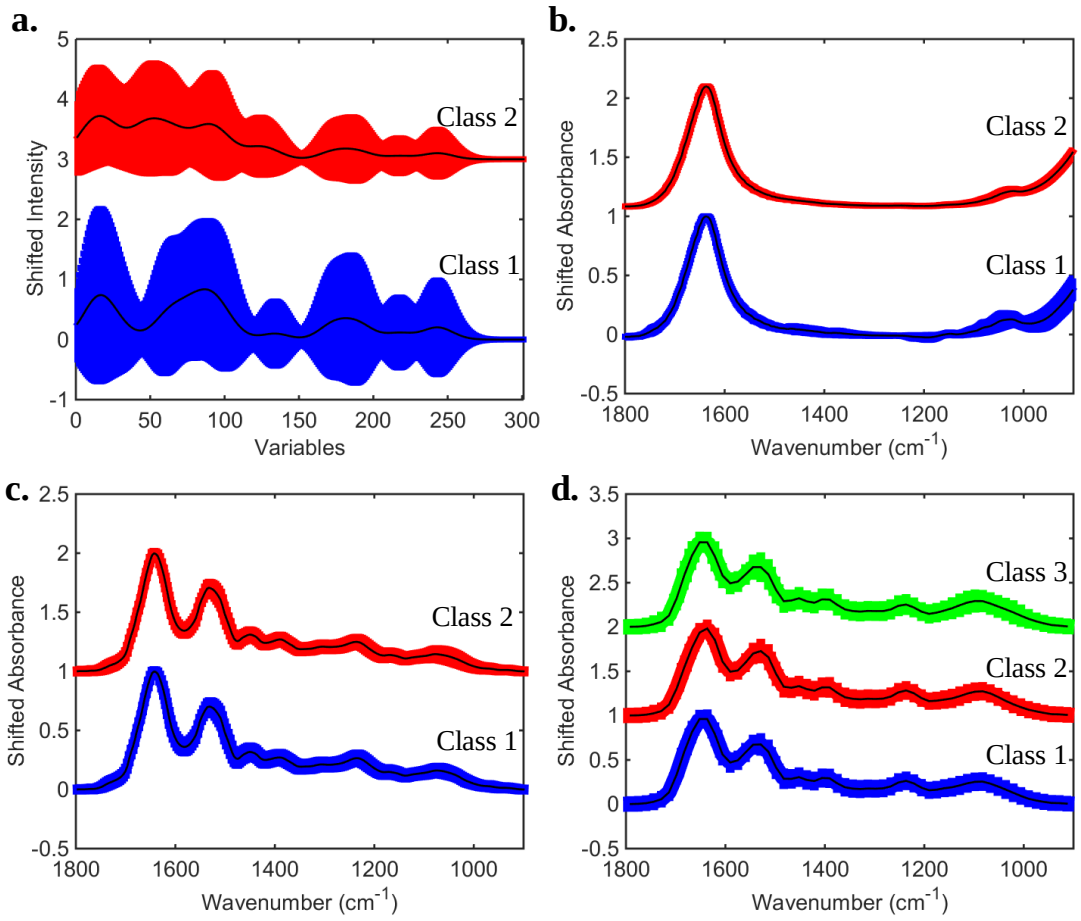


Figure 8.2. Mean and standard-deviation (shaded area) for (a) dataset 1, (b) dataset 2, (c) dataset 3, and (d) dataset 4.

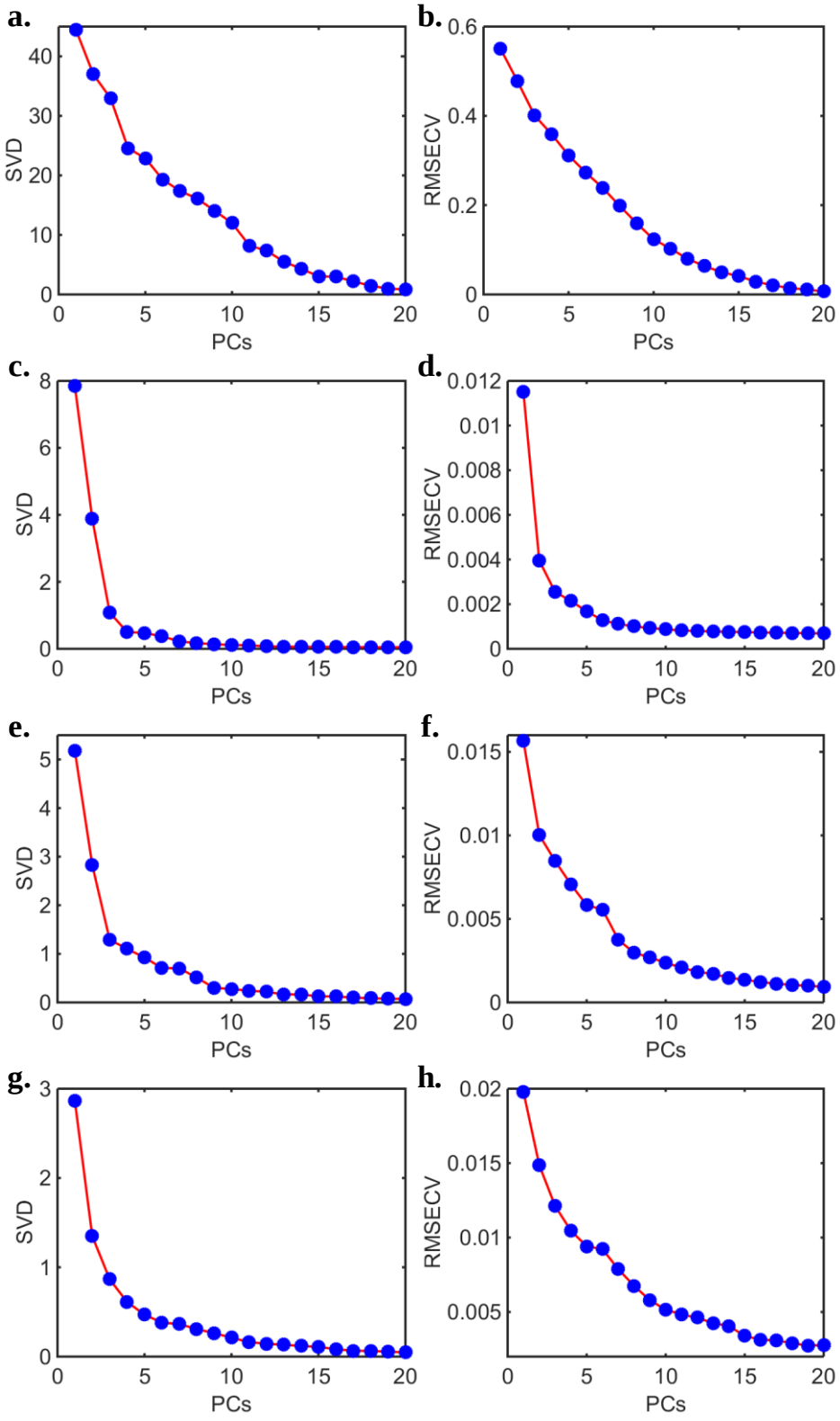


Figure 8.3. Singular value decomposition (SVD) for (a) dataset 1, (c) dataset 2, (e) dataset 3 and (g) dataset4; root mean square error of cross-validation (RMSECV) of PCA for (b) dataset 1, (d) dataset 2, (f) dataset 3 and (h) dataset 4 varying the number of principal components (PCs).

Table 8.1. Figures of merit calculated for the external validation set in datasets 1–4. PPV stands for positive predictive value, NPV for negative predictive value, YOU for Youden's index, and m_p stands for average misclassification probability.

Dataset 1	Accuracy	Sensitivity	Specificity	PPV	NPV	YOU	m_p
PCA-LDA	44.4%	44.4%	44.4%	44.4%	44.4%	-11.1%	0.520
PCA-QDA	88.9%	77.8%	100%	100%	81.8%	77.8%	0.113
PCA-SVM	94.4%	88.9%	100%	100%	90.0%	88.9%	0.152
Dataset 2							
PCA-LDA	86.9%	97.6%	76.2%	80.4%	97.0%	73.8%	0.328
PCA-QDA	100%	100%	100%	100%	100%	100%	0.212
PCA-SVM	97.6%	95.2%	100%	100%	95.5%	95.2%	0.500
Dataset 3							
PCA-LDA	68.1%	80.0%	59.5%	58.5%	80.6%	39.5%	0.319
PCA-QDA	88.9%	90.0%	88.1%	84.4%	92.5%	78.1%	0.276
PCA-SVM	100%	100%	100%	100%	100%	100%	0.244
Dataset 4							
PCA-LDA							
Class 1	94.6%	94.7%	94.4%	97.3%	89.5%	89.2%	0.265
Class 2	89.3%	83.8%	100%	100%	76.0%	83.8%	0.217
Class 3	89.3%	91.9%	84.2%	91.9%	84.2%	76.1%	0.299
PCA-QDA							
Class 1	76.8%	100%	27.8%	74.5%	100%	27.8%	0.500
Class 2	73.2%	100%	21.1%	71.2%	100%	21.1%	0.434
Class 3	75.0%	62.2%	100%	100%	57.6%	62.2%	0.217
PCA-SVM							
Class 1	98.2%	97.4%	100%	100%	94.7%	97.4%	0.447
Class 2	100%	100%	100%	100%	100%	100%	0.468
Class 3	73.2%	59.5%	100%	100%	55.9%	59.5%	0.303

SVM was applied to the PCA scores by means of PCA-SVM generating also a good prediction response (accuracy = 94.4%). Although SVM fitting and prediction are better than QDA in terms of accuracy, sensitivity and specificity; its average misclassification probability is slightly higher (m_p = 0.152). A robustness test was then performed by adding white Gaussian noise to the spectra in 6 different levels of signal-to-noise ratio (S/N) measured in decibels (dB). S/N values of 50 dB, 45 dB, 40 dB, 35 dB, 30 dB and 25 dB were tested. As can be seen in Figure 8.4a, by adding noise to the spectra, the predictive performance in terms of overall accuracy remained constant for PCA-QDA and PCA-SVM models. For PCA-LDA, the addition of noise at 25 dB improved the accuracy to 50%. This phenomenon could happen due to the poor-fitting of

the LDA model for dataset 1 (sensitivity and specificity of 44.4%), since in this case the model response might not be entirely reliable on the signal quality.

For dataset 2 (*Cryptococcus* fungi specimens), PCA-QDA also had a better performance than PCA-LDA. According to Figure 8.2b, class 1 has a clear higher variance for the variables in the range of 900-1200 cm^{-1} (phosphodiester, polysaccharides, glycogen and PO_2^- symmetric stretching in DNA/RNA (Movasaghi *et al.*, 2008)) in comparison with class 2. PCA-QDA achieved perfect class segregation (accuracy = 100%), while PCA-LDA achieved fair results with accuracy at 86.9%. All models were built using 8 PCs determined by SVD and RMSECV values (Figure 8.3c and d) (cumulative variance of 99.8%). Average misclassification probabilities of 0.328 and 0.212 were found for LDA and QDA models, respectively; confirming the higher trustworthiness of PCA-QDA over PCA-LDA for this dataset (Table 8.1). PCA-SVM also achieved good classification results, with an accuracy of 97.6% in the external validation set. However, the average misclassification probability was found at 0.500, which indicates that this model is not stable. The negative predictive value (NPV) for PCA-SVM indicates that the presence of misclassification is present only in the negative samples (*Cryptococcus neoformans*), a possible overfitting sign. Robustness was again evaluated by adding white Gaussian noise to the spectra set. The PCA-QDA was the only model that remained stable with noise, while the other two models (PCA-LDA and PCA-SVM) had an accentuated decrement of accuracy after S/N of 40 dB (Figure 8.4b). As expected by the misclassification probabilities values, the performance of PCA-SVM when the spectra were perturbated by noise was even worse than using PCA-LDA, since its accuracy dropped to 50% at 25 dB.

Dataset 3 is composed of IR spectra of normal brain tissue samples (class 1) and glioblastoma brain tissue samples (class 2) (Figure 8.2c). Both classes seem to have similar spectral profiles and standard-deviations. PCA-SVM classified the data with 100% accuracy (misclassification probability of 0.244) using 10 PCs selected by SVD and RMSECV values (Figure 8.3e and f) (cumulative variance of 99.4%). The second best classification performance was found using PCA-QDA (accuracy = 88.9%, misclassification probability of 0.276) and, for last, PCA-LDA (accuracy = 68.1, misclassification probability of 0.319). The three models are stable until S/N 35 dB, but after this point, all the classifiers tend to lose their classification performance converging

to accuracies of 54.2% (PCA-LDA), 58.3% (PCA-QDA) and 62.5% (PCA-SVM) at 25 dB (Figure 8.3c).

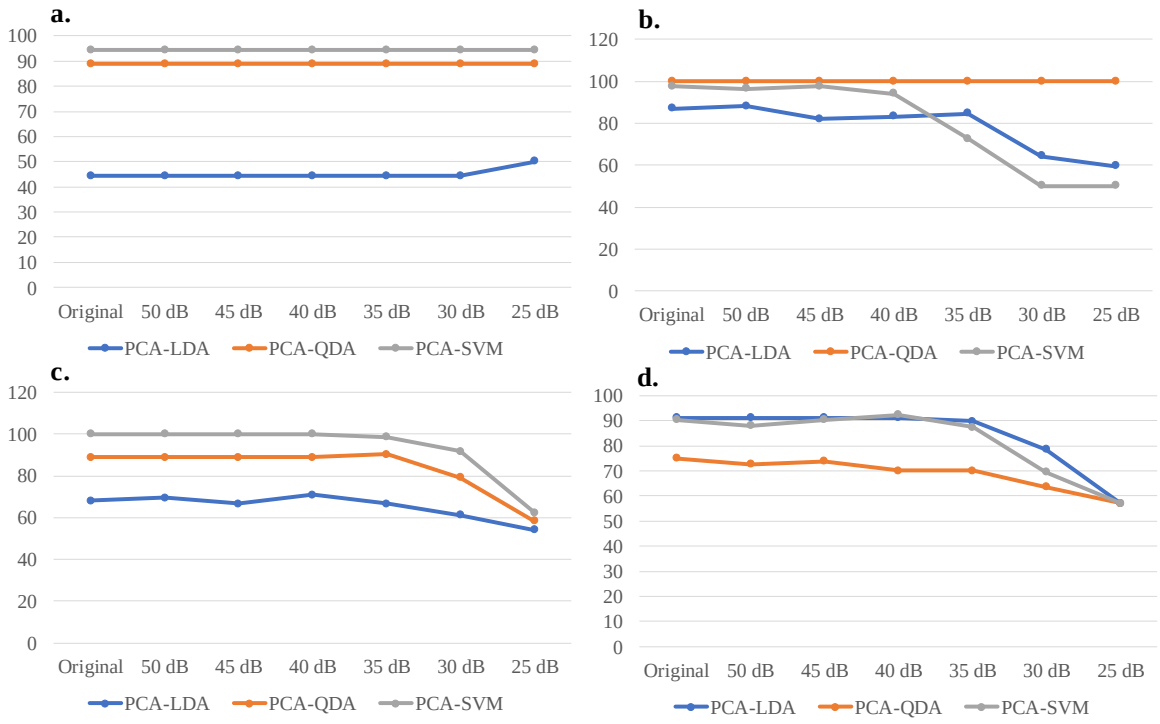


Figure 8.4: Overall accuracy in percentage for PCA-LDA, PCA-QDA and PCA-SVM models in (a) dataset 1, (b) dataset 2, (c) dataset 3 and (d) dataset 4, by adding white Gaussian noise to the spectra datasets in the following levels of signal-to-noise ratio: 50 dB, 45 dB, 40 dB, 35 dB, 30 dB and 25 dB.

Dataset 4 is composed of 3 classes of samples measured by ATR-FTIR. The average spectra with standard-deviation for class 1 (SHE cells contaminated with B[*a*]P), class 2 (SHE cells contaminated with 3-MCA) and class 3 (SHE cells contaminated with Ant) are depicted in Figure 8.2d. The variance among the classes seem to be evenly distributed, according to the similar standard-deviation observed in Figure 8.2d. PCA-LDA was applied using 10 PCs selected by SVD and RMSECV values (Figure 8.3g and h) (cumulative variance of 98.9%), generating an overall accuracy of 91.1% (average misclassification probability = 0.260). This model had the best classification performance in comparison with PCA-QDA and PCA-SVM, which seem to be overfitted according to the small sensitivity and specificity values observed between the classes (Table 8.1). PCA-QDA achieved an overall accuracy of 75.0% (average misclassification probability of 0.384) and PCA-SVM with an overall accuracy of 90.4% (average misclassification

probability of 0.406). By applying noise to the data (Figure 8.4d), the model performance for PCA-LDA remained constant until 35 dB, then quickly dropped afterwards. For PCA-SVM, the model maintained overall accuracy around 90% until 40 dB, followed by a quickly dropping at 35 dB; and for PCA-QDA, the overall accuracy decreased steadily until 25 dB. At 25 dB, all models converged to the same accuracy of 57%.

The mean misclassification probability and uncertainty propagation estimate based on Eq. 8.9 for SVM models are compared in Figure 8.5. An exponential trend is observed between the two parameters (Figure 8.5a), where the uncertainty propagation is proportional to the misclassification probability. A linear relationship between the two parameters is depicted in Figure 8.5b by the application of a natural logarithm function, where an R^2 of 0.971 is found; indicating that the classification uncertainty by bootstrap behaves similar to that one found using RBF functions (Allegrini & Oliveiri, 2016).

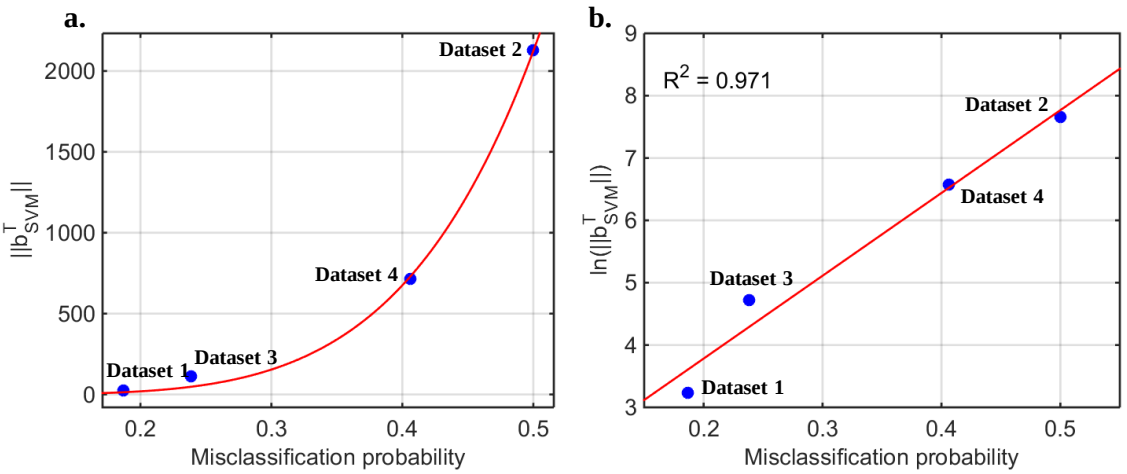


Figure 8.5. (a) Mean misclassification probability using bootstrap *versus* norm of uncertainty propagation coefficients ($\mathbf{b}^{\mathrm{T}}_{\mathrm{SVM}}$) calculated for SVM models with the training samples of datasets 1–4; and (b) mean misclassification probability using bootstrap *versus* natural logarithm of the norm of uncertainty propagation coefficients ($\mathbf{b}^{\mathrm{T}}_{\mathrm{SVM}}$) calculated for SVM models with the training samples of datasets 1–4 (linear equation: $y = 13.3x + 1.13$).

8.4 Conclusion

Misclassification probabilities were determined for PCA-LDA, PCA-QDA and PCA-SVM models applied to 4 different datasets (1 simulated and 4 real data). Uncertainty estimations were calculated by bootstrapping in order to obtain confidence intervals for misclassification probability calculations, presented herein as a new quality parameter to indicate model trustworthiness for these three classifiers. A correlation between the misclassification probability and model robustness was observed by adding white Gaussian noise to the spectral datasets, in which models with higher misclassification probabilities were more susceptible to error. Therefore, the misclassification probability can be used as a new figure of merit to assess model quality in classification applications, containing information of the model uncertainty and being also used to evaluate model robustness.

CHAPTER 9 | STANDARDIZATION OF COMPLEX BIOLOGICALLY-DERIVED SPECTROCHEMICAL DATASETS

This chapter is published in Nature Protocols (IF 11.334). It is a protocol demonstrating how to standardise biospectroscopy datasets acquired in different laboratories or centres in order to have uniform classification results:

- Morais CLM, Paraskevaidi M, Cui L, Fullwood NJ, Martin I, Lima KMG, Martin-Hirsch PL, Sreedhar H, Trevisan J, Walsh MJ, Zhang D, Zhu YG, Martin FL. Standardization of complex biologically derived spectrochemical datasets. Nat. Protoc. **2019**; 14: 1546–1577. https://doi.org/10.1038/s41596-019-0150-x

Abstract: Spectroscopic techniques such as Fourier-transform infrared (FTIR) spectroscopy are used to study interactions of light with biological materials. This interaction forms the basis of many analytical assays used in disease screening/diagnosis, microbiological studies, and forensic/environmental investigations. Advantages of spectrochemical analysis are its low cost, minimal sample preparation, non-destructive nature and substantially accurate results. However, an urgent need exists for repetition and validation of these methods in large-scale studies and across different research groups, which would bring the method closer to clinical and/or industrial implementation. For this to succeed, it is important to understand and reduce the effect of random spectral alterations caused by inter-individual, inter-instrument and/or inter-laboratory variations, such as variations in air humidity and CO_2 levels, and aging of instrument parts. Thus, it is evident that spectral standardization is critical to the widespread adoption of these spectrochemical technologies. By using calibration transfer procedures, in which the spectral response of a secondary instrument is standardized to resemble the spectral response of a primary instrument, different sources of variation can be normalized into a single model using computational-based methods, such as direct standardization (DS) and piecewise direct standardization (PDS); therefore, measurements performed under different conditions can generate the same result, eliminating the need for a full recalibration. Here, we have constructed a protocol for model standardization using different transfer technologies described for FTIR spectrochemical applications. This is a critical step toward the construction of a practical spectrochemical analysis model for daily routine analysis, where uncertain and random variations are present.

Author contribution: C.L.M.M. performed the experiments, data analysis and wrote the manuscript draft.

Camilo L. M. Morais, PhD candidate — Prof. Francis L. Martin, Supervisor

9.1 Introduction

Vibrational spectroscopy has shown great promise as an analytical tool for the investigation of numerous sample types with wide applications in diverse sectors, such as biomedicine, pharmaceutics or environmental sciences (Baker *et al.*, 2014; Baker *et al.*, 2018; Eliasson & Matousek, 2007; Llabjnai *et al.*, 2009; Melin *et al.*, 2000). Fourier-transform infrared (FTIR) spectroscopy is one of the preferred techniques for identification of biomolecules through the study of their characteristic vibrational movements. Another commonly used approach is Raman spectroscopy, which provides complementary spectral information to IR. Raman spectroscopy exploits the inelastic scattering of light whereas IR studies light absorption. Both methods have their benefits and drawbacks. A limitation of IR, for instance, is that water generates undesired peaks at the region of interest, which can mask important biological information, and therefore extra sample preparation and/or spectral processing is necessary. On the contrary, Raman spectroscopy has an inherently weak signal and fluorescence interference, which can, however, be addressed by optimizing the experimental settings or by applying enhancement techniques to increase the Raman signal. For the purposes of this protocol we have used FTIR spectroscopy to demonstrate our standardization model.

Using chemometric approaches, the system is trained to recognize unique spectral features within a sample, so that when unknown samples are introduced an accurate classification is feasible. Alterations in the measurement parameters could interfere with the spectral signature and produce random variations. Therefore, a crucial step is spectral correction, or standardization, which would provide comparable results and allow system transferability. The idea is that non-biological variations, such as those arising from different users, locations or instruments, will no longer affect the classification result; therefore any collected data could be imported into a central database and handled for further exploration or diagnostic purposes. Several groups and companies worldwide are developing spectrochemical approaches for diagnosis, discrimination and monitoring of diseases, as well as for other uses. Combination of multiple datasets would facilitate the conduction of large-scale studies which are still lacking in the field of bio-spectroscopy.

9.1.1 Sensor-based Technologies

Sensor-based technologies are an integral part of daily life ranging from locating sensor-based technology, such as global positioning system (GPS) (Hofmann-Wellenhof *et al.*, 2012), to image biosensors, such as X-rays (Kalita & Misra, 2000; Lagleyre *et al.*, 2009; Lee *et al.*, 2009; Morris & Perkins, 2012) and γ-rays (Jagust *et al.*, 2007; Schrevens *et al.*, 2004; Zhou *et al.*, 2009), which are used extensively for medical applications. Other powerful approaches that make use of sensor-based technologies toward medical disease examination and diagnostics include circular dichroism (CD) spectroscopy (Greenfield, 2006; Micsonai *et al.*, 2015; Miles & Wallace, 2016; Wallace *et al.*, 2004), ultraviolet (UV) or visible spectroscopy (Brown *et al.*, 2009; Yang *et al.*, 2016), fluorescence (Shahzad *et al.*, 2009; Shahzad *et al.*, 2010; Shin *et al.*, 2010; Sierón *et al.*, 2013; World Health Organization, 2005), nuclear magnetic resonance (NMR) spectroscopy (Chan *et al.*, 2016; Frisoni *et al.*, 2010; Gowda *et al.*, 2008; Möller-Hartmann *et al.*, 2002; Palmnas & Vogel, 2013) and ultrasound (US) (Menon *et al.*, 2009; Morris & Perkins, 2012; Navani *et al.*, 2015; Patil & Dasgupta, 2012; Smith-Bindman *et al.*, 1998).

Over the last two decades, optical biosensors employing vibrational spectroscopy, particularly IR spectroscopy, have seen tremendous progress in biomedical and biological research. A number of studies using the above-mentioned methods have focused on cancer investigation with malignancies such as brain (Bury *et al.*, 2019a; Gajjar *et al.*, 2013; Hands *et al.*, 2014; Hands *et al.*, 2016), breast (Backhaus *et al.*, 2010; Lane & Seo, 2012; Walsh *et al.*, 2012), oesophagus (Maziak *et al.*, 2007; Wang *et al.*, 2003), skin (Hammondy *et al.*, 2005; McIntosh *et al.*, 1999; McIntosh *et al.*, 2001; Mordechai *et al.*, 2004; Mostaço-Guidolin *et al.*, 2009), colorectal (Kondepati *et al.*, 2007; Rigas *et al.*, 1990; Yao *et al.*, 2014), lung (Akalin *et al.*, 2015; Großerueschkamp *et al.*, 2015; Lewis *et al.*, 2010), ovarian (Gajjar *et al.*, 2013; Mehrotra *et al.*, 2010; Owens *et al.*, 2014; Theophilou *et al.*, 2016), endometrial (Gajjar *et al.*, 2013; Paraskevaidi *et al.*, 2018c; Taylor *et al.*, 2011), cervical (Gajjar *et al.*, 2014; Podshyvalov *et al.*, 2005; Walsh *et al.*, 2007; Wood *et al.*, 1996) and prostate (Baker *et al.*, 2009; Derenne *et al.*, 2011; Gazi *et al.*, 2006; Theophilou *et al.*, 2015) cancer being some of them. Non-cancerous diseases have also been examined, namely neurodegenerative disorders (Carmona *et al.*, 2013; Carmona *et al.*, 2015; Paraskevaidi *et al.*, 2017b; Paraskevaidi *et al.*, 2018b), HIV/AIDS (Sitole *et al.*, 2014), diabetes (Coopman *et al.*, 2017; Scott *et al.*, 2010; Varma *et al.*, 2016), rheumatoid arthritis (Canvin *et al.*, 2003; Lechowicz *et al.*, 2016), cardiovascular

diseases (Oemrawsingh *et al.*, 2014; Wang *et al.*, 2002), malaria (Khoshmanesh *et al.*, 2014; Martin *et al.*, 2017; Roy *et al.*, 2017), alkaptonuria (Markus *et al.*, 2001), cystic fibrosis (Grimard *et al.*, 2004), thalassemia (Aksoy *et al.*, 2012), prenatal disorders (Graça *et al.*, 2013; Hasegawa *et al.*, 2010), macular degeneration (Semoun *et al.*, 2009; Theelen *et al.*, 2009), atherosclerosis (Peters *et al.*, 2017; Wang *et al.*, 2002) and osteoarthritis (Afara *et al.*, 2017; Bi *et al.*, 2007; David-Vaudey *et al.*, 2005).

9.1.2 Limitations

Spectrochemical approaches are advantageous when compared with traditional molecular methods as they provide a holistic status of the sample under interrogation, thus generating typical spectral regions widely known as “fingerprint regions”. These methods have also been shown to be rapid, inexpensive and non-destructive while they also improve diagnostic performance and eliminate subjective diagnosis (*e.g.*, histopathological diagnosis), where inter- and intra-observer variability are present (Trevisan *et al.*, 2012). However, like any other analytical method, vibrational spectroscopy also comes with some limitations. For instance, prior to FTIR studies, optimization of instrumental settings, sample preparation and operation mode also needs to be conducted in order to improve the spectral quality and molecular sensitivity (Baker *et al.*, 2014; Chan & Kazarian, 2016; Pilling & Gardner, 2016). Overall, the above-mentioned barriers can be overcome after careful consideration of the experimental design.

A considerable limitation that is yet under-investigated in the field of spectrochemical techniques is associated with the difficulties entailed in data conformation and system standardization. Currently, there are multiple pilot studies showing promising results but an approach towards standardization for biological applications is lacking. Random variation between studies can originate from differences in instrumentation, operators, and environmental conditions, such as room temperature and humidity.

The main objective of this article is to present a protocol for model standardization which can be applied in FTIR spectrochemical techniques to rule out the chance of random spectral alterations. Inter-individual, inter-instrument, inter-sample and/or inter-laboratory variations can be a source of unwanted, non-biological alterations, thus leading

to incorrect conclusions. However, for a method to become reliable and clinically translatable, it is important that measurements performed under different conditions generate comparable results. The aim of the spectral standardization model presented here is to expedite multi-centre studies with large numbers of samples; this would bring these spectrochemical techniques closer to clinical implementation and facilitate life-changing decisions. We describe a protocol that has four main components: (i) sample preparation, (ii) spectral acquisition, (iii) data pre-processing and (iv) model standardization. The current protocol has an in-depth insight obtained from cross-laboratory collaborations with leading experts in the field. This article offers a step-by-step procedure, which can be implemented by a non-specialist in spectrochemical studies. For further information about instrumental and software options, spectral acquisition steps and data analysis for a range of different analytical systems the reader is directed towards additional protocols (Baker *et al.*, 2014; Beckonert *et al.*, 2007; Butler *et al.*, 2016; Felten *et al.*, 2015; Harmsen *et al.*, 2017; Kong *et al.*, 2011; Martin *et al.*, 2010; Sreedhar *et al.*, 2015; Yang *et al.*, 2015).

9.1.3 Applications

Spectrochemical approaches, in combination with computational analysis, have been proven to be effective for biomedical research through facilitating the diagnosis, classification, prognosis, treatment stratification and modulation or monitoring of a disease and treatment. However, these techniques are widely applicable to other fields as well, namely food industry (Osborne & Fearn, 2000; Qu *et al.*, 2015; Song *et al.*, 2013; Varriale *et al.*, 2007), toxicology (Harrigan *et al.*, 2004; Melin *et al.*, 2000; Penido *et al.*, 2016; Ryder, 2002), microbiology (Carmona *et al.*, 2005; Cui *et al.*, 2018; Choo-Smith *et al.*, 2001; Helm *et al.*, 1991; Lasch & Naumann, 2015; Maquelin *et al.*, 2002), forensics (Ali *et al.*, 2008; Day *et al.*, 2004; Hargreaves & Matousek, 2016; Lewis *et al.*, 1995; Macleod & Matousek, 2008), pharmacy (Eliasson & Matousek, 2007; Melin *et al.*, 2000; Vergote *et al.*, 2002), environmental and plant science (Comino *et al.*, 2018; Heys *et al.*, 2017; Lohr *et al.*, 2017), as well as defence and security (Eliasson *et al.*, 2007; Golightly *et al.*, 2009; Liu *et al.*, 2007). Applications of standardization algorithms vary according to the spectral technique and sample matrix studied, and have been mostly applied to Raman and Fourier-transform near-infrared (FT-NIR) spectroscopy. Table 9.1 summarizes some standardization applications.

Table 9.1. Examples of applications involving standardization techniques.

Sample matrix	Spectroscopic technique	Aim	Ref.
Tissue	Raman	Standardization of various perturbations on Raman spectra for diagnosis of breast cancer based on snap frozen tissues	(Sattlecker *et al.*, 2011)
	Raman	Standardization of spectra acquired in 3 different sites for analysing oesophageal samples based on snap frozen tissues	(Isabelle *et al.*, 2016)
Cells	Raman	Standardization of spectra acquired with 4 different instruments for classification of three different cultured spore species	(Guo *et al.*, 2017)
Biofluids	FT-NIR	Standardization of spectra acquired with 3 different instruments for measuring haematocrit in the blood of grazing cattle	(Luo *et al.*, 2017)
	LC-MS	Standardization of spectra acquired with 2 different instruments for mapping rendition times and matching metabolite features of subjects diagnosed with small cell lung cancer based on blood serum and plasma samples analysis	(Vaughan *et al.*, 2012)
Pharmaceutical materials	Raman	Standardization of spectra acquired with 5 different instruments for analysing various pharmaceutical excipients, active pharmaceutical ingredients (APIs) and common contaminants	(Rodriguez *et al.*, 2011)
	FT-NIR	Standardization of spectra acquired with 2 different instruments for simultaneous determination of rifampicin and isoniazid in pharmaceutical formulations	(de Andrade *et al.*, 2018)
	FT-NIR	Standardization of spectra acquired with 2 different instruments for predicting content of 654 pharmaceutical tablets	(Yu *et al.*, 2016)
Food	FT-NIR	Standardization of spectra acquired with 3 different instruments for predicting parameters in corn samples	(Ni *et al.*, 2019)
	FT-NIR	Standardization of spectra acquired with 2 different instruments for predicting vitamin C in navel orange	(Hu & Xia, 2011)
	FT-NIR	Standardization of spectra recorded in 4 different labs for determining moisture, proteins and oil content in soy seeds	(Forina *et al.*, 1995)
	FT-NIR	Standardization of spectra acquired by a benchtop and portable instrument for determining total soluble solid contents in single grape berry	(Xiao *et al.*, 2017)
	UV-Vis	Standardization of visible spectra acquired with 3 different instruments for measuring pH of Sala mango	(Yahaya *et al.*, 2015)
Plant	FT-NIR	Standardization of spectra acquired with 2 different instruments for predicting baicalin contents in radix scutellariae samples	(Ni *et al.*, 2019)
	FT-NIR	Standardization of spectra acquired by 2 different instruments and in three physical states (powder, filament and intact leaf) for determining total sugars, reducing sugars and nicotine in tobacco leaf samples	(Bin *et al.*, 2017)
	NMR	Standardization of spectra acquired with 3 different instruments for authenticity control of sunflower lecithin	(Monakhova & Diehl, 2016)
Cosmetic	CD spectroscopy	Standardization of spectra acquired between standard and real-world samples for determining Pb^{2+} in cosmetic samples	(Zuo *et al.*, 2017)
Inorganic substances	FT-IR	Standardization of interferogram spectra acquired with 2 instruments for classifying acetone and SF_6 samples	(Koehler *et al.*, 2000)
Fuel	FT-IR	Standardization of spectra acquired with 2 different instruments for predicting density of crude oil samples	(Rodrigues *et al.*, 2017)

9.1.4 Model Transferability

Transferability models have been previously developed, however this is still an under-investigated field, especially for biomedical applications. These models use computer-based methods to standardize spectral data generated across different experimental settings (*e.g.*, different instruments, operators or laboratories). An inclusive standardization protocol that could be implemented in a range of different spectrochemical approaches is of great need. Differences are present even between identical instruments; for instance, changes in signal intensity caused by replacement, alignment or ageing of optical and spectrometer components, natural variations in optics and detectors construction, changes in measurement conditions (temperature and humidity), changes in physical constitution of the sample (particle size and surface texture) and operator discrepancies could all lead to wavenumber shifts and artefacts in the spectra. In all of these cases, prediction errors of the estimated group categories (*e.g.*, whether the sample is classified as healthy or cancerous) can become very large, especially when the whole spectrum is used in the model. Standardization techniques aim to generate a uniform spectral response under differing conditions, ensuring the interchangeability of results obtained in different situations, without having to perform a full calibration for each situation.

Previous standardization methods include the use of simple slope and bias correction (Brouckaert *et al.*, 2018; Wang *et al.*, 1991), direct standardization (DS) (de Andrade *et al.*, 2018; Khaydukova *et al.*, 2017; Morais & Lima, 2015; Panchuk *et al.*, 2017; Zamora-Rojas *et al.*, 2012), piecewise direct standardization (PDS) (Barreiro *et al.*, 2008; Sulub *et al.*, 2008; Wang *et al.*, 1991; Zhang *et al.*, 2003), piecewise linear discriminant analysis (PLDA) (Koehler *et al.*, 2000), guided model reoptimization (GMR) (Zhang *et al.*, 2003), back-propagation neural network (BNN) (Koehler *et al.*, 2000), generalized least squares weighting (GLSW) (Martens *et al.*, 2003), model updating (MU) (Feudale *et al.*, 2002; Woody *et al.*, 2004), orthogonal signal correction (OSC) (Greensill *et al.*, 2001; Sjöblom *et al.*, 1998), orthogonal projections to latent structures (OPLS) (Rodrigues *et al.*, 2017), wavelet hybrid direct standardization (WHDS) (Sulub *et al.*, 2008), maximum likelihood PCA (MLPCA) (Andrews *et al.*, 1997), Shenk and Westerhaus method (SW) (Bouveresse *et al.*, 1994; Shenk & Westerhaus, 1991), positive matrix factorization (PMF) (Paatero & Tapper, 1994; Xie & Hopke, 1999), artificial neural networks (ANN) drift correction (Goodacre *et al.*, 1997), transfer *via* extreme learning machine auto-encoder method (TEAM) (Chen *et al.*, 2016), calibration transfer based on the maximum margin criterion (CTMMC) (Hu *et al.*, 2012), calibration transfer based on canonical correlation analysis (CTCCA) (Fan *et al.*, 2008) and

calibration methods, such as wavenumber offset correction, instrument response correction and baseline correction (Isabelle *et al.*, 2016). In this protocol, we use direct standardization (DS) and piecewise direct standardization (PDS), because they are the most common methods for spectral standardization.

Direct standardization. DS is one of the most used methods for data standardization. It was initially proposed to correct relatively large spectral differences between data collected from the same sample measured by two different instruments (Wang *et al.*, 1991). In DS, the entire spectrum from a new secondary response (*e.g.*, a different instrument) is transformed to resemble the spectrum from the primary source (*e.g.*, original instrument) (de Andrade *et al.*, 2018). This is performed based on a linear relationship between the data acquired under different circumstances (Feudale *et al.*, 2002):

$$\mathbf{S}_1 = \mathbf{S}_2\mathbf{F} \tag{9.1}$$

where $\mathbf{S}_1$ represents the data acquired for the primary response; $\mathbf{S}_2$ represents the data acquired for the secondary response; and **F** is the transformation matrix that maintains the relationship between $\mathbf{S}_1$ and $\mathbf{S}_2$.

The transformation matrix **F** is estimated in a least-squares sense by (Wang *et al.*, 1995):

$$\mathbf{F} = \mathbf{S}_2^{+}\mathbf{S}_1 \tag{9.2}$$

where $\mathbf{S}_2^{+}$ is the pseudo-inverse of $\mathbf{S}_2$, calculated by:

$$\mathbf{S}_2^{+} = (\mathbf{S}_2^{\mathrm{T}}\mathbf{S}_2)^{-1}\mathbf{S}_2^{\mathrm{T}} \tag{9.3}$$

in which T stands for the matrix transpose operation.

Then, when samples are measured under the secondary system, the signals generated **X** are transformed to resemble the primary system response by (Feudale *et al.*, 2002):

$$\widehat{\mathbf{X}}^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}\mathbf{F} \tag{9.4}$$

where $\widehat{\mathbf{X}}$ is the standardized response for **X**.

Problems related to different background information between instruments can affect the standardization procedure. To correct for this, the standardization process is usually adapted with the background correction method (Wang *et al.*, 1995), in which the transformation matrix described in Eq. 9.2 is calculated with a background correction factor ($\mathbf{F}_\mathrm{b}$) and an additive background correction vector $\mathbf{b}_\mathrm{s}$ as follows:

$$\mathbf{S}_1 = \mathbf{S}_2\mathbf{F}_\mathrm{b} + \mathbf{1}\mathbf{b}_\mathrm{s}^\mathrm{T} \quad (9.5)$$

where $\mathbf{1}$ is an all-ones vector and $\mathbf{b}_\mathrm{s}$ is obtained by:

$$\mathbf{b}_\mathrm{s} = \mathbf{s}_{1\mathrm{m}} - \mathbf{F}_\mathrm{b}^\mathrm{T}\mathbf{s}_{2\mathrm{m}} \quad (9.6)$$

in which $\mathbf{s}_{1\mathrm{m}}$ is the mean vector of $\mathbf{S}_1$ and $\mathbf{s}_{2\mathrm{m}}$ is the mean vector of $\mathbf{S}_2$.

One of the key steps for DS is the selection of the number of samples to transfer (called "transfer samples"). These are samples' spectra from the primary system ($\mathbf{S}_1$) that will be used to transform the signal obtained using the secondary system ($\mathbf{S}_2$). The transfer samples are obtained from a same cohort of samples (*e.g.*, plasma samples) measured in the two instruments (primary and secondary systems). Usually, the procedure for selecting transfer samples is based on sample selection techniques, such as Kennard-Stone (KS) algorithm (Kennard & Stone, 1969) or leverage (Wang *et al.*, 1991). Subsequently, the number of transfer samples is evaluated using a validation set through an arbitrary cost function. For quantification applications, a common cost function is the root-mean-square error of prediction, while for classification one can use the misclassification rate.

A disadvantage of DS is that each transformed variable is calculated using the whole spectrum, which carries a high risk of overfitting. The estimation of **F** in Eq. (9.2) is an ill-conditioned problem, because the number of variables (*e.g.*, wavenumber) may be much larger than the number of standard samples.

Piecewise direct standardization. PDS is another standardization procedure commonly employed for system transferability. It is based on DS, however it uses windows (*e.g.*, wavenumber portions) to make the standardization process more suitable for smaller regions of the data. When compared to DS, PDS is calculated by using the transformation matrix **F** with most of its off-diagonal elements set to zero (Wang *et al.*, 1991). With this, PDS fits minor spectral modifications not covered by DS. PDS is the technique of preference for correcting smaller spectral variations, such as small wavelengths shift, intensity variations, and bands enlargement and reduction (Wang *et al.*, 1991). In addition, an advantage of PDS compared to DS is that the local rank of each window will be smaller than the rank of the whole data matrix, which means that the number of standard samples can be smaller, and indeed good results have been obtained with very few samples.

One disadvantage of PDS is the need of an additional optimization process, because in addition to the number of transfer samples, PDS also needs a window size optimization, which

might lead to a risk of overfitting. In this protocol, window size optimization is made using a cost function expressed as the misclassification rate calculated for each window size tested, being evaluated using a validation set where the window with smaller misclassification is selected for final model construction.

9.2 Experimental Design

Any study using vibrational spectroscopy, follows these general steps: careful experimental design, protocol optimisation and development of experimental procedure document, sample collection and preparation, spectral collection, pre-processing of the derived information and lastly the use of chemometrics for exploratory, classification and standardization purposes. FTIR spectroscopy is described in more detail in this study, however, the standardization protocol described here can be adapted to a range of techniques, including attenuated total reflection (ATR-FTIR), transmission and transflection FTIR, near-IR (NIR), UV-visible, NMR spectroscopy and mass spectrometry (MS). Nevertheless, intrinsic features of each technique should be taken into consideration before standardization and the protocol may change depending on the application of interest.

A number of biological samples can be analyzed with the above-mentioned analytical methods such as tissues, cytological materials or biological fluids. Sample type and preparation may differ depending on the technique that is employed each time. For instance, IR spectroscopy is limited by water interference at the fingerprint region that can mask the signal of the analyte close to the water peak. This could be addressed with an extra step of sample drying, in contrast to Raman spectroscopy, for example, where water does not generate a signal in this region.

Typical steps for sample preparation, acquisition of spectra and data pre-processing are briefly presented here. However, the main focus of this protocol is placed on the calibration transfer and standardization procedures. Readers are directed to additional literature for more detailed information regarding sample format and preparation[4,98-100,105,175-177], suitability of substrates (Baker *et al.*, 2014; Butler *et al.*, 2016), instrumentation settings (Aebersold & Mann, 2003; Baker *et al.*, 2014; Butler *et al.*, 2016; Martin *et al.*, 2010; Palonpon *et al.*, 2013; Pence & Mahadevan-Jansen, 2016; Sreedhar *et al.*, 2015) or available software packages (Table 9.2) and manufacturers (Baker *et al.*, 2014; Butler *et al.*, 2016) .

Table 9.2. Software packages for data standardization.

Software	Website	Description	Availability
PLS_Toolbox	http://www.eigenvector.com/	MATLAB toolbox for chemometric analysis. Contains standardization routines using DS, PDS, double window PDS, spectral subspace transformation, GLSW, OSC, and alignment of matrices.	Commercial
Unscrambler® X	http://www.camo.com/	Software for multivariate data analysis and design of experiments. Contains standardization routines using interpolation, bias and slope correction, and PDS.	Commercial
OPUS	https://www.bruker.com/	Spectral acquisition software with data processing features. Contains a standardization routine using PDS.	Commercial
Pirouette®	https://infometrix.com/	Chemometrics modelling software. Contains standardization routines using DS and PDS.	Commercial

9.2.1 Experimental Design: Sampling

Sample preparation. Biological samples have been studied extensively with spectrochemical techniques for disease research. Tissue specimens can be analysed fresh, snap-frozen or formalin-fixed, paraffin-embedded (FFPE). Fresh or snap-frozen histology sections are preferable as they are devoid of contaminants whereas FFPE treatment contributes to characteristic peaks, hindering the biological information. FFPE tissues can be deparaffinized either by chemical methods (*e.g.*, incubation in xylene, hexane or Histo-Clear solutions) (Baker *et al.*, 2014), which can alter tissue structures and be inefficient for the complete wax removal (Ibrahim *et al.*, 2017), or by applying chemometrics (*e.g.*, digital dewaxing) (Byrne *et al.*, 2016; Tfayli *et al.*, 2009), which keeps the tissue intact but might introduce artefacts due to over- or under-estimation of the wax contribution (Ibrahim *et al.*, 2017).

Fixatives, such as ethanol, methanol or formalin, are often used for the preservation of cytological material, also generating strong peaks and interfering with the spectra; thus, a washing step is crucial before spectroscopic interrogation. Fixation in tissue or cells for preservation purposes generates protein cross-linking which can cause changes in the spectra, especially on the Amide I peak (Meade *et al.*, 2010). Alternatively, cells can be studied live after washing from residual medium.

Preparation and pre-treatment of biological fluids depend on the sample type. Some of the biofluids that have been previously used in spectroscopic studies include blood (whole

blood, plasma or serum), urine, sputum, saliva, tears, cerebrospinal fluid (CSF), synovial fluid, ascitic fluid or amniotic fluid (Baker *et al.*, 2016; Bonifacio *et al.*, 2015; Mitchell *et al.*, 2014). An initial centrifugation step should precede analysis in cases where the cells present in these fluids are not the focus of the study; the supernatant could then be kept for further analysis. In blood-based studies, the user should also consider the anticoagulant of preference (*e.g.*, EDTA, citrate or heparin) as it could generate unwanted spectral peaks (Bonifacio *et al.*, 2014; Lovergne *et al.*, 2016; Paraskevaidi *et al.*, 2017a). Careful planning of experiments as well as consistence throughout a study are of great importance for the generation of robust results. Care should be taken to generate samples that are stable, since the spectral differences between the data collected under different situations (*e.g.*, different instruments or temperature) should be directly related to the difference between the systems and not a change caused by chemical or physical degradation of the samples. Optimal sample thickness, suitability of substrates and sample formats can differ from one analytical technique to another and thus the user should decide and tailor these according to the study's objective (a list with appropriate substrates is given in the Materials-Equipment section). Another consideration is the number of freeze-thaw cycles and long-term storage as these could compromise the integrity of the samples (Lovergne *et al.*, 2016; Mitchell *et al.*, 2005). Preferably, FFPE tissue samples should be analysed after thorough dewaxing and freeze-thaw cycles or long-term storage avoided since these could result in many confounding factors for analysis.

Spectral acquisition. Depending on the study's objective, FTIR spectral information can be collected using either point spectra or imaging.

FTIR spectra can be collected in different operational modes, namely ATR-FTIR, transmission or transflection. Instrument parameters such as resolution, aperture size, interferometer mirror velocity and co-additions have to be optimised before acquisition of spectra to achieve high SNR (Baker *et al.*, 2014; Martin *et al.*, 2010). Metal surfaces can also be used to increase the IR signal in a technique known as surface-enhanced IR absorption (SEIRA) (Glassford *et al.*, 2013; Kundu *et al.*, 2008). As water interference can mask biological information in IR spectra, the user can purge the spectrometer with dry air or nitrogen gas to reduce the internal humidity of the instrument, or use computational analysis to remove the water signature. In addition, samples should be dried until all water content evaporates; however, drying of a sample is not without consequences, since chemical changes may occur such as loss of volatile compounds. A background sample is collected regularly to account for any changes in the atmospheric or instrument conditions.

For analysing homogenous samples (*e.g.*, biofluids), measurements can be performed by acquiring spectra on different regions of the centre of a drop and across its borders. In transmission measurements, the sample can be measured raw or diluted. Usually, 10 spectra are collected per sample. A higher number of spectral replicas can be performed to decrease the standard-deviation (SD) between measurements, since the SD is proportion to $1/\sqrt{n}$, where n is the number of replicas. For heterogeneously distributed samples (*e.g.*, tissues), spectra should be acquired covering the sample surface as uniformly as possible, to ensure that all sources of variation in the sample are stored in the spectral data. Sample replicas are also recommended at least as triplicates. For precision estimation, at least six replicates at three levels should be performed. The minimum number of samples for analysis can be estimated using a power test at an 80% power (Jones *et al.*, 2003). Further details regarding sampling methodologies for analysing biological materials using FT-IR spectroscopy can be found in our previous protocols (Baker *et al.*, 2014; Martin *et al.*, 2010).

9.2.2 Experimental Design: Data Quality Evaluation

Before processing, the data can be assessed to identify the presence of anomalous behaviours or biased patterns. This can be made initially by visual inspection (*e.g.*, identification of very anomalous spectra) followed by Hotelling T^2 *versus* Q residuals charts using only the mean-centred spectra. PCA residuals (Beebe *et al.*, 1998) can be explored to identify biased patterns, in which heteroscedastic distributions are signs of biased experimental measurements; while homoscedastic distributions are associated with good sampling. SNR can be estimated by dividing the power (P) of signal by the power of noise, that is $\mathrm{SNR} = P_{signal}/P_{noise} = \left(A_{signal}/A_{noise}\right)^2$, where A is the amplitude; or by the inverse of the coefficient of variation, when only non-negative variables are measured. Collinearity can be evaluated by calculation of the condition number, which is a matrix calculation that measures how sensitive the result is to perturbations in the input data (*i.e.*, spectra) and to roundoff errors made during the solution process. This value is naturally high for spectral data (high collinearity).

9.2.3 Experimental Design: Pre-processing

Data pre-processing is used to maximise the SNR. This process is fundamental for correcting physical interferences, such as light scattering, different sample thickness, different optical paths and instrumental noise. Therefore, the pre-processing step has fundamental importance to highlight the signal of interest, reduce interferences and possibly correct anomalous samples.

For standardization applications, the pre-processing step is also important for reducing differences between the different systems that are used. Before any additional pre-processing, the spectrum should be trunctated to the biofingerprint region (*e.g.*, 900-1800 cm^{-1}) before analysis. This region contains the main absorptions from biochemical compounds and it suffers only minor effects of environmental variability, such as air humidity (free νO-H = 3650–3600 cm^{-1}, hydrogen-bonded νO-H = 3400 – 3300 cm^{-1}) and air CO_2 ($\nu_s CO_2$ = 2350 cm^{-1}) (Pavia *et al.*, 2008). Table 9.3 summarizes the main pre-processing techniques for correcting noise in biologically-derived datasets.

Figure 9.1 shows the effect of a pre-processing approach employed for a blood plasma dataset acquired under different experimental conditions (*i.e.*, different systems and operators). In this Figure, the reduction of the spectral differences between the systems is evident after data pre-processing (Savitzky-Golay smoothing, MSC, baseline correction and normalization).

After pre-processing (Table 9.3), a scaling step should be done, because most classification methods require all the variables (*e.g.*, wavenumbers) in the dataset to be at the same scale in order to work properly.

For spectral data, mean-centring (also referred as "standardization" by Hastie *et al.* (2009)) is a very reasonable approach, after which all variables in the dataset will have zero mean. When data contain values represented by different scales (*e.g.*, after data fusion using both IR and Raman spectra), block-scaling should be used, where each block of data (*i.e.*, data from each instrumental technique) would have the same sum-of-squares (normally after mean-centring).

Table 9.3. Main pre-processing used for biologically-derived datasets.

Pre-processing	Interfering	Technique	Advantage	Disadvantage	Optimization
Savitzky-Golay smoothing (Savitzky & Golay, 1964)	Instrumental noise	ATR-FTIR, FTIR, NIR, Raman, NMR, UV-Vis	Corrects spectral noise without changing the shape of data significantly	The polynomial order and window size for polynomial fit affects the result	The polynomial function should have an order similar to the spectral data (*e.g.*, 2nd order polynomial function for IR data) and the window size should be an odd number and not too small (keeping the noise) or too large (changing the spectral shape)
Multiplicative scatter correction (MSC) (Geladi *et al.*, 1985)	Light scattering (Mie scattering), different pressure over the sample when using ATR or probe, different lengths of optical path	ATR-FTIR, FTIR, NIR, Raman, NMR, UV-Vis	Corrects light scattering maintaining the same spectral shape and signal scale	Need of a reference spectrum representative of all measurements	The reference spectrum is regularly set as the average spectrum across all training samples
Standard normal variate (SNV) (Barnes *et al.*, 1989)	Light scattering (Mie scattering), different pressure over the sample when using ATR or probe, different lengths of optical path	ATR-FTIR, FTIR, NIR, Raman, NMR, UV-Vis	Corrects light scattering maintaining the same spectral shape	Creates negative signals since the data are centralized to zero (y-scale)	--
Spectral differentiation (Savitzky & Golay, 1964)	Light scattering (Mie scattering), different pressure over the sample when using ATR or probe, different lengths of optical path, background absorption interfering	ATR-FTIR, FTIR, NIR, Raman, NMR, UV-Vis	Corrects light scattering and baseline problems; highlights smaller spectral differences	Changes the signal scale, shifts the data and increases noise	The order of the derivative function should be used carefully to avoid increased noise (usually 1st or 2nd order differentiation is preferred). The differentiation can be coupled to Savitzky-Golay smoothing
Baseline correction (Brereton, 2003)	Background absorption interfering	ATR-FTIR, FTIR, NIR, Raman, NMR, UV-Vis, MS	Corrects the baseline maintaining the same spectral shape	--	There are many methods for baseline correction (*e.g.*, rubber band, automatic weighted least squares, Whittaker filter). The method chosen should be maintained consistent for all systems used
Normalization (Trevisan *et al.*, 2012)	Different sample thickness and concentration	ATR-FTIR, FTIR, Raman	Avoids influence of non-desired signals among the samples	The normalization might hide signal differences between samples at important bands, such as Amide I and Amide II; and also may introduce non-linearities	--

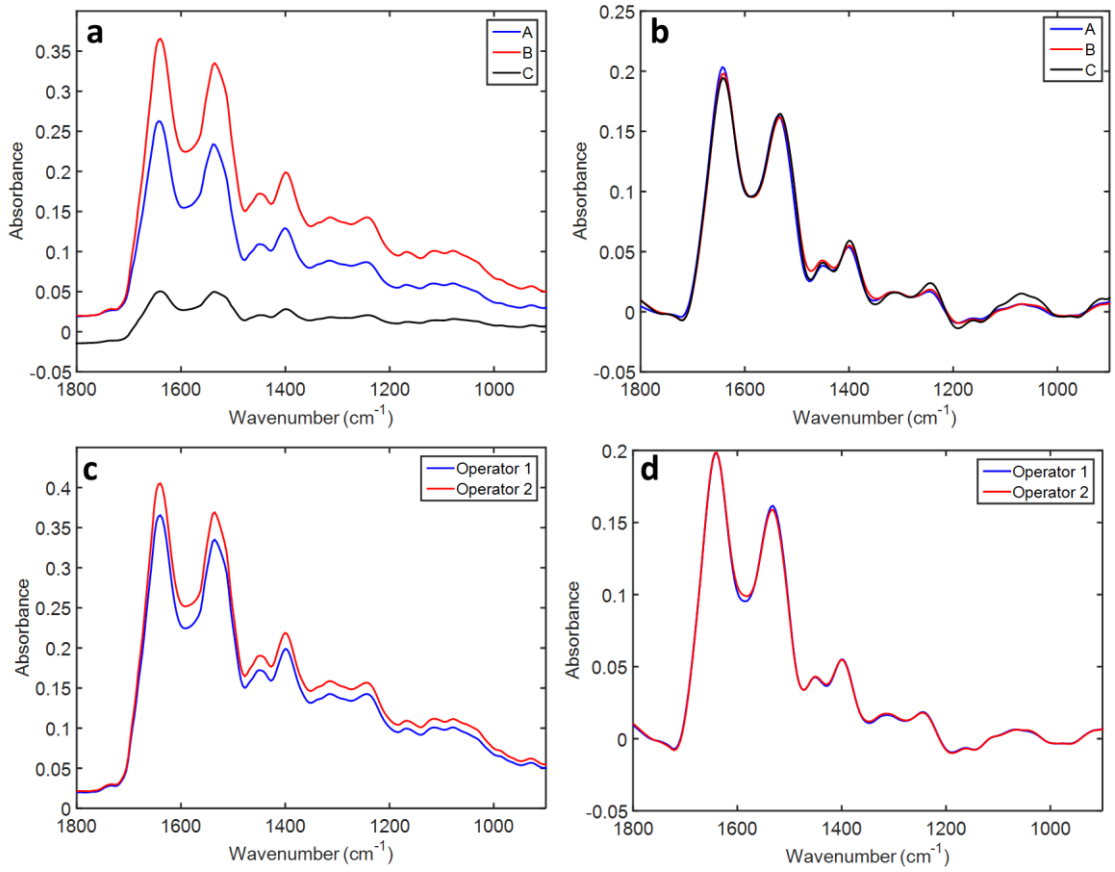


Figure 9.1. IR spectra of healthy control (absence of disease) samples varying ATR-FTIR instruments and operators. Average (a) raw and (b) pre-processed IR spectra for healthy control samples measured across three different ATR-FTIR spectrometers in the same institute (A, B and C). Average (c) raw and (d) pre-processed IR spectra for healthy control samples across two different operators (Operator 1 and 2).

Another important aspect of pre-processing is the order in which each step is applied. Pre-processing should be employed in a logical order so that the next pre-processing step is not affected by the previous one. For example, pure spectral differentiation cannot be employed before smoothing, since the spectral differentiation will increase the original noise. Therefore, smoothing should be applied before differentiation. Albeit, Savitzky-Golay routine incorporates smoothing and spectral differentiation so, in practical terms, these can be performed together. To summarise, the suggested order of pre-processing is as follows:

1. Spectral Truncation
2. Smoothing

3. Light scattering correction
4. Baseline correction
5. Normalization
6. Scaling

Further details about these pre-processing steps are provided in “Procedure: Data pre-processing” section. When using different instruments but same type of sample, the pre-processing steps should be the same for the data acquired under different circumstances.

9.2.4 Experimental Design: Data Analysis

Sample splitting. Sample splitting is fundamental for constructing a predictive chemometric model. It consists of a data analysis step performed before construction of a chemometric model, in which a portion of the samples are assigned to a training set, while the remaining samples are assigned to a validation and/or test set. The training set is used for model construction, the validation set for model optimization, and the test set for final model evaluation. The process of dividing the samples in three sets can be performed manually or by computer-based methodologies. Manual splitting can generate biased results, therefore we recommend a computational-based split instead. Some examples of these include random selection, leverage (Wang *et al.*, 1991) or the KS algorithm (Kennard & Stone, 1969). KS works based on Euclidian distance calculation by firstly assigning the sample with the maximum distance to all other samples to the calibration set, and then by selecting the samples which are as far away as possible from the selected samples to this set, until the designed number of selected samples is reached. This ensures that the calibration model will contain samples that uniformly cover the complete sample space, where no or minimal extrapolation of the remaining samples are necessary; avoiding problems of manual or random selection, such as non-reproducibility and non-representative selection. Usually, the dataset is split with 70% of the samples assigned for training, 15% for validation and 15% for test. In this case, the test set is dependent on the initial group of samples measured, and it is not a regular independent test set where a new set of similar samples are measured.

Exploratory analysis. Exploratory analysis is an important tool to provide an initial assessment of the data. Using exploratory analysis, the analyst can see the clustering patterns

and then draw conclusions related to the nature of samples, outliers and experimental errors. One of the most common techniques for exploratory analysis is principal component analysis (PCA), in which the original data are decomposed into a few principal components (PCs) responsible for most of the variance within the original dataset. The PCs are orthogonal to each other and are generated in a decreasing order of explained variance, so that the first PC represents most of the original data variance, followed by the second PC and so on (Bro & Smilde, 2014). Mathematically the decomposition takes the form:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{9.7}$$

where **X** represents the pre-processed data (*e.g.*, pre-processed samples' spectra); **T** are the scores; **P** are the loadings; and **E** are the residuals.

The PCA scores represent the variance in the sample direction and they are used to assess similarities/dissimilarities among the samples, thus detecting clustering patterns. The PCA loadings represent the variance in the variable (*e.g.*, wavenumber) direction and they are used to detect which variables show the highest importance for the pattern observed on the scores. The PCA loadings are commonly employed as a tool for searching spectral markers that distinguish different biological classes (Martin *et al.*, 2007). The PCA residuals represent the difference between the decomposed and original data and can be used to identify experimental errors. Ideally, the PCA residuals should be random and close to zero, representing a heteroscedastic distribution. Otherwise, they can indicate experimental bias according to a homoscedastic distribution.

For standardization applications, PCA is a fast, intuitive and reliable tool to observe if there are differences between the spectra acquired by different systems. Ideally, if the same sample is measured under different conditions (different laboratories, instrument manufacturers or user operators) their PCA scores should be random and completely superposed. If a discrimination pattern is observed on the PCA scores, then it is indicative that the data need standardization. Figure 9.2 illustrates a PCA scores plot from the same samples (blood plasma of healthy controls) measured using three IR instruments before (Fig. 9.2a) and after (Fig. 9.2b) PDS. Even though the samples in Fig. 9.2a are pre-processed, three different clusters are still evident. After PDS the samples measured using different systems are normalized into a single cluster.

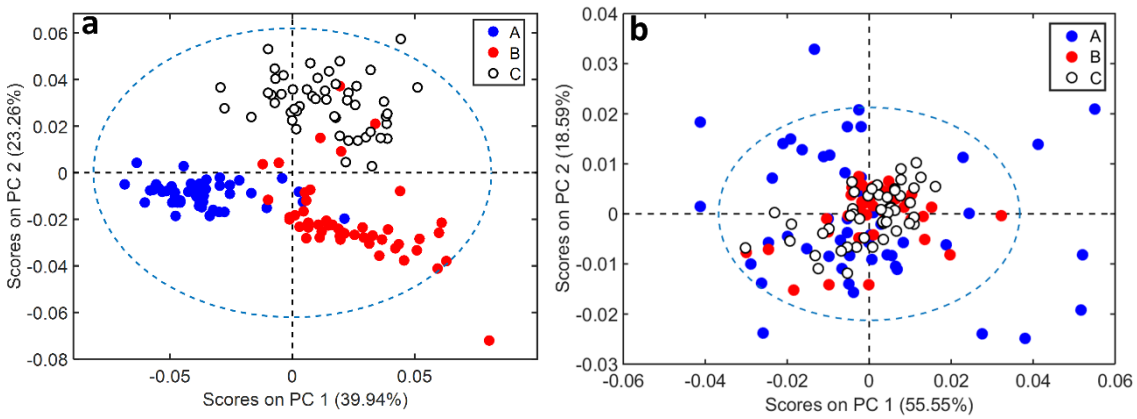


Figure 9.2. PCA scores for healthy control (absence of disease) samples varying ATR-FTIR instruments before and after standardization. (a) PCA scores for healthy control samples across three different ATR-FTIR spectrometers in the same institute (A, B and C) after pre-processing but before PDS; (b) PCA scores for healthy control samples across three different ATR-FTIR spectrometers in the same institute (A, B and C) after PDS (model built with 55 transfer samples and window size of 23 wavenumbers). The dotted blue circle shows 95 % confidence ellipse (two-sided). Each measurement observation (circle) corresponds to the data acquired from a unique operator.

Outlier detection. Outlier detection is important to prevent samples, which differ from the original dataset, from affecting the results using predictive models. Outliers can be attributed to experimental errors, such as inconsistent sample preparation or spectral acquisition, or to larger experimental noise, such as Johnson noise, shot noise, flicker noise and environmental noise. These samples can have large leverage for classification, masking the real signal from the samples of interest; therefore, it is advised that they be removed from the dataset used to train the predictive model.

To detect outliers, techniques such as Jack-knife (Martens & Martens, 2000), Z-score (Rousseeuw & Hubert, 2011) or *K*-modes clustering (Jiang *et al.*, 2016) can be utilised among others (Domingues *et al.*, 2018). One of the most popular and visually intuitive technique for detecting outliers is the Hotelling T^2 *vs* Q residual test (Bakeev, 2010). In this test, a chart is created using the Hotelling T^2 values in *x*-axis and the Q residuals in the *y*-axis, generating a scatter plot. The Hotelling T^2 represents the sum of the normalized squared scores, which is the distance from the multivariate mean to the projection of the sample onto the PCs (Kuligowski *et al.*, 2012). The Q residuals represent the sum of squares of each sample in the

error matrix, thus measuring the residues between a sample and its projection onto the PCs (Kuligowski *et al.*, 2012). All samples far from the origin of this graph are considered outliers and should be removed one at a time, as the PCA is highly influenced by the samples that are included in the model. Samples with high values in both Hotelling T^2 and Q residuals are the worst outliers; while samples with high values in only one of these axis are the second worst outliers. Appendix C illustrates an example for outlier detection. Squared confidence limits can be draw based on this graph; however, this can hinder outlier detection. For example, if the confidence limits is set at a 95% level, certain amount of data-points (5%) should be statistically outside these boundaries.

Classification. Classification techniques are employed for sample discrimination. Using chemometric analysis, one can distinguish classes of samples based on their spectral features and then make further predictions based on these. The prediction capability of a classification model should be evaluated with external samples (unknown samples) through the calculation of figures of merit, including accuracy (proportion of samples correctly classified considering true positives and true negatives), sensitivity (proportion of positives that are correctly identified) and specificity (proportion of negatives that are correctly identified) (Morais & Lima, 2017).

There are many types of classification techniques for spectral data. Table 9.4 summarizes the main classification techniques employed for bio-spectroscopy applications, along with their advantages and disadvantages.

When employing classification techniques, one must follow a parsimony order (Seasholtz & Kowalski, 1993), where the simplest algorithms should be used first, reducing the need for more complex algorithms which would require more optimization steps. An order for using these classification algorithms is: LDA>PLS-DA>QDA>KNN>SVM>ANN>Random forests>Deep learning approaches, from the simplest to the most complex.

Table 9.4. Classification techniques.

Classification Technique	Advantage	Disadvantage
Linear discriminant analysis (LDA) (Dixon & Brereton, 2009)	Simplicity, fast calculation	Needs data reduction, does not account for classes having different variance structures, greatly affected by classes having different sizes
Quadratic discriminant analysis (QDA) (Dixon & Brereton, 2009)	Fast calculation, accounts for classes having different variance structures, not much affected by classes having different sizes	Needs data reduction, higher risk of overfitting
Partial least squares discriminant analysis (PLS-DA) (Brereton & Lloyd, 2014)	Fast calculation, high accuracy	Greatly affected by classes having different sizes, needs optimization of the number of latent variables (LVs)
K-Nearest Neighbours (KNN) (Cover & Hart, 1967)	Simplicity, non-parametric, suitable for large datasets	Time consuming, needs optimization of the distance calculation method and *k* value, highly sensitive to the "curse of dimensionality"[199]
Support vector machines (SVM) (Cortes & Vapnik, 1995)	Non-linear classification nature, high accuracy	High complexity, high risk of overfitting, needs optimization of kernel function and SVM parameters, time consuming
Artificial neural networks (ANN) (Abraham, 2005)	Non-linear classification nature, ability to work with incomplete knowledge, high accuracy	High computational cost, needs optimization of the number of neurons and layers, no interpretability ("black box" model)
Random forests (Fawagreh *et al.*, 2014)	Non-linear classification nature, high accuracy, relatively low computational cost	High risk of overfitting, needs optimization of the number of trees, no interpretability ("black box" model)
Deep learning approaches (LeCun *et al.*, 2015)	Non-linear classification nature, native feature extraction (e.g., in convolutional neural networks (CNN)), local spatial coherence (CNN), high accuracy	High computational cost, needs hyperparameter optimization, needs large datasets, time consuming, no interpretability ("black box" model)

Classification algorithms can be coupled to feature extraction and feature selection techniques in order to reduce data collinearity/redundancy, thus reducing the risk of overfitting in the classifier training, and speeding up such training, as there are less variables involved. An additional benefit of such a feature extraction/selection step is to provide spectral markers identification as a "side-effect" (depending on the feature extraction/selection method applied). For feature extraction, the most popular technique is PCA. In this case, a PCA is firstly applied to the data, and then the PCA scores are used as the input variables (instead of the wavenumbers data points) for the classification techniques mentioned above (Morais & Lima, 2017). PLS-DA is also a feature extraction technique (Brereton & Lloyd, 2014), and normally it performs better than a PCA followed by LDA, as the scores from a PCA does not necessarily describe the difference between the samples, but rather the variance in the data. In PLS-DA, a partial least squares (PLS) model is applied to

the data in an interactive process reducing the original variables to a few number of LVs, where a LDA is used for classifying the groups (Hibbert, 2016). Other discriminant classifiers, in particular QDA, also could be used in this classification step to circumvent problems observed with LDA. For feature selection, there are many techniques commonly employed in biological datasets, including genetic algorithm (GA) (McCall, 2005) and successive projections algorithm (SPA) (Soares *et al.*, 2013). The variables (*e.g.*, wavenumbers) selected by these techniques are used as input variables for the classification models described in Table 9.2. An important advantage of GA is its relatively low-computational cost compared to SPA and reduction of data collinearity. Furthermore, GA-based techniques are intuitive and simple to understand in the algorithmic sense but they also have a non-deterministic nature and require optimization of many parameters. SPA's advantage relies on its deterministic nature, minor parameter optimization and reduction of data collinearity, however, it is very time consuming. For hyperspectral imaging, feature selection can also be performed by Minimum Redundancy Maximum Relevance (MRMR) algorithm (Kamandar & Ghassemian, 2010), where the selection process is based on maximizing the relevance of extracted features and simultaneously minimize redundancy between them.

Standardization. Data standardization should be employed when a primary classification model is built and new data comes to be predicted from a secondary system (different laboratory or instrument manufacturers), or when there is a change in instrument components (*e.g.*, laser, gratings, etc.) or when the data of the chemometric model are acquired under different circumstances (different analysts, days, instrumental settings, etc.). As previously mentioned, the most common and reliable methods for data standardization are the DS and PDS algorithms. These methods can be found in a few software packages (described in Table 9.3).

Figure 9.3 summarises the standardization protocol using DS applied to spectra acquired under different conditions. The first step consists of applying KS algorithm for selecting the number of transfer samples from the primary system as well as the number of training samples for the secondary systems, which is ideally 70% of the dataset. Thereafter, the DS transform generation algorithm is employed to estimate the transform matrix. The validation set of the secondary system is then used with the classification model of the

primary system to evaluate the optimum number of transfer samples. This optimization step is repeated depending on the number of transfer samples from the primary system. After this number is defined, the validation set of the secondary system is finally standardized and the final classification model is subsequently applied. This procedure is realized with a certain number of samples measured in all instruments being standardized. This procedure should be realized in as similar manner as possible to reduce spectral differences. After the model is standardized and proper validated, new external samples can be measured in any of the instruments and predicted by the standardized classification model.

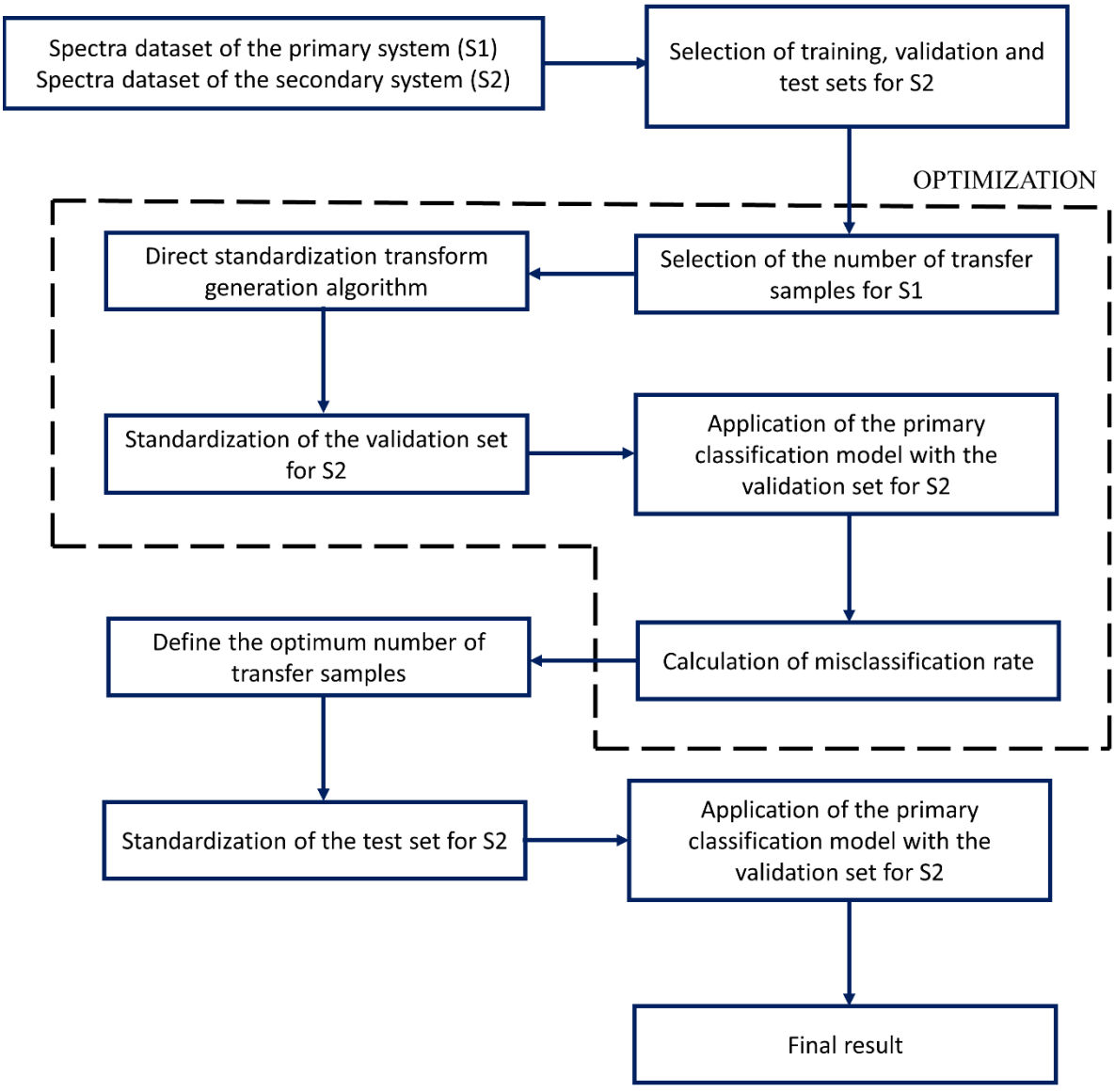


Figure 9.3. Flowchart for standardization using Direct Standardization (DS).

For PDS, an extra step is added after defining the number of transfer samples to estimate the optimum window size. The dashed region in Fig. 9.3 is repeated according to the window size.

For multi-laboratory studies the flowchart depicted in Fig. 9.4 illustrates how the standardization protocol should be employed.

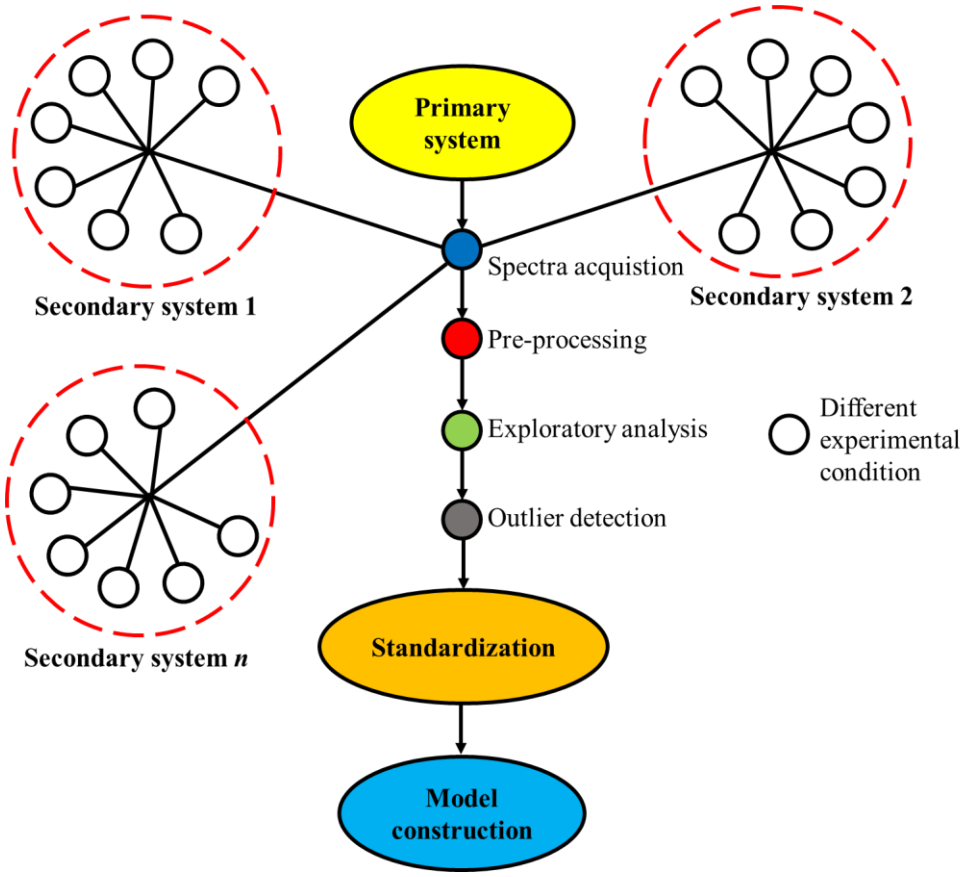


Figure 9.4. Flowchart for a standardization protocol using different experimental conditions.

In Fig. 9.4, spectra acquired under different experimental conditions are used for a global standardization model. A primary system should be designated and then all spectra from secondary systems are equally pre-processed, followed by an exploratory analysis to assess samples' similarities/dissimilarities, outlier detection, standardization by the method outlined in Figure 9.3; the final model construction follows last. With this, all sources of variations present in different systems can be included into a general chemometric model.

9.3 Materials

9.3.1 Reagents

- Biological samples (tissue, cells, biofluids)(see Reagent Setup).

▲ **CRITICAL** Human samples should be collected with appropriate local institutional review board for ethical approval and adhere to the Declaration of Helsinki principles. Similarly, for studies involving animals, all experiments should be performed in accordance with relevant guidelines and regulations. Ethical approval has to be obtained before any sample collection.

- Optimal cutting temperature (OCT) compound (Agar Scientific, cat. no. AGR1180)
- Liquid nitrogen (BOC, CAS no. 7727-37-9) **! CAUTION** Asphyxiation hazard; make sure room is well ventilated. Causes burns; wear face shield, gloves and protective clothing.
- Paraplast Plus paraffin wax (Thermo Fisher Scientific, cat. no. SKU502004)
- Isopentane (Fisher Scientific, cat. no. P/1030/08) **! CAUTION** Extremely flammable, irritant, aspiration hazard and toxic; use in a fume hood.
- Distilled water
- PBS (10×; MP Biomedicals, cat. no. 0919610)
- Virkon (Antec, DuPont, cat. no. A00960632)
- Trypsin–EDTA (0.05%, Sigma-Aldrich, Thermo Fisher Scientific cat. no. 25300054)

Anticoagulants

- EDTA (Thermo Fisher Scientific, BD Vacutainer, cat. no. 02-687-107)
- Sodium citrate (Thermo Fisher Scientific, BD Vacutainer)
- Lithium/sodium heparin (Thermo Fisher Scientific, BD Vacutainer)

Fixative and preservative agents

- Formalin, 10% (vol/vol; Sigma-Aldrich, cat. no. HT501128) **! CAUTION** Potential carcinogen, irritant and allergenic; use in a fume hood.
- Ethanol (Fisher Scientific, cat. no. E/0600DF/17)

- Methanol (Fisher Scientific, cat. no. A456-212) **! CAUTION** Toxic vapours; use in a fume hood.
- Acetone (Fisher Scientific, cat. no. A19-1) **! CAUTION** Acetone vapors may cause dizziness; use in a fume hood.
- ThinPrep (PreservCyt Solution, Cytyc Corp)
- SurePath (Becton Dickinson Diagnostics)

Dewaxing agents

- Xylene (Sigma-Aldrich, cat. no. 534056) **! CAUTION** Potential carcinogen, irritant and allergenic; use in a fume hood.
- Histo-Clear (Fisher Scientific, cat. no. HIS-010-010S) **! CAUTION** It is an irritant.
- Hexane (Fisher Scientific, cat. no. 10764371) **! CAUTION** Extremely flammable liquid, can cause skin irritation; use protective equipment as required; use in a fume hood.

9.3.2 Equipment

- Microtome (Thermo Fisher Scientific, cat. no. 902100A; or cat. no. 956651)
- Wax dispenser (Electrothermal, cat. no. MH8523B)
- Sectioning bath (Electrothermal, cat. no. MH8517)
- Centrifuge (Thermo Fisher Scientific, cat. no. 75002410)
- Desiccator (Thermo Fisher Scientific, cat. no. 5311-0250)
- Desiccant (Sigma-Aldrich, cat. no. 13767)
- Laser power meter (Coherent, cat. no. 1098293)
- Spectrometer
- Computer system

Substrates

▲ **CRITICAL** Substrate should be carefully chosen depending on the spectrochemical approach and the experimental mode that will be used. For more details about the choice of substrate see ref (Baker *et al.*, 2014; Butler *et al.*, 2016).

- Low-E slides (Kevley Technologies, CFR)
- BaF_2 slides (Photox Optical Systems)
- CaF_2 slides (Crystran, cat. no. CAFP10-10-1)
- Silicon multi-well plate (Bruker Optics)
- Glass slides (Fisher Scientific, cat. no. 12657956)
- Quartz slides (UQG Optics, cat. no. FQM-2521)
- Aluminum-coated slides (EMF, cat. no. AL134)
- Mirrored stainless steel (Renishaw, cat. no. A-9859-1825-01)

9.3.3 Reagent Setup

Tissue. For FFPE tissue, the excised specimen is immersed in fixative (*e.g.*, formalin), dehydrated in ethanol, cleared in xylene and embedded in paraffin wax. Specimens can then be stored indefinitely at room temperature. For snap-frozen tissue, the specimen is immersed in OCT, followed by cooling of isopentane with liquid N_2.

▲ **CRITICAL** Snap-frozen tissue should be thawed before analysis. Spectroscopic analysis should be performed directly after excision in case of fresh tissue to avoid sample degradation.

Cells. Cells can be treated with a suitable fixative or preservative solution or studied alive.

▲ **CRITICAL** In case cells are fixed or stored in a preservative solution, a number of washing steps using centrifugation should be followed prior to spectroscopic analysis to remove unwanted signature. If cells are studied alive, optimum living conditions (*e.g.*, growth medium, temperature and pH) should be maintained; washing of live cells from medium is also necessary.

Biofluids. Biofluids can be collected in designated, sterile tubes using standard operating procedures to achieve uniformity of performance. Preparation of biofluids depends on the sample type and the experiment's objective. If cellular material is not directly studied, it should be removed from the biofluid before storage. Biofluids can be analysed right after their collection or stored at a -80°C freezer.

▲ **CRITICAL** If biofluids have been stored in a freezer, it is essential that they are fully thawed before acquiring aliquots for spectroscopic analysis.

▲ **CRITICAL** Users are advised to store biofluids in smaller, single-use aliquots at -80°C to avoid repeated freeze-thaw cycles.

9.3.4 Equipment Setup

The user can choose from a range of different instrumental setups and spectral acquisition modes. General information about FTIR systems is provided below. For more details about equipment setup see refs. (Baker *et al.*, 2014; Butler *et al*, 2016; Martin *et al.*, 2010).

The FTIR spectrometer can be left on for long periods of time. Before spectral acquisition, the user should check the interferogram signal for amplitude and position and keep a record of the measurements.

▲ **CRITICAL** For detectors that require a prior cooling step using liquid nitrogen (*e.g.*, mercury cadmium telluride (MCT) detectors), the signal should be allowed to stabilize for approximately 10 min before data collection.

▲ **CRITICAL** In case that the interferogram signal deviates from the last measurement, re-alignment or part replacement may be required.

Software. Software for spectral acquisition is typically provided by the manufacturer. Software packages for spectral analysis and data standardization are provided in Table 9.3.

9.4 Procedure

Sample preparation

1| Prepare the biological samples for spectrochemical analysis using the following steps: option A for FFPE tissue samples, option B for snap-frozen or fresh tissue samples, option C for cells and option D for biofluids.

▲ **CRITICAL** Sample preparation is briefly presented in this protocol. More details about sample preparation can be found in refs. (Baker *et al.*, 2014; Butler *et al*, 2016; Martin *et al.*, 2010).

(A) Tissue (FFPE) ● TIMING 1-1.5 h

(i) Obtain FFPE tissue blocks.

(ii) Section the whole tissue block using a microtome to obtain tissue sections at desired thickness (2-10 μm).

▲ **CRITICAL STEP** Cooling of the tissue on an ice block for 10 min prior to sectioning, hardens the wax and allows easier cutting.

(iii) Float the tissue ribbons in a warm H_2O bath (40-44°C) and then deposit onto the substrate of choice.

(iv) Allow the tissue sections to dry either at room temperature (30 min) or in a 60°C oven (10 min).

▲ **CRITICAL STEP** The tissue slide may be dried in the oven for longer periods of time, depending on the type of tissue, to ensure optimal, initial melting of the wax.

(v) Dewax the samples by performing three sequential immersions in a dewaxing reagent such as fresh xylene, Histo-Clear solution or hexane (each immersion should last at least 5 min).

▲ **CRITICAL STEP** Thorough dewaxing is important for eliminating all spectral peaks attributed to paraffin.

(vi) Immerse the tissue slide in acetone or ethanol (5 min) to remove the xylene and then left to air-dry.

■ **PAUSE POINT** Slides can be stored in a desiccator at room temperature for at least 1 year.

(B) Tissue (Snap-frozen or fresh) ● TIMING 2 h + drying time (3 h for FTIR only)

▲ **CRITICAL** Snap-frozen tissue can be stored at -80°C for several months.

▲ **CRITICAL** For fresh tissue, proceed to step 1B(ii).

(i) Acquire snap-frozen tissue from freezer and place onto a cryostat (30 min) to allow the tissue to reach the cryostat's temperature (-20°C).

(ii) Use a cryostat to obtain tissue sections at desired thickness (8-10 μm).

(iii) Deposit the tissue sections onto an appropriate substrate before spectra are collected (see a list of substrates in the Materials-Equipment section).

▲ **CRITICAL** For FTIR studies the tissue sections need to dry for at least 3 h to remove the H_2O interference from the IR spectra.

▲ **CRITICAL** Exposure to light should be minimised to prevent sample degradation due to oxidation.

(C) Cells (fixed or live) ● TIMING 30 min + desiccation time (3 h for FTIR only)

▲ **CRITICAL** If you are working with fixed cells, do step 1C(i) and then proceed to step 1C(iii). If you are working with live cells, proceed to step 1C(ii)

(i) Wash fixed cells to remove the fixative or preservative solution as these chemicals cause spectral interference in the fingerprint region. Three sequential washes with distilled H_2O or PBS have been shown to remove unwanted peaks.

(ii) Detach cultured cells from the growth substrate adding 2-3 mL of fresh warm trypsin/EDTA solution to the side wall of the flask; gently swirl the contents to cover the cell layer. Wash with warmed sterile PBS to remove the medium and trypsin (×3 times; gentle centrifuge at 300 *g* for 7 min).

▲ **CRITICAL STEP** All reagents should be warmed to 37°C to reduce the shock to cells and maintain morphology.

(iii) After the final wash, resuspend the remaining cell pellet in distilled H_2O (~50-100 μL) and mount onto a substrate of choice; allow sample to dry before analysis.

▲ **CRITICAL STEP** The final suspension of cells (~50-100 μL) should be evenly deposited on the slide either by cytospinning or by micro-pipetting. For cytospinning, take a maximum volume of 200 μL of cells in suspension (spin-fixed cells at 800 *g* for 5 min). After spinning, leave the slide to air-dry.

▲ **CRITICAL** For FTIR studies the sample needs to dry for at least 3 h.

(D) Biofluids (frozen or fresh) ● TIMING 5 min + thawing (20 min) + drying (1-1.5 h)

▲ **CRITICAL** If biofluids are analysed fresh, immediately after collection, continue to step 1D(ii).

(i) Acquire biofluids from the -80°C freezer and allow them to fully thaw.

(ii) Mix or gently vortex the sample before obtaining the desired volume for analysis.

▲ **CRITICAL STEP** Only a small amount of the biofluid is typically required for spectroscopic studies (1-100 μL). However, this depends and should be tailored according to the study and experimental design. For instance, in case a substrate is used for experiments in the ATR mode, a larger volume is preferred as it allows spectral acquisition from multiple locations of the blood spot. On the contrary, if no substrate is used, such as in the case of the direct deposition of the sample on the ATR crystal, smaller volumes can also be used.

(iii) Deposit the biological fluid onto an appropriate substrate.

▲ **CRITICAL STEP** For ATR-FTIR spectroscopic studies, an alternative option is to deposit the sample directly on the ATR crystal instead of a substrate if the instrumentation setting allows (*i.e.*, if crystal is facing upwards). However, if the sample is sufficiently thick (>2-3 μm) to avoid substrate interference, then the use of a holding substrate is advantageous as it allows measurements from multiple locations as well as longer storage.

▲ **CRITICAL STEP** For FTIR studies the sample needs to dry adequately before spectroscopic analysis (50 μL dry within approximately 1 h at room temperature). Drying can be sped up by using a gentle stream of air over the sample at a specific flow rate (in a sterile laminal flow hood).

Spectral acquisition for FTIR spectroscopy ● TIMING 2 - 5 min per spectrum

▲ **CRITICAL** Spectrochemical information can be collected as follows for FTIR spectroscopy.

▲ **CRITICAL** Spectral acquisition is briefly presented in this protocol. More details can be found in refs. (Baker *et al.*, 2014; Butler *et al*, 2016; Martin *et al.*, 2010).

2 | Optimise the settings before each new study to increase the SNR (see 'Experimental design: spectral acquisition').

▲ **CRITICAL STEP** Some of the parameters that need to be adjusted include the resolution, spectral range, co-additions, aperture size, interferometer mirror velocity, and interferogram zero-filling.

▲ **CRITICAL STEP** To improve reproducibility and decrease differences between the data collected by different operators, the spectral resolution should be set constant, since it can cause major differences between data collected across different experimental setups.

▲ **CRITICAL STEP** The pressure applied on the sample in the ATR mode affects the signal intensity (*i.e*., absorbance) between data collected by different instruments and operators. Thus, the pressure applied on the sample should be as similar as possible across different experimental setups to reduce differences between the spectra collected. Depending on the sampling mode that has been chosen (ATR-FTIR, transmission or transflection), deposit the sample onto the appropriate holding substrate.

3 | Acquire a background spectrum to account for atmospheric changes.

▲ **CRITICAL STEP** This should be done before every sample.

4 | Load the sample and visualise the region of interest; information can then be acquired either as point map or as image maps.

▲ **CRITICAL** Typically, 5-25 point spectra are collected per sample while for image maps the step size should be the same or smaller than the selected aperture size divided by two. Sampling can be performed with 6 replicates in 3 levels.

■ **PAUSE POINT** Save the acquired data in a database until further analysis.

Data quality evaluation ● TIMING 15 min – 4 h (depending on the size of the dataset)

5 | Evaluate the raw data using quality tests to identify anomalous spectra or biased patterns before applying pre-processing. This can be made by visual inspection of the collected spectra followed by Hotelling T^2 *versus* Q residuals charts (see Appendix C) using only the mean-centred data, and analysis of PCA residuals. Samples far from the origin of the Hotelling T^2 *versus* Q residuals chart should be removed, and PCA residuals should be random and close to zero. Further instructions about data quality evaluation can be found at "Experimental Design: data quality evaluation" section.

Data pre-processing ● TIMING 15 min – 4 h (depending on the size of the dataset)

▲ **CRITICAL** Steps 6-11 below can be modified depending on the nature of the dataset. Table 9.1 provides more details about these pre-processing steps. In case of an ATR-FTIR dataset where samples were acquired and analysed under different experimental conditions, the pre-processing method should follow this order:

6 | **Cutting at biofingerprint region (900-1800 cm^{-1}).** Truncate the spectra to the biofingerprint region, to eliminate atmospheric interference present in other regions of the spectra.

7 | **Savitzky-Golay smoothing for removing spectral-noise.** Window size varies according to the size of the spectra dataset (*e.g.*, wavenumber). The window size should be an odd number, since a central data point is required for the smoothing process. Try different window sizes from 3 to 21 and observe how the spectra change (in shape) and how the noise is reduced. Use the smallest window that removes the noise considerably whilst maintaining the original spectral shape. Using a spectral resolution of 4 cm^{-1}, the biofingerprint region (900-1800 cm^{-1}) usually contains 235 wavenumbers. In that case, a window size of 5 points should be used. The polynomial

order for Savitzky-Golay fitting should be 2nd order for IR spectroscopy due to the band shape.

8 | **Light scattering correction using either multiplicative scatter correction (MSC), SNV or 2nd derivative.** First try using MSC or SNV, as MSC maintains the spectral scale and both methods maintain the original spectral shape. If the results are not satisfactory (*e.g.*, classification accuracy < 75%), try using the 2nd derivative spectra.

9 | **Perform baseline correction using automatic weighted least squares or rubber band baseline correction.** If spectral differentiation is applied as light scattering correction method, baseline correction is not necessary.

10 |**Normalization** Normalize the spectrum to the amide I peak or amide II peak, or perform a vector normalization (2-Norm, length = 1) to correct different scales across spectra (*e.g.*, due to different sample thicknesses when using FTIR in transmission mode).

11 |**Scaling** Mean-centre the data for each variable, and divide this value by the variable standard deviation. In case of data fusion, block-scaling should be used.

Data analysis

Exploratory analysis. ● TIMING 1h – 4 d (depending on the data size)

12 |Determine whether a standardisation procedure is necessary by performing PCA. The PCA scores plot (PC1 *vs* PC2) should generate a unique clustering pattern for the same type of sample. If two or more clusters are observed for the same type of sample measured under different experimental conditions, then a standardisation procedure is necessary (see Figure 9.2).

Outlier detection. ● TIMING 1h – 1 d (depending on the data size)

13 |Apply PCA to the dataset and then estimate the Q residuals and Hotelling T^2 values. Use the chart of Q residuals *versus* Hotelling T^2 to identify outliers. The outliers (*e.g.*, cosmic rays, artefacts, low signal spectra and substrate only (non-tissue) spectra) should be removed from the data set before proceeding to the next steps.

Sample split. ● TIMING 1 – 4 h (depending on the data size)

14 |Separate the samples that will be used for the training and the test sets. Sample split should be performed before construction of standardization of multivariate

classification models. The samples can be split into training (70%) and test (30%) sets, using a cross-validated model; or split into training (70%), validation (15%) and test (15%) sets without using cross-validation. To maintain consistency and account for a well-balanced training model, KS algorithm should be employed to separate the samples into each set. KS algorithm is freely available at https://doi.org/10.6084/m9.figshare.7607420.v1.

Standardization. ● TIMING 1h – 4 d (depending on the data size)

▲ **CRITICAL** Standardization methods should be employed in the following order: DS > PDS (DS should be done before PDS), since the latter is more complex and requires an additional optimization step (window size optimization). The data from the secondary response should be separated into training (70%), validation (15%) and test (15%) sets using KS algorithm. The number of transfer samples should be firstly optimized using the validation set from the secondary response. Then, when employing PDS, the window size should be optimized according to the size of the dataset.

15 |Use DS to vary the number of transfer samples from 10-100% of the training set from the primary system. Use the validation set from the secondary instrument to find the optimum number of transfer samples using the misclassification rate as cost function.

16 |Perform PDS using the optimum number of samples found with DS. Test different window sizes using the validation set from the secondary system with the misclassification rate as cost function. The window size should vary from 3-29 for a spectral set with resolution of 4 cm^{-1} in the biofingerprint region (235 variables).

Model construction. ● TIMING 1h – 4 d (depending on the data size)

▲ **CRITICAL** Feature extraction (*e.g.*, by means of PCA) or feature selection (*e.g.*, by means of GA or SPA) should be employed to reduce data collinearity and speed up data processing and analysis time. PLS-DA is already a feature extraction method, thus the performance of prior feature extraction is not necessary in this case. The classification technique employed must follow a parsimony order: LDA>PLS-DA>QDA>KNN>SVM>ANN>Random forests>Deep learning approaches.

17 |Apply the feature extraction or selection technique. The optimization of the number of PCs during PCA can be performed using an external validation set (15% of the

original dataset) or using cross-validation (leave-one-out for small dataset [ppl samples] or venetian blinds [sample splitting: 10] for large datasets [>20 samples]). GA should be realized three-times starting from different initial populations and the best result using an external validation set (15% of the original dataset) should be used. Cross-over probability should be set for 40% and mutation probability should be set for 1-10% according to the size of the dataset.

18 |The classification method should be employed using optimization with an external validation set or cross-validation, especially for selecting the number of latent variables of PLS-DA and the kernel parameters for SVM. The kernel function for SVM should be RBF kernel, due to its adaptation to different data distributions. To avoid overfitting, cross-validation should be always performed during model construction to estimate the best RBF parameters.

9.5 Troubleshooting

Spectral acquisition: Spectral resolution, spectral range, SNR and signal aperture should be optimized during experimental setup. Operators using different systems should try to keep these parameters constant to reduce spectral differences.

Data pre-processing: To reduce spectral differences, the same data pre-processing should be applied for spectra acquired in different systems.

Standardization: To improve the prediction capability of the classification model, the primary system used should be the one with highest spectral resolution and smallest noise, since all data from the secondary systems will be standardized to this pattern.

9.6 Timing

Sample preparation:

Step 1(A) Tissue (FFPE): 1-1.5 h

1(B) Tissue (Snap-frozen or fresh): 2 h + drying time (3 h)

1(C) Cells (fixed or live): 30 min + desiccation time (3 h)

1(D) Biofluids (frozen or fresh): 5 min + thawing (20 min) + drying (1-1.5 h)

Steps 2-4, Spectral acquisition: 1 s – 5 min per spectrum (depending on the instrument and spectral acquisition configurations)

Step 5, Data quality evaluation: **15 min – 4 h (depending on the size of the dataset)**

Steps 6-11, Data pre-processing: 15 min – 4 h

Data analysis:

Step 12, Exploratory analysis: 1 h – 4 d

Step 13, Outlier detection: 1 h – 1 d

Step 14, Sample split: 1- 4h (depending on sample size)

Step 15-16, Standardization: 1 h – 4 d

Step 17-18, Model construction: 1 h – 4 d

9.7 Anticipated Results

To illustrate how this protocol can be used in practice, we conducted a pilot study to evaluate the effect of different instrument manufacturers and operators towards spectral acquisition of healthy controls and ovarian cancer samples based on blood plasma (5 healthy controls with 10 spectra per sample; 5 ovarian cancers with 10 spectra per sample) for a binary classification model using ATR-FTIR spectroscopy. All specimens were collected with ethical approval obtained at Royal Preston Hospital UK (16/EE/0010). Table 9.5 summarizes the experimental conditions in which the experiments were performed.

Instrument A and B were Bruker Tensor 27 with an HELIOS ATR attachment while instrument C was an ATR-FTIR Thermo Scientific Nicolet iS10. The spectra were collected for the same types of samples within three different days (operator 1: instrument A in day 1, instrument B in day 3, and instrument C in day 2; operator 2: instrument A in day 2,

instrument B in day 1, and instrument C in day 3) and across two different laboratories (instrument A and B in laboratory 1 and instrument C in laboratory 2). Each operator prepared the samples individually from the same bulk, and measured them individually. Spectral acquisition times were around 30 s for instruments A and B, and 40 s for instrument C.

Table 9.5. Experimental conditions for pilot study.

Instrument	Operator	Spectral range	Number of co-additions	Spectral resolution	Room temperature	Air humidity
A	1	4000-400 cm^{-1}	32	4 cm^{-1}	23.0ºC	23%
	2	4000-400 cm^{-1}	32	4 cm^{-1}	23.4ºC	26%
B	1	4000-400 cm^{-1}	32	4 cm^{-1}	24.0ºC	26%
	2	4000-400 cm^{-1}	32	4 cm^{-1}	24.9ºC	24%
C	1	4000-400 cm^{-1}	48	4 cm^{-1}	22.5ºC	28%
	2	4000-400 cm^{-1}	48	1 cm^{-1}	22.8ºC	26%

9.7.1 Effect of Different Instruments

Three different ATR-FTIR spectrometers were used to analyse the samples. Data were pre-processed by truncating at the biological fingerprint region (900-1800 cm^{-1}), followed by Savitzky-Golay smoothing (window of 15 points, 2nd order polynomial function), MSC, baseline correction using automatic weighted least squares and vector normalization (2-Norm, length = 1). Each data set (A, B and C) was pre-processed individually. The raw and pre-processed spectra for healthy controls and ovarian cancer samples are depicted in Appendix C, Figure C1.1. All spectra collected by the three instrument maintained the same spectral shape, indicating that the chemical information stayed the same; however, large differences between the absorbance intensity were observed between instrument C and the others (A, B), being caused due to different pressures applied on the sample in the ATR module. The pressure applied to keep the sample in contact with the ATR crystal directly affects the spectral signal intensity, which for instrument A and B (same manufactures) were somewhere controlled by a contra weight, while for instrument C the pressure was set based on a mechanical screw on the device, thus being biased by the operator usage. The absorbance intensity variation between A and B is observed for this same reason, but in a minor scale. Outlier detection was performed using a Hotelling T^2 *versus* Q residual test (Appendix C, Figure C1.2).

(i) Classification. Classification was performed using PCA-LDA (10 PCs, explained variance of 99.21%). Fig. 9.5a depicts the discriminant function (DF) score plot for PCA-LDA using only the primary system (ATR-FTIR A). As observed, there is an almost perfect separation between the samples from the two classes (accuracy = 100%, sensitivity = 100%, specificity = 100%). However, when the spectra acquired using instruments B and C are predicted using the model for A, the results decreased significantly (accuracy = 66.7%, sensitivity = 83.2%, specificity = 48.9%) (Fig. 9.5b), necessitating the use of a standardization procedure.

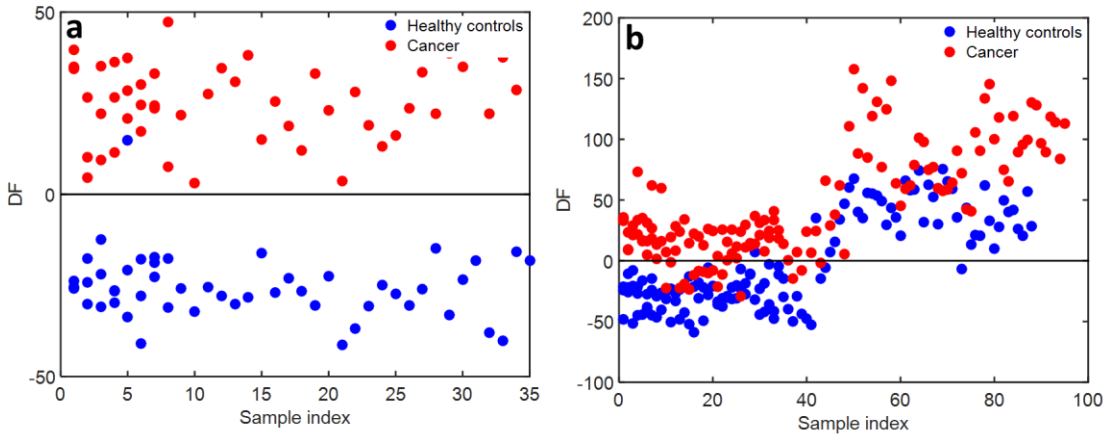


Figure 9.5. Discriminant function (DF) plots using PCA-LDA to discriminate healthy control (absence of disease) samples from ovarian cancer samples varying the instrument. (a) DF plot of the PCA-LDA model for the primary system; (b) DF plot of the PCA-LDA model for the primary system predicting the samples from the secondary systems. Sample index represents the number of samples' spectra.

(i) Standardization. Standardization was employed using both DS and PDS in order to compare the two methods. The number of transfer samples for DS was optimized according to the misclassification rate obtained for the validation set using the secondary system (Fig. 9.6a). An optimum number corresponding to 80% of the samples in the training set of the primary system (55 transfer samples) was obtained, resulting to a misclassification rate of 22.2% in the validation set of the secondary system. This improved the accuracy (77.8%) and specificity (80.0%). Sensitivity decreased to 75.0%, which is an acceptable value. The results after DS are better balanced than without standardization. Fig. 9.6b shows the DF plot for the

PCA-LDA model using the training of the primary system and prediction with the secondary system after DS.

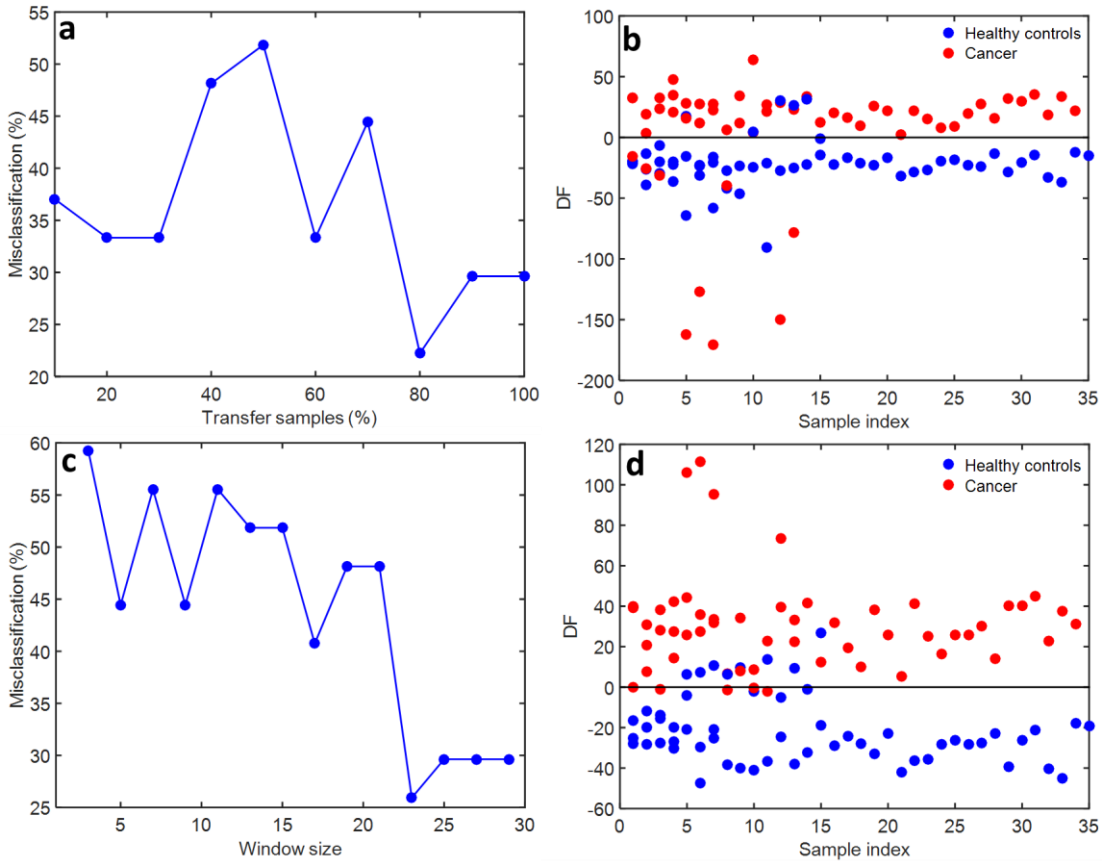


Figure 9.6. PCA-LDA results for DS and PDS standardisation models for spectra collected by the three different instruments. (a) Misclassification rate in % for the validation set of the secondary system varying the number of transfer samples in % from the primary system for DS optimization; (b) DF plot of the PCA-LDA model for the primary system predicting the validation set from the secondary system after DS; (c) Misclassification rate in % for the validation set of the secondary system varying the window size for PDS optimization; (d) DF plot of the PCA-LDA model for the primary system predicting the validation set from the secondary system after PDS. Transfer samples (%) refer to the percentage of training samples' spectra from the primary instrument that are used to transform the signal obtained using the secondary instrument.

PDS was also applied. The number of transfer samples was maintained as 55 (80% of the primary training set) and the window size was optimized by using the validation set of the secondary system. An optimum window size of 23 wavenumbers was selected with a misclassification rate of 25.9% (Fig. 9.6c). The accuracy, sensitivity and specificity using

PDS were 74.1%, 71.4% and 75.0%, respectively. The DS presented a slightly higher performance than PDS for this dataset. However, DS generated some outliers not observed before, while PDS did not. Thus, in general, PDS provided a better standardization of the data. The PCA-LDA DF plot after PDS is depicted in Fig. 9.6d.

9.7.2 Effect of Different Operators

The effect of different user operators acquiring spectra from the same samples using the same instruments was also evaluated. Similarly to before, data were pre-processed by cutting the biological fingerprint region (900-1800 cm^{-1}), followed by Savitzky-Golay smoothing (window of 15 points, 2^{nd} order polynomial function), MSC, baseline correction using automatic weighted least squares and vector normalization (2-Norm, length = 1). Each dataset was pre-processed individually. All raw and pre-processed spectra varying operators are depicted in Figures C1.4 and C1.5 (Appendix C). Outlier detection was performed using a Hotelling T^2 *versus* Q residual test (Figure C1.7, Appendix C). The PCA scores plots for the pre-processed spectra are depicted in Figure C1.6, Appendix C. The main difference between the operators was observed for instrument C (Figure C1.5, Appendix C), since the spectral resolutions used by them were different, which can cause major data distortion.

(i) Classification. Classification was performed using PCA-LDA (10 PCs, explained variance of 98.62%). Fig. 9.7a depicts the DF score plot for PCA-LDA using only the primary system (Operator 1). There is a significant separation between the samples from the two classes (accuracy = 88.4%, sensitivity = 77.3%, specificity = 100%). When the spectra acquired by Operator 2 are predicted using the model for Operator 1, the results decreased (accuracy = 75.6%, sensitivity = 66.7%, specificity = 84.6%) (Fig. 9.7b), which again necessitates the use of a standardization procedure.

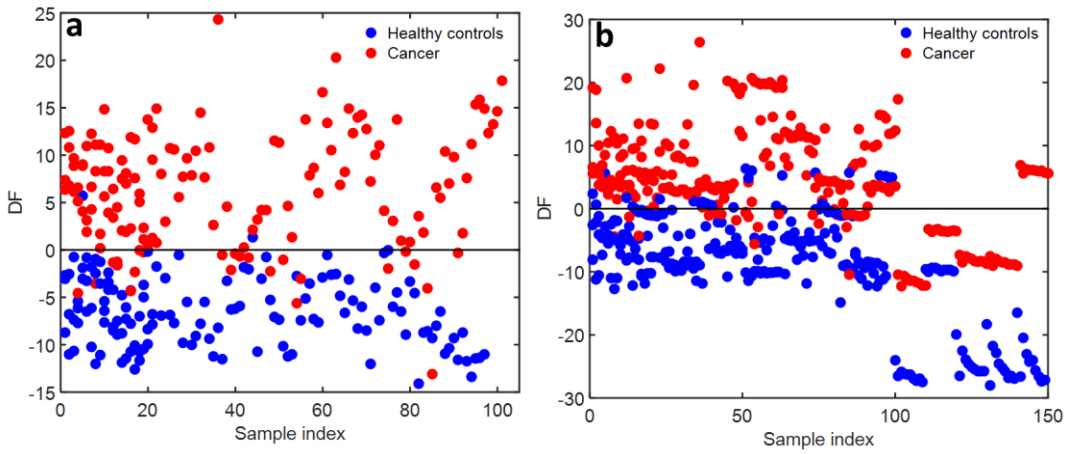


Figure 9.7. Discriminant function (DF) plots using PCA-LDA to discriminate healthy control (absence of disease) samples from ovarian cancer samples varying the operator. (a) DF plot of the PCA-LDA model for the primary system (Operator 1); (b) DF plot of the PCA-LDA model for the primary system predicting the samples from the secondary system (Operator 2).

(i) Standardization. DS and PDS were employed as standardization methods. The number of transfer samples for DS was optimized according to the misclassification rate obtained for the validation set using the secondary system (Operator 2) (Fig. 9.8a). An optimum number of 59 transfer samples (30% of the samples in the training set of the primary system [Operator 1]) was obtained, resulting in a misclassification rate of 17.8% in the validation set of the secondary system. This improved the accuracy (82.2%), sensitivity (69.6%) and specificity (95.5%) compared to the results without DS. Fig.9. 8b shows the DF plot for the PCA-LDA model using the training of the primary system and prediction with the secondary system after DS.

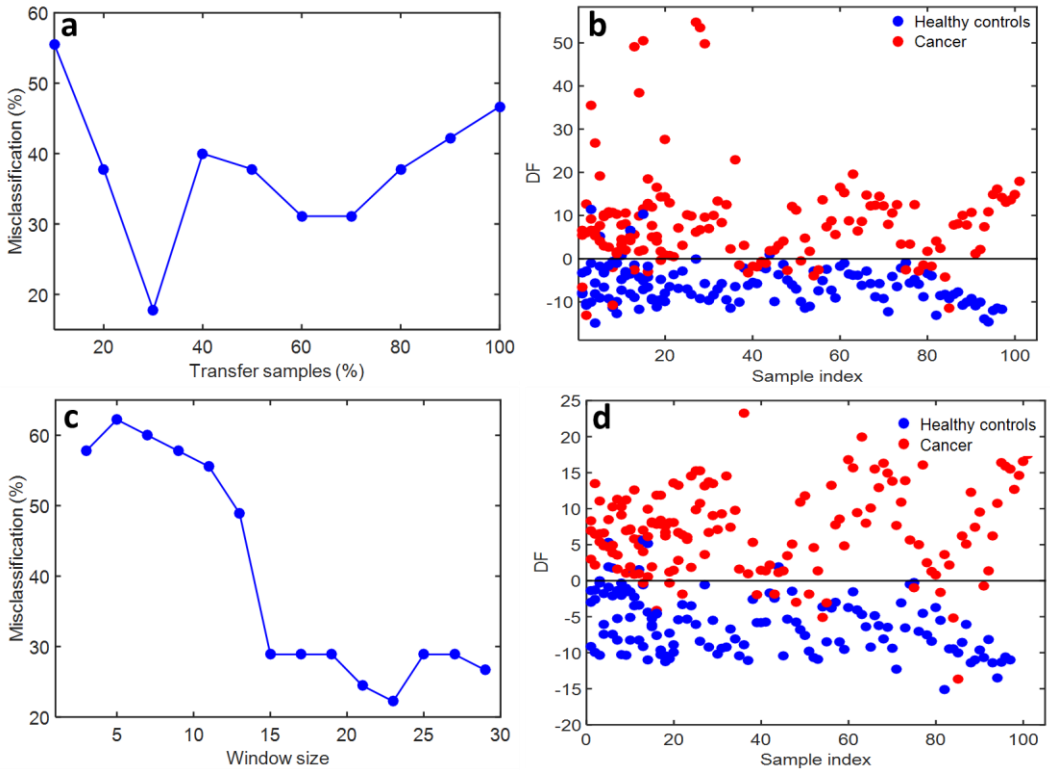


Figure 9.8. PCA-LDA results for DS and PDS standardisation models for spectra collected by two different operators. (a) Misclassification rate in % for the validation set of the secondary system (Operator 2) varying the number of transfer samples in % from the primary system (Operator 1) for DS optimization; (b) DF plot of the PCA-LDA model for the primary system predicting the validation set from the secondary system after DS; (c) Misclassification rate in % for the validation set of the secondary system varying the window size for PDS optimization; (d) DF plot of the PCA-LDA model for the primary system predicting the validation set from the secondary system after PDS.

The number of transfer samples was maintained as 59 for PDS; and the window size was optimized by using the validation set of the secondary system. An optimum window size of 23 wavenumbers was selected with a misclassification rate of 22.2% (Fig. 9.8c). The accuracy, sensitivity and specificity using PDS were 77.8%, 100% and 54.5%, respectively. Although DS obtained an average better classification performance than PDS for this dataset, it also generated some outliers as mentioned before. For this reason, the results after PDS seem better standardized. The PCA-LDA DF plot after PDS is depicted in Fig. 9.8d.

CHAPTER 10 | TTWD-DA: A MATLAB TOOLBOX FOR DISCRIMINANT ANALYSIS BASED ON TRILINEAR THREE-WAY DATA

This chapter is published in Chemometrics and Intelligent Laboratory Systems (IF 2.786). It demonstrates a new user-friendly graphical interface to classify trilinear 3D data:

- Morais CLM, Lima KMG, Martin FL. TTWD-DA: A MATLAB toolbox for discriminant analysis based on trilinear three-way data. Chemom. Intell. Lab. Syst. **2019**; 188: 46–53. https://doi.org/10.1016/j.chemolab.2019.03.007

Abstract: Three-way trilinear data is increasingly used in chemical and biochemical applications. This type of data is composed of three-way structures representing two different signal responses and one sample dimension distributed among a 3D structure, such as the data represented by fluorescence excitation-emission matrices (EMMs), spectral-pH responses, spectral-kinetic responses, spectral-electric potential responses, among others. Herein, we describe a new MATLAB toolbox for classification of trilinear three-way data using discriminant analysis techniques (linear discriminant analysis [LDA], quadratic discriminant analysis [QDA], and partial least squares discriminant analysis [PLS-DA]), termed "TTWD-DA". These discrimination techniques were coupled to multivariate deconvolution techniques by means of parallel factor analysis (PARAFAC) and Tucker3 algorithm. The toolbox is based on a user-friendly graphical interface, where these algorithms can be easily applied. Also, as output, multiple figures of merit are automatically calculated, such as accuracy, sensitivity and specificity. This software is freely available online.

Author contribution: C.L.M.M. developed the software, performed the data analysis and wrote the manuscript.

Camilo L. M. Morais, PhD candidate

Prof. Francis L. Martin, Supervisor

10.1 Introduction

Molecular fluorescence spectroscopy is an analytical technique based on the fluorescence capacity of a sample, where a beam of high energy light (*e.g.*, in the ultraviolet region) is incident on a sample which, after excitation to a higher electronic state, will rapidly lose energy through internal conversion and return to the lowest vibrational state of the lowest electronic excited state. The molecule remains in this excited vibronic level for a short period of time known as fluorescence lifetime and then returns to the fundamental electronic state emitting a photon with energy lower that the one used for excitation. This process is called emission. The excitation and emission spectra can be combined by computer software generating a three-way data structure termed excitation-emission (EEM) matrix (Bachmann *et al.*, 2006; Santos *et al.*, 2017). The advantages of molecular fluorescence spectroscopy are its high sensitivity and relatively low-cost instrumentation (Santos *et al.*, 2017). In addition, the EEM data generated is contemplated by the "second-order advantage" (Booksh & Kowalski, 1994), a property that allows concentrations and spectral profiles of the components of a sample to be extracted in the presence of unknown interferences using second-order chemometric methods (Calimag-Williams *et al.*, 2014; Li *et al.*, 2011).

EEM data is an example of trilinear three-way array, in which a three-way structure representing two different signal responses and one sample dimension are distributed among a 3D structure. This type of data, mainly characterized by fluorescence EEM spectroscopy, also can be generated by combinations of different instrumental responses, such as spectral-pH, spectral-kinetic and spectral-electric potential responses. Common second-order algorithms for decomposition of trilinear three-way data are the parallel factor analysis (PARAFAC) (Bro, 1997) and Tucker3 algorithm (Tucker, 1966). Both PARAFAC and Tucker3 decompose the three-way data into factors containing scores (information pertaining to the sample's variability) and two different loadings, one for the 1st mode (*e.g.*, emission) and another for the 2nd mode (e.g., excitation) profiles (Bro, 1997; Bro *et al.*, 2009). The difference between these techniques is that the Tucker3 method also generates a core array containing the scores and loadings weights for each factor generated (Bro *et al.*, 2009; Gallo, 2015; Tucker, 1966). Both PARAFAC and Tucker3 significantly reduce the dataset, speeding up computational processing time, solving problems of ill-conditioned data and removing

interference. The scores generated from these techniques can then be used as input variables for calibration and classification models.

Discriminant analysis (DA) is a supervised classification technique employed for differentiating classes based on a Mahalanobis distance calculation (Dixon & Brereton, 2009; Morais & Lima, 2018). DA can be divided into linear discriminant analysis (LDA) or quadratic discriminant analysis (QDA). In LDA, the variance structures of the classes being analysed are considered similar, therefore the discriminant function is calculated using a pooled variance-covariance matrix among the classes. However, in QDA, each class is considered to have a different variance structure; therefore, the discriminant function is calculated using the variance-covariance matrix for each class individually (Morais & Lima, 2018). This property increases the classification performance of QDA over LDA when classes exhibiting large within-category variances are being analysed.

Another common algorithm for discrimination of three-way data is the partial least squares discriminant analysis (PLS-DA), where the data is decomposed by partial least squares (PLS) followed by a linear discriminant function (Brereton & Lloyd, 2014). There are many applications for which chemometric techniques are employed for analysing three-way data, such as for assessing food quality (Azcarate *et al.*, 2015; Merás *et al.*, 2018; Sádecká *et al.*; 2018), detection of substances in the atmosphere (Pan, 2015), and differentiation of fungi (Costa *et al.*, 2017) using EEM spectroscopy; analysis of heavy metal ions using spectral-kinematic responses (da Silva & Oliveira, 1999); and evaluation of different juices colorants via spectral-pH responses (Marsili *et al.*, 2005). However, despite the possible advantages of QDA for complex datasets, the number of applications using this approach with fluorescence spectroscopy are fewer compared to LDA (Costa *et al.*, 2017; Morais & Lima, 2017; Stelzle *et al.*, 2013; Stelzle *et al.*, 2017). This is possibly the result of a lack of user-friendly or accessible algorithms for building QDA-based models towards analysing fluorescence data. Herein, a new user-friendly graphical user interface (GUI) was developed containing LDA and QDA routines combined with PARAFAC and Tucker3 for discrimination of fluorescence data. In addition, PLS-DA algorithm is also present for class discrimination. The software, named TTWD-DA (Trilinear Three-way Data – Discriminant Analysis) is free available and described hereafter.

10.2 Software

10.2.1 System Requirements and Installation

This software was developed in MATLAB R2014b environment (The MathWorks, Inc., USA). It makes use of MATLAB functions and lab-made routines, as well as the N-way toolbox for MATLAB version 3.30 (http://www.models.life.ku.dk/nwaytoolbox) (Andersson & Bro, 2000) for building PARAFAC and Tucker3 models. The software is an open-source toolbox for MATLAB users only. It is freely available under the University of Central Lancashire (UCLan) license using the following address:

https://uclanip.co.uk/discriminant_analysis_fluorescence_data/5af2ba83c6b8fb6d28d76291 . It has been tested on MATLAB R2014b version 8.4.0 only, but it should work in any subsequent version. The authors are not responsible for misfunctioning in older MATLAB versions. For installation, the download file should be unzipped and added to the path within MATLAB. The main GUI can be accessed by typing the command 'startup' on MATLAB command window. For usage instructions, please refer to this paper or to the manual present in the software webpage.

10.2.2 Theory

The following classification algorithms are included in the toolbox: PARAFAC-LDA, PARAFAC-QDA, Tucker3-LDA, Tucker3-QDA and PLS-DA. PARAFAC is a multivariate deconvolution approach of high-order data based on a trilinear system (Bro, 1997). It decomposes the three-way data $\underline{\mathbf{X}}$ as follows (Costa *et al.*, 2017):

$$\underline{\mathbf{X}} = \mathbf{A}(\mathbf{C}|\otimes|\mathbf{B})^{\mathrm{T}} + \underline{\mathbf{E}} \qquad (10.1)$$

where $\mathbf{A}$ is the PARAFAC scores matrix representing the sample direction; $\mathbf{B}$ is the PARAFAC loadings matrix representing the excitation direction; $\mathbf{C}$ is the PARAFAC loadings matrix representing the emission direction; $\underline{\mathbf{E}}$ is a residual three-way array; and $|\otimes|$ represents the Khatri-Rao product (Liu, 1999).

Tucker3 is another multivariate deconvolution method for higher-order data also known as "3-way principal component analysis (PCA)" (Henrion, 1994). It decomposes the three-way data $\underline{\mathbf{X}}$ as follows (Gallo, 2015):

$$\underline{\mathbf{X}} = \mathbf{AG}(\mathbf{C} \otimes \mathbf{B})^{\mathrm{T}} + \underline{\mathbf{E}} \tag{10.2}$$

where $\mathbf{A}$ is the Tucker3 scores matrix representing the sample direction; $\mathbf{B}$ is the Tucker3 loadings matrix representing the excitation direction; $\mathbf{C}$ is the Tucker3 loadings matrix representing the emission direction; $\underline{\mathbf{E}}$ is a residual three-way array; $\mathbf{G}$ is the core matrix; and $\otimes$ represents the Kronecker product (Van Loan, 2000).

After these decompositions, the scores matrix from PARAFAC and Tucker3 are used as input variables for LDA and QDA algorithms. LDA and QDA classification scores can be calculated in a non-Bayesian form using the Mahalanobis distance as follows (Dixon & Brereton, 2009; Morais & Lima, 2018):

$$L_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}} \mathbf{C}_{pooled}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_k) \tag{10.3}$$

$$Q_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}} \mathbf{C}_k^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_k) \tag{10.4}$$

where L_{ik} is the LDA classification score for sample i of class k; Q_{ik} is the QDA classification score for sample i of class k; $\mathbf{x}_i$ is the vector containing the classification variables for sample i (*e.g.*, scores from PARAFAC or Tucker3); $\bar{\mathbf{x}}_k$ is the mean vector for class k; $\mathbf{C}_{pooled}$ is the pooled covariance matrix; and $\mathbf{C}_k$ is the variance-covariance matrix of class k. $\mathbf{C}_{pooled}$ and $\mathbf{C}_k$ are calculated as:

$$C_{pooled} = \frac{1}{n} \sum_{k=1}^{K} n_k \mathbf{C}_k \tag{10.5}$$

$$\mathbf{C}_k = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k) (\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}} \tag{10.6}$$

in which n is the number of objects in the training set; K is the number of classes; and n_k is the number of objects in class k.

PLS-DA performs a partial least squares (PLS) decomposition of the reshaped spectral array [$\underline{\mathbf{X}}(n \times m \times k) \rightarrow \mathbf{X}(n \times m * k)$] followed by a linear discriminant classifier (Brereton & Lloyd, 2014). PLS decomposition takes the form (Brereton & Lloyd, 2014):

$$\mathbf{X} = \mathbf{TP} + \mathbf{E} \quad (10.7)$$

$$\mathbf{y} = \mathbf{Tq} + \mathbf{f} \quad (10.8)$$

where **T** is a common scores matrix; **P** are the spectral loadings; **E** are the spectral residuals; **y** is the response vector (*e.g.*, 0 or 1); **q** is the response loadings; and **f** the response residuals. This decomposition can be performed in an interactive process according to the number of selected components, as described by Brereton and Lloyd (2014). After the model is built, it is possible to predict the value of **y** for the original training data or future test samples as follows (Brereton & Lloyd, 2014):

$$\mathbf{b} = \mathbf{W}(\mathbf{PW})^{-1}\mathbf{q} \quad (10.9)$$

$$\hat{\mathbf{y}} = \mathbf{Xb} \quad (10.10)$$

where **b** are PLS coefficients; **W** is a weight matrix; and $\hat{\mathbf{y}}$ is the predicted response vector.

10.2.3 Figures of Merit

Different quality parameters are used to evaluate the performance of LDA- and QDA-based models. These figures of merit were: correction classification rate (CC%), accuracy (AC), sensitivity (SENS), specificity (SPEC) and F-score. The CC% represents the percentage of samples correctly classified considering their true classes; the AC represents the total number of samples correctly classified considering true and false negatives; the SENS represents the proportion of positives that are correctly identified; the SPEC represents the proportion of negatives that are correctly identified; and, the F-score represents the overall classification performance considering imbalanced data (Morais & Lima, 2017). These parameters are calculated as follows:

$$\mathrm{CC\%} = 100 - \frac{(\varepsilon_1 - \varepsilon_2)}{N} \times 100 \quad (10.11)$$

$$\mathrm{AC(\%)} = \left(\frac{\mathrm{TP+TN}}{\mathrm{TP+FP+TN+FN}}\right) \times 100 \quad (10.12)$$

$$\mathrm{SENS(\%)} = \left(\frac{\mathrm{TP}}{\mathrm{TP+FN}}\right) \times 100 \quad (10.13)$$

$$\mathrm{SPEC}(\%) = \left(\frac{\mathrm{TN}}{\mathrm{TN+FP}}\right) \times 100 \tag{10.14}$$

$$\mathrm{F-score} = \frac{2\times\mathrm{SENS}\times\mathrm{SPEC}}{\mathrm{SENS+SPEC}} \tag{10.15}$$

where TP stands for true positive, TN for true negative, FP for false positive, FN for false negative; and ε_1 and ε_2 represents the number of errors in the test set for class 1 and 2, respectively.

10.2.4 Software Overview

The main GUI features of TTWD-DA are depicted in Figure 10.1.

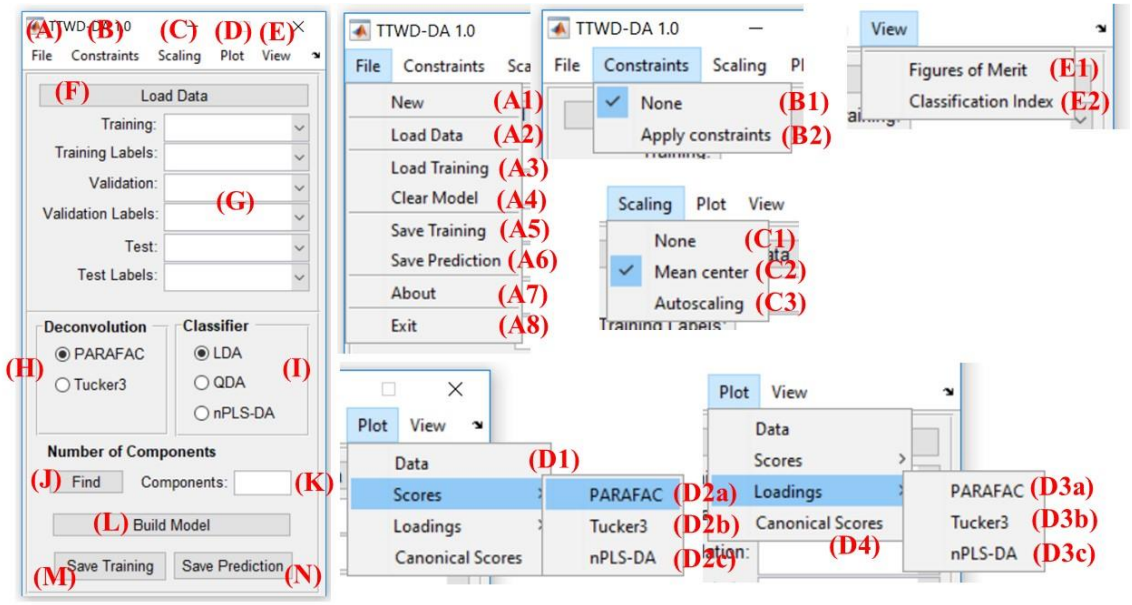


Figure 10.1. EEM-DA main interface overview. Insets (A)-(N) refer to the text.

The main classification interface (Figure 10.1) contains four menu options (A, B, C and D). Menu (A) enables the user to open a new software window (A1); to load data from a .mat file (A2); load a pre-built training model in order to make new data predictions (A3); clear the training model (A4); save the training model in order to make further data predictions (A5); save prediction results into a .mat file (A6); to obtain information about the software (A7); and, exit (A8). Menu (B) contains constraint options: (B1) no constraints (default); (B2) constraints for PARAFAC and Tucker3 algorithms, which includes

orthogonality, nonnegativity, unimodality and nonnegativity, L1 fitting, and L1 fitting and nonnegativity (these are applied for each mode individually using a new window with options that appears after clicking on B2). (C) Scaling options: (C1) no scaling; (C2) mean-centring scaling (default); and, (C3) autoscaling. Menu (D) contains plotting options: (D1) three-way data plotting, including profiles in mode 1 and 2; (D2a) PARAFAC scores; (D2b) Tucker3 scores; (D2c) PLS-DA scores; (D3a) PARAFAC loadings; (D3b) Tucker3 loadings; (D3c) PLS-DA loadings and coefficients; and, (D4) canonical scores and predicted class. Menu (E) contains viewing options, including figures of merit (E1), which contains correct classification rates, accuracy, sensitivity, specificity and F-score; and, the predicted classification indexes (E2) using the chemometric method selected to build the model. The button (F) loads the data (same in A2); in the region (G), the user chooses the training, validation and test sets with their respective classes labels (the use of a validation set is optional, but recommended for optimization of the number of components); in the region (H), the user chooses the multivariate deconvolution method (PARAFAC or Tucker3); region (I) contains the type of discriminant analysis technique (LDA, QDA or PLS-DA); in button (J), the user can use singular value decomposition (SVD) (Morais & Lima, 2017) in order to select the number of components for PARAFAC and Tucker3, or training and validation misclassification errors for selecting the number of latent variables for PLS-DA; in (K), the user has to insert the number of components for PARAFAC or PLS-DA algorithms; the button (L) calculates the discriminant analysis model; in (M), the user can save a file to use as a training model for further predictions (same in A5); and, in (N), the user can export all prediction results in a .mat file, including PARAFAC scores and loadings; Tucker3 scores, loadings, and core matrix; PLS-DA scores and loadings; figures of merit; and the predicted class indexes for the samples in the training, validation and test set.

10.3 Test Dataset

The dataset tested herein is composed of fluorescence EEM data collected from cod (*Gadus morhua*) fillets. This dataset is publicly available at http://www.models.life.ku.dk/datasets by Andersen *et al.* (2003). Aqueous extracts containing fish muscle were measured in the range of 250–370 nm (resolution of 10 nm) for

excitation and 270–600 nm (resolution of 1 nm) for emission using a Perkin-Elmer LS50B spectrofluorimeter. The data were divided into 3 classes: class 1 containing 63 cod samples stored up to 1 week (0–7 days); class 2 containing 21 cod samples stored for 2 weeks (14 days); and, class 3 containing 21 cod samples stored for 3 weeks (21 days). The average EEM for each class are depicted in Figure 10.2. More details about the experimental procedure for data acquisition can be found at Andersen *et al.* (2003).

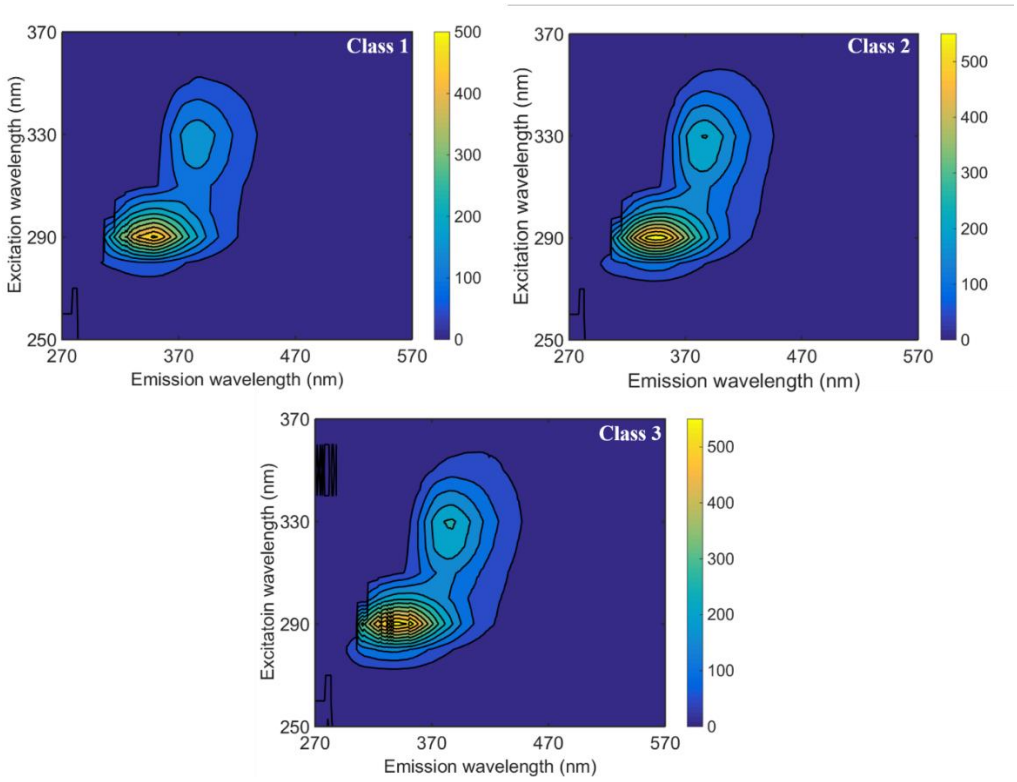


Figure 10.2. Average EEM for the test dataset.

10.4 Software Application

10.4.1 Before Loading the Data

Before loading the dataset into TTWD-DA, the dataset can be pre-processed and must be organized in a three-dimensional manner and separated into Training, Validation and Test or Training and Test sets. Pre-processing and sample splitting techniques are not covered by this software; thus, it should be performed separately employing other routines available elsewhere. Herein, the dataset is already pre-processed by removing Rayleigh and Raman

scatterings using the 'EEMscat' algorithm (Bahram *et al.*, 2006), which is of fundamental importance for EEM data; and the sample splitting was made with the Training (n = 59), Validation (n = 23) and Test (n = 23) sets separated using the Kennard-Stone algorithm (Kennard & Stone, 1969). Each three-way array size should be in the format: $n \times m \times k$, where n is the number of samples; m is the number of emission wavelengths; and k the number of excitation wavelengths. Figure 10.3 depicts these type of data in MATLAB.

10.4.2 Loading the Data

To load the data, the user should select the .mat file containing the three-way array for analysis and select the Training set, Training Labels, Validation set, Validation Labels, Test set and Test Labels (Figure 10.3). Only previously saved .mat files can be used as input.

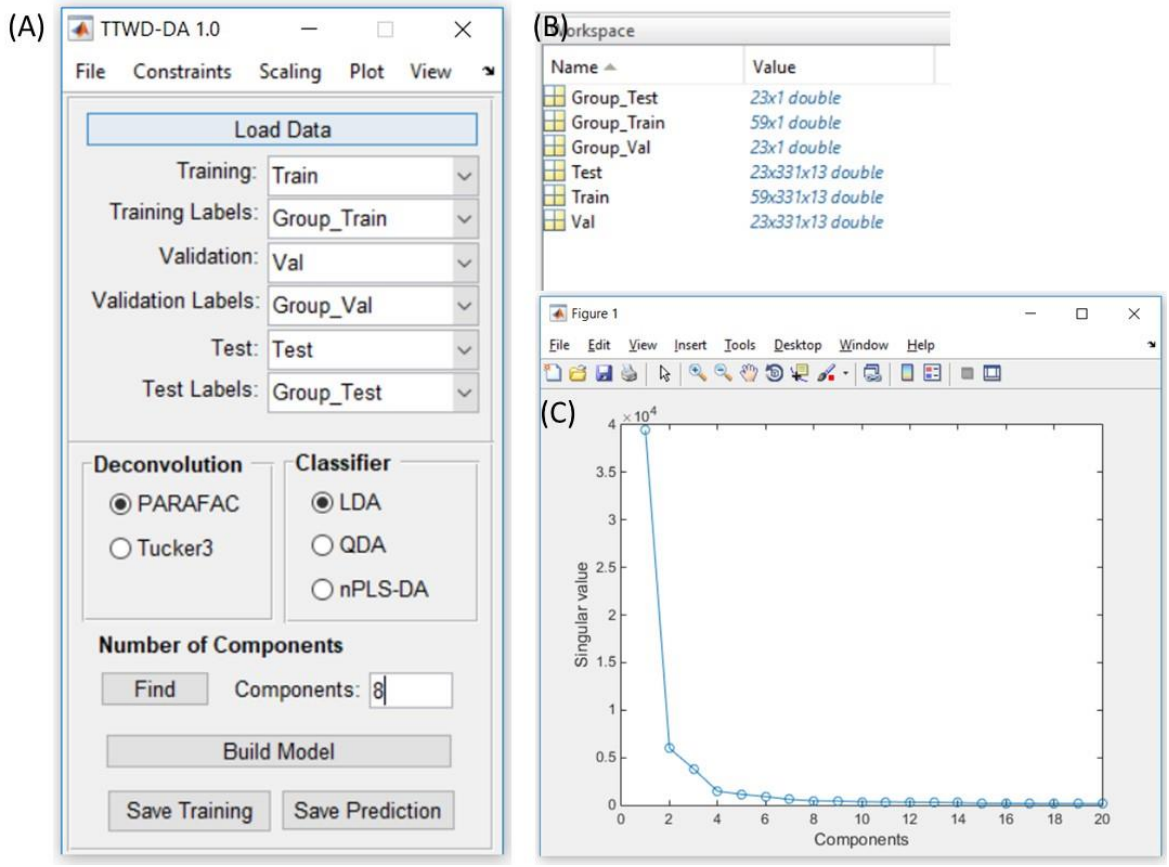


Figure 10.3. A) Main intertace with the dataset loaded and the number of components selected; B) workspace variables containing the dataset used; C) singular values varying the number of components.

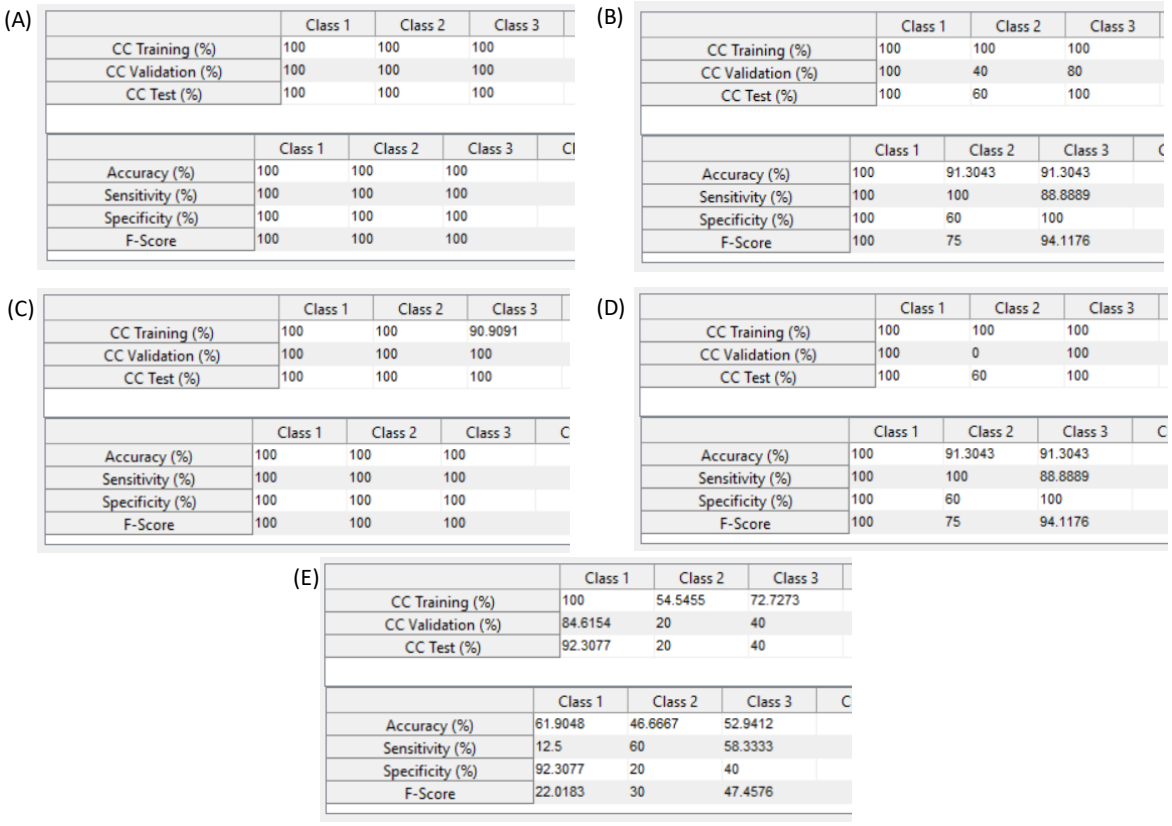
10.4.3 Model Construction

The user must select the deconvolution method (either PARAFAC or Tucker3) and type of discriminant technique to be employed (either LDA or QDA). PLS-DA can be chosen as feature extraction and discriminant method combined. Herein, all of them are tested. Next, the number of components for data deconvolution should be selected by clicking on "Find" button. This is performed based on a SVD model of the unfolded three-way array for PARAFAC and Tucker3 options, where the number of components (*i.e.*, factors) should be selected as the minimum singular value before it becomes constant while varying the components; and based on the training and validation misclassification errors for PLS-DA, where the number of components (*i.e.*, latent variables) that provides the minimum error should be selected. The number of components ≤10 should be preferred to avoid addition of random noise. However, this can be optimized by using the validation set. For this test dataset, 8 components were selected based on SVD (Figure 10.3).

The number of classes for which there is capability to analyse within this toolbox varies from 2 to 10. The software is limited by 10 classes for two reasons: (1) the use of >10 classes for classification implies the need for >10 components; (2) for a multi-class system, the classification is performed on a binary basis of one-against-the-others; thus, the size of the second relative class is enlarged by K-1 times, where K is the number of classes. Such a difference in size might greatly affect the classifier performance. In addition, the user has the option to include constraints in either PARAFAC or Tucker3 models by selecting the menu "Constraints > Apply constraints". The user can choose between orthogonality, nonnegativity, unimodality and nonnegativity, L1 fitting, and L1 fitting and nonnegativity to be applied independently in each mode of the three-way data array. Finally, the model is built by clicking in "Build Model". The data was mean-centred (default option) before analysis in the menu "Scaling".

10.4.4 Results

After the model is built, a new window appears showing the correct classification rates for each dataset and the figures of merit for the test set (Figure 10.4). This window also can be accessed by clicking on View > Figures of Merit.



(A)

	Class 1	Class 2	Class 3
CC Training (%)	100	100	100
CC Validation (%)	100	100	100
CC Test (%)	100	100	100

	Class 1	Class 2	Class 3
Accuracy (%)	100	100	100
Sensitivity (%)	100	100	100
Specificity (%)	100	100	100
F-Score	100	100	100

(B)

	Class 1	Class 2	Class 3
CC Training (%)	100	100	100
CC Validation (%)	100	40	80
CC Test (%)	100	60	100

	Class 1	Class 2	Class 3
Accuracy (%)	100	91.3043	91.3043
Sensitivity (%)	100	100	88.8889
Specificity (%)	100	60	100
F-Score	100	75	94.1176

(C)

	Class 1	Class 2	Class 3
CC Training (%)	100	100	90.9091
CC Validation (%)	100	100	100
CC Test (%)	100	100	100

	Class 1	Class 2	Class 3
Accuracy (%)	100	100	100
Sensitivity (%)	100	100	100
Specificity (%)	100	100	100
F-Score	100	100	100

(D)

	Class 1	Class 2	Class 3
CC Training (%)	100	100	100
CC Validation (%)	100	0	100
CC Test (%)	100	60	100

	Class 1	Class 2	Class 3
Accuracy (%)	100	91.3043	91.3043
Sensitivity (%)	100	100	88.8889
Specificity (%)	100	60	100
F-Score	100	75	94.1176

(E)

	Class 1	Class 2	Class 3
CC Training (%)	100	54.5455	72.7273
CC Validation (%)	84.6154	20	40
CC Test (%)	92.3077	20	40

	Class 1	Class 2	Class 3
Accuracy (%)	61.9048	46.6667	52.9412
Sensitivity (%)	12.5	60	58.3333
Specificity (%)	92.3077	20	40
F-Score	22.0183	30	47.4576

Figure 10.4. Figures of merit for A) PARAFAC-LDA, B) PARAFAC-QDA, C) Tucker3-LDA, D) Tucker3-QDA, E) PLS-DA.

Cod fillets are used in this test dataset. The samples were divided into 3 classes according to their storage time (class 1 - relatively new samples stored up to 1 week; class 2 - samples stored for 2 weeks; and, class 3 - relatively old samples stored for 3 weeks). Freshness is an important parameter to assess fish quality, since the fish retains its original characteristics closer to the harvest and the aging process leads to changes such as microbiological growth and alterations in biochemical, chemical and physical properties (Morais & Lima, 2017; Nilsen *et al.*, 2002). For this dataset, the CC% for LDA-based methods were much higher compared to the QDA-based methods in the training (n=59),

validation (*n*=23) and test (*n*=23) sets. PARAFAC, Tucker3 and PLS-DA models were built using 8 components based on SVD. The model with best correct classification overall was the PARAFAC-LDA, showing 100% correct classification for all classes in the training, validation and test sets (Figure 10.4A). The predictive classification performance of PARAFAC-QDA was inferior than PARAFAC-LDA, in which the accuracy and F-score for PARAFAC-QDA were equal to 91.3–100% and 75.0–100%, respectively; and, for PARAFAC-LDA they were both equal to 100%. The same trend was observed for Tucker3-LDA, where the accuracy and F-score were equal to 100%, compared to 91.3–100% and 75.0–100% in Tucker3-QDA. QDA-based models perform better than LDA in systems containing different variance structures (Morais & Lima, 2018), however it has an inferior performance compared to LDA for datasets with small number of samples (Wu *et al.*, 1996). Comparing the variance among the three classes in dataset 3 (Figure 10.5), classes 1 and 2 have similar variance structures, whereas class 3 exhibits a different pattern with lower variance in the region of 330 nm in the excitation direction. The main disadvantage of QDA in relation to LDA is that QDA is more affected by classes having a small number of samples, since the variance structures of the classes are not well represented, which can lead to overfitting problems. Therefore, QDA usually achieves better classification performance when the number of samples in the dataset is relatively large (Wu *et al.*, 1996).

In comparison with the LDA- and QDA-based models, PLS-DA generated the poorer discriminant performance, with accuracies ranging from 46.7-61.9% and F-scores ranging from 30.0-47.5%. Class 1 seems to be well fitted in PLS-DA, with good correct classification values; however, for class 2 and 3, the prediction performance is greatly affected. Figure 10.6 shows the PLS-DA canonical scores of latent variables 1 and 2, and the predicted class values. In this figure, it is clearly shown that class 2 and 3 are mixed together, while class 1 distinguishes from them.

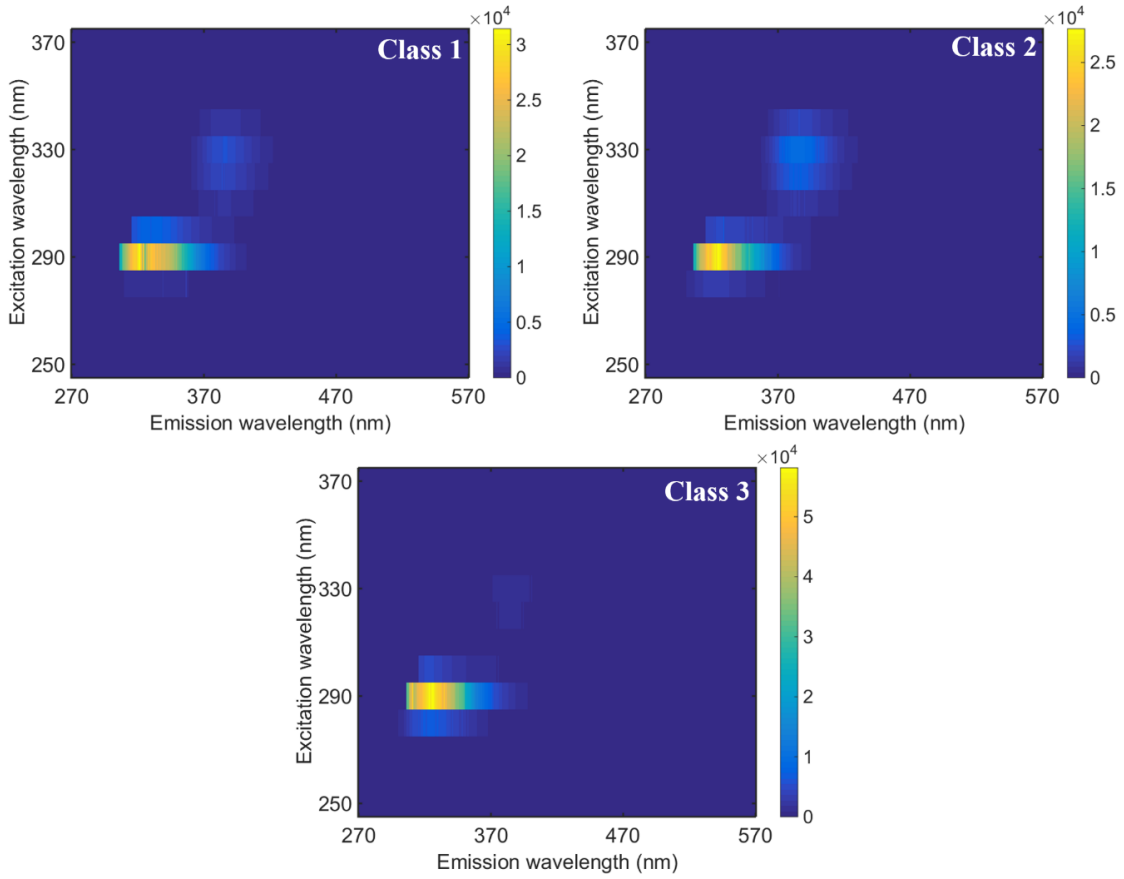


Figure 10.5. Variance calcualted for the test dataset.

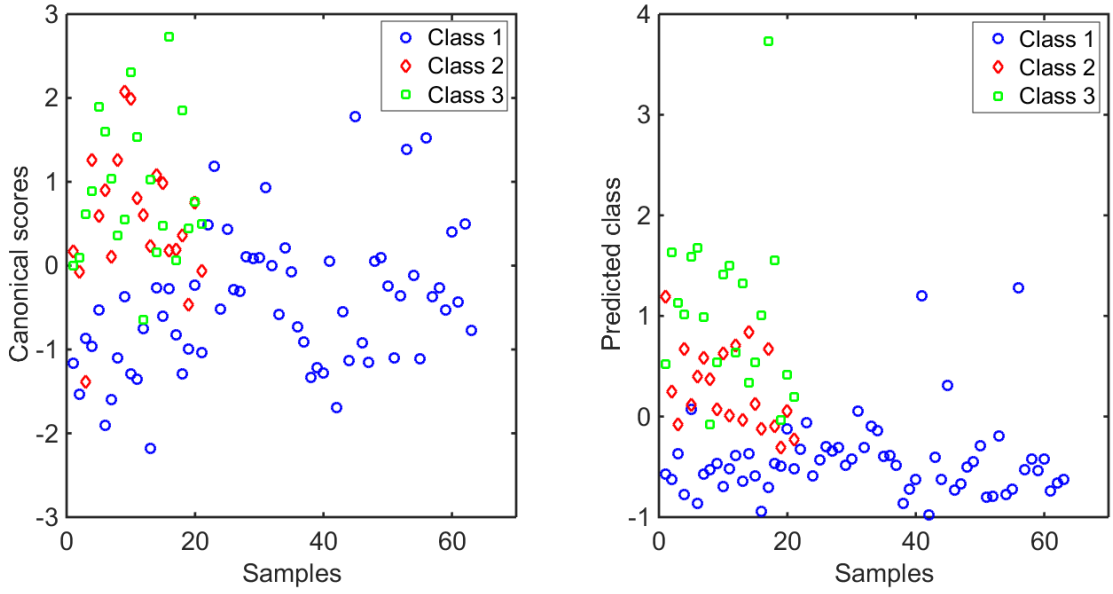


Figure 10.6. PLS-DA canonical scores (left) and predicted class (right).

After the model is built, the measured and predicted sample indexes for each class can be viewed in the Menu: View > Classification Index (Figure 10.6). These results can be saved in a .xls file by clicking on “Export”. Also, all the results and matrices generated during analysis can be saved by clicking on “Save Prediction” in the main window, or in the Menu: File > Save Prediction. The training model also can be saved by clicking on “Save Training” in the main window, or in the Menu: File > Save Training for further predictions of new test sets.

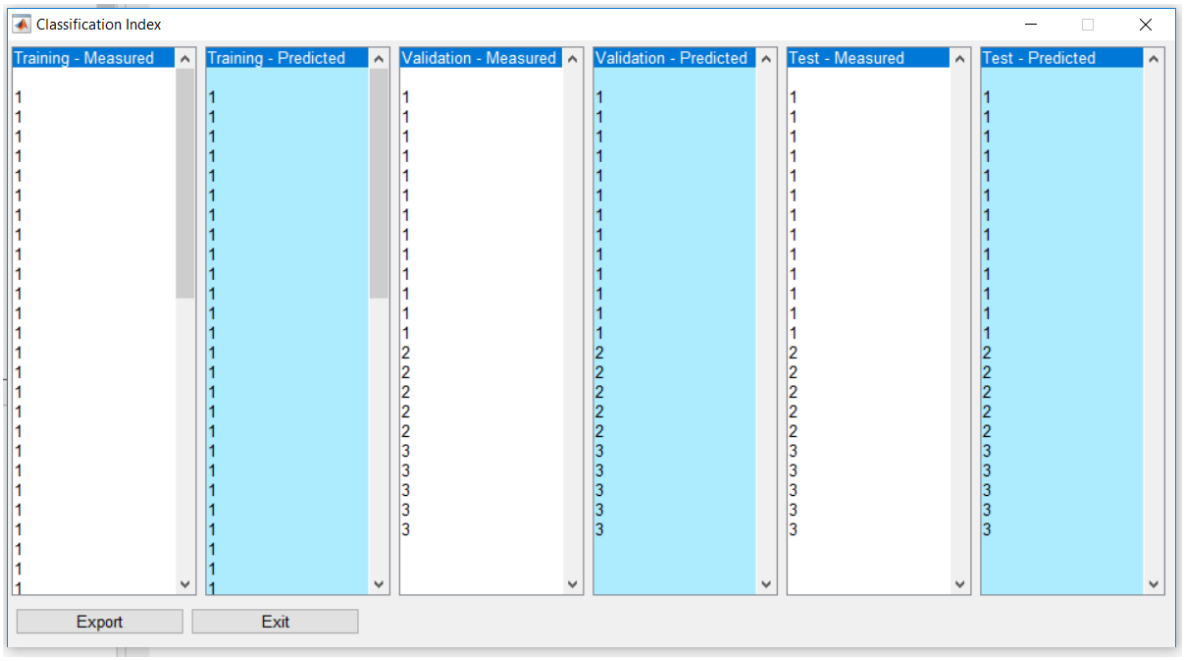


Figure 10.7. Classification indexes predicted by the toolbox for the PARAFAC-LDA model.

10.5 Conclusion

TTWD-DA is a user-friendly GUI for building discriminant analysis models (LDA, QDA and PLS-DA) for three-way data. The software makes use of PARAFAC and Tucker-3 algorithms as multivariate deconvolution techniques, followed by LDA and QDA discrimination functions; or PLS-DA as joined feature extraction and discrimination techniques. Parameters such as accuracy, sensitivity and specificity are automatically calculated. The software is based on MATLAB environment, being open source and freely available online. It can be applied in any three-way array, in particular fluorescence EEM data. There is room for evolving the software by adding new classification algorithms and pre-processing options, thus having the potential to be a standard tool for analysing trilinear three-way data.

10.6 Independent Testing

TTWD-DA was independently tested by Prof. Héctor C. Goicoechea at the Laboratorio de Desarrollo Analítico y Quimiometría-LADAQ, Cátedra de Química Analítica I, Facultad de Bioquímica y Ciencias Biológicas, Universidad Nacional del Litoral, CONICET, Ciudad Universitaria 3000 Santa Fe, Argentina (hgoico@fbcb.unl.edu.ar). It was reported that the software worked correctly in a user-friendly fashion: “This program allows implementation of classification using second-order data applying PARAFAC o Tuker3 (as compression tools) followed by LDA or QDA. I installed the files provided by the authors and used it not only with the data provided by them, but also with data generated in our lab. The program works as described in the user manual in a user friendly way”.

CHAPTER 11 | DISCUSSION AND CONCLUSIONS

Vibrational spectroscopy is a fast, low-cost and reliable sensor-based technique that when applied to investigate biological materials provides vast benefits towards quick biochemical profiling of biological-derived samples. IR and Raman spectroscopy are robust budget-omic tools capable of generating non-destructive spectrochemical fingerprints for samples with minimum or no sample preparation and acting as substitute or complementary techniques to traditional methods of analysis (Baker *et al.*, 2014; Butlet *et al.*, 2016; Martin *et al.*, 2010). However, these techniques present complex overlapping spectral features that require the use of computational-aided methods in order to extract meaningful information from the experimental data and allow the analyst to draw significant conclusions about the samples.

Chemometric techniques (Beebe *et al.*, 1998) are employed in biospectroscopy datasets in four phases: (1) pre-processing, where the raw spectral data undergo a series of pre-treatments such as smoothing, baseline correction and normalisation to enhance the analyte signal and mitigate physical effects that may mask the signal of interest; (2) sample selection, where the experimental pre-processed data are split into training and test subsets, the first composed of representative spectra used to build the chemometric model and the latter composed of external samples used to validated it; (3) model construction, where the training data are used for feature extraction or selection techniques responsible for mining relevant features in the dataset, hence, removing redundant information and reducing the data complexity, followed by classification techniques where discriminant rules are defined based on distinguishing features within the spectral data so the spectra are assigned to pre-defined classes based on their spectrochemical signature; and (4) validation, where the results obtained blindly for the test set are used to calculate statistical metrics in order to assess the true model performance.

Although there are a variety of chemometric methods used in steps (1) to (4), there are still some obstacles that hinder real implementation of biospectroscopy for routine applications. Experimental protocols for sample preparation and measurement are well defined (Baker *et al.*, 2014; Butlet *et al.*, 2016; Martin *et al.*, 2010), however little has been

made regarding the spectral data analysis, specially towards sample selection, processing of digital images, standardisation and, finally, the production of chemometric protocols to analyse biospectroscopy data. This thesis was produced to address these issues, where new techniques for sample selection and processing of digital images, as well as chemometric protocols for spectral data analysis and standardisation are provided.

Chapter 2 is a protocol demonstrating how to perform all the steps from (1) to (4) in order to build classification models based on the spectral data acquired from biological materials. The protocol encompasses most of classical and new chemometric techniques available along with software suggestions, examples, and step-by-step procedures. The aim of this protocol is to provide a solid support for students who need to build chemometric models for classification analysis of biospectroscopy data.

Chapter 3 and 4 report a new algorithm for sample selection, called the MLM algorithm, that was created to improve the sample selection process. Sample selection (or data splitting) is the name given to the step where the experimental spectral data are split into training and validation or test sets. Commonly, there are two methods that can be employed for this: random selection, where these samples are split into training and validation/test randomly; or using the Kennard-Stone (KS) algorithm, where a set of samples as far away as possible from the others in an Euclidian-distance space are assigned to the training set and the remaining to the validation or test sets. At the end, this procedure includes the samples with the maximum variation from the mean in the training set and the samples closer to the mean to the validation or test sets. However, these two common methodologies have problems for biological applications. Random selection introduces a high-degree of extrapolation and overfitting, since not necessarily most source of variance within the dataset will be included in the trained model; and the KS method, apart from the risk of overoptimistic results in the validation/test sets since these are closer to the class mean, it does not account for extreme samples that may appear in future predictions due to the random behavior of the biological medium. The MLM algorithm combines both the random selection and KS into one single method. Initially, the samples are divided into training and test or validation sets using the Euclidean-distance approach of the KS algorithm; then, a random mutation factor is employed transferring some of the training samples to the validation/test

set and some of the validation/test samples to the training set. We have demonstrated in Chapter 3 by using 6 real-world datasets and 1000 simulations that the MLM algorithm provides a better classification ability than the random selection or KS method alone, since the classification models built after MLM are more robust and have a lower risk of overfitting demonstrated by the better predictive performance in the external test sets. Chapter 4 is a complement to Chapter 3, where we provide a protocol showing how to use the MLM algorithm in a step-by-step procedure.

Chapters 5, 6 and 7 demonstrate new applications of chemometric techniques for the analysis of hyperspectral images of biological samples. Chapter 5 is a classic example of the application of Raman microspectroscopy imaging to distinguish WHO grade I and grade II meningioma tumours based on formalin-fixed paraffin-embedded (FFPE) tissue analysis. Determining the meningioma WHO tumour grade is critical for patient diagnosis and treatment (Yeo *et al.*, 2019). In Chapter 5, 90 meningioma brain tissue samples (66 WHO grade I, 24 WHO grade II) were analysed in order to distinguish these two tumour grades in a fast and analyst-independent fashion. MCR-ALS was employed for image decomposition, where a single component was extract from the image datasets and used to build concentration distribution maps allowing to identify WHO grade II tumour regions within the tissue along with their specific Raman signature, where biomarkers such as phospholipids, amide III and amide I were obtained; and principal component analysis quadratic discriminant analysis (PCA-QDA) and successive projections algorithm quadratic discriminant analysis (SPA-QDA) were employed to systematically classify the samples into WHO grade I and grade II, where both techniques achieved an accuracy of 96.2% (85.7% sensitivity and 100% specificity) for WHO grade prediction in the test set.

In Chapter 6, we propose a new algorithm for exploratory analysis of hyperspectral images multisets called the three-dimensional principal component analysis algorithm (3D-PCA) algorithm. Ten sample images (5 healthy controls, 5 ovarian cancer patients) were used to exemplify the algorithm, where 3D-PCA provided an almost perfect separation between the samples based on the image data. In addition to the excellent sample segregation, the main outcome from this study was the speed of the 3D-PCA method in comparison with

classical PCA. 3D-PCA took approximately 1 min to analyse the whole set of 10 hyperspectral images, while classical PCA or similar methods would take much longer.

Chapter 7 is an extension of Chapter 6, where 3D-PCA was coupled to discriminant analysis techniques in order to quickly classify hyperspectral images multisets. For this, The 3D-PCA scores are used as input variables for linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) (3D-PCA-LDA and 3D-PCA-QDA) showing a much superior performance than traditional PCA-LDA and PCA-QDA applied to the hyperspectral data in the classical way, where the hyperspectral data-cube is unfolded into a series of spectra. 3D-PCA-LDA and 3D-PCA-QDA were used to classify 38 samples (20 benign controls, 18 ovarian cancer patients) with accuracy of 100% (100% sensitivity, 100% specificity); in comparison to 64% accuracy (20% sensitivity and 100% specificity) using the traditional PCA-LDA and PCA-QDA with the unfolded data. We believe that the reason for this improvement is that 3D-PCA works with the whole hyperspectral data structure without unfolding, thus maintaining the spatial relationship between the pixel positions, while in the unfolding procedure for classical PCA-LDA and PCA-QDA this relationship is lost. In this way, 3D-PCA accounts for the spectral information (as the unfolded method), but also the spatial information (pixel positions and their neighboring relationship) into a single method.

Chapter 8 focus on the step (4), model validation. Although most studies report accuracy, sensitivity and specificity as classification metrics; these metrics do not bring information of the model uncertainty and robustness. We found a way inspired by previous work done with partial least squares discriminant analysis (PLS-DA) (de Almeida *et al.*, 2013; Rocha & Sheen, 2016) and artificial neural networks (ANN) (Allegrini & Olivieri, 2016) to quantify uncertainty and estimate robustness in classification-based models for spectral data. Uncertainty is measured by the misclassification probability rate, a measure between 0 and 1 that inform the *a posteriori* probability of failure of a model; that is, a high misclassification probability rate indicates that the model is probably overfitted and will "struggle" to provide good future predictions; while a low misclassification probability indicates that the model has lower uncertainty and therefore is more stable and will provide more trustworthy predictions in the future. This was evaluated with 3 real-world datasets and 1 simulated dataset. The results indicated that support vector machines (SVM) models tend

to be more susceptible to overfitting in comparison with LDA and QDA, and that models with lower misclassification probabilities will have a better predictive ability in the future when the data are subject to random noise variations, hence, the misclassification probability is also a measure of the model robustness.

The last issue regarding data analysis, which is standardisation, is addressed in Chapter 9. Chapter 9 is a protocol showing how to standardise biospectral datasets acquired by different operators, by different instruments or in different laboratories in order to have the same or close to the same result. This is because factors such as air humidity, CO_2 level, ageing of instrument parts, or even random environmental noise may affect the spectral response for samples being analysed under different conditions, hence, the chemometric model response may be affected and the model accuracy will decrease. Standardisation techniques allow to "standardise" the spectral response so physical variations in the spectral profile can be corrected mathematically. The protocol show in a step-by-step fashion how to standardise and analyse a given biological-derived spectral dataset from the start.

Finally, Chapter 10 demonstrates a new user-friendly graphical interface developed in MATLAB to classify trilinear 3D data. Trilinear 3D data are spectrochemical data distributed in 3D dimensions, such as chromatography, fluorescence or imaging data. The software was developed to facilitate the use of LDA, QDA and PLS-DA algorithms to classify datasets in a simple and straightforward way. The software was independently tested and is freely available online under a UCLan license.

The novel chemometric approaches developed during this PhD enrich the computational data analysis framework and provide support for the community interested in working with biospectroscopy, making this field one step closer to real-world implementation, where the successful development of these techniques may allow to trial biospectroscopy in clinical settings.

Future perspectives for the field

Biospectroscopy is a science that continues to advance by using more sophisticated methods of data analysis and bigger cohort of samples every day. With the increase data complexity, machine learning approaches will be more often used; and studies will have much bigger cohorts of patients, from hundreds of patients in nowadays studies to thousands of patients in future investigations. In addition, real biomarking profiling will be attempted by combining spectroscopy data with other type of omics data such as mass spectrometry or genetic profiling, in order to provide anticipated and customised disease diagnosis.

All these will require the development of advanced chemometric tools. Initially, algorithms and protocols to work with combined data (data-fusion) and better and faster imaging processing approaches will be required. Non-liner machine learning approaches will be developed using better forward feature selection methods for class separation. Ultimately, biospectroscopy will become a common omics tool to analyse biological materials and will start being used in routine applications.

REFERENCES

Abraham A. Artificial Neural Networks. In: Sydenham PH, Thorn R (Ed.) Handbook of Measuring System Design. John Wiley & Sons: **2005**.

Abramczyk H, Brozek-Pluska B. Raman imaging in biochemical and biomedical applications. Diagnosis and treatment of breast cancer. Chem. Rev. **2013**; 113(8): 5766–5781.

Andersen CM, Bro R. Practical aspects of PARAFAC modeling of fluorescence excitation-emission data. J. Chemometr. **2003**; 17(4): 200–215.

Andersson CA, Bro R. The N-way Toolbox for MATLAB. Chemometr. Intell. Lab. Syst. **2000**; 52(1): 1–4.

Azcarate SM, Gomes AA, Alcaraz MR, de Araújo MCU, Camiña JM, Goicoechea HC. Modeling excitation–emission fluorescence matrices with pattern recognition algorithms for classification of Argentine white wines according grape variety. Food Chem. **2015**; 184: 214–219.

Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature **2003**; 422(6928): 198–207.

Afara IO, Prasadam I, Arabshahi Z, Xiao Y, Oloyede A. Monitoring osteoarthritis progression using near infrared (NIR) spectroscopy. Sci. Rep. **2017**; 7: 11463.

Aguayo JB, Blackband SJ, Schoeniger J, Mattingly MK, Hintermann M. Nuclear magnetic resonance imaging of a single cell. Nature **1986**; 322: 190–191.

Akalin A, Mu X, Kon MA, Ergin A, Remiszewski SH, Thompson CM, Raz DJ, Diem M, Bird B, Miljković M. Classification of malignant and benign tumors of the lung by infrared spectral histopathology (SHP). Lab. Invest. **2015**; 95(4): 406–421.

Aksoy C, Guliyev A, Kilic E, Uckan D, Severcan F. Bone marrow mesenchymal stem cells in patients with beta thalassemia major: molecular analysis with attenuated total reflection-Fourier transform infrared spectroscopy study as a novel method. Stem Cells Dev. **2012**; 21(11): 2000–2011.

Ali EM, Edwards HG, Hargreaves MD, Scowen IJ. Raman spectroscopic investigation of cocaine hydrochloride on human nail in a forensic context. Anal. Bioanal. Chem. **2008**; 390(4): 1159–1166.

Allegrini F, Olivieri AC. Sensitivity, Prediction Uncertainty, and Detection Limit for Artificial Neural Network Calibrations. Anal. Chem. **2016**; 88(15): 7807–7812.

Almeida MR, Logrado LPL, Zacca JJ, Correa DN, Poppi RJ. Raman hyperspectral imaging in conjunction with independent component analysis as a forensic tool for explosive analysis: The case of an ATM explosion. Talanta **2017**; 174: 628–632.

Alsberg BK, Hagen OJ. How octave can replace Matlab in chemometrics. Chemom. Intell. Lab. Syst. **2006**; 84(1–2): 195–200.

Amigo JM, Babamoradi H, Elcoroaristizabal S. Hyperspectral image analysis. A tutorial. Anal. Chim. Acta **2015**; 896: 34–51.

Amigo JM, Martí I, Gowen A. Chapter 9 – Hyperspectral Imaging and Chemometrics: A Perfect Combination for the Analysis of Food Structure, Composition and Quality. In: Marini F (Ed.) Chemometrics in Food Chemistry. Elsevier: **2013**, pp. 343–370.

Andrews DT, Wentzell PD. Applications of maximum likelihood principal component analysis: incomplete data sets and calibration transfer. Anal. Chim. Acta **1997**; 350(3): 341–352.

Bachmann L, Zezell DM, Ribeiro AC, Gomes L, Ito AS. Fluorescence Spectroscopy of Biological Tissues – A Review. Appl. Spectrosc. Rev. **2006**; 41(6): 575–590.

Backhaus J, Mueller R, Formanski N, Szlama N, Meerpohl HG, Eidt M, Bugert P. Diagnosis of breast cancer with infrared spectroscopy from serum samples. Vib. Spectrosc. **2010**; 52(2): 173–177.

Bahram M, Bro R, Stedmon C, Afkhami A. Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation. J. Chemometr. **2006**; 20(3–4): 99–105.

Bakeev KA. Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries. 2nd edn. John Wiley & Sons: Chichester, **2010**.

Baker MJ, Byrne HJ, Chalmers J, Gardner P, Goodacre R, Henderson A, Kazarian SG, Martin FL, Moger J, Stone N, Sulé-Suso J. Clinical applications of infrared and Raman spectroscopy: state of play and future challenges. Analyst **2018**; 143: 1735–1757.

Baker MJ, Gazi E, Brown MD, Shanks JH, Clarke NW, Gardner P. Investigating FTIR based histopathology for the diagnosis of prostate cancer. J. Biophotonics **2009**; 2(1–2): 104–113.

Baker MJ, Hussain SR, Lovergne L, Untereiner V, Hughes C, Lukaszewski RA, Thiéfin G, Sockalingum GD. Developing and understanding biofluid vibrational spectroscopy: a critical review. Chem. Soc. Rev. **2016**; 45: 1803–1818.

Baker MJ, Trevisan J, Bassan P, Bhargava R, Butler HJ, Dorling KM, Fielden PR, Fogarty SW, Fullwood NJ, Heys KA, Hughes C, Lasch P, Martin-Hirsch PL, Obinaju B, Sockalingum GD, Sulé-Suso J, Strong RJ, Walsh MJ, Wood BR, Gardner P, Martin FL. Using Fourier transform IR spectroscopy to analyze biological materials. Nat. Protoc. **2014**; 9(8): 1771–1791.

Ballabio D, Consonni V. Classification tools in chemistry. Part 1: linear models. PLS-DA. Anal. Methods **2013**; 5: 3790–3798.

Ballabio D, Grisoni F, Todeschini R. Multivariate comparison of classification performance measures. Chemometr. Intell. Lab. Syst. **2018**; 174: 33–44.

Baranska M, Roman M, Dobrowlski JC, Schulz H, Baranski R. Recent Advances in Raman Analysis of Plants: Alkaloids, Carotenoids, and Polyacetylenes. Curr. Anal. Chem. **2013**; 9(1): 108–127.

Barnes R, Dhanoa MS, Lister SJ. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl. Spectrosc. **1989**; 43(5): 772–777.

Barreiro P, Herrero D, Hernández N, Gracia A, León L. Calibration Transfer Between Portable and Laboratory NIR Spectrophotometers. Acta Hort. **2008**; 802: 373–378.

Bassan P, Byrne HJ, Bonnier F, Lee J, Dumas P, Gardner P. Resonant Mie scattering in infrared spectroscopy of biological materials – understanding the 'dispersion artefact'. Analyst **2009**; 134: 1586–1593.

Bassan P, Kohler A, Martens H, Lee J, Byrne HJ, Dumas P, Gazi E, Brown M, Clarke N, Gardner P. Resonant Mie scattering (RMieS) correction of infrared spectra from highly scattering biological samples. Analyst **2010**; 135: 268–277.

Bassan P, Mellor J, Shapiro J, Williams KJ, Lisanti MP, Gardner P. Transmission FT-IR Chemical Imaging on Glass Substrates: Applications in Infrared Spectral Histopathology. Anal. Chem. **2014**; 86(3): 1648–1653.

Beckonert O, Keun HC, Ebbels TM, Bundy J, Holmes E, Lindon JC, Nicholson JK. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. Nat. Protoc. **2007**; 2: 2692–2703.

Beebe KR, Pell RJ, Seasholtz MB. Chemometrics: A Practical Guide. John Wiley & Sons, Inc.: New York, **1998**.

Bin J, Li X, Fan W, Zhou JH, Wang CW. Calibration transfer of near-infrared spectroscopy by canonical correlation analysis coupled with wavelet transform. Analyst **2017**; 142(12): 2229–2238.

Bittner LK, Schonbichler SA, Bonn GK, Huck CW. Near Infrared Spectroscopy (NIRS) as a Tool to Analyze Phenolic Compounds in Plants. Curr. Anal. Chem. **2013**; 9(3): 417–423.

Bi X, Yang X, Bostrom MP, Bartusik D, Ramaswamy S, Fishbein KW, Spencer RG, Camacho NP. Fourier transform infrared imaging and MR microscopy studies detect compositional and structural changes in cartilage in a rabbit model of osteoarthritis. Anal. Bioanal. Chem. **2007**; 387(5): 1601–1612.

Bonifacio A, Cervo S, Sergo V. Label-free surface-enhanced Raman spectroscopy of biofluids: fundamental aspects and diagnostic applications. Anal. Bioanal. Chem. **2015**; 407(27): 8265–8277.

Bonifacio A, Dalla Marta S, Spizzo R, Cervo S, Steffan A, Colombatti A, Sergo V. Surface-enhanced Raman spectroscopy of blood plasma and serum using Ag and Au nanoparticles: a systematic study. Anal. Bioanal. Chem. **2014**; 406(9–10): 2355–2365.

Booksh KS, Kowalski BR. Theory of Analytical Chemistry. Anal. Chem. **1994**; 66(15): 782A–791A.

Bouveresse E, Massart DL, Dardenne P. Calibration transfer across near-infrared spectrometric instruments using Shenk's algorithm: effects of different standardisation samples. Anal. Chim. Acta **1994**; 297(3): 405–416.

Brereton RG. Chemometrics: Data Analysis for the Laboratory and Chemical Plant. John Wiley & Sons Ltd: Chichester, **2003**.

Brereton RG, Jansen J, Lopes J, Marini F, Pomerantsev A, Rodionova O, Roger JM, Walczak B, Tauler R. Chemometrics in analytical chemistry-part I: history, experimental design and data analysis tools. Anal. Bioanal. Chem. **2017**; 409(25): 5891–5899.

Brereton RG, Lloyd GR. Partial least squares discriminant analysis: taking the magic away. J. Chemometr. **2014**; 28(4): 213–225.

Brereton RG, Lloyd GR. Support Vector Machines for classification and regression. Analyst **2010**; 135: 230–267.

Bro R, Harshman RA, Sidiropoulos ND, Lundy ME. Modeling multi-way data with linearly dependent loadings. J. Chemometr. **2009**; 23(7–8): 324–340.

Bro R. PARAFAC. Tutorial and applications. Chemometr. Intell. Lab. Syst. **1997**; 38(2): 149–171.

Bro R, Smilde AK. Principal component analysis. Anal. Methods **2014**; 6: 2812–2831.

Brouckaert D, Uyttersprot JS, Broeckx W, De Beer T. Calibration transfer of a Raman spectroscopic quantification method from at-line to in-line assessment of liquid detergent compositions. Anal. Chim. Acta **2017**; 971: 14–25.

Brown CD, Green RL. Critical factors limiting the interpretation of regression vectors in multivariate calibration. Trends Anal. Chem. **2009**; 28(4): 506–514.

Brown CD, Wentzell PD. Hazards of digital smoothing filters as a preprocessing tool in multivariate calibration. J. Chemometr. **1999**; 13(2): 133–152.

Brown JQ, Vishwanath K, Palmer GM, Ramanujam N. Advances in quantitative UV–visible spectroscopy for clinical and pre-clinical application in cancer. Curr. Opin. Biotechnol. **2009**; 20(1): 119–131.

Bunchberger AR, DeLaney K, Johnson J, Li L. Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights. Anal. Chem. **2018**; 90(1): 240–265.

Buitrago MF, Skidmore AK, Groen TA, Hecker CA. Connecting infrared spectra with plant traits to identify species. ISPRS J. Photogramm. Remote Sens. **2018**; 139: 183–200.

Bunaciu AA, Aboul-Enein HY, Fleschin Ş. Vibrational Spectroscopy in Clinical Analysis. Appl. Spectrosc. Rev. **2014**; 50(2): 176–191.

Bury D, Faust G, Paraskevaidi M, Ashton KM, Dawson TP, Martin FL. Phenotyping Metastatic Brain Tumors Applying Spectrochemical Analyses: Segregation of Different Cancer Types. Anal. Lett. **2019a**; 52(4): 575–587.

Bury D, Morais CLM, Ashton KM, Dawson TP, Martin FL. Ex Vivo Raman Spectrochemical Analysis Using a Handheld Probe Demonstrates High Predictive Capability of Brain Tumour Status. Biosensors **2019b**; 9(2): 49.

Bury D, Morais CLM, Paraskevaidi M, Ashton KM, Dawson TP, Martin FL. Spectral classification for diagnosis involving numerous pathologies in a complex clinical setting: A neuro-oncology example. Spectrochim. Acta A Mol. Biomol. Spectrosc. **2019c**; 206: 89–96.

Butler HJ, Ashton L, Bird B, Cinque G, Curtis K, Dorney J, Esmonde-White K, Fullwood NJ, Gardner B, Martin-Hirsch PL, Walsh MJ, McAinsh MR, Stone N, Martin FL. Using Raman spectroscopy to characterize biological materials. Nat. Protoc. **2016**; 11(4): 664–687.

Byrne HJ, Knief P, Keating ME, Bonnier F. Spectral pre and post processing for infrared and Raman spectroscopy of biological tissues and cells. Chem. Soc. Rev. **2016**; 45: 1865–1878.

Cacuci DG, Ionescu-Bujor M. Sensitivity and Uncertainty Analysis, Data Assimilation, and Predictive Best-Estimate Model Calibration. In: Cacuci DG (Ed.) Handbook of Nuclear Engineering. Springer: Boston, **2010**, pp. 1913–2051.

Caja J, Gómez E, Maresca P. Optical measuring equipments. Part I: Calibration model and uncertainty estimation. Precision Engineering **2015**; 40: 298–304.

Calimag-Williams K, Knobel G, Goicoechea HC, Campiglia AD. Achieving second order advantage with multi-way partial least squares and residual bi-linearization with total synchronous fluorescence data of monohydroxy–polycyclic aromatic hydrocarbons in urine samples. Anal. Chim. Acta **2014**; 811: 60–69.

Callery EL, Morais CLM, Paraskevaidi M, Brusic V, Vijayadurai P, Anantharachagan A, Martin FL, Rowbottom AW. New approach to investigate Common Variable Immunodeficiency patients using spectrochemical analysis of blood. Sci. Rep. **2019;** 9: 7239.

Canvin JMG, Bernatsky S, Hitchon CA, Jackson M, Sowa MG, Mansfield JR, Eysel HH, Mantsch HH, El-Gabalawy HS. Infrared spectroscopy: shedding light on synovitis in patients with rheumatoid arthritis. Rheumatology **2003**; 42(1): 76–82.

Carmona P, Molina M, Calero M, Bermejo-Pareja F, Martínez-Martín P, Toledano A. Discrimination analysis of blood plasma associated with Alzheimer's disease using vibrational spectroscopy. J. Alzheimers Dis. **2013**; 34(4): 911–920.

Carmona P, Molina M, López-Tobar E, Toledano A. Vibrational spectroscopic analysis of peripheral blood plasma of patients with Alzheimer's disease. Anal. Bioanal. Chem. **2015**; 407(25): 7747–7756.

Carmona P, Monzón M, Monleón E, Badiola JJ, Monreal J. In vivo detection of scrapie cases from blood by infrared spectroscopy. J. Gen. Virol. **2005**; 86: 3425–3431.

Chan AW, Mercier P, Schiller D, Bailey R, Robbins S, Eurich DT, Sawyer MB, Broadhurst D. 1H-NMR urinary metabolomic profiling for diagnosis of gastric cancer. Br. J. Cancer **2016**; 114(1): 59–62.

Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. **2011**; 27: 1–27.

Chan KLA, Kazarian SG. Attenuated total reflection Fourier-transform infrared (ATR-FTIR) imaging of tissues and live cells. Chem. Soc. Rev. **2016**; 45: 1850–1864.

Chen WR, Bin J, Lu HM, Zhang ZM, Liang YZ. Calibration transfer via an extreme learning machine auto-encoder. Analyst **2016**; 141: 1973–1980.

Choo-Smith LP, Maquelin K, van Vreeswijk T, Bruining HA, Puppels GJ, Ngo Thi NA, Kirschner C, Naumann D, Ami D, Villa AM, Orsini F, Doglia SM, Lamfarraj H, Sockalingum GD, Manfait M, Allouch P, Endtz HP. Investigating microbial (micro)colony heterogeneity by vibrational spectroscopy. Appl. Environ. Microbiol. **2001**; 67(4): 1461–1469.

Comino F, Aranda V, García-Ruiz R, Ayora-Cañada MJ, Domínguez-Vidal A. Infrared spectroscopy as a tool for the assessment of soil biological quality inagricultural soils under contrasting management practices. Ecol. Indicators **2018**; 87: 117–126.

Coopman R, Van de Vyver T, Kishabongo AS, Katchunga P, Van Aken EH, Cikomola J, Monteyne T, Speeckaert MM, Delanghe JR. Glycation in human fingernail clippings using ATR-FTIR spectrometry, a new marker for the diagnosis and monitoring of diabetes mellitus. Clin. Biochem. **2017**; 50(1–2): 62–67.

Cordella CBY, Bertrand D. SAISIR: A new general chemometric toolbox. Trends Anal. Chem. **2014**; 54: 75–82.

Cortes C, Vapnik V. Support-vector networks. Mach. Learn. **1995**; 20: 273–297.

Costa FSL, Silva PP, Morais CLM, Arantes TD, Milan EP, Theodoro RC, Lima KMG. Attenuated total reflection Fourier transform-infrared (ATR-FTIR) spectroscopy as a new technology for discrimination between *Cryptococcus neoformans* and *Cryptococcus gattii*. Anal. Methods **2016**; 8: 7107–7115.

Costa FSL, Silva PP, Morais CLM, Theodoro RC, Arantes TD, Lima KMG. Comparison of multivariate classification algorithms using EEM fluorescence data to distinguish Cryptococcus neoformans and Cryptococcus gattii pathogenic fungi. Anal. Methods **2017**; 9: 3968–3976.

Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory **1967**; 13(1): 21–27.

Cozzolino D. Use of Infrared Spectroscopy for In-Field Measurement and Phenotyping of Plant Properties: Instrumentation, Data Analysis, and Examples. Appl. Spectrosc. **2014**; 49(7): 564–584.

Cui L, Yang K, Li HZ, Zhang H, Su JQ, Paraskevaidi M, Martin FL, Ren B, Zhu YG. Functional Single-Cell Approach to Probing Nitrogen-Fixing Bacteria in Soil

Communities by Resonance Raman Spectroscopy with 15N2 Labeling. Anal. Chem. **2018**; 90(8): 5082–5089.

da Silva AC, Soares SFC, Insausti M, Galvão RKH, Band BSF, de Araújo MCU. Two-dimensional linear discriminant analysis for classification of three-way chemical data. Anal. Chim. Acta **2016**; 938: 53–62.

da Silva JCGE, Oliveira CJS. Parafac decomposition of three-way kinetic-spectrophotometric spectral matrices corresponding to mixtures of heavy metal ions. Talanta **1999**; 49(4): 889–897.

David-Vaudey E, Burghardt A, Keshari K, Brouchet A, Ries M, Majumdar S. Fourier Transform Infrared Imaging of focal lesions in human osteoarthritic cartilage. Eur. Cell Mater. **2005**; 10: 51–60.

Day JS, Edwards HGM, Dobrowski SA, Voice AM. The detection of drugs of abuse in fingerprints using Raman spectroscopy I: latent fingerprints. Spectrochim. Acta A Mol. Biomol. Spectrosc. **2004**; 60(3): 563–568.

de Almeida MR, Correa DN, Rocha WFC, Scafi FJO, Poppi RJ. Discrimination between authentic and counterfeit banknotes using Raman spectroscopy and PLS-DA with uncertainty estimation. Microchem. J. **2013**; 109: 170–177.

de Andrade EWV, Morais CLM, Costa FSL, Lima KMG. A Multivariate Control Chart Approach for Calibration Transfer between NIR Spectrometers for Simultaneous Determination of Rifampicin and Isoniazid in Pharmaceutical Formulation. Curr. Anal. Chem. **2018**; 14(5): 488–494.

De Bruyne S, Speeckaert MM, Delanghe JR. . Applications of mid-infrared spectroscopy in the clinical laboratory setting. Crit. Rev. Clin. Lab. Sci. **2018**; 55(1): 1–20.

De Gussem K, De Gelder J, Vandenabeele P, Moens L. The Biodata toolbox for MATLAB. Chemom. Intell. Lab. Syst. **2009**; 95(1): 49–52.

de Jong S. SIMPLS: An alternative approach to partial least squares regression. Chemom. Intell. Lab. Syst. **1993**; 18(3): 251–263.

de Juan A, Tauler R, Dyson R, Marcolli C, Rault M, Meeder M. Spectroscopic imaging and chemometrics: a powerful combination for global and local sample analysis. Trends Anal. Chem. **2004**; 23(1): 70–79.

de Juan A, Tauler R. Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications. Crit. Rev. Anal. Chem. **2006**; 36(3–4): 163–176.

de Lima FA, Gobinet C, Sockalingum G, Garcia SB, Manfait M, Untereiner V, Piot O, Bachmann L. Digital de-waxing on FTIR images. Analyst **2017**; 142: 1358–1370.

Derenne A, Gasper R, Goormaghtigh E. The FTIR spectrum of prostate cancer cells allows the classification of anticancer drugs according to their mode of action. Analyst **2011**; 136: 1134–1141.

Desroches J, Jermyn M, Pinto M, Picot F, Tremblay MA, Obaid S, Marple E, Urmey K, Trudel D, Soulez G, Guiot MC, Wilson BC, Petrecca K, Leblond F. A new method using

Raman spectroscopy for in vivo targeted brain cancer tissue biopsy. Sci. Rep. **2018**; 8: 1792.

Diem M, Mazur A, Lenau K, Schubert J, Bird B, Miljković M, Krafft C, Popp J. Molecular pathology via IR and Raman spectral imaging. J. Biophotonics **2013**; 6(11–12): 855–886.

Dixon SJ, Brereton RG. Comparison of performance of five common classifiers represented as boundary methods: Euclidean Distance to Centroids, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Learning Vector Quantization and Support Vector Machines, as dependent on data structure. Chemometr. Intell. Lab. Syst. **2009**; 95(1): 1–17.

Domingues R, Filippone M, Michiardi P, Zouaoui J. A comparative evaluation of outlier detection algorithms: Experiments and analyses. Pattern Recognit. **2018**; 74: 406–421.

Duan P, Liu B, Morais CLM, Zhao J, Li X, Tu J, Yang W, Chen C, Long M, Feng X, Martin FL, Xiong C. 4-Nonylphenol effects on rat testis and sertoli cells determined by spectrochemical techniques coupled with chemometric analysis. Chemosphere **2019**; 218: 64–75.

Duchesne C, Liu JJ, MacGregor JF. Multivariate image analysis in the process industries: A review. Chemometr. Intell. Lab. Syst. **2012**; 117: 116–128.

Eliasson C, Matousek P. Noninvasive authentication of pharmaceutical products through packaging using spatially offset Raman spectroscopy. Anal. Chem. **2007**; 79(4): 1969–1701.

Eylenbosch D, Bodson B, Baeten V, Pierna JAF. NIR hyperspectral imaging spectroscopy and chemometrics for the discrimination of roots and crop residues extracted from soil samples. J. Chemometr. **2018**; 32(1): e2982.

Fan W, Liang Y, Yuan D, Wang J. Calibration model transfer for near-infrared spectra based on canonical correlation analysis. Anal. Chim. Acta **2008**; 623(1): 22–29.

Fawagreh K, Gaber MM, Elyan R. Random forests: from early developments to recent advancements. Syst. Sci. Control Eng. **2014**; 2(1): 602–609.

Felten J, Hall H, Jaumot J, Tauler R, de Juan A, Gorzsás A. Vibrational spectroscopic image analysis of biological material using multivariate curve resolution–alternating least squares (MCR-ALS). Nat. Protoc. **2015**; 10: 217–240.

Ferrés M, Platikanov S, Tsakovski S, Tauler R. Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. J. Chemometr. **2015**; 29(10): 528–536.

Feudale RN, Woody NA, Tan H, Myles AJ, Brown SD, Ferré J. Transfer of multivariate calibration models: a review. Chemometr. Intell. Lab. Syst. **2002**; 64(2): 181–192.

Forina M, Drava G, Armanino C, Boggia R, Lanteri S, Leardi R, Corti P, Conti P, Giangiacomo R, Galliena C, Bigoni R, Quartari I, Serra C, Ferri D, Leoni O, Lazzeri L. Transfer of calibration function in near-infrared spectroscopy. Chemometr. Intell. Lab. Syst. **1995**; 27(2): 189–203.

Frisoni GB, Fox NC, Jack CR Jr, Scheltens P, Thompson PM. The clinical use of structural MRI in Alzheimer disease. Nat. Rev. Neurol. **2010**; 6(2): 67–77.

Gajjar K, Ahmadzai AA, Valasoulis G, Trevisan J, Founta C, Nasioutziki M, Loufopoulos A, Kyrgiou M, Stasinou SM, Karakitsos P, Paraskevaidis E, Da Gama-Rose B, Martin-Hirsch PL, Martin FL. Histology Verification Demonstrates That Biospectroscopy Analysis of Cervical Cytology Identifies Underlying Disease More Accurately than Conventional Screening: Removing the Confounder of Discordance. PLoS One **2014**; 9(1): e82416.

Gajjar K, Heppenstall LD, Pang W, Ashton KM, Trevisan J, Patel II, Llabjani V, Stringfellow HF, Martin-Hirsch PL, Dawson T, Martin FL. Diagnostic segregation of human brain tumours using Fourier-transform infrared and/or Raman spectroscopy coupled with discriminant analysis. Anal. Methods **2013**; 5: 89–102.

Gallo M. Tucker3 Model for Compositional Data. Communications in Statistics – Theory and Methods **2015**; 44(21): 4441–4453.

Gazi E, Baker M, Dwyer J, Lockyer NP, Gardner P, Shanks JH, Reeve RS, Hart CA, Clarke NW, Brown MD. A correlation of FTIR spectra derived from prostate cancer biopsies with gleason grade and tumour stage. Eur. Urol. **2006**; 50(4): 750–760.

Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. Anal. Chim. Acta **1986**; 185: 1–17.

Geladi P, MacDougall D, Martens H. Linearization and scatter-correction for near-infrared reflectance spectra of meat. Appl. Spectrosc. **1985**; 39(3): 491–500.

Gemperline PJ, Miller KH, West TL, Weinstein JE, Hamilton JC, Bray JT. Principal component analysis, trace elements, and blue crab shell disease. Anal. Chem. **1992**; 64(9): 523A–532A.

Glassford SE, Byrne B, Kazarian SG. Recent applications of ATR FTIR spectroscopy and imaging to proteins. Biochim. Biophys. Acta **2013**; 1834(12): 2849–2858.

Golightly RS, Doering WE, Natan MJ. Surface-Enhanced Raman Spectroscopy and Homeland Security: A Perfect Match? ACS Nano **2009**; 3(10): 2859–2869.

Goodacre R, Timmins ÉM, Jones A, Kell DB, Maddock J, Heginbothom ML, Magee JT. On mass spectrometer instrument standardization and interlaboratory calibration transfer using neural networks. Anal. Chim. Acta **1997**; 348(1–3): 511–532.

Gowda GA, Zhang S, Gu H, Asiago V, Shanaiah N, Raftery D. Metabolomics-based methods for early disease diagnostics. Expert Rev. Mol. Diagn. **2008**; 8(5): 617–633.

Graça G, Moreira AS, Correia AJ, Goodfellow BJ, Barros AS, Duarte IF, Carreira IM, Galhano E, Pita C, Almeida Mdo C, Gil AM. Mid-infrared (MIR) metabolic fingerprinting of amniotic fluid: a possible avenue for early diagnosis of prenatal disorders? Anal. Chim. Acta **2013**; 764: 24–31.

Greenfield NJ. Using circular dichroism spectra to estimate protein secondary structure. Nat. Protoc. **2006**; 1: 2876–2890.

Greensill CV, Wolfs PJ, Spiegelman CH, Walsh KB. Calibration Transfer between PDA-Based NIR Spectrometers in the NIR Assessment of Melon Soluble Solids Content. Appl. Spectrosc. **2001**; 55(5): 647–653.

Grimard V, Li C, Ramjeesingh M, Bear CE, Goormaghtigh E, Ruysschaert JM. Phosphorylation-induced conformational changes of cystic fibrosis transmembrane conductance regulator monitored by attenuated total reflection-Fourier transform IR spectroscopy and fluorescence spectroscopy. J. Biol. Chem. **2004**; 279(7): 5528–5536.

Großerueschkamp F, Kallenbach-Thieltges A, Behrens T, Brüning T, Altmayer M, Stamatis G, Theegarten D, Gerwert K. Marker-free automated histopathological annotation of lung tumour subtypes by FTIR imaging. Analyst **2015**; 140: 2114–2120.

Guo S, Heinke R, Stöckel S, Rösch P, Bocklitz T, Popp J. Towards an improvement of model transferability for Raman spectroscopy in biological applications. Vib. Spectrosc. **2017**; 91: 111–118.

Hammody Z, Sahu RK, Mordechai S, Cagnano E, Argov S. Characterization of Malignant Melanoma Using Vibrational Spectroscopy. Sci. World J. **2005**; 5: 173–182.

Hands JR, Clemens G, Stables R, Ashton KM, Brodbelt A, Davis C, Dawson TP, Jenkinson MD, Lea RW, Walker C, Baker MJ. Brain tumour differentiation: rapid stratified serum diagnostics via attenuated total reflection Fourier-transform infrared spectroscopy. J. Neurooncol. **2016**; 127(3): 463–472.

Hands JR, Dorling KM, Abel P, Ashton KM, Brodbelt A, Davis C, Dawson T, Jenkinson MD, Lea RW, Walker C, Baker MJ. Attenuated total reflection fourier transform infrared (ATR-FTIR) spectral discrimination of brain tumour severity from serum samples. J. Biophotonics **2014**; 7(3–4): 189–199.

Hargreaves MD, Matousek P. Threat detection of liquid explosive precursor mixtures by Spatially Offset Raman Spectroscopy (SORS). Proc. SPIE 7486, Optics and Photonics for Counterterrorism and Crime Fighting V **2009**; 74860B.

Harmsen S, Wall MA, Huang R, Kircher MF. Cancer imaging using surface-enhanced resonance Raman scattering nanoparticles. Nat. Protoc. **2017**; 12(7): 1400–1414.

Harrigan GG, LaPlante RH, Cosma GN, Cockerell G, Goodacre R, Maddox JF, Luyendyk JP, Ganey PE, Roth RA. Application of high-throughput Fourier-transform infrared spectroscopy in toxicology studies: contribution to a study on the development of an animal model for idiosyncratic toxicity. Toxicol. Lett. **2004**; 146(3): 197–205.

Harrington PB. Support Vector Machine Classification Trees. Anal. Chem. **2015**; 87(21): 11065–11071.

Hasegawa J, Nakamura M, Matsuoka R, Mimura T, Ichizuka K, Sekizawa A, Okai T. Evaluation of placental function using near infrared spectroscopy during fetal growth restriction. J. Perinat. Med. **2010**; 38(1): 29–32.

Hastie T, Tibshinari R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd edn. Springer: New York, **2009**.

Helm D, Labischinski H, Schallehn G, Naumann D. Classification and identification of bacteria by Fourier-transform infrared spectroscopy. J. Gen. Microbiol. **1991**; 137(1): 69–79.

Henrion R. N-way principal component analysis theory, algorithms and applications. Chemometr. Intell. Lab. Syst. **1994**; 25(1): 1–23.

Heys KA, Shore RF, Pereira MG, Martin FL. Vibrational biospectroscopy characterizes biochemical differences between cell types used for toxicological investigations and identifies alterations induced by environmental contaminants. Environ. Toxicol. Chem. **2017**; 36(11): 3127–3137.

Hibbert DB. Vocabulary of concepts and terms in chemometrics (IUPAC Recommendations 2016). Pure Appl. Chem. **2016**; 88(4): 407–443.

Hofmann-Wellenhof B, Lichtenegger H, Collins J. Global positioning system: theory and practice. Springer Science & Business Media, **2012**.

HORIBA. Raman Spectroscopy for Analysis and Monitoring. < https://static.horiba.com/fileadmin/Horiba/Technology/Measurement_Techniques/Molecular_Spectroscopy/Raman_Spectroscopy/Raman_Academy/Raman_Tutorial/Raman_bands.pdf>. Accessed on 01 February 2020.

Hu Q, Lü X, Lu W, Chen Y, Liu H. An extensive study on Raman spectra of water from 253 to 753 K at 30 MPa: A new insight into structure of water. J. Mol. Spectrosc. **2013**; 292: 23–27.

Hu R, Xia J. Calibration transfer of near infrared spectroscopy based on DS algorithm. 2011 International Conference on Electric Information and Control Engineering **2011**; 3062–3065.

Hu Y, Peng S, Bi Y, Tang L. Calibration transfer based on maximum margin criterion for qualitative analysis using Fourier transform infrared spectroscopy. Analyst **2012**; 137: 5913–5918.

Ibrahim O, Maguire A, Meade AD, Flint S, Toner M, Byrne HJ, Lyng FM. Improved protocols for pre-processing Raman spectra of formalin fixed paraffin preserved tissue sections. Anal. Methods **2017**; 9: 4709–4717.

Isabelle M, Dorney J, Lewis A, Lloyd GR, Old O, Shepherd N, Rodriguez-Justo M, Barr H, Lau K, Bell I, Ohrel S, Thomas G, Stone N, Kendall C. Multi-centre Raman spectral mapping of oesophageal cancer tissues: a study to assess system transferability. Faraday Discuss. **2016**; 187: 87–103.

Jacyna J, Kordalewska M, Markuszewski MJ. Design of Experiments in metabolomics-related studies: An overview. J. Pharm. Biomed. Anal. **2019**; 164: 598–606.

Jagust W, Reed B, Mungas D, Ellis W, Decarli C. What does fluorodeoxyglucose PET imaging add to a clinical diagnosis of dementia? Neurology **2007**; 69(9): 871–877.

Jarvis RM, Broadhurst D, Johnson H, O'Boyle NM, Goodacre R. PYCHEM: a multivariate analysis package for python. Bioinformatics **2006**; 22(20): 2565–2566.

Jarvis RM, Goodacre R. Discrimination of Bacteria Using Surface-Enhanced Raman Spectroscopy. Anal. Chem. **2004**; 76(1): 40–47.

Jaumot J, de Juan A, Tauler R. MCR-ALS GUI 2.0: New features and applications. Chemom. Intell. Lab. Syst. **2015**; 140: 1–12.

Jayson GC, Kohn EC, Kitchener HC, Ledermann JA. Ovarian cancer. Lancet **2014**; 384(9951): 1376–1388.

Jiang F, Liu G, Du J, Sui Y. Initialization of K-modes clustering using outlier detection techniques. Inf. Sci. **2016**; 332: 167–183.

Jing R, Sun J, Wang Y, Li M, Pu X. PML: A parallel machine learning toolbox for data classification and regression. Chemom. Intell. Lab. Syst. **2014**; 138: 1–6.

Jin H, Lu Q, Chen X, Ding H, Gao H, Jin S. The use of Raman spectroscopy in food processes: A review. Appl. Spectrosc. **2015**; 51(1): 12–22.

Jones S, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg. Med. J. **2003**; 20(5): 453–458.

Kalita J, Misra U. Comparison of CT scan and MRI findings in the diagnosis of Japanese encephalitis. J. Neurol. Sci. **2000**; 174(1): 3–8.

Kamandar M, Ghassemian H. Maximum relevance, minimum redundancy feature extraction for hyperspectral images. 2010 18th Iranian Conference on Electrical Engineering **2010**; 254–259.

Kandpal LM, Cho BK, Tewari J, Gopinathan N. Raman spectral imaging technique for API detection in pharmaceutical microtablets. Sens. Actuators B Chem. **2018**; 260: 213–222.

Karoui R, Downey G, Blecker C. Mid-infrared spectroscopy coupled with chemometrics: a tool for the analysis of intact food systems and the exploration of their molecular structure-quality relationships - a review. Chem. Rev. **2010**; 110(10): 6144–6168.

Kato H, Faria TN, Stannard B, Roberts CT Jr, LeRoith D. Role of tyrosine kinase activity in signal transduction by the insulin-like growth factor-I (IGF-I) receptor. Characterization of kinase-deficient IGF-I receptors and the action of an IGF-I-mimetic antibody (alpha IR-3). J. Biol. Chem. **1993**; 268(4): 2655–2661.

Kelly JG, Trevisan J, Scott AD, Carmichael PL, Pollock HM, Martin-Hirsch PL, Martin FL. Biospectroscopy to metabolically profile biomolecular structure: a multistage approach linking computational analysis with biomarkers. J. Proteome Res. **2011**; 10(4): 1437–1448.

Kendall C, Isabelle M, Bazant-Hegemark F, Hutchings J, Orr L, Babrah J, Baker R, Stone N. Vibrational spectroscopy: a clinical tool for cancer diagnostics. Analyst **2009**; 134: 1029–1045.

Kennard RW, Stone LA. Computer Aided Design of Experiments. Technometrics **1969**; 11(1): 137–148.

Khaydukova M, Medina-Plaza C, Rodriguez-Mendez ML, Panchuk V, Kirsanov D, Legin A. Multivariate calibration transfer between two different types of multisensor systems. Sensors Actuators B Chem. **2017**; 246: 994–100.

Khoshmanesh A, Dixon MWA, Kenny S, Tilley L, McNaughton, Wood BR. Detection and Quantification of Early-Stage Malaria Parasites in Laboratory Infected Erythrocytes by Attenuated Total Reflectance Infrared Spectroscopy and Multivariate Analysis. Anal. Chem. **2014**; 86(9): 4379–4386.

Kiefer W, Popp J, Lankers M, Trunk M, Hartmann I, Urlaub E, Musick J. Raman-Mie scattering from single laser trapped microdroplets. J. Mol. Struct. **1997**; 408–409: 113–120.

Kirsch M, Schackert G, Salzer R, Krafft C. Raman spectroscopic imaging for in vivo detection of cerebral brain metastases. Anal. Bioanal. Chem. **2010**; 398(4): 1707–1713.

Koehler FW, Small GW, Combs RJ, Knapp RB, Kroutil RT. Calibration Transfer Algorithm for Automated Qualitative Analysis by Passive Fourier Transform Infrared Spectrometry. Anal. Chem. **2000**; 72(7): 1690–1698.

Kondepati VR, Keese M, Mueller R, Menegold BC, Backhaus J. Application of near-infrared spectroscopy for the diagnosis of colorectal cancer in resected human tissue specimens. Vib. Spectrosc. **2007**; 44(2): 236–242.

Kong L, Zhang P, Wang G, Yu J, Setlow P, Li YQ. Characterization of bacterial spore germination using phase-contrast and fluorescence microscopy, Raman spectroscopy and optical tweezers. Nat. Protoc. **2011**; 6(5): 625–639.

Kroonenberg PM, Basford KE, Gemperline PJ. Grouping three-mode data with mixture methods: the case of the diseased blue crabs. J. Chemometr. **2004**; 18(11): 508–518.

Kuligowski J, Quintás G, Herwig C, Lendl B. A rapid method for the differentiation of yeast cells grown under carbon and nitrogen-limited conditions by means of partial least squares discriminant analysis employing infrared micro-spectroscopic data of entire yeast cells. Talanta **2012**; 99: 566–573.

Kundu J, Le F, Nordlander P, Halas NJ. Surface enhanced infrared absorption (SEIRA) spectroscopy on nanoshell aggregate substrates. Chem. Phys. Lett. **2008**; 452(1–3): 115–119.

Lagleyre S, Sorrentino T, Calmels MN, Shin YJ, Escudé B, Deguine O, Fraysse B. Reliability of high-resolution CT scan in diagnosis of otosclerosis. Otol. Neurotol. **2009**; 30(8): 1152–1159.

Lane R, See SS. Attenuated total reflectance Fourier transform infrared spectroscopy method to differentiate between normal and cancerous breast cells. J. Nanosci. Nanotechnol. **2012**; 12(9): 7395–7400.

Lasch P, Naumann D. Infrared Spectroscopy in Microbiology. Encyclopedia Anal. Chem. **2015**; 102–131.

Lechowicz L, Chrapek M, Gaweda J, Urbaniak M, Konieczna I. Use of Fourier-transform infrared spectroscopy in the diagnosis of rheumatoid arthritis: a pilot study. Mol. Biol. Rep. **2016**; 43(12): 1321–1326.

LeCun Y, Bengio Y, Hinton G. Deep learning. Nature **2015**; 521: 436–444.

Lee SS, Kim AY, Yang SK, Chung JW, Kim SY, Park SH, Ha HK. Crohn disease of the small bowel: comparison of CT enterography, MR enterography, and small-bowel follow-through as diagnostic techniques. Radiology **2009**; 251(3): 751–761.

Lewis IR, Daniel NW Jr, Chaffin NC, Griffiths PR, Tungol MW. Raman spectroscopic studies of explosive materials: towards a fieldable explosives detector. Spectrochim. Acta A Mol. Biomol. Spectrosc. **1995**; 51(12): 1985–2000.

Lewis PD, Lewis KE, Ghosal R, Bayliss S, Lloyd AJ, Wills J, Godfrey R, Kloer P, Mur LA. Evaluation of FTIR spectroscopy as a diagnostic tool for lung cancer using sputum. BMC Cancer **2010**; 10: 640.

Li-Chan ECY. The applications of Raman spectroscopy in food science. Trends Food Sci. Tech. **1996**; 7(11): 361–370.

Li HD, Xu QS, Liang YZ. libPLS: An integrated library for partial least squares regression and linear discriminant analysis. Chemom. Intell. Lab. Syst. **2018**; 176: 34–43.

Liland KH, Kohler A, Afseth NK. Model-based pre-processing in Raman spectroscopy of biological samples. J. Raman Spectrosc. **2016**; 47(6): 643–650.

Lindon JC, Tranter GE, Koppenaal DW. Encyclopedia of Spectroscopy and Spectrometry. 3rd edn. Academic Press: Oxford, **2017**.

Liu HB, Zhong H, Karpowicz N, Chen Y, Zhang XC. Terahertz Spectroscopy and Imaging for Defense and Security Applications. Proceedings of the IEEE **2007**; 95(8): 1514–1527.

Liu S. Matrix results on the Khatri-Rao and Tracy-Singh products. Linear Algebra Appl. **1999**; 289(1–3): 267–277.

Livermore LJ, Isabelle M, Bell IM, Scott C, Walsby-Tickle J, Gannon J, Plaha P, Vallance C, Ansorge O. Rapid intraoperative molecular genetic classification of gliomas using Raman spectroscopy. Neurooncol. Adv. **2019**; 1(1): vdz008.

Li YN, Wu HL, Qing XD, Nie CC, Li SF, Yu YJ, Zhang SR, Yu RQ. The maintenance of the second-order advantage: Second-order calibration of excitation–emission matrix fluorescence for quantitative analysis of herbicide napropamide in various environmental samples. Talanta **2011**; 85(1): 325–332.

Llabjani V, Jones KC, Thomas GO, Walker LA, Shore RF, Martin FL. Polybrominated diphenyl ether-associated alterations in cell biochemistry as determined by attenuated total reflection Fourier-transform infrared spectroscopy: a comparison with DNA-reactive and/or endocrine-disrupting agents. Environ. Sci. Technol. **2009**; 43(9): 3356–3364.

Logiurato F. Relativistic Derivations of de Broglie and Planck-Einstein Equations. J. Mod. Phys. **2014**; 5: 1–7.

Lohr D, Tillmann P, Druege U, Zerche S, Rath T, Meinken E. Non-destructive determination of carbohydrate reserves in leaves of ornamental cuttings by near-infrared spectroscopy (NIRS) as a key indicator for quality assessments. Biosys. Eng. **2017**; 158: 51–63.

Lohumi S, Kim MS, Qin J, Cho BK. Improving Sensitivity in Raman Imaging for Thin Layered and Powdered Food Analysis Utilizing a Reflection Mirror. Sensors **2019**; 19(12): 2698.

Lorenz B, Wichmann C, Stöckel S, Rösch P, Popp J. Cultivation-Free Raman Spectroscopic Investigations of Bacteria. Trends Microbiol. **2017**; 25(5): 413–424.

Lovergne L, Bouzy P, Untereiner V, Garnotel R, Baker MJ, Thiéfin G, Sockalingum GD. Biofluid infrared spectro-diagnostics: pre-analytical considerations for clinical applications. Faraday Discuss. **2016**; 187: 521–537.

Lui H, Zhao J, McLean D, Zeng H. Real-time Raman spectroscopy for in vivo skin cancer diagnosis. Cancer Res. **2012**; 72(10): 2491–2500.

Luo X, Ikehata A, Sashida K, Piao S, Okura T, Terada Y. Calibration transfer across near infrared spectrometers for measuring hematocrit in the blood of grazing cattle. J. Near Infrared Spec. **2017**; 25(1): 15–25.

Macleod NA, Matousek P. Emerging non-invasive Raman methods in process control and forensic applications. Pharm. Res. **2008**; 25(10): 2205–2215.

Maitra I, Morais CLM, Lima KMG, Ashton KM, Date RS, Martin FL. Attenuated total reflection Fourier-transform infrared spectral discrimination in human bodily fluids of oesophageal transformation to adenocarcinoma. Analyst **2019**; 144: 7447–7456.

Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J. Thorac. Oncol. **2010**; 5(9): 1315–1316.

Maquelin K, Kirschner C, Choo-Smith LP, van den Braak N, Endtz HP, Naumann D, Puppels GJ. Identification of medically relevant microorganisms by vibrational spectroscopy. J. Microbiol. Methods **2002**; 51(3): 255–271.

Marini F, Bucci R, Magrì AL, Magrì AD. Artificial neural networks in chemometrics: History, examples and perspectives. Microchem. J. **2008**; 88(2): 178–185.

Marini F. Classification Methods in Chemometrics. Curr. Anal. Chem. **2010**; 6(1): 72–79.

Markus APJA, Swinkels DW, Jakobs BS, Wevers RA, Trijbels JMF, Willems HL. New technique for diagnosis and monitoring of alcaptonuria: quantification of homogentisic acid in urine with mid-infrared spectrometry. Anal. Chim. Acta **2001**; 429(2): 287–292.

Marsili NR, Lista A, Band BS, Goicoechea HC, Olivieri AC. Evaluation of complex spectral-pH three-way arrays by modified bilinear least-squares: determination of four different dyes in interfering systems. Analyst **2005**; 130: 1291–1298.

Martens H, Høy M, Wise BM, Bro R, Brockhoff PB. Pre-whitening of data by covariance-weighted pre-processing. J. Chemometr. **2003**; 17(3): 153–165.

Martens H, Martens M. Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). Food Qual. Prefer. **2000**; 11(1–2): 5–16.

Martin FL, German MJ, Wit E, Fearn T, Ragavan N, Pollock HM. Identifying variables responsible for clustering in discriminant analysis of data from infrared microspectroscopy of a biological sample. J. Comput. Biol. **2007**; 14(9): 1176–1184.

Martin FL, Kelly JG, Llabjani V, Martin-Hirsch PL, Patel II, Trevisan J, Fullwood NJ, Walsh MJ. Distinguishing cell types or populations based on the computational analysis of their infrared spectra. Nat. Protoc. **2010**; 5(11): 1748–1760.

Martin M, Perez-Guaita D, Andrew DW, Richards JS, Wood BR, Heraud P. The effect of common anticoagulants in detection and quantification of malaria parasitemia in human red blood cells by ATR-FTIR spectroscopy. Analyst **2017**; 142: 1192–1199.

Mayo DW, Miller FA, Hannah RW. Course Notes on the Interpretation of Infrared and Raman Spectra. John Wiley & Sons: Hoboken, **2003**.

Maziak DE, Do MT, Shamji FM, Sundaresan SR, Perkins DG, Wong PT. Fourier-transform infrared spectroscopic study of characteristic molecular structure in cancer cells of esophagus: an exploratory study. Cancer Detect. Prev. **2007**; 31(3): 244–253.

McCall J. Genetic algorithms for modelling and optimisation. J. Comput. Appl. Math. **2005**; 184(1): 205–222.

McIntosh LM, Jackson M, Mantsch HH, Stranc MF, Pilavdzic D, Crowson AN. Infrared spectra of basal cell carcinomas are distinct from non-tumor-bearing skin components. J. Invest. Dermatol. **1999**; 112(6): 951–956.

McIntosh LM, Summers R, Jackson M, Mantsch HH, Mansfield JR, Howlett M, Crowson AN, Toole JW. Towards non-invasive screening of skin lesions by near-infrared spectroscopy. J. Invest. Dermatol. **2001**; 116(1): 175–181.

Meade AD, Clarke C, Draux F, Sockalingum GD, Manfait M, Lyng FM, Byrne HJ. Studies of chemical fixation effects in human cell lines using Raman microspectroscopy. Anal. Bioanal. Chem. **2010**; 396(5): 1781–1791.

Mehrotra R, Tyagi G, Jangir DK, Dawar R, Gupta N. Analysis of ovarian tumor pathology by Fourier Transform Infrared Spectroscopy. J. Ovarian Res. **2010**; 3: 27.

Mehta K, Atak A, Sahu A, Srivastava S, Krishna C M. An early investigative serum Raman spectroscopy study of meningioma. Analyst **2018**; 143: 1916–1923.

Meksiarun P, Ishigaki M, Huck-Pezzei VAC, Huck CW, Wongravee K, Sato H, Ozaki Y. Comparison of multivariate analysis methods for extracting the paraffin component from the paraffin-embedded cancer tissue spectra for Raman imaging. Sci. Rep. **2017**; 7: 44890.

Melin AM, Perromat A, Déléris G. Pharmacologic application of Fourier transform IR spectroscopy: in vivo toxicity of carbon tetrachloride on rat liver. Biopolymers **2000**; 57(3): 160–168.

Mendel J. Statistical methods in analytical chemistry. J. Chem. Educ. **1949**; 26(10): 534.

Menon U, Gentry-Maharaj A, Hallett R, Ryan A, Burnell M, Sharma A, Lewis S, Davies S, Philpott S, Lopes A, Godfrey K, Oram D, Herod J, Williamson K, Seif MW, Scott I, Mould T, Woolas R, Murdoch J, Dobbs S, Amso NN, Leeson S, Cruickshank D, McGuire A, Campbell S, Fallowfield L, Singh N, Dawnay A, Skates SJ, Parmar M, Jacobs I. Sensitivity and specificity of multimodal and ultrasound screening for ovarian cancer, and stage distribution of detected cancers: results of the prevalence screen of the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). Lancet Oncol. **2009**; 10(4): 327–340.

Merás ID, Manzano JD, Rodríguez DA, Peña AM. Detection and quantification of extra virgin olive oil adulteration by means of autofluorescence excitation-emission profiles combined with multi-way classification. Talanta **2018**; 178: 751–762.

Micsonai A, Wien F, Kernya L, Lee YH, Goto Y, Réfrégiers M, Kardos J. Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. Proc. Natl. Acad. Sci. U.S.A. **2015**; 112(24): E3095–E3103.

Miles AJ, Wallace BA. Circular dichroism spectroscopy of membrane proteins. Chem. Soc. Rev. **2016**; 45: 4859–4872.

Mitchell AL, Gajjar KB, Theophilou G, Martin FL, Martin-Hirsch PL. Vibrational spectroscopy of biofluids for disease screening or diagnosis: translation from the laboratory to a clinical setting. J. Biophotonics **2014**; 7(3–4): 153–165.

Mitchell BL, Yasui Y, Li CI, Fitzpatrick AL, Lampe PD. Impact of freeze-thaw cycles and storage time on plasma samples used in mass spectrometry based biomarker discovery projects. Cancer Inform. **2005**; 1: 98–104.

Mobaraki N, Amigo JM. HYPER-Tools. A graphical user-friendly interface for hyperspectral image analysis. Chemom. Intell. Lab. Syst. **2018**; 172: 174–187.

Möller-Hartmann W, Herminghaus S, Krings T, Marquardt G, Lanfermann H, Pilatus U, Zanella FE. Clinical application of proton magnetic resonance spectroscopy in the diagnosis of intracranial mass lesions. Neuroradiology **2002**; 44(5): 371–381.

Monakhova YB, Diehl BWK. Transfer of multivariate regression models between high-resolution NMR instruments: application to authenticity control of sunflower lecithin. Magn. Reson. Chem. **2016**; 54(9): 712–717.

Morais CLM, Costa FSL, Lima KMG. Variable selection with a support vector machine for discriminating *Cryptococcus* fungal species based on ATR-FTIR spectroscopy. Anal. Methods **2017**; 9: 2964–2970.

Morais CLM, Lima KMG. Comparing unfolded and two-dimensional discriminant analysis and support vector machines for classification of EEM data. Chemom. Intell. Lab. Syst. **2017**; 170: 1–12.

Morais CLM, Lima KMG. Determination and analytical validation of creatinine content in serum using image analysis by multivariate transfer calibration procedures. Anal. Methods **2015**; 7: 6904–6910.

Morais CLM, Lima KMG, Martin FL. TTWD-DA: A MATLAB toolbox for discriminant analysis based on trilinear three-way data. Chemom. Intell. Lab. Syst. **2019a**; 188: 46–53.

Morais CLM, Lima KMG, Martin FL. Uncertainty estimation and misclassification probability for classification models based on discriminant analysis and support vector machines. Anal. Chim. Acta **2019b**; 1063: 40–46.

Morais CLM, Lima KMG. Principal Component Analysis with Linear and Quadratic Discriminant Analysis for Identification of Cancer Samples Based on Mass Spectrometry. J. Braz. Chem. Soc. **2018**; 29(3): 472–481.

Morais CLM, Martin FL, Lima KMG. A computational protocol for sample selection in biological-derived infrared spectroscopy datasets using Morais-Lima-Martin (MLM) algorithm. Protocol Exchange, **2018a**. DOI: 10.1038/protex.2018.141.

Morais CLM, Paraskevaidi M, Cui L, Fullwood NJ, Isabelle M, Lima KMG, Martin-Hirsch PL, Sreedhar H, Trevisan J, Walsh MJ, Zhang D, Zhu YG, Martin FL. Standardization of complex biologically derived spectrochemical datasets. Nat. Protoc. **2019c**; 14: 1546–1577.

Morais CLM, Santos MCD, Lima KMG, Martin FL. Improving data splitting for classification applications in spectrochemical analyses employing a random-mutation Kennard-Stone algorithm approach. Bioinformatics **2019d**; 35(24): 5257–5263.

Morais CLM, Martin-Hirsch PL, Martin FL. A three-dimensional principal component analysis approach for exploratory analysis of hyperspectral data: identification of ovarian cancer samples based on Raman microspectroscopy imaging of blood plasma. Analyst **2019e**; 144: 2312–2319.

Morais CLM, Shore RF, Pereira MG, Martin FL. Assessing Binary Mixture Effects from Genotoxic and Endocrine Disrupting Environmental Contaminants Using Infrared Spectroscopy. ACS Omega **2018b**, 3(10): 13399–13412.

Mordechai S, Sahu RK, Hammody Z, Mark S, Kantarovich K, Guterman H, Podshyvalov A, Goldstein J, Argov S. Possible common biomarkers from FTIR microspectroscopy of cervical cancer and melanoma. J. Microsc. **2004**; 215: 86–91.

Morris P, Perkins A. Diagnostic imaging. Lancet **2012**; 379(9825): 1525–1533.

Mostaço-Guidolin LB, Mukarami LS, Nomizo A, Bachmann L. Fourier Transform Infrared Spectroscopy of Skin Cancer Cells and Tissues. Appl. Spectrosc. Rev. **2009**; 44(5): 438–455.

Movasaghi Z, Rehman S, ur Rehman I. Fourier Transform Infrared (FTIR) Spectroscopy of Biological Tissues. Appl. Spectrosc. Rev. **2008**; 43(2): 134–179.

McCall J. Genetic algorithms for modelling and optimisation. J. Comput. Appl. Math. **2006**; 184(1): 205–222.

Miller LM, Dumas P. From structure to cellular mechanism with infrared microspectroscopy. Curr. Opin. Struct. Biol. **2010**; 20(5): 649–656.

Naes T, Isaksson T, Fearn T, Davies T. A User-Friendly Guide to Multivariate Calibration an Classification. NIR Publications: Chichester, **2002**.

Nasdala L, Smith DC, Kaindl R, Ziemann MA. Raman spectroscopy: Analytical perspectives in mineralogical research. In: Beran A, Libowitzky E (Ed.) Spectroscopic methods in mineralogy. Eötvös University Press: Budapest, **2004**, pp. 281-343.

Naumann D, Helm D, Labischinski H. Microbiological characterizations by FT-IR spectroscopy. Nature **1991**; 351: 81–82.

Navani N, Nankivell M, Lawrence DR, Lock S, Makker H, Baldwin DR, Stephens RJ, Parmar MK, Spiro SG, Morris S, Janes SM, Lung-BOOST trial investigations. Lung cancer diagnosis and staging with endobronchial ultrasound-guided transbronchial needle aspiration compared with conventional approaches: an open-label, pragmatic, randomised controlled trial. Lancet Respir. Med. **2015**; 3(4): 282–289.

Neves ACO, Morais CLM, Mendes TPP, Vaz BG, Lima KMG. Mass spectrometry and multivariate analysis to classify cervical intraepithelial neoplasia from blood plasma: an untargeted lipidomic study. Sci. Rep. **2018**; 8: 3954.

Neves ACO, Silva PP, Morais CLM, Miranda CG, Crispim JCO, Lima KMG. ATR-FTIR and multivariate analysis as a screening tool for cervical cancer in women from northeast Brazil: a biospectroscopic approach. RSC Adv. **2016**; 6: 99648–99655.

Ni L, Han M, Luan S, Zhang L. Screening wavelengths with consistent and stable signals to realize calibration model transfer of near infrared spectra. Spectrochim. Acta A Mol. Biomol. Spectrosc. **2019**; 206: 350–358.

Nilsen H, Esaiassen M, Heia K, Sigernes F. Visible/Near-Infrared Spectroscopy: A New Tool for the Evaluation of Fish Freshness? J. Food Sci. **2002**; 67(5): 1821–1826.

Nørgaard L, Saudland A, Wagner J, Nielsen JP, Munck L, Engelsen SB. Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. Appl. Spectrosc. **2000**; 54(3): 413–419.

Obinaju BE, Martin FL. Novel biospectroscopy sensor technologies towards environmental health monitoring in urban environments. Environ. Pollut. **2013**; 183: 46–53.

Oemrawsingh RM, Cheng JM, García-García HM, van Geuns RJ, de Boer SP, Simsek C, Kardys I, Lenzen MJ, van Domburg RT, Regar E, Serruys PW, Akkerhuis KM, Boersma E; ATHEROREMO-NIRS Investigators. Near-infrared spectroscopy predicts cardiovascular outcome in patients with coronary artery disease. J. Am. Coll. Cardiol. **2014**; 64(23): 2510–2518.

Ofner J, Kamilli KA, Eitenberger E, Friedbacher G, Lendl B, Held A, Lohninger H. Chemometric Analysis of Multisensor Hyperspectral Images of Precipitated Atmospheric Particulate Matter. Anal. Chem. **2015**; 87(18): 9413–9420.

Olmos V, Marro M, Loza-Alvarez P, Raldúa D, Prats E, Padrós F, Piña B, Tauler R, de Juan A. Combining hyperspectral imaging and chemometrics to assess and interpret the effects of environmental stressors on zebrafish eye images at tissue level. J. Biophotonics **2018**; 11(3): e201700089.

Osborne BG. Near-infrared spectroscopy in food analysis. Encyclopedia Anal. Chem. **2000**; 5: 4069–4082.

Owens GL, Gajjar K, Trevisan J, Fogarty SW, Taylor SE, Da Gama-Rose B, Martin-Hirsch PL, Martin FL. Vibrational biospectroscopy coupled with multivariate analysis extracts potentially diagnostic features in blood plasma/serum of ovarian cancer patients. J. Biophotonics **2014**; 7(3–4): 200–209.

Paatero P, Tapper U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics **1994**; 5(2): 111–126.

Palmnas MSA, Vogel HJ. The Future of NMR Metabolomics in Cancer Therapy: Towards Personalizing Treatment and Developing Targeted Drugs? Metabolites **2013**; 3(2): 373–396.

Palonpon AF, Ando J, Yamakoshi H, Dodo K, Sodeoka M, Kawata S, Fujita K. Raman and SERS microscopy for molecular imaging of live cells. Nat. Protoc. **2013**; 8: 677–692.

Panchuk V, Kirsanov D, Oleneva E, Semenov V, Legin A. Calibration transfer between different analytical methods. Talanta **2017**; 170: 457–463.

Pan YL. Detection and characterization of biological and other organic-carbon aerosol particles in atmosphere using fluorescence. J. Quant. Spectrosc. Radiat. Transfer **2015**; 150: 12–35.

Paraskevaidi M, Ashton KM, Stringfellow HF, Wood NJ, Keating PJ, Rowbottom AW, Martin-Hirsch PL, Martin FL. Raman spectroscopic techniques to detect ovarian cancer biomarkers in blood plasma. Talanta **2018a**; 189: 281–288.

Paraskevaidi M, Martin-Hirsch PL, Martin FL. ATR-FTIR Spectroscopy Tools for Medical Diagnosis and Disease Investigation. Kumar CSSR (Ed.) Nanotechnology Characterization Tools for Biosensing and Medical Diagnosis. Springer: Berlin, **2017a**, pp. 163–211.

Paraskevaidi M, Morais CLM, Freitas DLD, Lima KMG, Mann DMA, Allsop D, Martin-Hirsch PL, Martin FL. Blood-based near-infrared spectroscopy for the rapid low-cost detection of Alzheimer's disease. Analyst **2018b**; 143: 5959–5964.

Paraskevaidi M, Morais CLM, Lima KMG, Snowden JS, Saxon JA, Richardson AMT, Jones M, Mann DMA, Allsop D, Martin-Hirsch PL, Martin FL. Differential diagnosis of Alzheimer’s disease using spectrochemical analysis of blood. Proc. Natl. Acad. Sci. U.S.A. **2017b**; 114(38): E7929–E7938.

Paraskevaidi M, Morais CLM, Raglan O, Lima KMG, Paraskevaidis E, Martin-Hirsch PL, Kyrgiou M, Martin FL. Aluminium foil as an alternative substrate for the spectroscopic interrogation of endometrial cancer. J. Biophotonics **2018c**; 11(7): e201700372.

Paraskevaidi M, Morais CLM, Lima KMG, Ashton KM, Stringfellow HF, Martin-Hirsch PL, Martin FL. Potential of mid-infrared spectroscopy as a non-invasive diagnostic test in urine for endometrial or ovarian cancer. Analyst **2018d**; 143(13): 3156–3163.

Parikh KS, Shah TP. Support Vector Machine – a Large Margin Classifier to Diagnose Skin Illnesses. Procedia Technol. **2016**; 23: 369–375.

Pasquini C. Near infrared spectroscopy: A mature analytical technique with new perspectives - A review. Anal. Chim. Acta **2018**; 1026: 8–36.

Patil P, Dasgupta B. Role of diagnostic ultrasound in the assessment of musculoskeletal diseases. Ther. Adv. Musculoskelet Dis. **2012**; 4(5): 341–355.

Pavia DL, Lampman GM, Kriz GS, Vyvyan JA. Introduction to Spectroscopy. Cengage Learning: Belmont, **2008**.

Pence I, Mahadevan-Jansen A. Clinical instrumentation and applications of Raman spectroscopy. Chem. Soc. Rev. **2016**; 45(7): 1958–1979.

Penido CAFO, Pacheco MTT, Lednev IK, Silveira L Jr. Raman spectroscopy in forensic analysis: identification of cocaine and other illegal drugs of abuse. J. Raman Spectrosc. **2016**; 47(1): 28–38.

Pérez NF, Ferré J, Boqué R. Calculation of the reliability of classification in discriminant partial least-squares binary classification. Chemom. Intell. Lab. Syst. **2009**; 95(2): 122–128.

Peters AS, Backhaus J, Pfützner A, Raster M, Burgard G, Demirel S, Böckler D, Hakimi M. Serum-infrared spectroscopy is suitable for diagnosis of atherosclerosis and its clinical manifestations. Vib. Spectrosc. **2017**; 92: 20–26.

Pierna JAF, Vermeulen P, Amand O, Tossens A, Dardenne P, Baeten V. NIR hyperspectral imaging spectroscopy and chemometrics for the detection of undesirable substances in food and feed. Chemometr. Intell. Lab. Syst. **2012**; 177: 233–239.

Pilling M, Gardner P. Fundamental developments in infrared spectroscopic imaging for biomedical applications. Chem. Soc. Rev. **2016**; 45: 1935–1957.

Podshyvalov A, Sahu RK, Mark S, Kantarovich K, Guterman H, Goldstein J, Jagannathan R, Argov S, Mordechai S. Distinction of cervical cancer biopsies by use of infrared microspectroscopy and probabilistic neural networks. Appl. Opt. **2005**; 44(18): 3725–3734.

Pomerantsev AL. Acceptance areas for multivariate classification derived by projection methods. J. Chemometr. **2008**; 22(11–12): 601–609.

Pomerantsev AL, Rodionova OY. Multiclass partial least squares discriminant analysis: Taking the right way—A critical tutorial. J. Chemometr. **2018**; 32(8): e3030.

Porro-Muñoz D, Duin RPW, Talavera I, Orozco-Alzate M. Classification of three-way data by the dissimilarity representation. Signal Process. **2011**; 91(11): 2520–2529.

Prats-Montalbán JM, de Juan A, Ferrer A. Multivariate image analysis: A review with applications. Chemometr. Intell. Lab. Syst. **2011**; 107(1): 1–23.

Prieto N, Pawluczyk O, Dugan MER, Aalhus JL. A Review of the Principles and Applications of Near-Infrared Spectroscopy to Characterize Meat, Fat, and Meat Products. Appl. Spectrosc. **2017**; 71(7): 1403–1426.

Pu YY, Feng YZ, Sun DW. Recent Progress of Hyperspectral Imaging on Quality and Safety Inspection of Fruits and Vegetables: A Review. Compr. Rev. Food Sci. F. **2015**; 14(2): 176–188.

Quintelas C, Mesquita DP, Lopes JA, Ferreira EC, Sousa C. Near-infrared spectroscopy for the detection and quantification of bacterial contaminations in pharmaceutical products. Int. J. Pharm. **2015**; 492(1–2): 199–206.

Qu JH, Liu D, Cheng JH, Sun DW, Ma J, Pu H, Zeng XA. Applications of near-infrared spectroscopy in food safety evaluation and control: a review of recent research advances. Crit. Rev. Food. Sci. Nutr. **2015**; 55(13): 1939–1954.

Radovic M, Ghalwash M, Filipovic N, Obradovic Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC Bioinformatics **2017**; 18: 9.

Reisner LA, Cao A, Pandya AK. An integrated software system for processing, analyzing, and classifying Raman spectra. Chemom. Intell. Lab. Syst. **2011**; 105(1): 83–90.

Reusch W. Virtual Textbook of Organic Chemistry, **1999**. < https://www2.chemistry.msu.edu/faculty/reusch/VirtTxtJml/intro1.htm>. Accessed on 01 February 2020.

Rigas B, Morgello S, Goldman IS, Wong PT. Human colorectal cancers display abnormal Fourier-transform infrared spectra. Proc. Natl. Acad. Sci. U.S.A. **1990**; 87(20): 8140–8144.

Rocha WFC, Sheen DA. Classification of biodegradable materials using QSAR modelling with uncertainty estimation. SAR QSAR Environ. Res. **2016**; 27(10): 799–811.

Rodrigues RRT, Rocha JTC, Oliveira MSL, Dias JCM, Müller EI, Castro EVR, Filgueiras PR. Evaluation of calibration transfer methods using the ATR-FTIR technique to predict density of crude oil. Chemom. Intell. Lab. Syst. **2017**; 166: 7–13.

Rodriguez JD, Westenberger BJ, Buhse LF, Kauffman JF. Standardization of Raman spectra for transfer of spectral libraries across different instruments. Analyst **2011**; 136(20): 4232–4240.

Rodriguez-Saona LE, Khambaty FM, Fry FS, Calvery EM. Rapid Detection and Identification of Bacterial Strains By Fourier Transform Near-Infrared Spectroscopy. J. Agric. Food Chem. **2001**; 49(2): 574–579.

Rossel RAV. ParLeS: Software for chemometric analysis of spectroscopic data. Chemom. Intell. Lab. Syst. **2008**; 90(1): 72–83.

Rousseeuw PJ, Hubert M. Robust statistics for outlier detection. Wiley Interdiscip. Rev. Data Min. **2011**; 1(1): 73–79.

Roy S, Perez-Guaita D, Andrew DW, Richards JS, McNaughton D, Heraud P, Wood BR. Simultaneous ATR-FTIR Based Determination of Malaria Parasitemia, Glucose and Urea in Whole Blood Dried onto a Glass Slide. Anal. Chem. **2017**; 89(10): 5238–5245.

Ryder AG. Classification of narcotics in solid mixtures using principal component analysis and Raman spectroscopy. J. Forensic Sci. **2002**; 47(2): 275–284.

Sádecká J, Jakubíková M, Májek P. Fluorescence spectroscopy for discrimination of botrytized wines. Food Control **2018**; 88: 75–84.

Sakudo A. Near-infrared spectroscopy for medical applications: Current status and future perspectives. Clin. Chim. Acta **2016**; 455: 181–188.

Santos MCD, Morais CLM, Nascimento YM, Araujo JMG, Lima KMG. Spectroscopy with computational analysis in virological studies: A decade (2006–2016). Trends Anal. Chem. **2017**; 97: 244–256.

Santos MCD, Nascimento YM, Monteiro JD, Alves BEB, Melo MF, Paiva AAP, Pereira HWB, Medeiros LG, Morais IC, Fagundes Neto JC, Fernandes JV, Araújo JMG, Lima KMG. ATR-FTIR spectroscopy with chemometric algorithms of multivariate classification in the discrimination between healthy vs. dengue vs. chikungunya vs. zika clinical samples. Anal. Methods **2018**; 10: 1280–1285.

Sattlecker M, Stone N, Smith J, Bessant C. Assessment of robustness and transferability of classification models built for cancer diagnostics using Raman spectroscopy. J. Raman Spectrosc. **2011**; 42(5): 897–903.

Savitzky A, Golay MJ. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. **1964**: 36(8): 1627–1639.

Schmitt J, Flemming HC. FTIR-spectroscopy in microbial and material analysis. Int. Biodeterior. Biodegratation **1998**; 41(1): 1–11.

Schrevens L, Lorent N, Dooms C, Vansteenkiste J. The role of PET scan in diagnosis, staging, and management of non-small cell lung cancer. Oncologist **2004**; 9(6): 633–643.

Scott DA, Renaud DE, Krishnasamy S, Meriç P, Buduneli N, Çetinkalp Ş, Liu KZ. Diabetes-related molecular signatures in infrared spectra of human saliva. Diabetol. Metab. Syndr. **2010**; 2: 48.

Scotter C. Use of near infrared spectroscopy in the food industry with particular reference to its applications to on/in-line food processes. Food Control **1990**; 1(3): 142–149.

Seasholtz MB, Kowalski B. The parsimony principle applied to multivariate calibration. Anal. Chim. Acta **1993**; 277(2): 165–177.

Semoun O, Guigui B, Tick S, Coscas G, Soubrane G, Souied EH. Infrared features of classic choroidal neovascularisation in exudative age-related macular degeneration. Br. J. Ophthalmol. **2009**; 93(2): 182–185.

Shahzad A, Knapp M, Edetsberger M, Puchinger M, Gaubitzer E, Köhler G. Diagnostic Application of Fluorescence Spectroscopy in Oncology Field: Hopes and Challenges. Appl. Spectrosc. Rev. **2010**; 45(1): 92–99.

Shahzad A, Köhler G, Knapp M, Gaubitzer E, Puchinger M, Edetsberger M. Emerging applications of fluorescence spectroscopy in medical microbiology field. J. Transl. Med. **2009**; 7: 99.

Shenk JS, Westerhaus MO. Populations Structuring of Near Infrared Spectra and Modified Partial Least Squares Regression. Crop Sci. **1991**; 31(6): 1548–1555.

Shin D, Vigneswaran N, Gillenwater A, Richards-Kortum R. Advances in fluorescence imaging techniques to detect oral cancer and its precursors. Future Oncol. **2010**; 6(7): 1143–1154.

Sieroń A, Sieroń-Stołtny K, Kawczyk-Krupka A, Latos W, Kwiatek S, Straszak D, Bugaj AM. The role of fluorescence diagnosis in clinical practice. Onco. Targets Ther. **2013**; 6: 977–982.

Siqueira LFS, Araújo Júnior RF, de Araújo AA, Morais CLM, Lima KMG. LDA vs. QDA for FT-MIR prostate cancer tissue classification. Chemom. Intell. Lab. Syst. **2017**; 162: 123–129.

Siqueira LFS, Lima KMG. A decade (2004 – 2014) of FTIR prostate cancer spectroscopy studies: An overview of recent advancements. Trends Anal. Chem. **2016a**; 82: 208–221.

Siqueira LFS, Lima KMG. MIR-biospectroscopy coupled with chemometrics in cancer studies. Analyst **2016b**; 141: 4833–4847.

Siqueira LFS, Morais CLM, Araújo Júnior RF, de Araújo AA, Lima KMG. SVM for FT-MIR prostate cancer classification: An alternative to the traditional methods. J. Chemometr. **2018**; 32(12): e3075.

Sitole L, Steffens F, Krüger TPJ, Meyer D. Mid-ATR-FTIR Spectroscopic Profiling of HIV/AIDS Sera for Novel Systems Diagnostics in Global Health. OMICS **2014**; 18(8): 513–523.

Sjöblom J, Svensson O, Josefson M, Kullberg H, Wold S. An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. Chemom. Intell. Lab. Syst. **1998**; 44(1–2): 229–244.

Skoog DA, Holler FJ, Crouch SR. Principles of Instrumental Analysis. 6th edn. Thomson Brooks/Cole: Belmont, **2007**.

Smith-Bindman R, Kerlikowske K, Feldstein VA, Subak L, Scheidler J, Segal M, Brand R, Grady D. Endovaginal ultrasound to exclude endometrial cancer and other endometrial abnormalities. JAMA **1998**; 280(17): 1510–1517.

Smith GP, McGoverin CM, Fraser SJ, Gordon KC. Raman imaging of drug delivery systems. Adv. Drug Deliv. Rev. **2015**; 89: 21–41.

Soares SFC, Gomes AA, Araujo MCU, Galvão Filho AR, Galvão RKH. The successive projections algorithm. Trends Anal. Chem. **2013**; 42: 84–98.

Song X, Li H, Al-Qadiri HM, Lin M. Detection of herbicides in drinking water by surface-enhanced Raman spectroscopy coupled with gold nanostructures. J. Food Meas. Charact. **2013**; 7(3): 107–113.

Sreedhar H, Varma VK, Nguyen PL, Davidson B, Akkina S, Guzman G, Setty S, Kajdacsy-Balla A, Walsh MJ. High-definition Fourier Transform Infrared (FT-IR) spectroscopic imaging of human tissue sections towards improving pathology. J. Vis. Exp. **2015**; 95: 52332.

Stelzle F, Knipfer C, Adler W, Rohde M, Oetter N, Nkenke E, Schmidt M, Tangermann-Gerk K. Tissue discrimination by uncorrected autofluorescence spectra: a proof-of-principle study for tissue-specific laser surgery. Sensors **2013**; 13(10): 13717–13731.

Stelzle F, Rohde M, Riemann M, Oetter N, Adler W, Tangermann-Gerk K, Schmidt M, Knipfer C. Autofluorescence spectroscopy for nerve-sparing laser surgery of the head and neck-the influence of laser-tissue interaction. Lasers Med. Sci. **2017**; 32: 1289–1300.

Stöckel S, Kirchhoff J, Neugebauer U, Rösch P, Popp J. The application of Raman spectroscopy for the detection and identification of microorganisms. J. Raman Spectrosc. **2016**; 47(1): 89–109.

Strola SA, Baritaux JC, Schultz E, Simon AC, Allier C, Espagnon I, Jary D, Dinten JM. Single bacteria identification by Raman spectroscopy. J. Biomed. Opt. **2014**; 19(11): 111610.

Sulub Y, LoBrutto R, Vivilecchia R, Wabuyele BW. Content uniformity determination of pharmaceutical tablets using five near-infrared reflectance spectrometers: a process analytical technology (PAT) approach using robust multivariate calibration transfer algorithms. Anal. Chim. Acta **2008**; 611(2): 143–150.

Takahashi Y, Wanibuchi M, Kimura Y, Akiyama Y, Mikami T, Mikuni N. Meningioma Originating from the Hypoglossal Canal: Case Report and Review of Literature. World Neurosurg. **2019**; 127: 525–529.

Taylor SE, Cheung KT, Patel II, Trevisan J, Stringfellow HF, Ashton KM, Wood NJ, Keating PJ, Martin-Hirsch PL, Martin FL. Infrared spectroscopy with multivariate analysis to interrogate endometrial tissue: a novel and objective diagnostic approach. Br. J. Cancer **2011**; 104(5): 790–797.

Tfayli A, Gobinet C, Vrabie V, Huez R, Manfait M, Piot O. Digital dewaxing of Raman signals: discrimination between nevi and melanoma spectra obtained from paraffin-embedded skin biopsies. Appl. Spectrosc. **2009**; 63(5): 564–570.

Theelen T, Berendschot TTJM, Hoyng CB, Boon CJF, Klevering BJ. Near-infrared reflectance imaging of neovascular age-related macular degeneration. Graefes Arch. Clin. Exp. Ophthalmol. **2009**; 247(12): 1625–1633.

Theophilou G, Lima KMG, Briggs M, Martin-Hirsch PL, Stringfellow HF, Martin FL. A biospectroscopic analysis of human prostate tissue obtained from different time periods points to a trans-generational alteration in spectral phenotype. Sci. Rep. **2015**; 5: 13465.

Theophilou G, Lima KMG, Martin-Hirsch PL, Stringfellow HF, Martin FL. ATR-FTIR spectroscopy coupled with chemometric analysis discriminates normal, borderline and malignant ovarian tissue: classifying subtypes of human cancer. Analyst **2016**; 141(2): 585–594.

Theophilou G, Morais CLM, Halliwell DE, Lima KMG, Drury J, Martin-Hirsch PL, Stringfellow HF, Hapangama DK, Martin FL. Synchrotron- and focal plane array-based Fourier-transform infrared spectroscopy differentiates the basalis and functionalis epithelial endometrial regions and identifies putative stem cell regions of human endometrial glands. Anal. Bioanal. Chem. **2018**; 410(18): 4541–4554.

Tian Z, Bing N, Xie L, Wang L, Yuan H. Raman Microscopy and Imaging in Pharmaceutical Applications. 2011 Third Int. Conf. Meas. Technol. Mechatronics Autom. **2011**; pp. 943–947.

Trevisan J, Angelov PP, Carmichael PL, Scott AD, Martin FL. Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives. Analyst **2012**; 137: 3202–3215.

Trevisan J, Angelov PP, Patel II, Najand GM, Cheung KT, Llabjani V, Pollock HM, Bruce SW, Pant K, Carmichael PL, Scott AD, FL Martin. Syrian hamster embryo (SHE) assay (pH 6.7) coupled with infrared spectroscopy and chemometrics towards toxicological assessment. Analyst **2010**; 135: 3266–3272.

Trevisan J, Angelov PP, Scott AD, Carmichael PL, Martin FL. IRootLab: a free and open-source MATLAB toolbox for vibrational biospectroscopy data analysis. Bioinformatics **2013**; 29(8): 1095–1097.

Tucker LR. Some mathematical notes on three-mode factor analysis. Psychometrika **1966**; 31: 279–311.

Türker-Kaya S, Huck CW. A Review of Mid-Infrared and Near-Infrared Imaging: Principles, Concepts and Applications in Plant Tissue Analysis. Molecules **2017**; 22(1): E168.

Utzinger U, Heintzelman DL, Mahadevan-Jansen A, Malpica A, Follen M, Richards-Kortum R. Near-Infrared Raman Spectroscopy for in Vivo Detection of Cervical Precancers. Appl. Spectrosc. **2001**; 55(8): 955–959.

van der Voet. Pseudo-degrees of freedom for complex predictive models: the example of partial least squares. J. Chemometr. **1999**; 13(3–4): 195–208.

Van Loan CF. The ubiquitous Kronecker product. J. Comput. Appl. Math. **2000**; 123(1–2): 85–100.

Varma VK, Kajdacsy-Balla A, Akkina S, Setty S, Walsh MJ. A label-free approach by infrared spectroscopic imaging for interrogating the biochemistry of diabetic nephropathy progression. Kidney Int. **2016**; 89(5): 1153–1159.

Varmuza K, Filzmoser P. Introduction to Multivariate Statistical Analysis in Chemometrics. 1st edn. CRC Press: Boca Raton, **2009**.

Varriale A, Rossi M, Staiano M, Terpetschnig E, Barbieri B, Rossi M, D'Auria S. Fluorescence Correlation Spectroscopy Assay for Gliadin in Food. Anal. Chem. **2007**; 79(12): 4687–4689.

Vaughan AA, Dunn WB, Allwood JW, Wedge DC, Blackhall FH, Whetton AD, Dive C, Goodacre R. Liquid chromatography-mass spectrometry calibration transfer and metabolomics data fusion. Anal. Chem. **2012**; 84(22): 9848–9857.

Vergote GJ, Vervaet C, Remon JP, Haemers T, Verpoort F. Near-infrared FT-Raman spectroscopy as a rapid analytical tool for the determination of diltiazem hydrochloride in tablets. Eur. J. Pharm. Sci. **2002**; 16(1–2): 63–67.

Wallace BA, Wien F, Miles AJ, Lees JG, Hoffmann SV, Evans P, Wistow GJ, Slingsby C. Biomedical applications of synchrotron radiation circular dichroism spectroscopy: identification of mutant proteins associated with disease and development of a reference database for fold motifs. Faraday Discuss **2004**; 126: 237–243.

Wallace RM. Analysis of Absorption Spectra of Multicomponent Systems. J. Phys. Chem. **1960**; 64(7): 899–901.

Walsh MJ, German MJ, Singh M, Pollock HM, Hammiche A, Kyrgiou M, Stringfellow HF, Paraskevaidis E, Martin-Hirsch PL, Martin FL. IR microspectroscopy: potential applications in cervical cancer screening. Cancer Lett. **2007**; 246(1–2): 1–11.

Walsh MJ, Kajdacsy-Balla A, Holton SE, Bhargava R. Attenuated total reflectance Fourier-transform infrared spectroscopic imaging for breast histopathology. Vib. Spectrosc. **2012**; 60: 23–28.

Wang J, Geng YJ, Guo B, Klima T, Lal BN, Willerson JT, Casscells W. Near-infrared spectroscopic characterization of human advanced atherosclerotic plaques. J. Am. Coll. Cardiol. **2002**; 39(8): 1305–1313.

Wang JS, Shi JS, Xu YZ, Duan XY, Zhang L, Wang J, Yang LM, Weng SF, Wu JG. FT-IR spectroscopic analysis of normal and cancerous tissues of esophagus. World J. Gastroenterol. **2003**; 9(9): 1897–1899.

Wang Y, Veltkamp DJ, Kowalski BR. Multivariate instrument standardization. Anal. Chem. **1991**; 63(23): 2750–2756.

Wang Z, Dean T, Kowalski BR. Additive Background Correction in Multivariate Instrument Standardization. Anal. Chem. **1995**; 67(14): 2379–2385.

Warrens MJ. Cohen's kappa is a weighted average. Stat. Methodol. **2011**; 8(6): 473–484.

Weber G. Enumeration of Components in Complex Systems by Fluorescence Spectrophotometry. Nature **1961**; 190: 27–29.

Wehrens R. Chemometrics with R: Multivariate Data Analysis in the Natural Sciences and Life Sciences. Springer: New York, 2011.

Wehrens R, Putter H, Buydens LMC. The bootstrap: a tutorial. Chemometr. Intell. Lab. Syst. **2000**; 54(1): 35–52.

Weiss R, Palatinszky M, Wagner M, Niessner R, Elsner M, Seidel M, Ivleva NP. Surface-enhanced Raman spectroscopy of microorganisms: limitations and applicability on the single-cell level. Analyst **2019**; 144: 943–953.

Witze ES, Old WM, Resing KA, Ahn NG. Mapping protein post-translational modifications with mass spectrometry. Nat. Methods **2007**; 4(10): 798–806.

Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. **2001**; 58(2): 109–130.

Wold S, Sjöström M. SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy. In: Kowalski BR (Ed.) Chemometrics: Theory and Application. American Chemical Society: Washington, **1977**, pp. 243–282.

Wood BR, Quinn MA, Burden FR, McNaughton D. An investigation into FTIR spectroscopy as a biodiagnostic tool for cervical cancer. Biospectroscopy **1996**; 2(3): 143–153.

Woody NA, Feudale RN, Myles AJ, Brown SD. Transfer of multivariate calibrations between four near-infrared spectrometers using orthogonal signal correction. Anal. Chem. **2004**; 76(9): 2595–2600.

World Health Organization. Fluorescence microscopy for disease diagnosis and environmental monitoring, **2005**.

Wu W, Mallet Y, Walczak B, Penninckx W, Massart DL, Heuerding S, Erni F. Comparison of regularized discriminant analysis linear discriminant analysis and quadratic discriminant analysis applied to NIR data. Anal. Chim. Acta **1996**; 329(3): 257–265.

Xiao H, Sun K, Sun Y, Wei K, Tu K, Pan L. Comparison of Benchtop Fourier-Transform (FT) and Portable Grating Scanning Spectrometers for Determination of Total Soluble Solid Contents in Single Grape Berry (Vitis vinifera L.) and Calibration Transfer. Sensors **2017**; 17(11): E2693.

Xie Y, Hopke PK. Calibration transfer as a data reconstruction problem. Anal. Chim. Acta **1999**; 384(2): 193–205.

Yahaya OKM, MatJafri MZ, Aziz AA, Omar AF. Visible spectroscopy calibration transfer model in determining pH of Sala mangoes. J. Instrum. **2015**; 10: T05002.

Yang H, Yang S, Kong J, Dong A, Yu S. Obtaining information about protein secondary structures in aqueous solution using Fourier transform IR spectroscopy. Nat. Protoc. **2015**; 10: 382–396.

Yang PW, Hsu IJ, Chang CW, Wang YC, Hsieh CY, Shih KH, Wong LF, Shih NY, Hsieh MS, Hou MTK, Lee JM. Visible-absorption spectroscopy as a biomarker to predict treatment response and prognosis of surgically resected esophageal cancer. Sci. Rep. **2016**; 6: 33414.

Yang Q, Zhang L, Wang L, Xiao H. MultiDA: Chemometric software for multivariate data analysis based on Matlab. Chemom. Intell. Lab. Syst. **2012**; 116: 1–8.

Yao H, Shi X, Zhang Y. The Use of FTIR-ATR Spectrometry for Evaluation of Surgical Resection Margin in Colorectal Cancer: A Pilot Study of 56 Samples. J. Spectrosc. **2014**; 2014: 4.

Yaseen T, Sun DW, Cheng JH. Raman imaging for food quality and safety evaluation: Fundamentals and applications. Trends Food Sci. Technol. **2017**; 62: 177–189.

Yeo Y, Park C, Lee JW, Kang Y, Ahn JM, Kang HS, Lee E. Magnetic resonance imaging spectrum of spinal meningioma. Clin. Imaging **2019**; 55: 100–106.

Yu B, Ji H, Kang Y. Standardization of near infrared spectra based on multi-task learning. Spectrosc. Lett. **2016**; 49(1): 23–29.

Zamora-Rojas E, Pérez-Marín D, De Pedro-Sanz E, Guerrero-Ginel JE, Garrido-Varo A. Handheld NIRS analysis for routine meat quality control: Database transfer from at-line instruments. Chemom. Intell. Lab. Syst. **2012**; 114: 30–35.

Zarnowiec P, Lechowicz Ł, Czerwonka G, Kaca W. Fourier Transform Infrared Spectroscopy (FTIR) as a Tool for the Identification and Differentiation of Pathogenic Bacteria. Curr. Med. Chem. **2015**; 22(14): 1710–1718.

Zhang L, Small GW, Arnold MA. Multivariate calibration standardization across instruments for the determination of glucose by Fourier transform near-infrared spectrometry. Anal. Chem. **2003**; 75(21): 5905–5915.

Zhang Y, Nayak TR, Hong H, Cai W. Biomedical applications of zinc oxide nanomaterials. Curr. Mol. Med. **2013**; 13(10): 1633–1645.

Zhou M, Johnson N, Gruner S, Ecklund GW, Meunier P, Byrn S, Glissmeyer M, Steinbock K. Clinical utility of breast-specific gamma imaging for evaluating disease extent in the newly diagnosed breast cancer patient. Am. J. Surg. **2009**; 197(2): 159–163.

Zhou Y, Liu CH, Sun Y, Pu Y, Boydston-White S, Liu Y, Alfano RR. Human brain cancer studied by resonance Raman spectroscopy. J. Biomed. Opt. **2012**; 17(11): 116021.

Zimmermann B, Bağcıoğlu M, Sandt C, Kohler A. Vibrational microspectroscopy enables chemical characterization of single pollen grains as well as comparative analysis of plant species based on pollen ultrastructure. Planta **2015**; 242: 1237–1250.

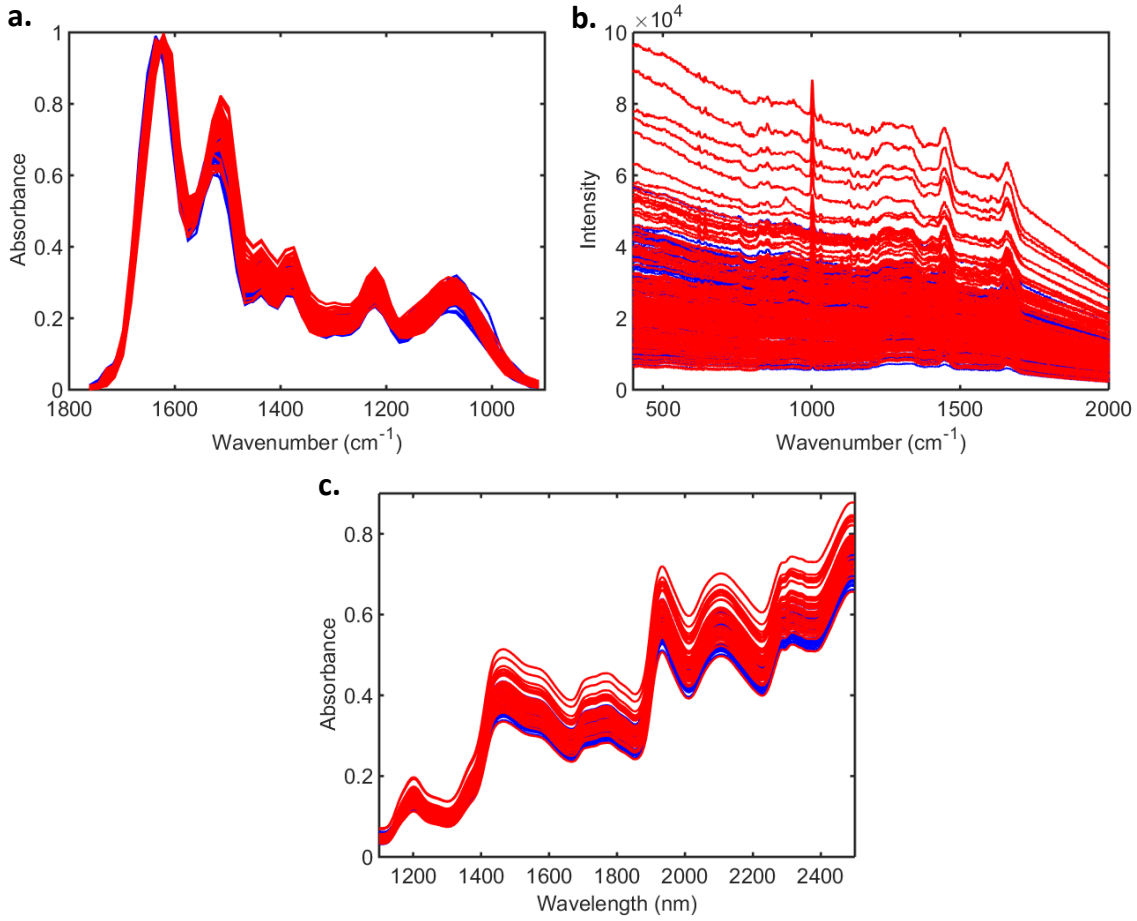
Zontov YV, Rodionova OY, Kucheryavskiy SV, Pomerantsev AL. DD-SIMCA – A MATLAB GUI tool for data driven SIMCA approach. Chemom. Intell. Lab. Syst. **2017**; 167: 23–28.

Zuo Q, Xiong S, Chen ZP, Chen Y, Yu RQ. A novel calibration strategy based on background correction for quantitative circular dichroism spectroscopy. Talanta **2017**; 174: 320–324.

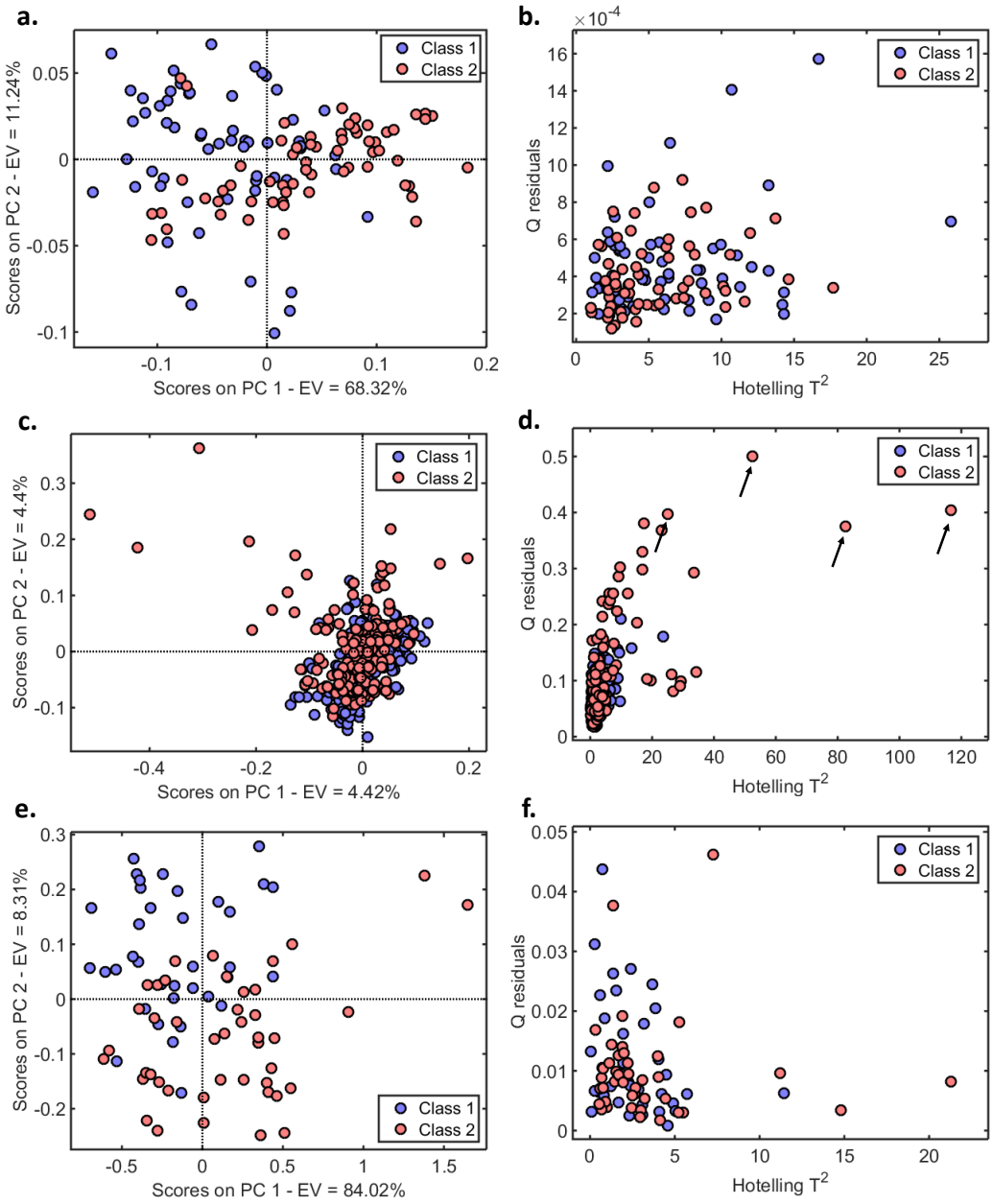
Zuzak KJ, Schaeberle MD, Lewis EN, Levin IW. Visible Reflectance Hyperspectral Imaging: Characterization of a Noninvasive, in Vivo System for Determining Tissue Perfusion. Anal. Chem. **2002**; 74(9): 2021–2028.

APPENDIX A – SUPPLEMENTARY MATERIAL FOR CHAPTER 2

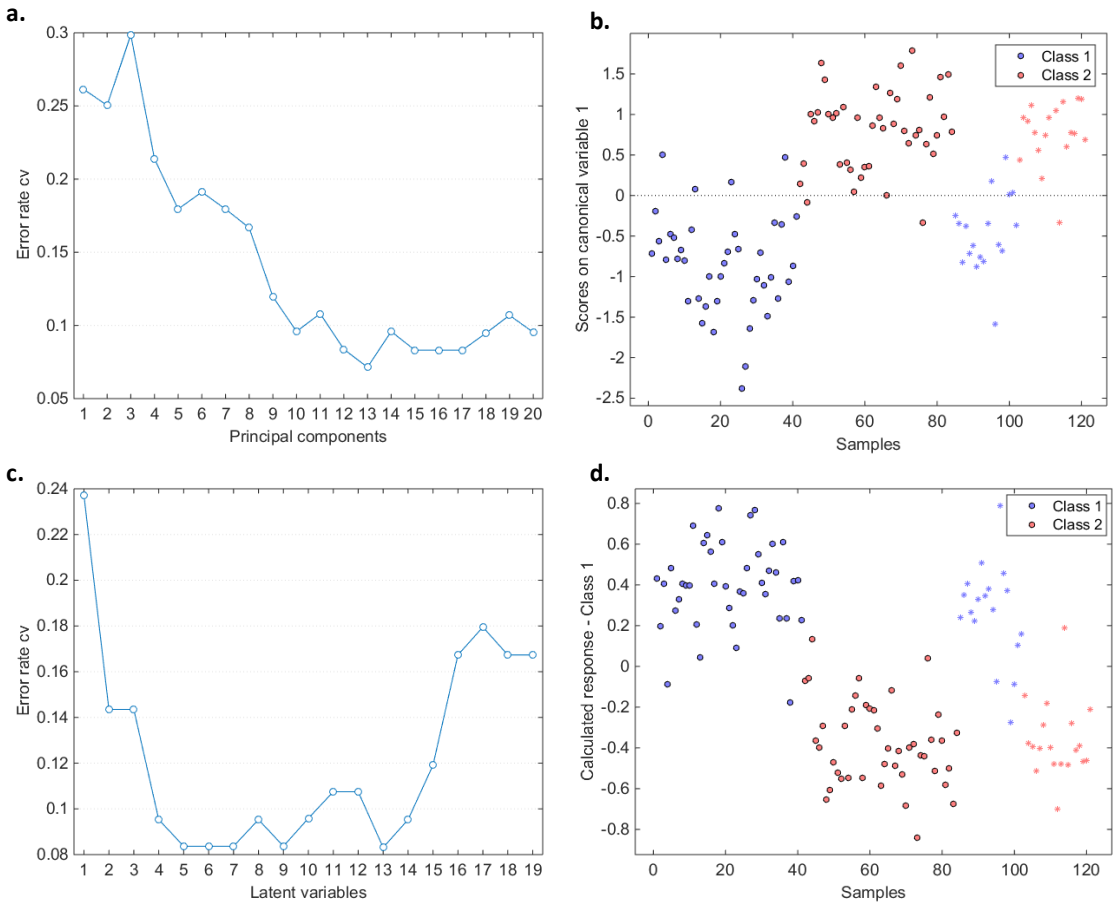
A1. Supplementary Information



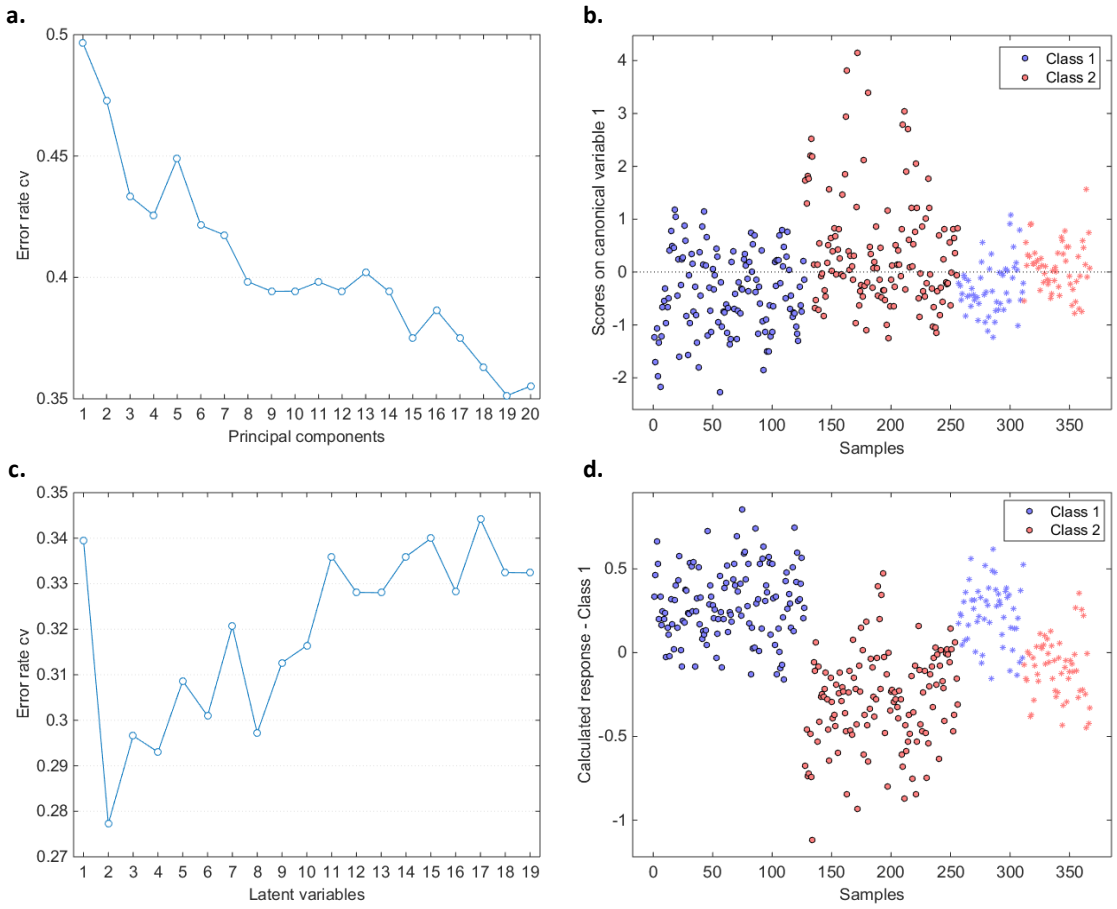
Supplementary Figure A1.1. Raw spectra for datasets 1–3. a) Raw spectra for dataset 1 (blue: class 1, red: class 2); (b) raw spectra for dataset 2 (blue: class 1; red: class 2); (c) raw spectra for dataset 3 (blue: class 1; red: class 2).



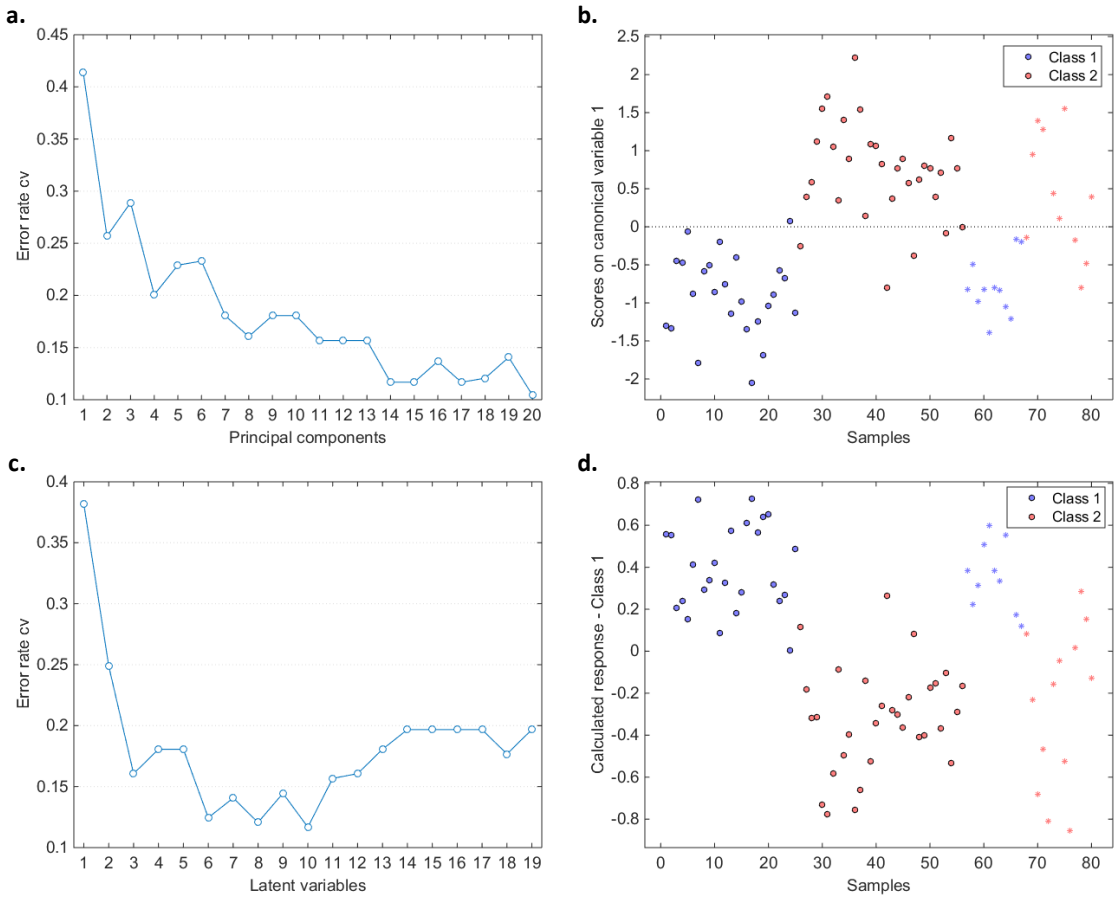
Supplementary Figure A1.2. PCA scores plot and Hotelling's T^2 *versus* Q residuals charts for datasets 1–3. a) PCA scores on PC1 *versus* PC2 for dataset 1; (b) Hotelling's T^2 *versus* Q residuals for dataset 1 (PCA model with 6 PCs (94.9% cumulative explained variance)); (c) PCA scores on PC1 *versus* PC2 for dataset 2; (d) Hotelling's T^2 *versus* Q residuals for dataset 2 (PCA model with 4 PCs (15.3% cumulative explained variance)), where the arrows indicate outliers (samples 244, 263, 264 and 297); (e) PCA scores on PC1 *versus* PC2 for dataset 3; (f) Hotelling's T^2 *versus* Q residuals for dataset 3 (PCA model with 3 PCs (95.0% cumulative explained variance)). EV stands for explained variance.



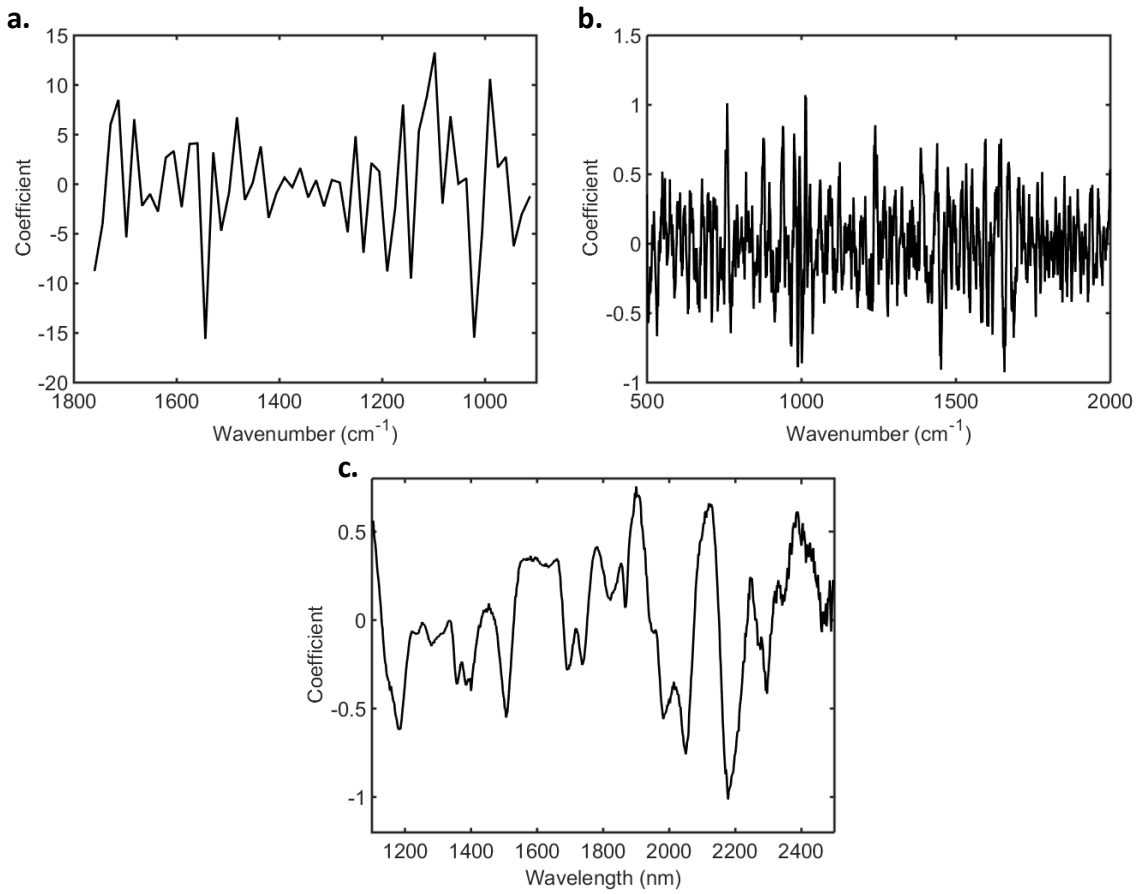
Supplementary Figure A1.3. PCA-LDA and PLS-DA results for dataset 1. (a) Cross-validation error rate varying the number of principal components in PCA-LDA; (b) scores on canonical variable 1 of PCA-LDA, where o = training samples and * = test samples; (c) Cross-validation error rate varying the number of latent variables in PLS-DA; (d) calculated PLS-DA response, where o = training samples and * = test samples.



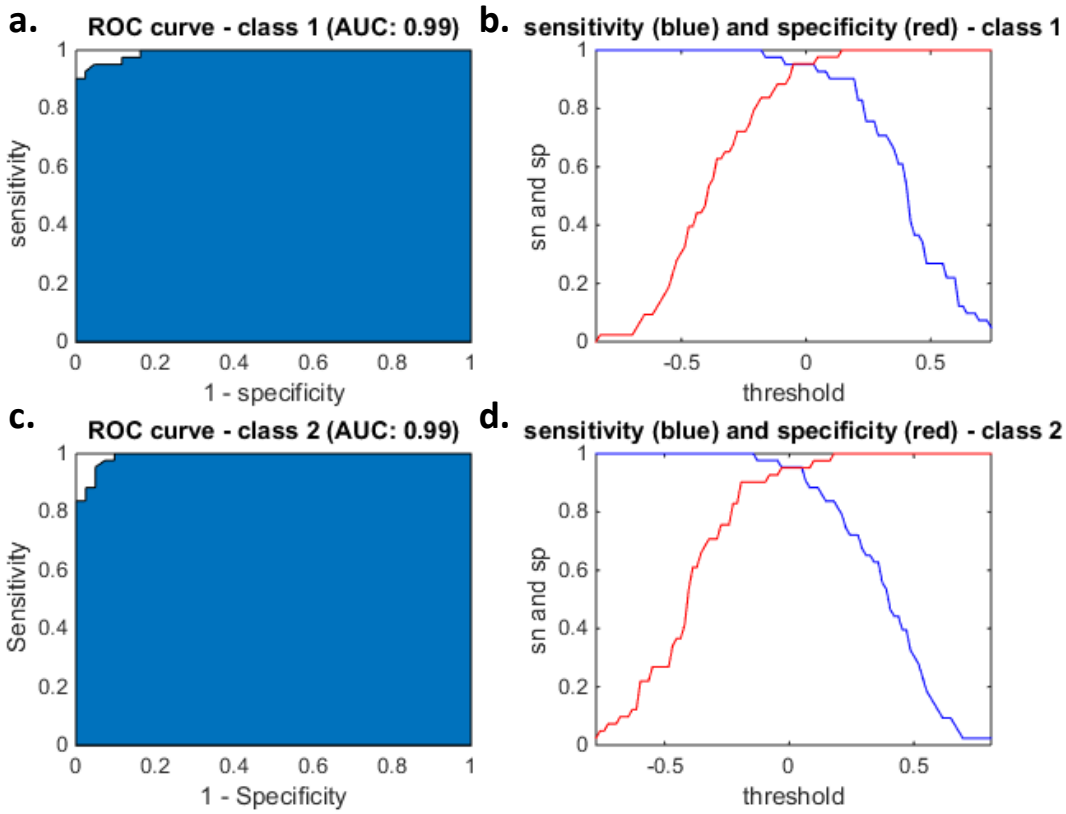
Supplementary Figure A1.4. PCA-LDA and PLS-DA results for dataset 2. (a) Cross-validation error rate varying the number of principal components in PCA-LDA; (b) scores on canonical variable 1 of PCA-LDA, where o = training samples and * = test samples; (c) Cross-validation error rate varying the number of latent variables in PLS-DA; (d) calculated PLS-DA response, where o = training samples and * = test samples.



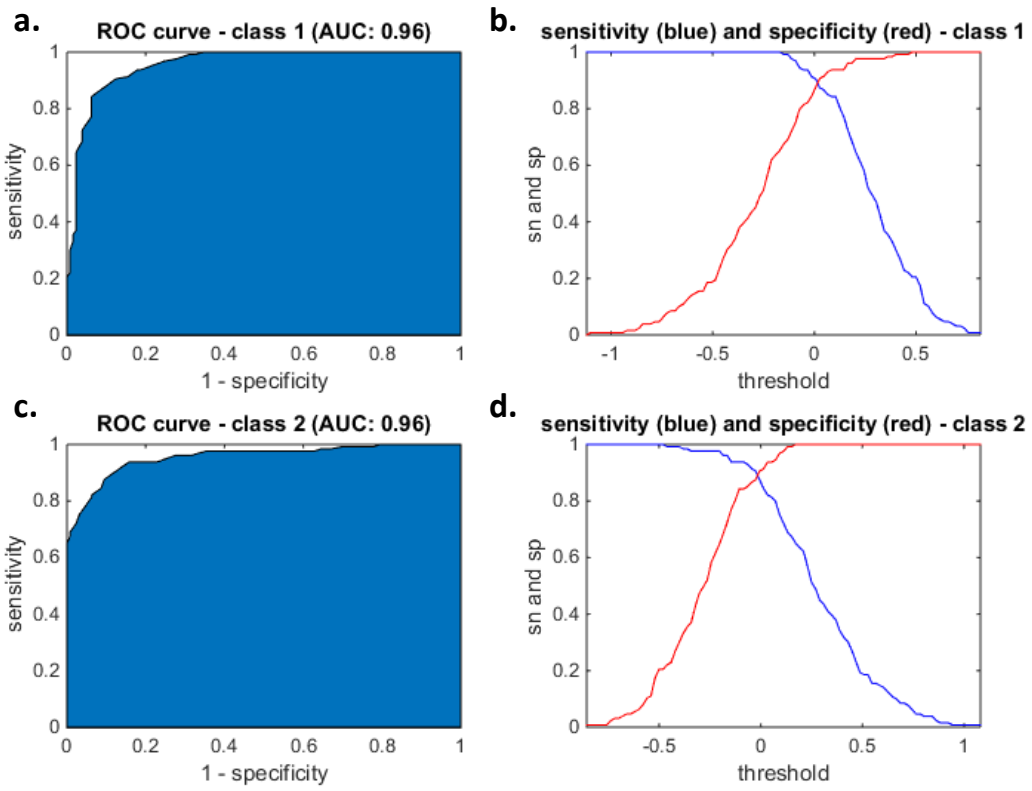
Supplementary Figure A1.5. PCA-LDA and PLS-DA results for dataset 3. (a) Cross-validation error rate varying the number of principal components in PCA-LDA; (b) scores on canonical variable 1 of PCA-LDA, where o = training samples and * = test samples; (c) Cross-validation error rate varying the number of latent variables in PLS-DA; (d) calculated PLS-DA response, where o = training samples and * = test samples.



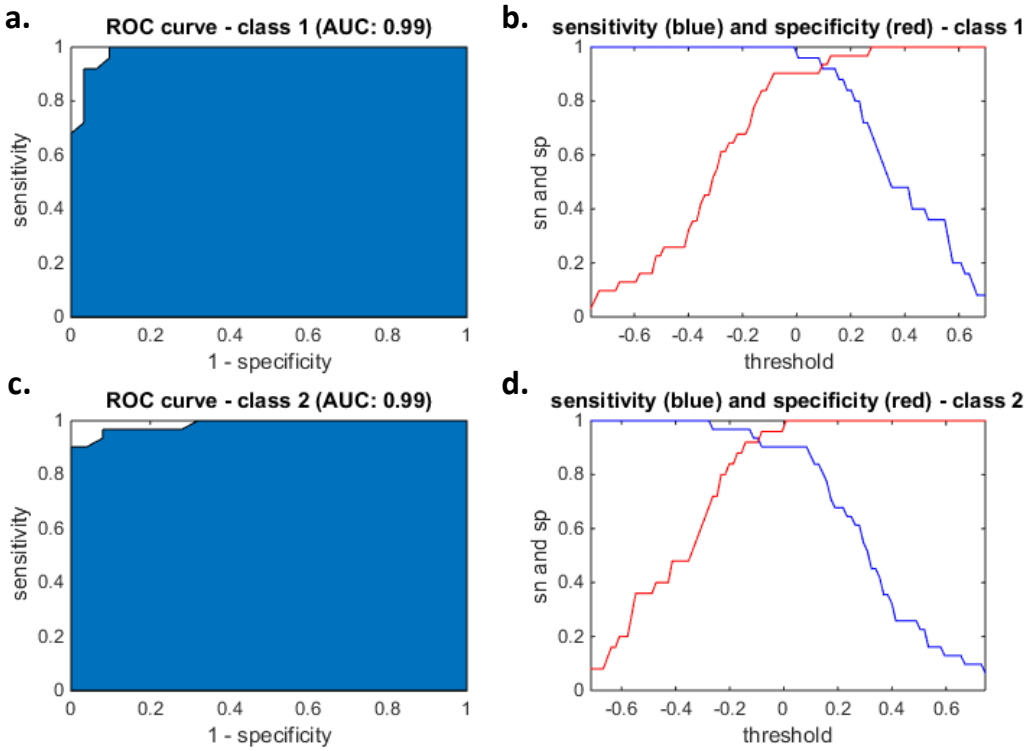
Supplementary Figure A1.6. PLS-DA regression coefficients for datasets 1–3. (a) Dataset 1; (b) dataset 2; and (c) dataset 3.



Supplementary Figure A1.7. Receiver operating characteristic (ROC) curves for dataset 1. (a) ROC curve for class 1; (b) classification threshold for class 1; (c) ROC curve for class 1; and (d) classification threshold for class 2. AUC stands for area under the curve, sn for sensitivity and sp for specificity.



Supplementary Figure A1.8. Receiver operating characteristic (ROC) curves for dataset 2. (a) ROC curve for class 1; (b) classification threshold for class 1; (c) ROC curve for class 1; and (d) classification threshold for class 2. AUC stands for area under the curve, sn for sensitivity and sp for specificity.



Supplementary Figure A1.9. Receiver operating characteristic (ROC) curves for dataset 3. (a) ROC curve for class 1; (b) classification threshold for class 1; (c) ROC curve for class 1; and (d) classification threshold for class 2. AUC stands for area under the curve, sn for sensitivity and sp for specificity.

A2. Supplementary Method – Protocol for Spectral Data Analysis: SHE Dataset

A. Software requirements

Required software (to download):

A1: MATLAB, version R2011a or above. *Free-trial* version available at https://www.mathworks.com/.

A2: IRootLab (version 0.17.8.22 or above) toolbox for MATLAB. *Free* version available at http://trevisanj.github.io/irootlab/.

A3: Classification Toolbox for MATLAB. *Free* version available at http://www.michem.unimib.it/download/matlab-toolboxes/classification-toolbox-for-matlab/.

A4: PCA Toolbox for MATLAB. *Free* version available at http://www.michem.unimib.it/download/matlab-toolboxes/pca-toolbox-for-matlab/.

A5: Automatic Outlier Detection Algorithm. *Free* version available at https://doi.org/10.6084/m9.figshare.7066610.v2.

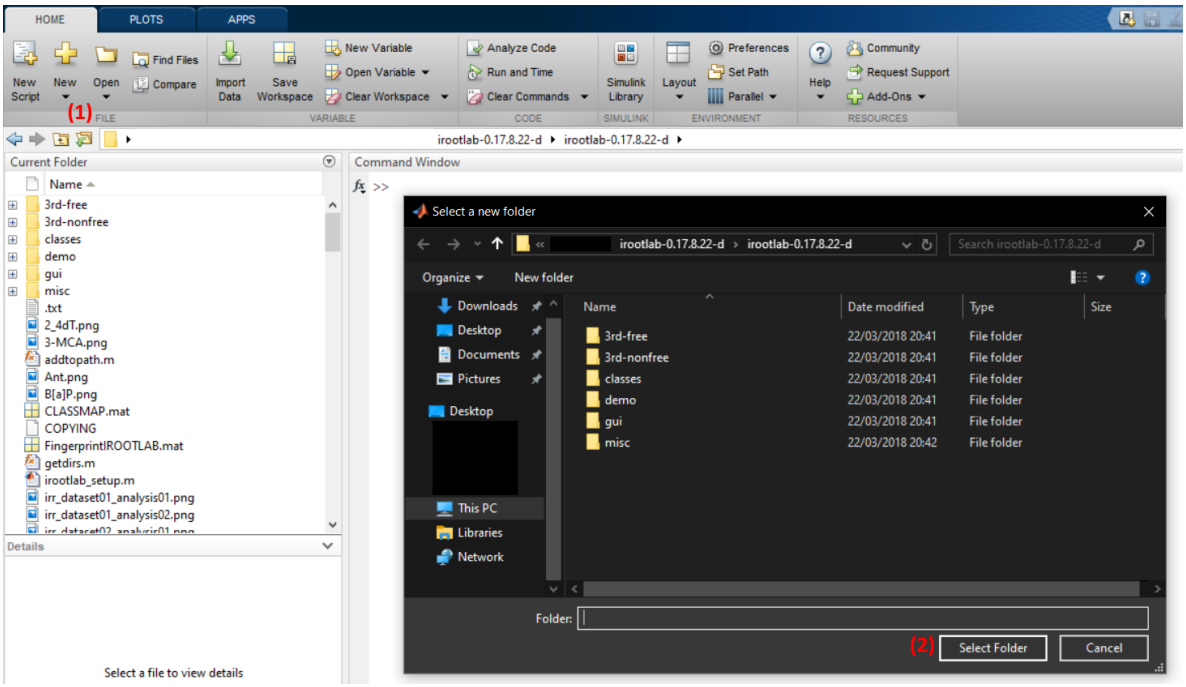
A6: MLM Sample Selection Algorithm. *Free* version available at https://doi.org/10.6084/m9.figshare.7393517.v2.

B. Loading the dataset

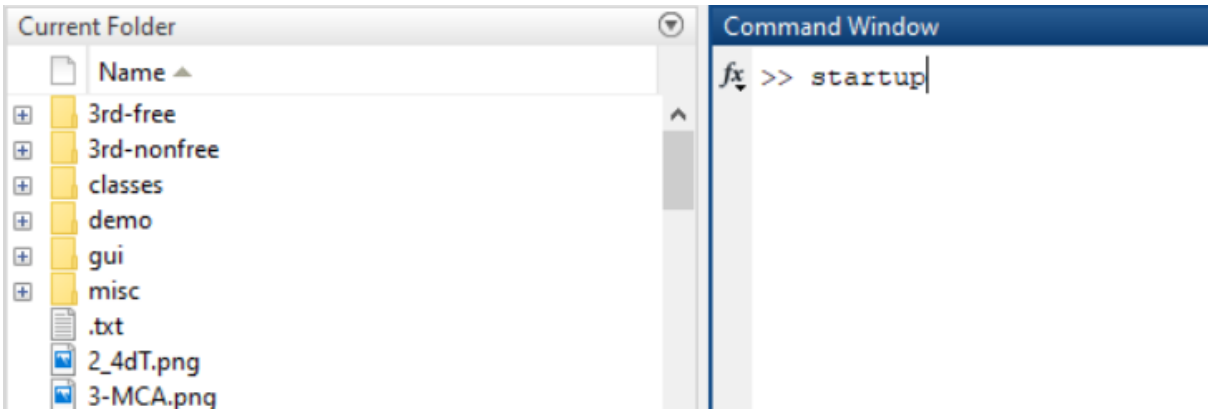
In this example, we will use a dataset available within the IRootLab toolbox for MATLAB. We will modify it to contain 2 classes only, and perform (C) data quality visualisation, (D) pre-processing, (E) exploratory analysis by PCA, (F) outlier detection, (G) sample selection, and (H) supervised classification by PCA-LDA/QDA.

To load the dataset, follow the instructions:

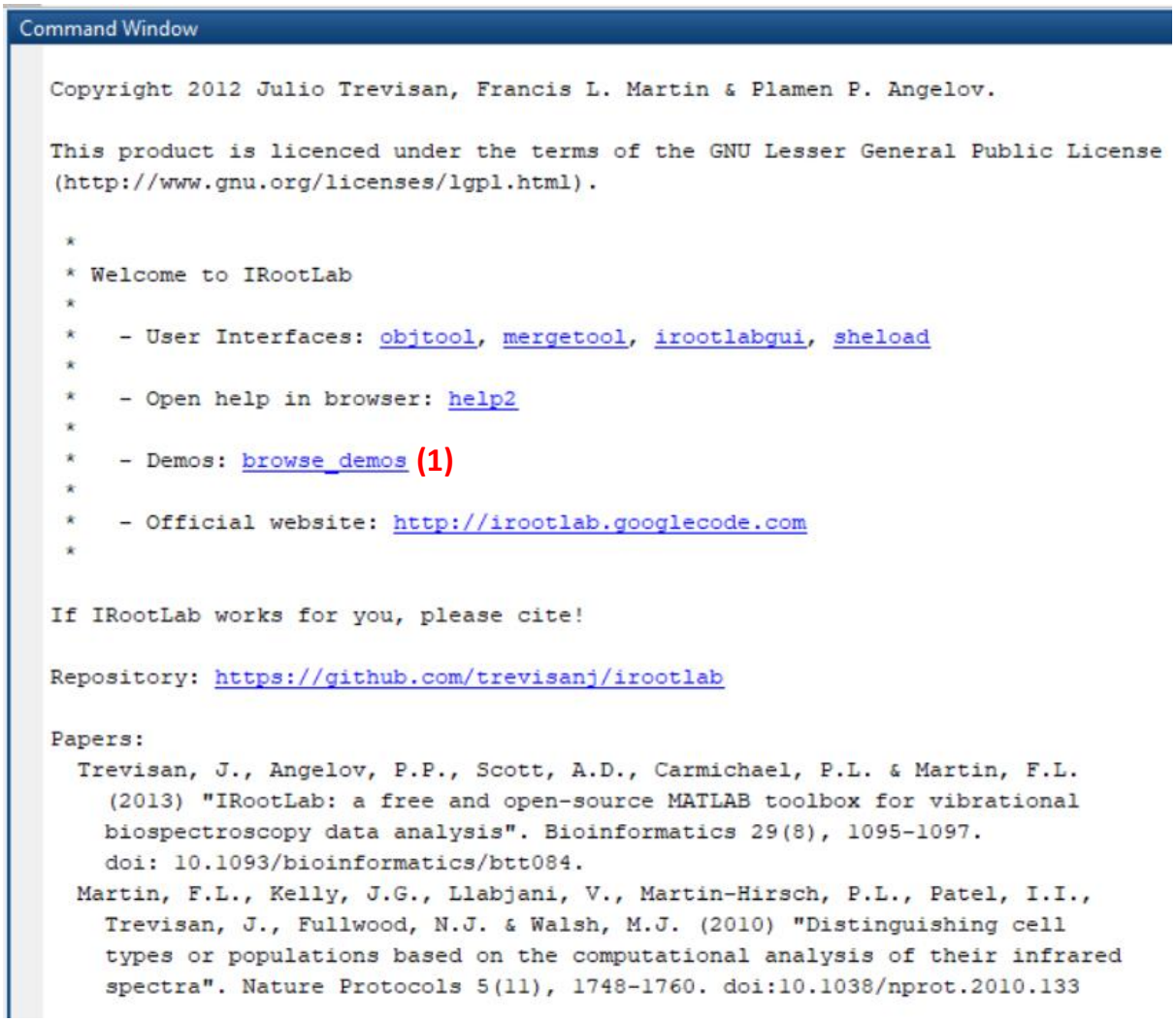
Step B1: Go to IRootLab folder ('irootlab-0.17.8.22-d') within MATLAB: Click on (1) and then navigate to the IRootLab folder. Click on (2) 'Select Folder' to select the folder.



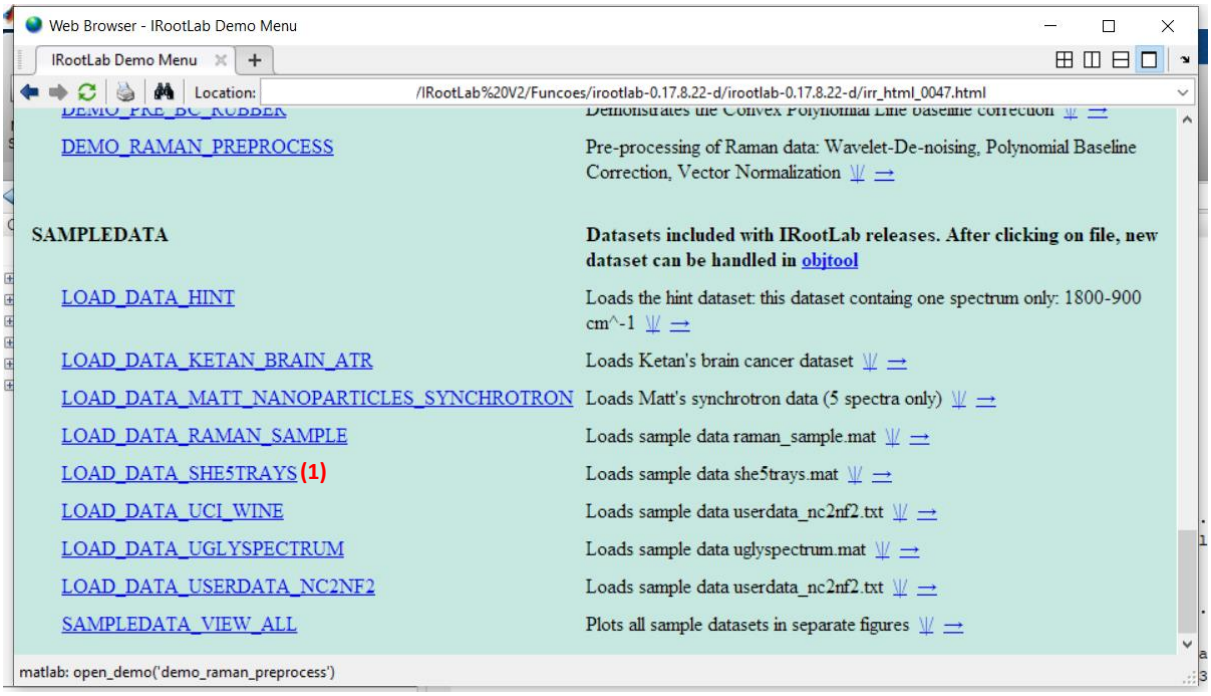
Step B2: Type 'startup' on the Command Window and press <ENTER>:



Step B3: Click on (1) 'browse_demos':



Step B4: Click on (1) 'LOAD_DATA_SHE5TRAYS' and close the window:

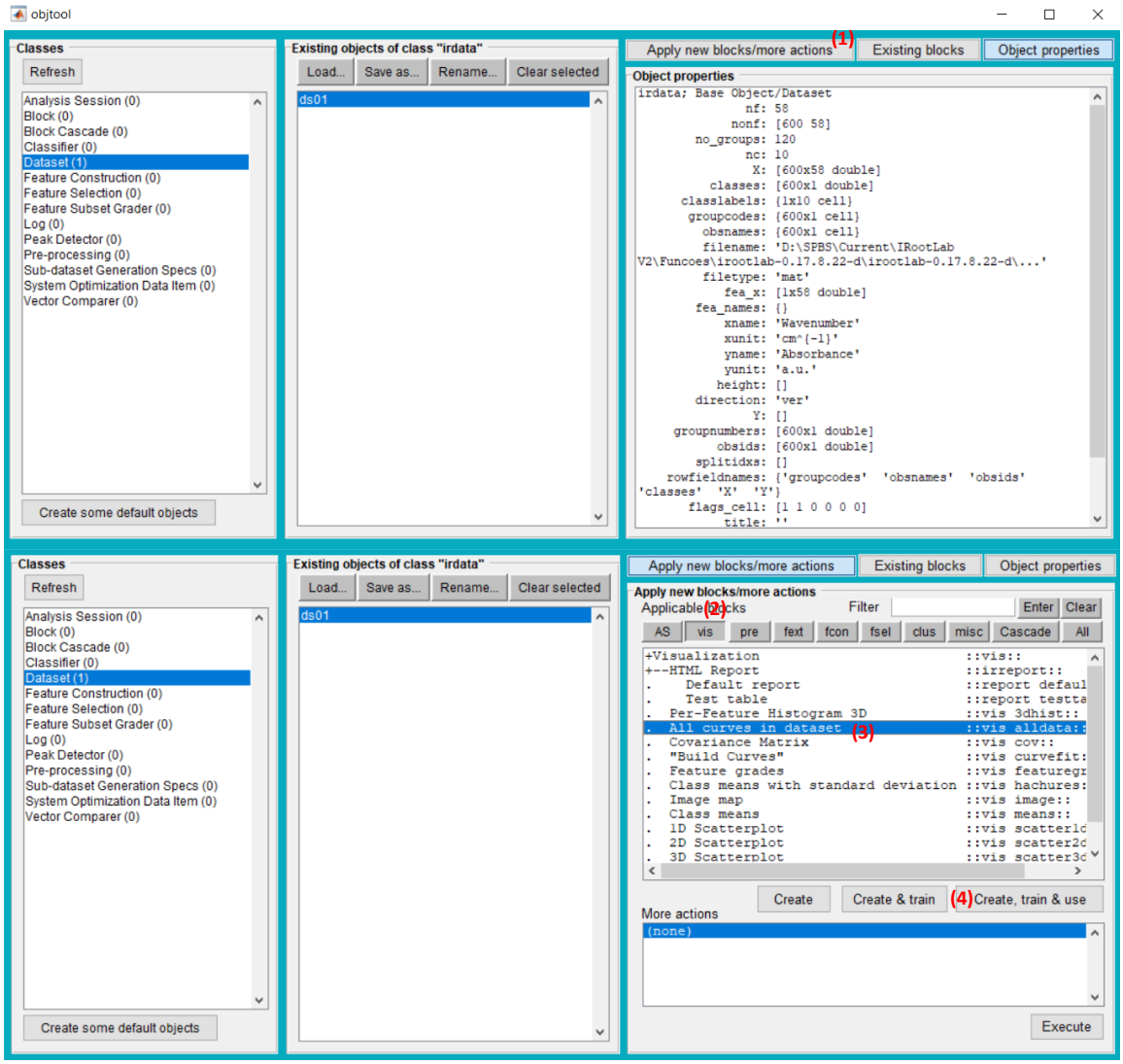


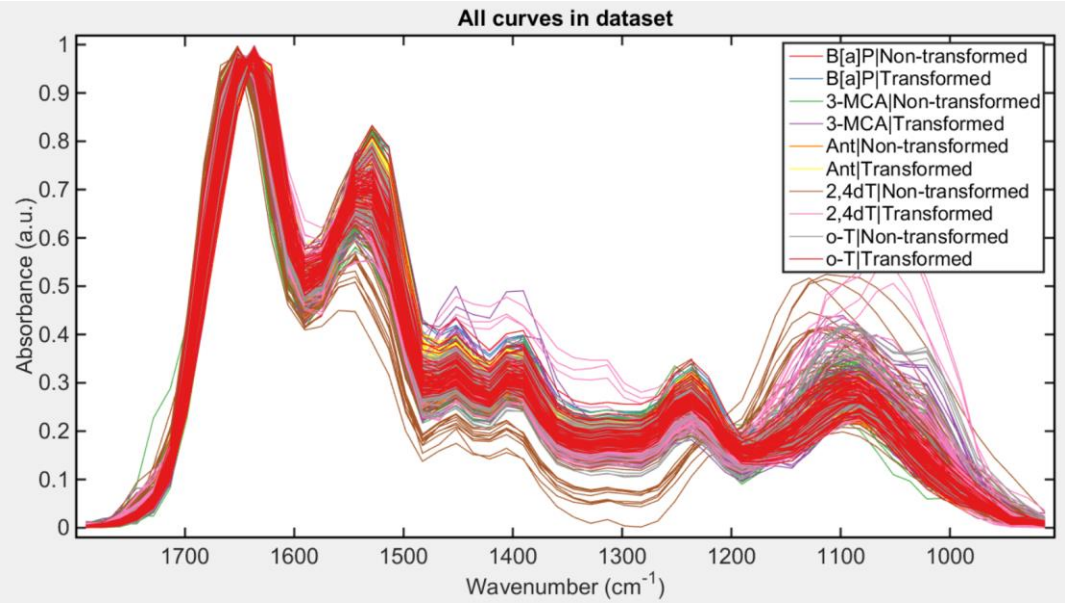
Step B5: Scroll up the Command Window page and click on (1) 'objtool':

C. Data quality visualization

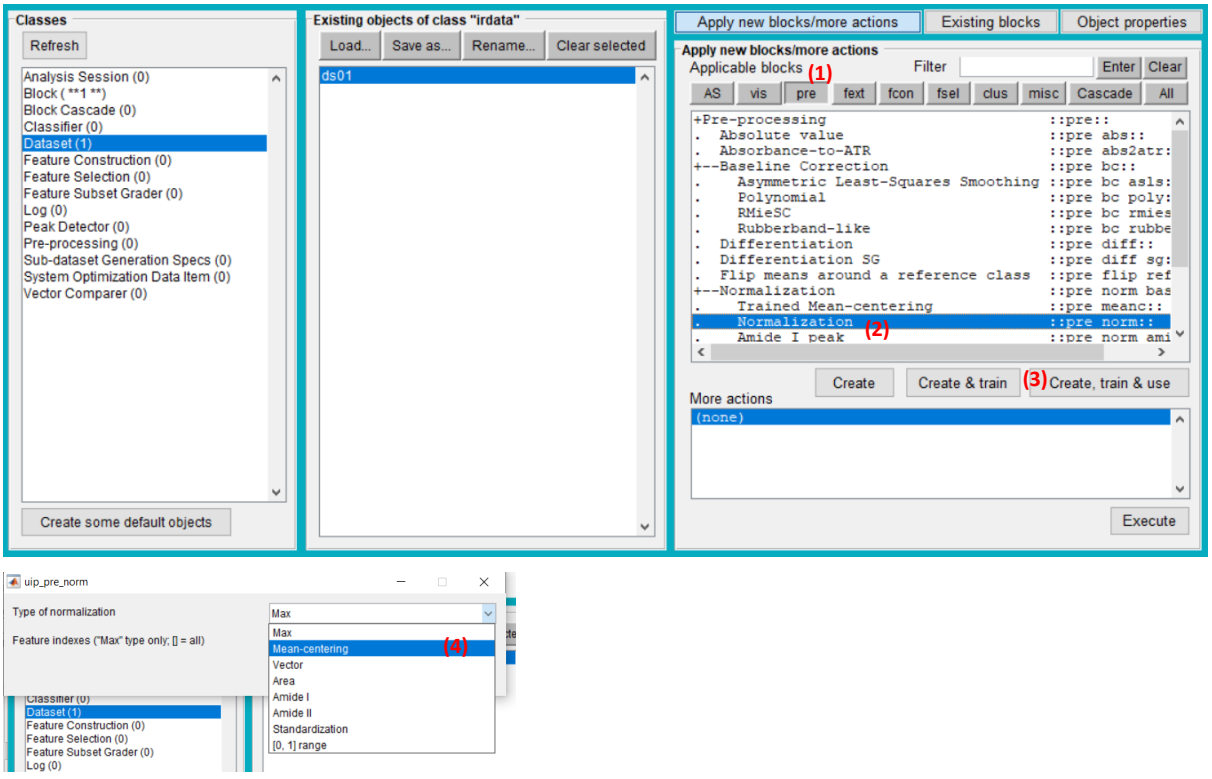
The original dataset contain 5 classes. To visualise the spectral data follow the steps:

Step C1: Click on (1) ‘Apply new blocks/more actions’, (2) ‘vis’, (3) ‘All curves in dataset’ and (4) ‘Create, train & use’.

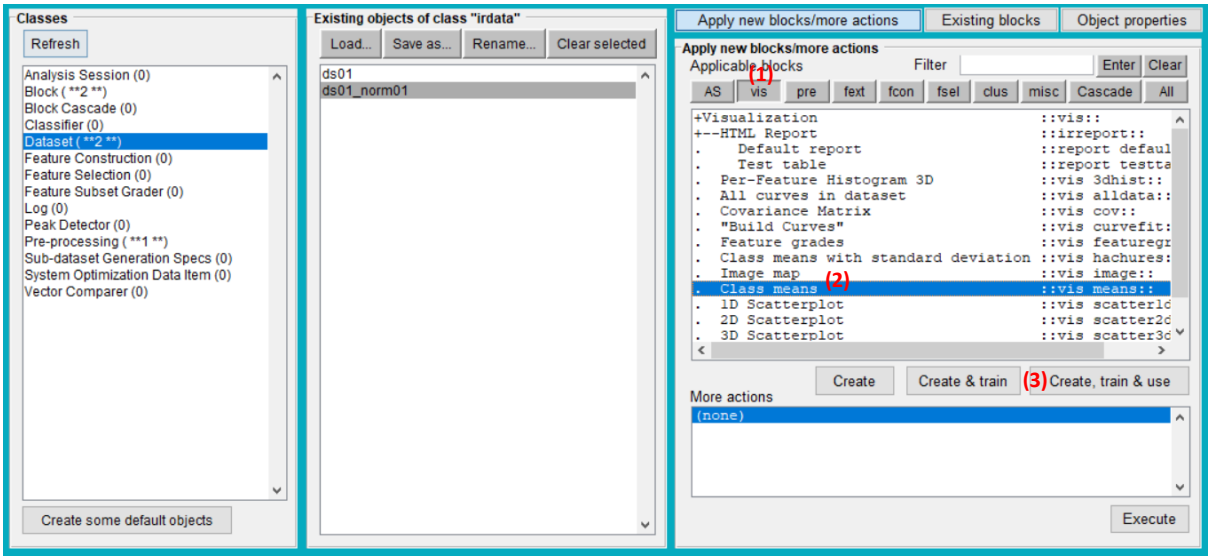


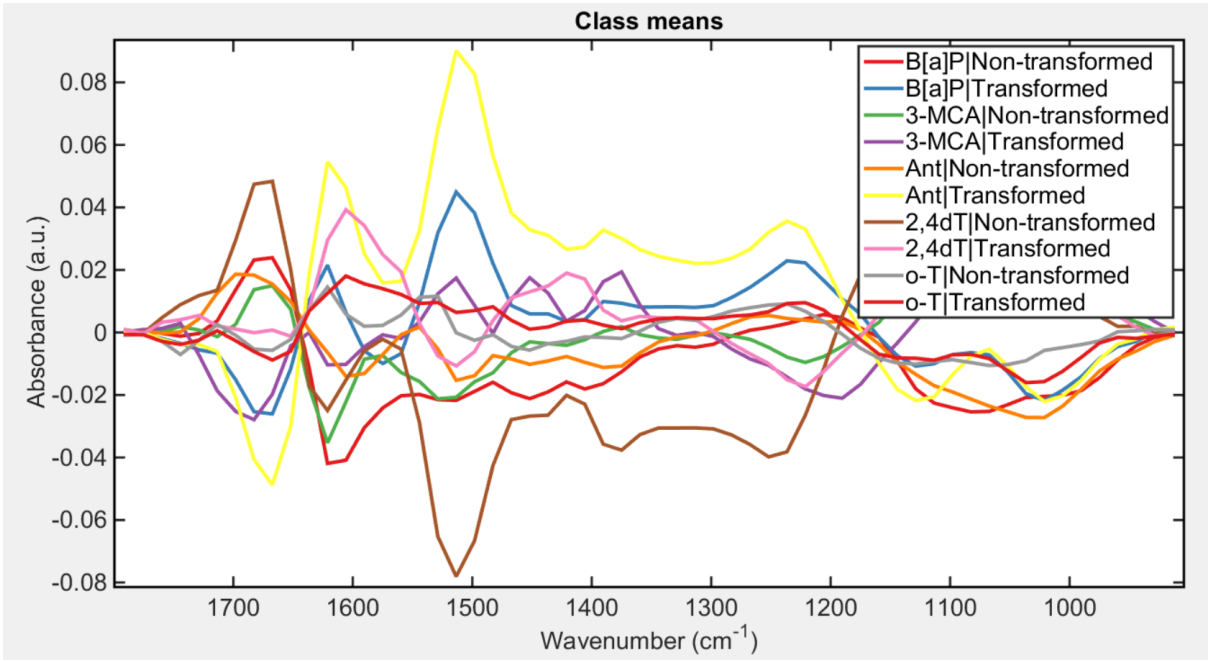


Step C2: Go to (1) 'pre', (2) 'Normalization', (3) 'Creat, train & use', (4) select 'Mean-centering' and press 'OK'.

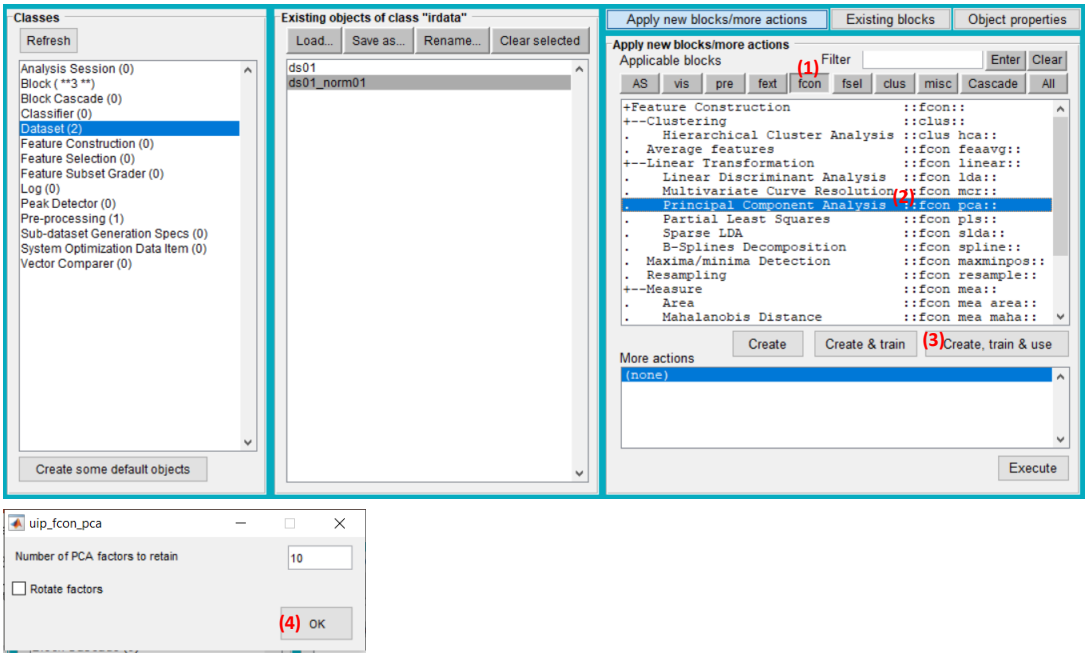


Step C3: Select ‘ds01_norm01’, go to (1) ‘vis’, (2) ‘Class means’, (3) ‘Creat, train & use’ and click in ‘OK’.

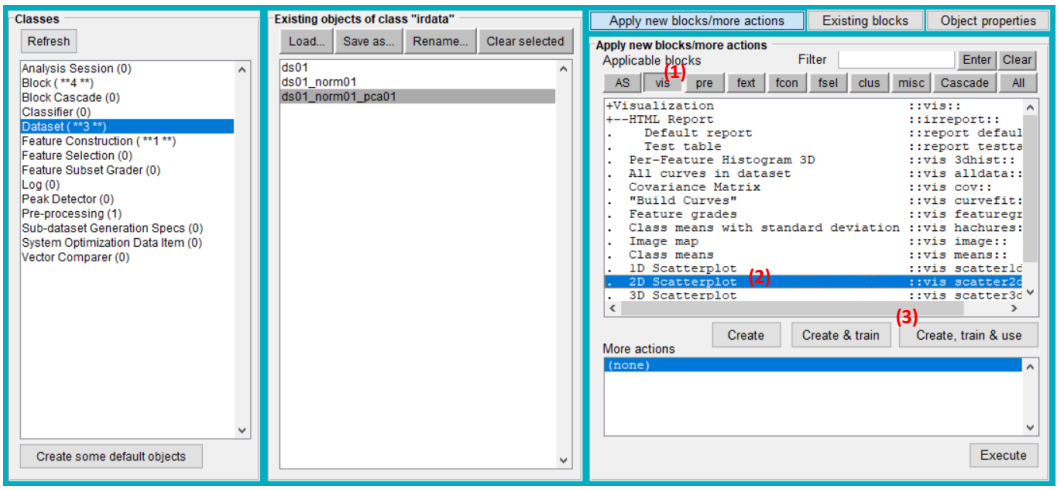


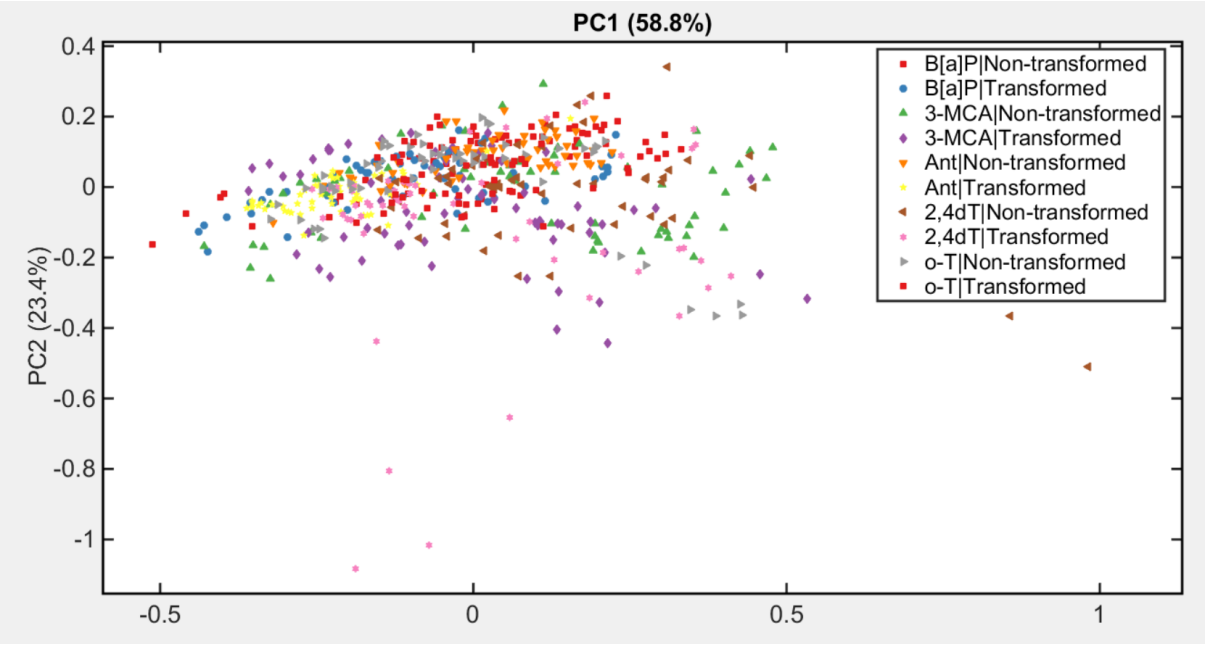


Step C4: Select ‘ds01_norm01’, go to (1) ‘fcon’, (2) ‘Principal Component Analysis’, (3) ‘Creat, train & use’ and (4) press ‘OK’.



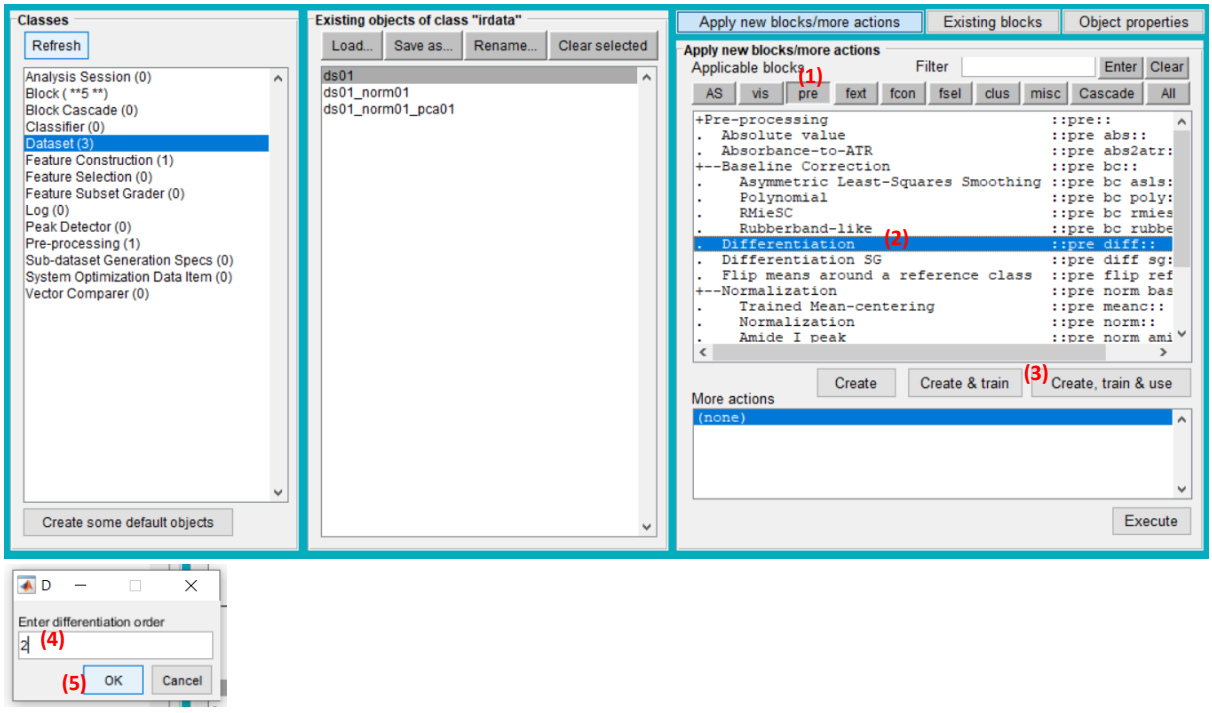
Step C5: Select ‘ds01_norm01_pca01’, go to (1) ‘vis’, (2) ‘2D Scatterplot’, (3) ‘Creat, train & use’, (4) input the PCs to plot and (5) press ‘OK’.



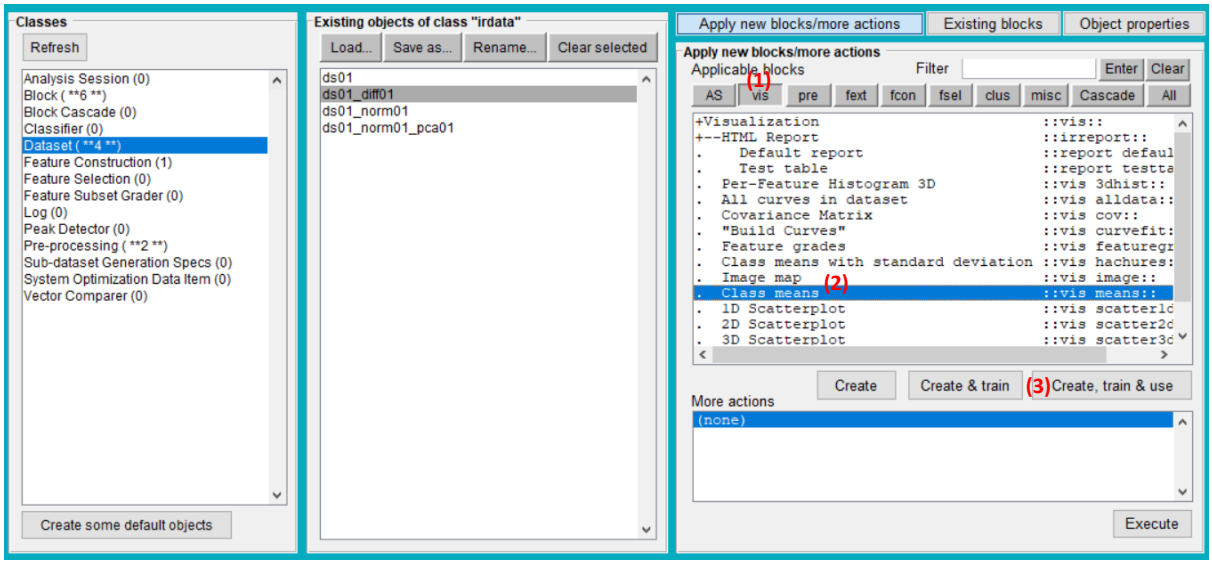


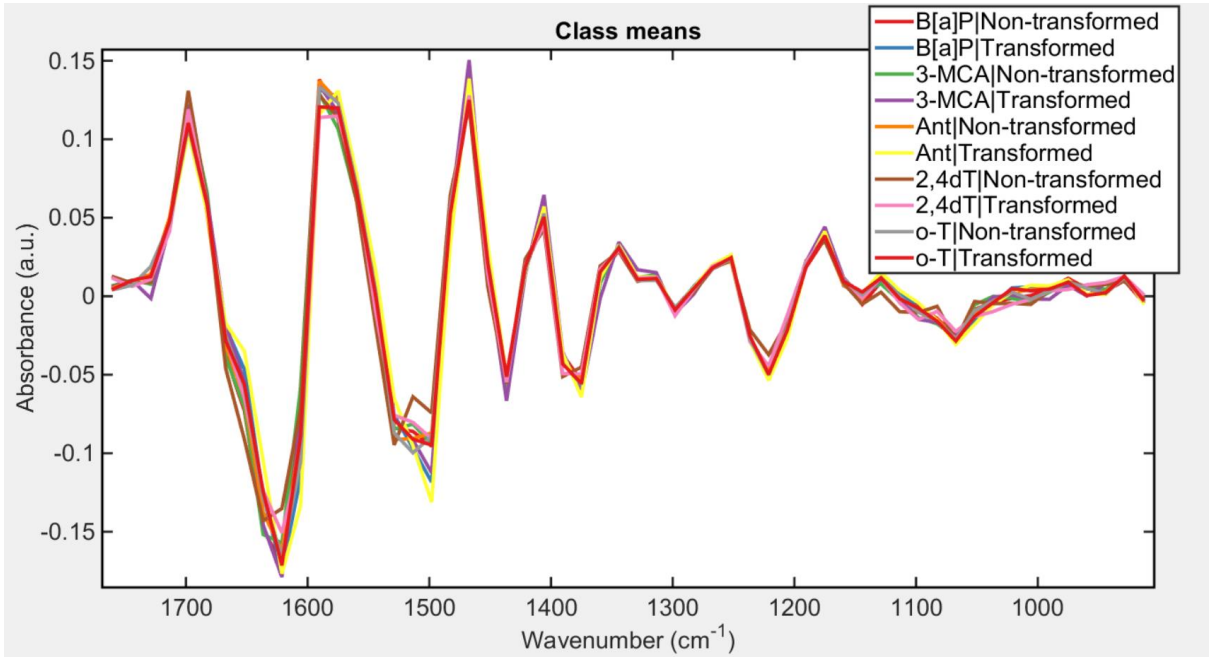
D. Pre-processing

Step D1: Select ‘ds01’, go to (1) ‘pre’, (2) ‘Differentiation’, (3) ‘Creat, train & use’, (4) insert the derivative order and (5) press ‘OK’.



Step D2: Select ‘ds01_diff01’, go to (1) ‘vis’, (2) ‘Class means’, (3) ‘Creat, train & use’ and (4) press ‘OK’.



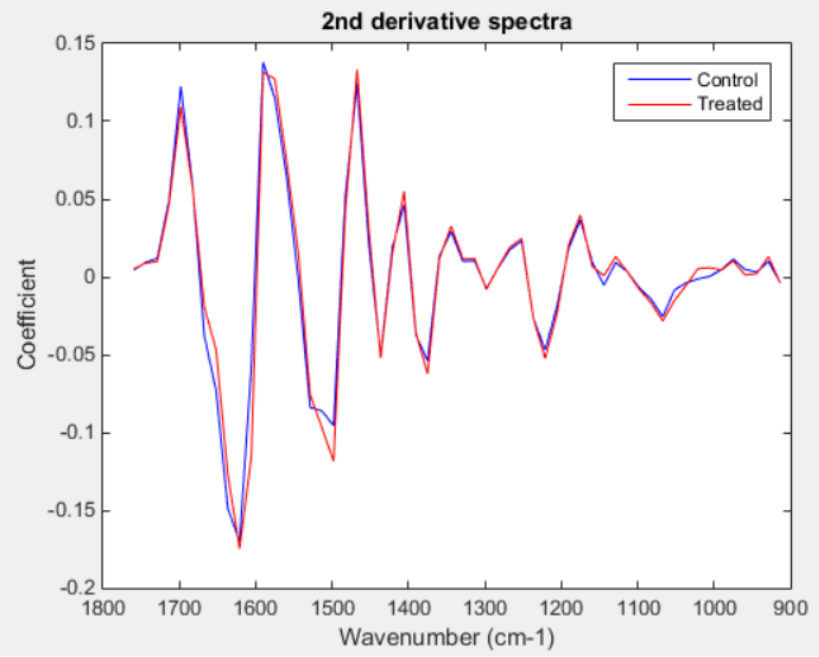


E. Exploratory analysis by PCA

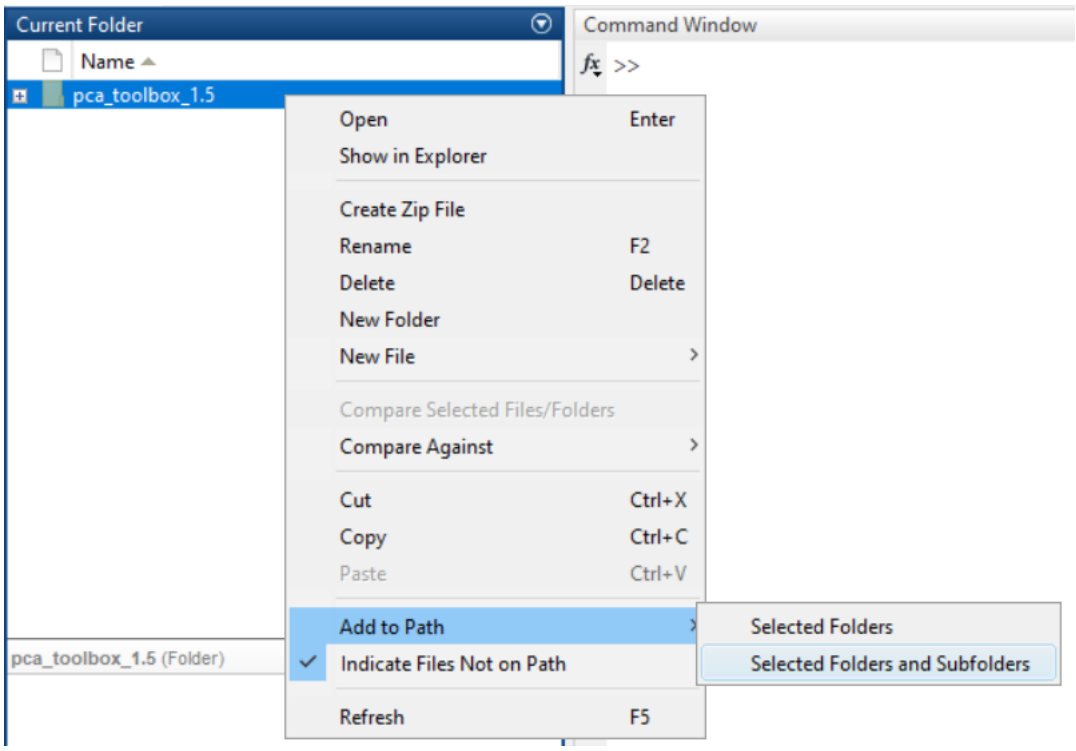
Only class 1 (B[a]P | Non-transformed, control) and class 2 (B[a]P | transformed, treated) will be used for analysis.

Step E1: Select only class 1 and 2 from the dataset. Type in the MATLAB Command Window:

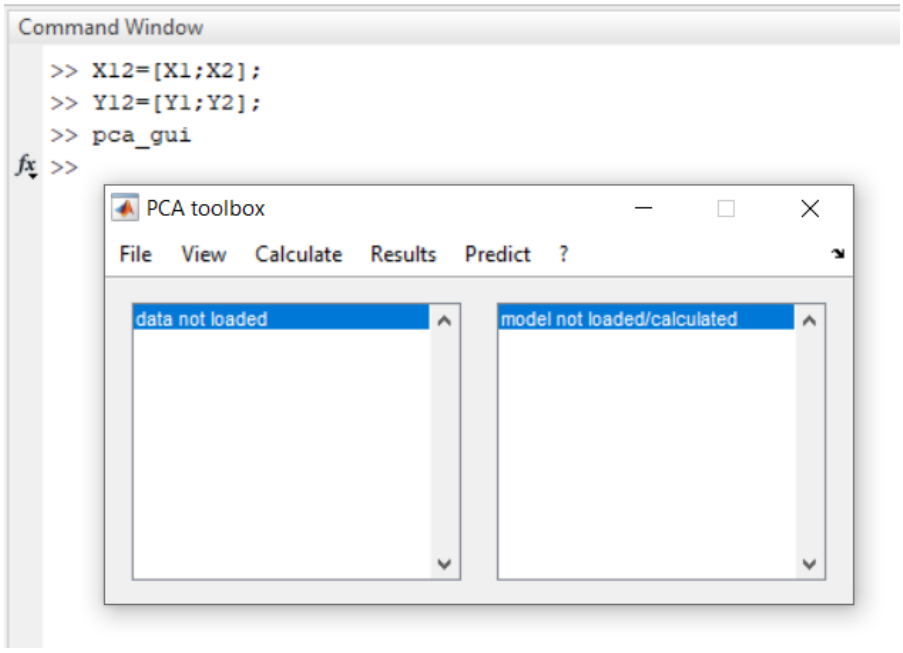
```
Command Window
>> cm = ds01_diff01.fea_x;
>> X = ds01_diff01.X;
>> Y = ds01_diff01.classes;
>> X1 = X(find(Y==0),:);
>> X2 = X(find(Y==1),:);
>> Y1 = ones(59,1);
>> Y2 = ones(62,1)+1;
>> figure,
>> plot(cm,mean(X1),'b');
>> hold on
>> plot(cm,mean(X2),'r');
>> xlabel('Wavenumber (cm-1)');
>> ylabel('Coefficient');
>> title('2nd derivative spectra');
>> legend({'Control','Treated'});
>> set(gca,'Xdir','reverse')
fx >>
```



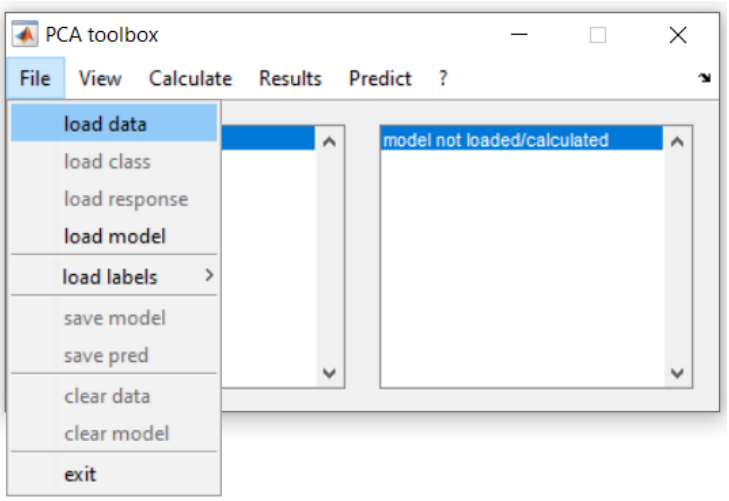
Step E2: Navigate within MATLAB to the PCA Toolbox for MATLAB folder (step A4). Right-click on the folder and select Add to Path > Select Folders and Subfolders.



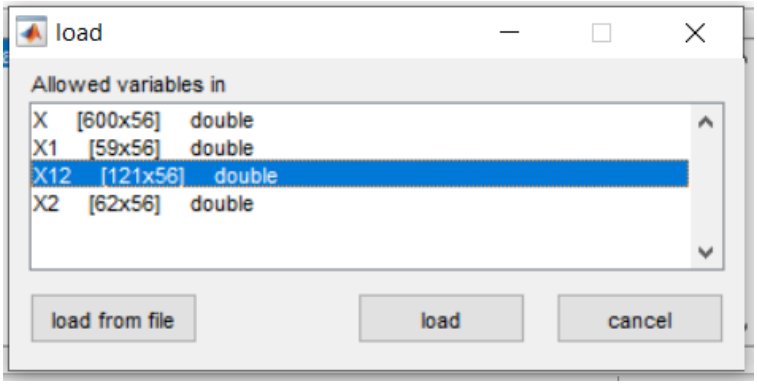
Step E3: Type in the Command Window:



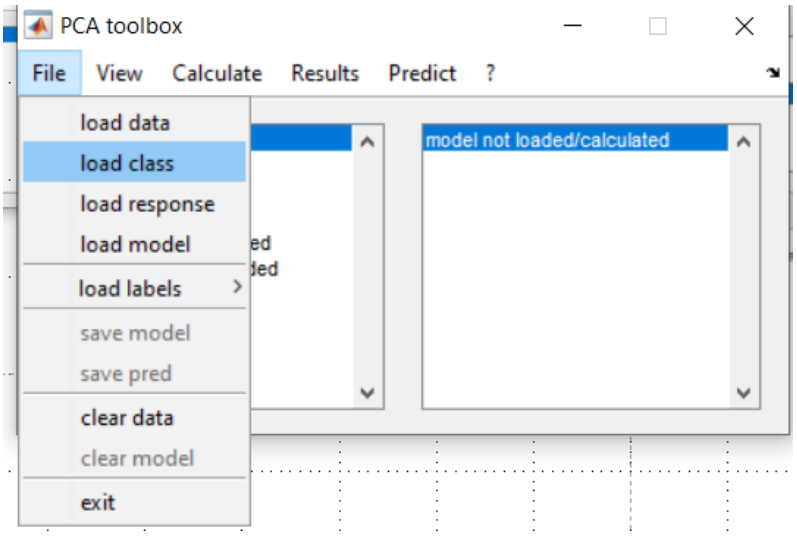
Step E4: Go to File > Load data.



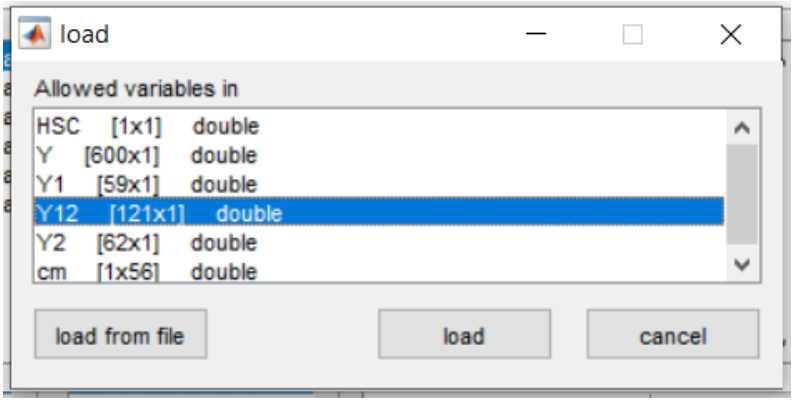
Step E5: Select the dataset (X12) and click on 'load'.



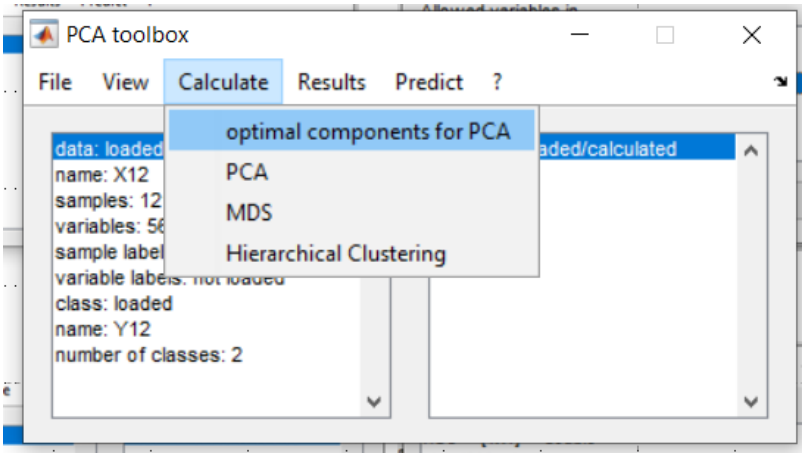
Step E6: Go to File > load class.



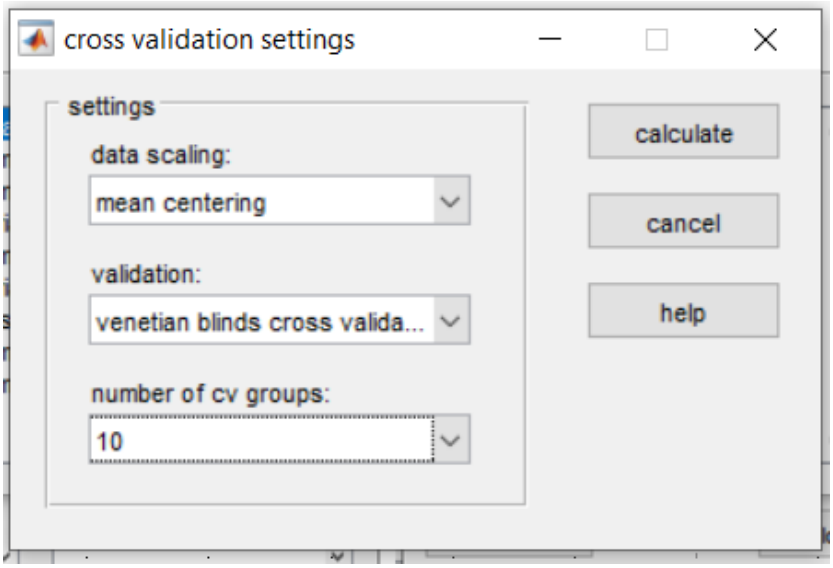
Step E7: Select the class labels (Y12).



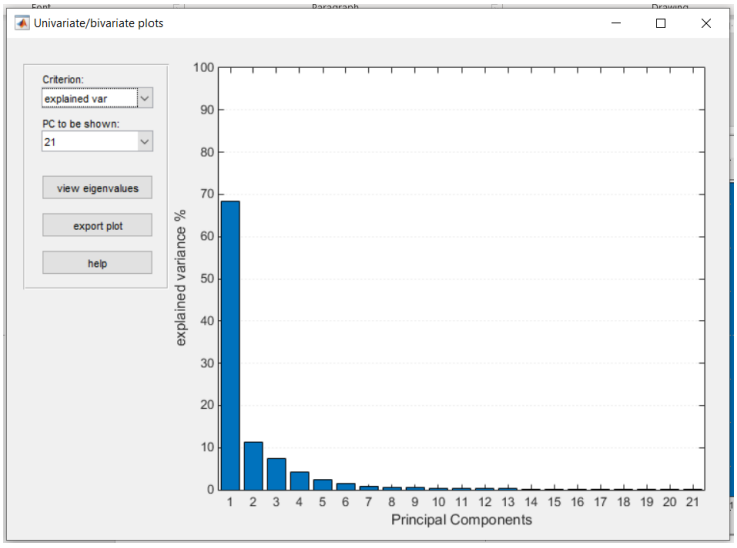
Step E8: Go to Calculate > optimal components for PCA.



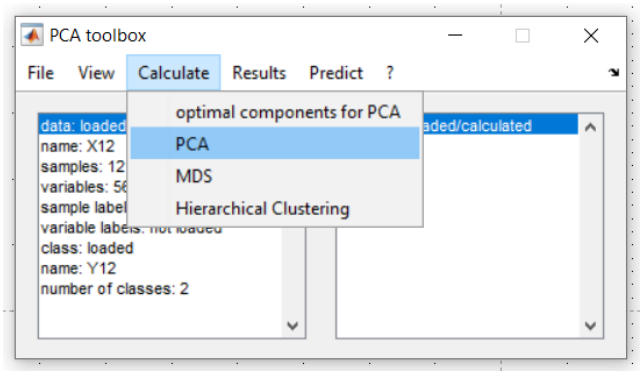
Step E9: Select the following settings and click on ‘calculate’.



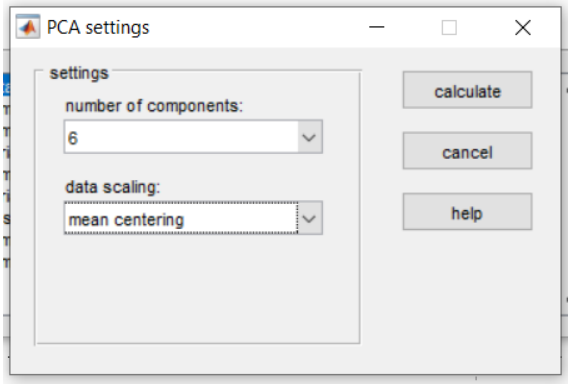
Step E10: Select the following settings and observe the number of Principal Components to choose (6 in this case).



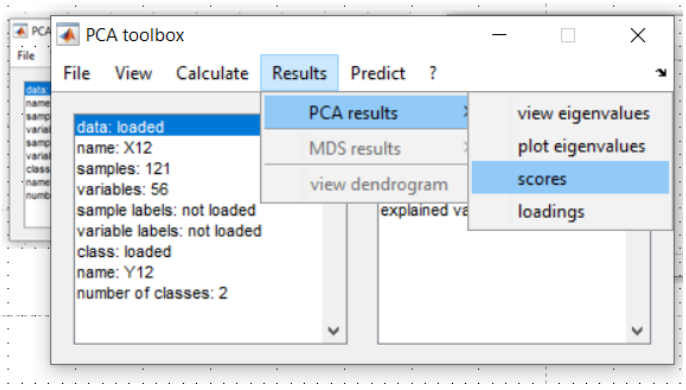
Step E11: Go to Calculate > PCA.



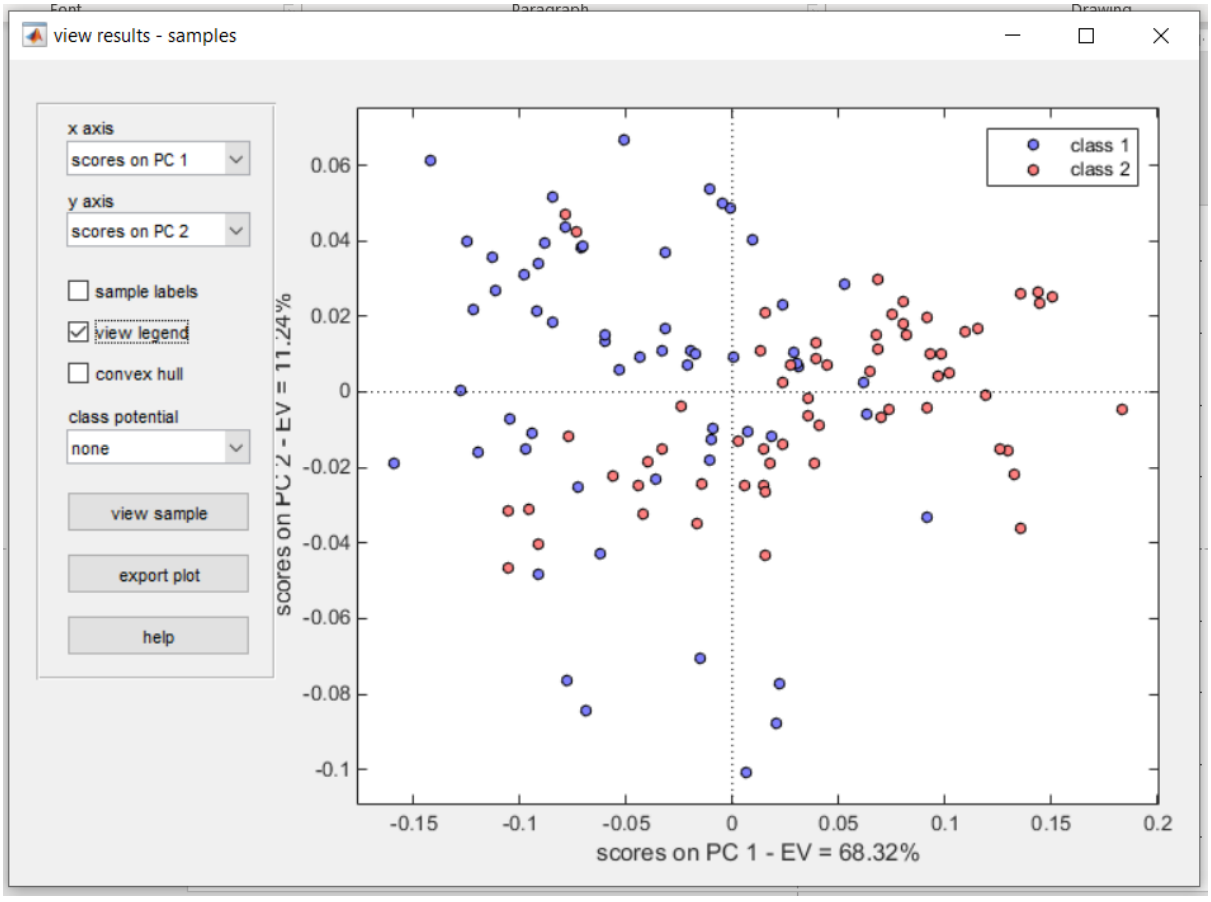
Step E12: Select the following settings and click on ‘calculate.



Step E13: Go to Results > PCA results > scores.



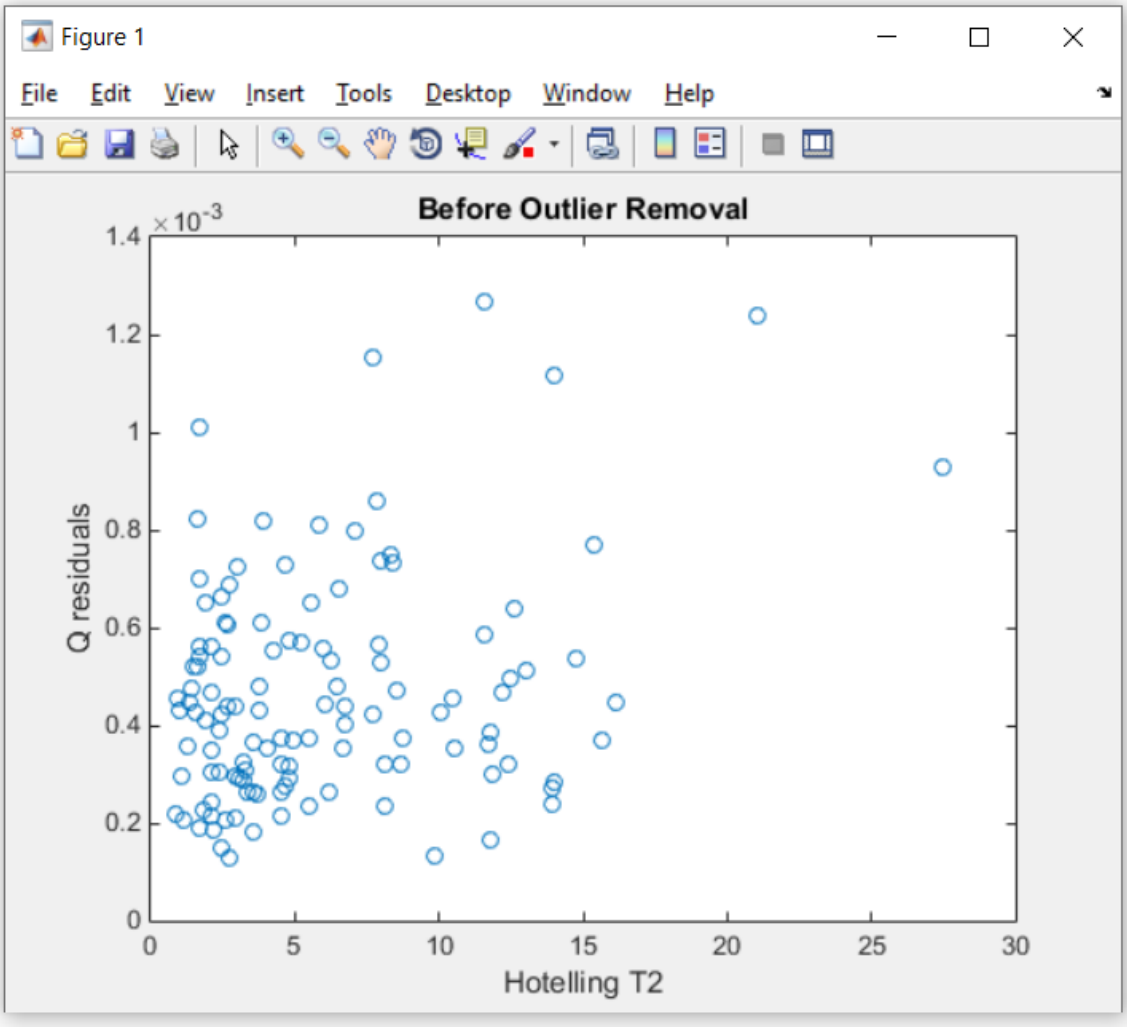
Step E14: Use the following settings and visualise the PCA scores. Note that class 1 is predominantly in the left-side of the graph and class 2 in the right-side of the graph, hence, PC1 (68.32% explained variance) influences the separation between class 1 and class 2.



F. Outlier detection

Step F1: Navigate within MATLAB to the folder containing the Outlier Detection Algorithm (step A5), so the files are shown in the Current Folder window. Type in the Command Window:

```
>> X12 = outlier(X12,6);
-----------------
Select the Hotelling T2 threshold:
```

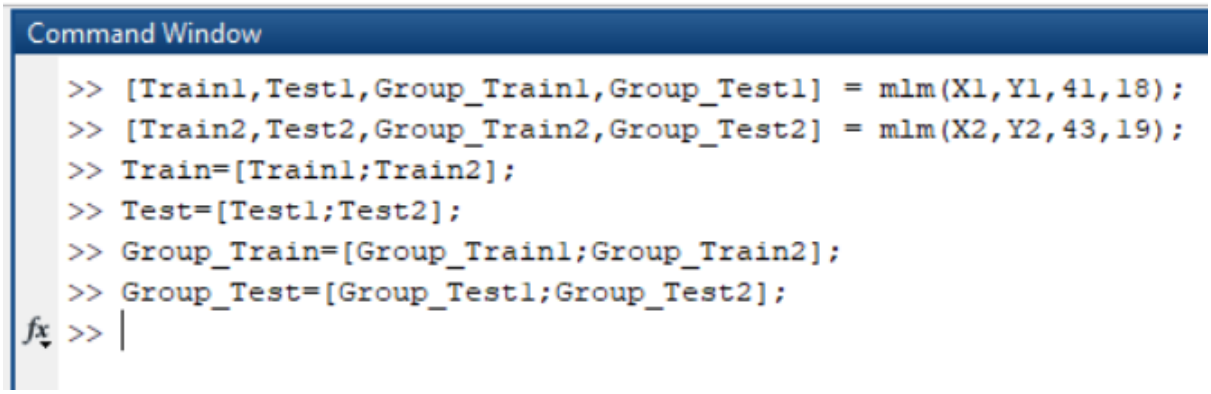


Observe that there is no outlier. All the samples are distributed close to the origin (0,0), where no sample is observed very far from the clustering containing the datapoints.

Step F2: Since no outlier is present, press <CTRL + C> to stop the MATLAB routine.

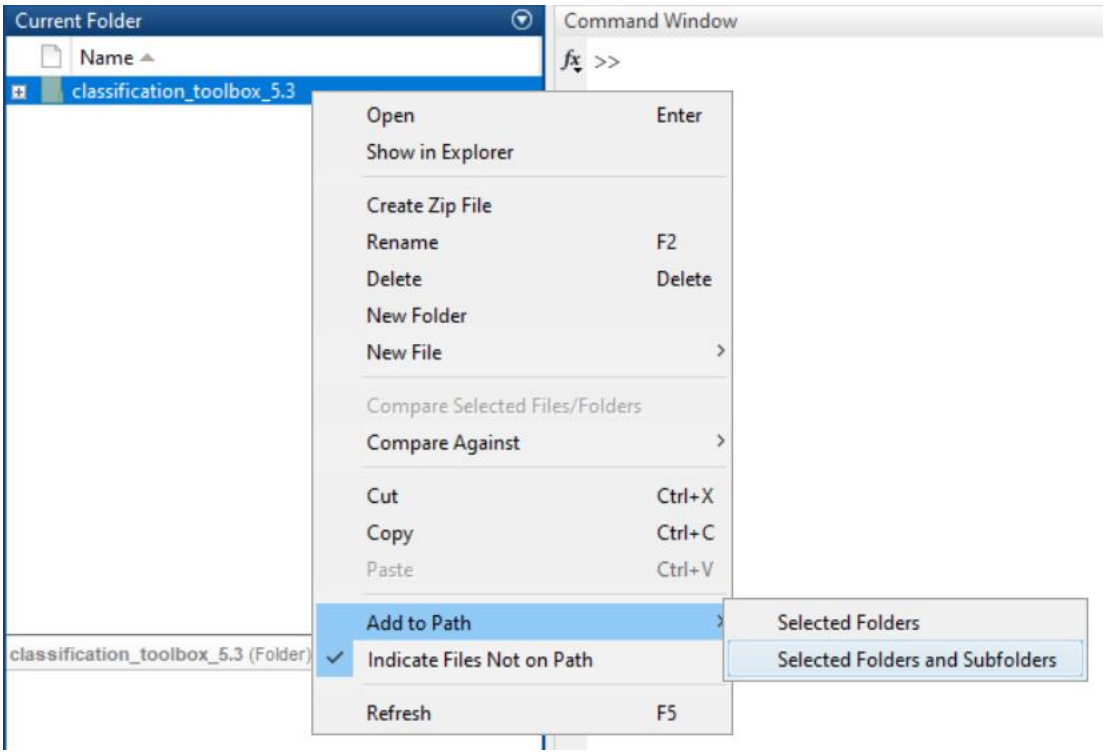
G. Sample selection

Step G1: Navigate within MATLAB to the folder containing the MLM algorithm (step A6), so the files are shown in the Current Folder window. Type in the Command Window:

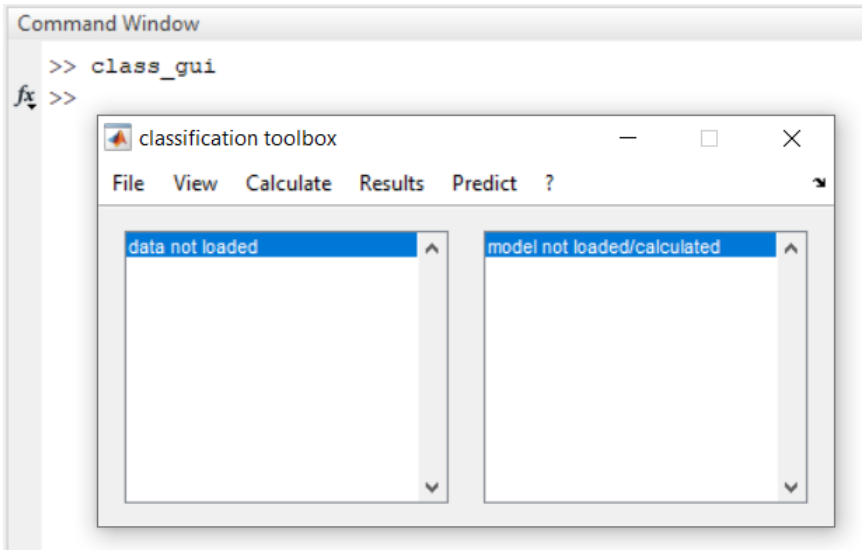


H. Supervised classification by PCA-LDA/QDA

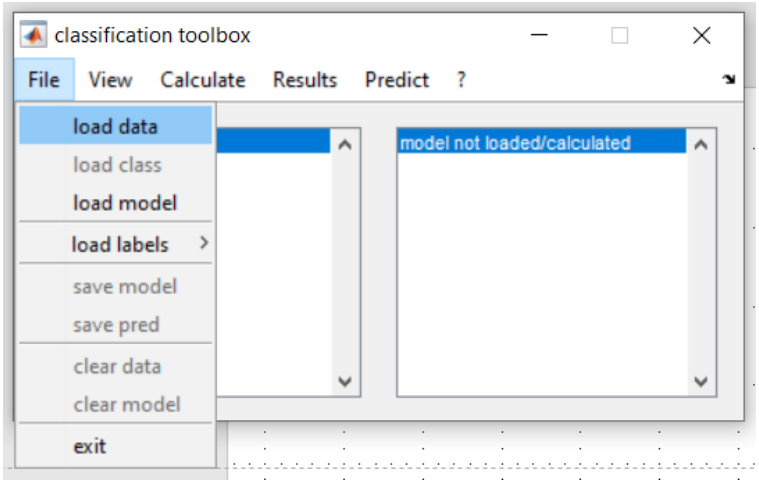
Step H1: Navigate within MATLAB to the Classification Toolbox for MATLAB folder (step A3). Right-click on the folder and select Add to Path > Select Folders and Subfolders.



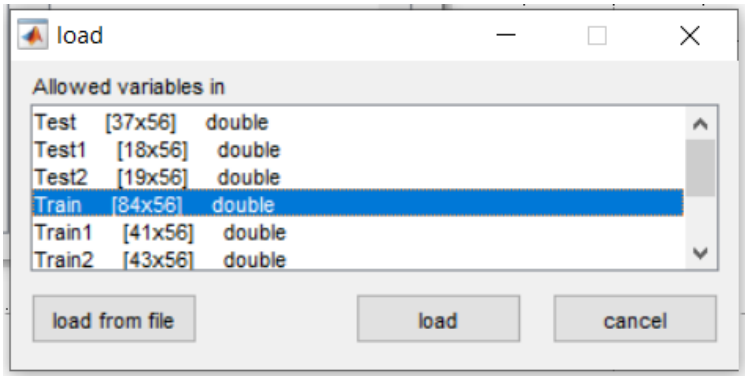
Step H2: Type ‘class_gui’.



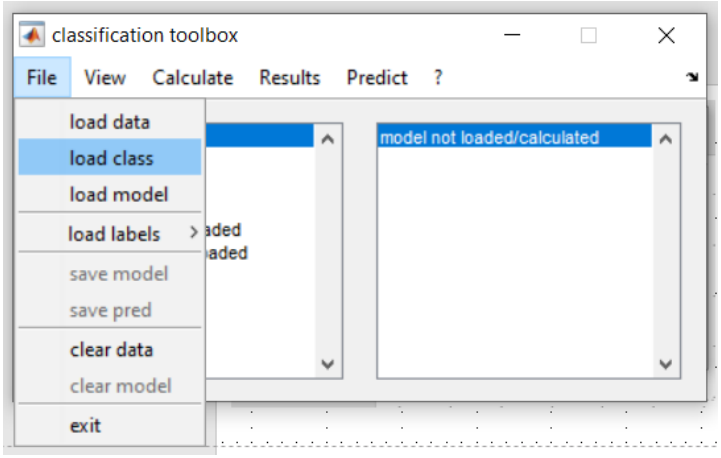
Step H3: Go to File > load data.



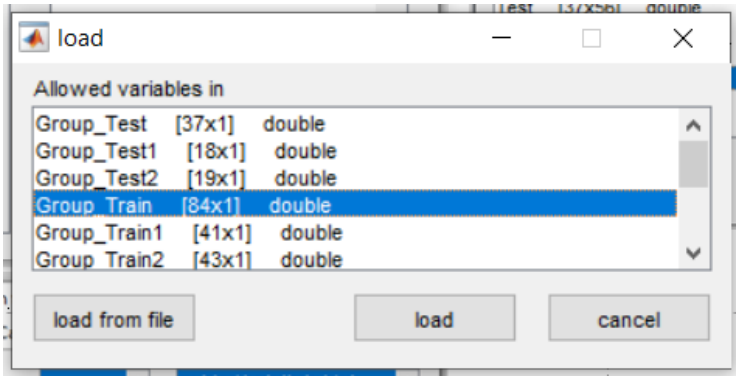
Step H4: Select the training data.



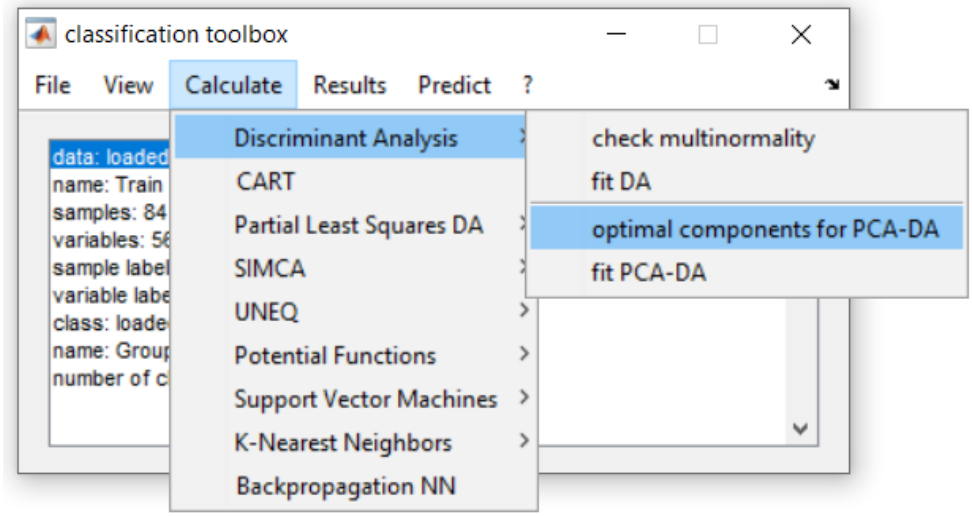
Step H5: Go to File > load class.



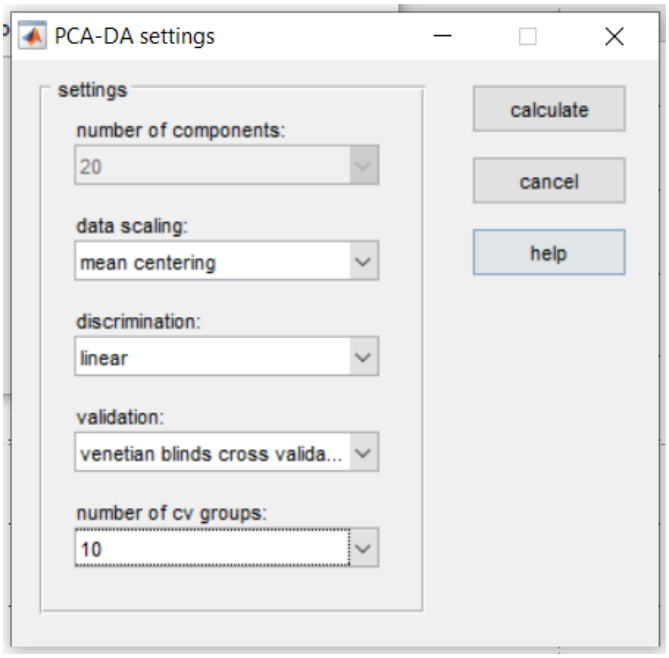
Step H6: Select the training class labels.



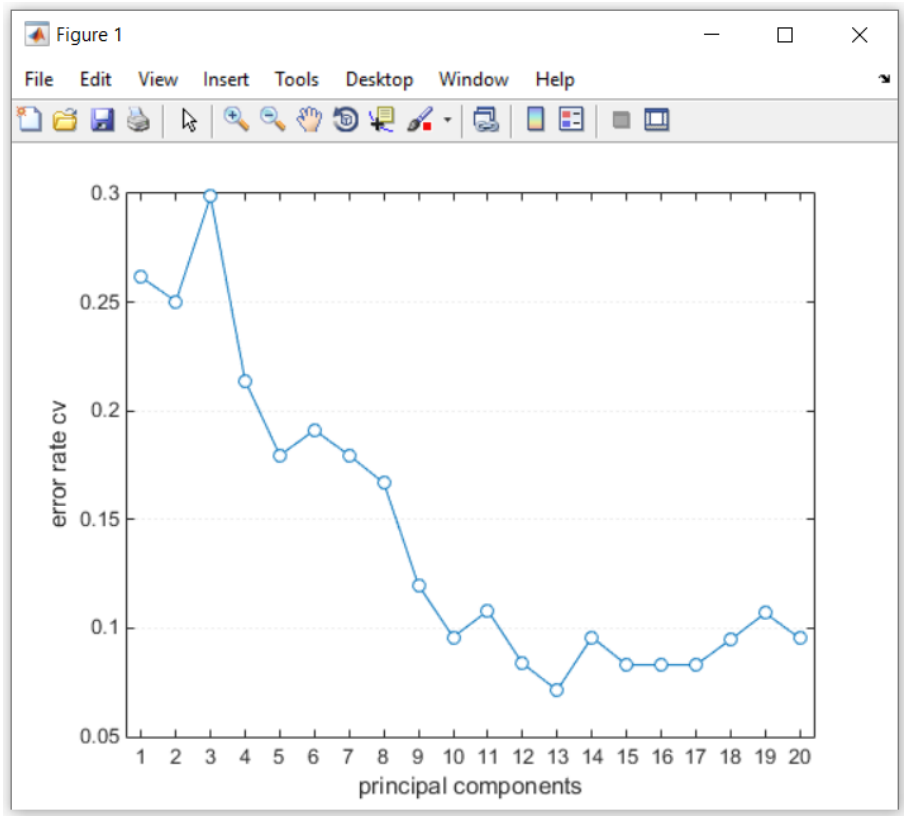
Step H7: Go to Calculate > Discriminant Analysis > optimal components for PCA-DA.



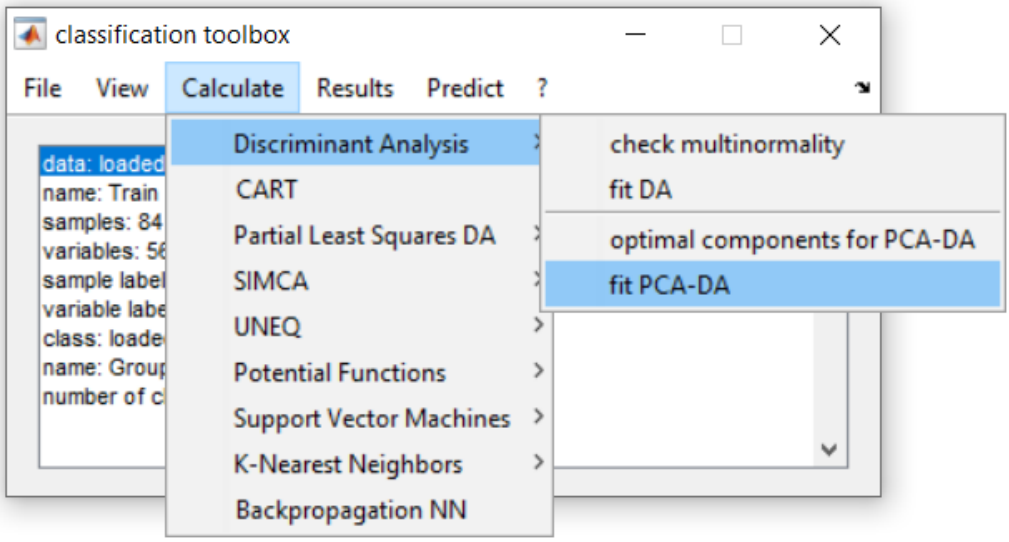
Step H8: Use the following settings and click on ‘calculate’. To perform a PCA-QDA model, change ‘linear’ to ‘quadratic’ in the ‘discrimination’ field.



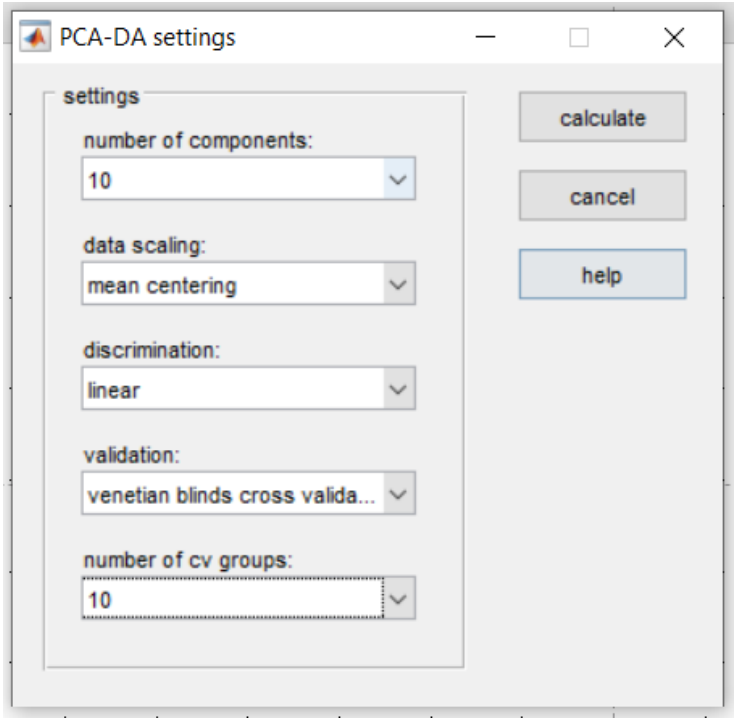
Step H9: Select the number of principal components to retain (10 in this case).



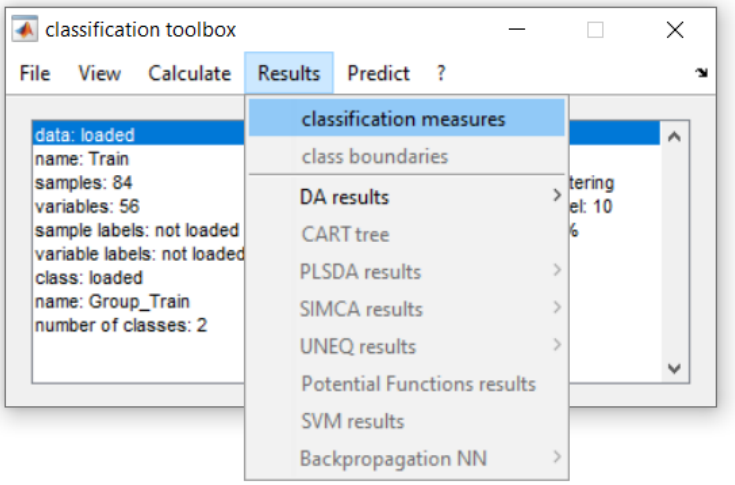
Step H10: Go to Calculate > Discriminant Analysis > fit PCA-DA.

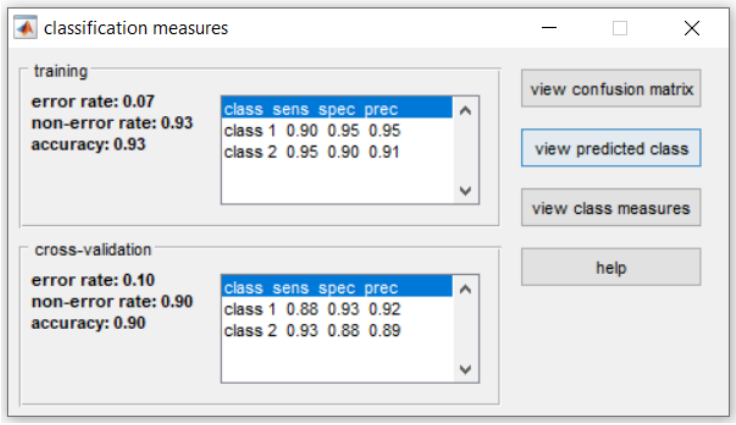


Step H11: Use the following settings and click on 'calculate'. To perform a PCA-QDA model, change 'linear' to 'quadratic' in the 'discrimination' field.

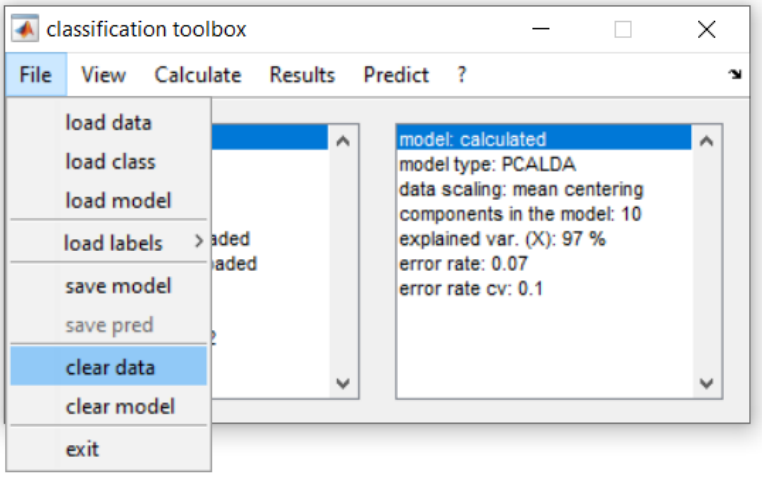


Step H12: Go to Results > classification measures, and observe the training and cross-validation metrics.

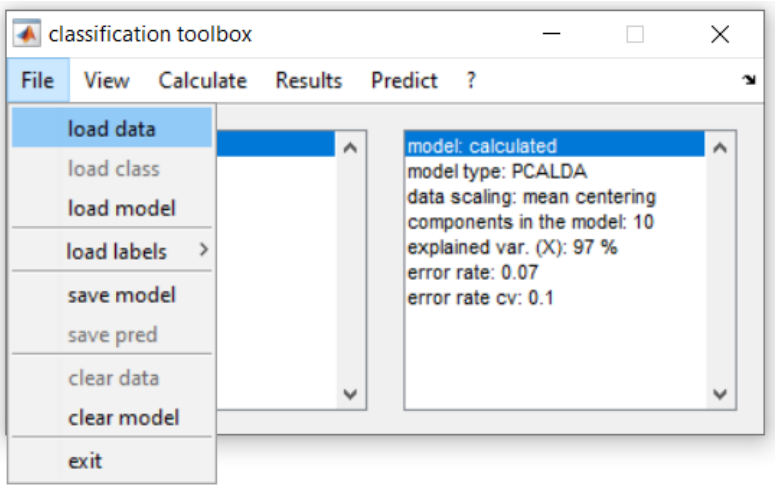




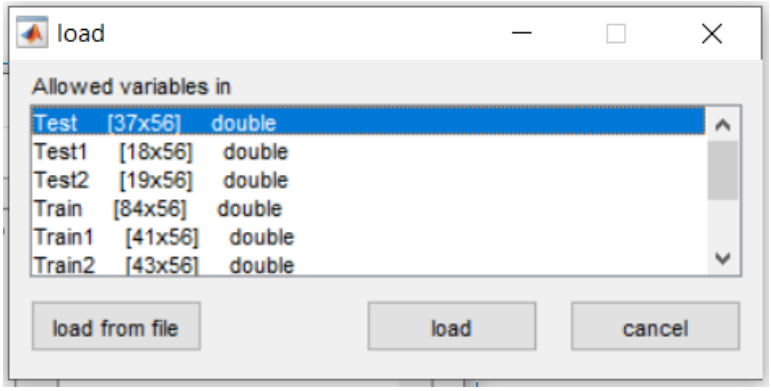
Step H13: Go to File > clear data.



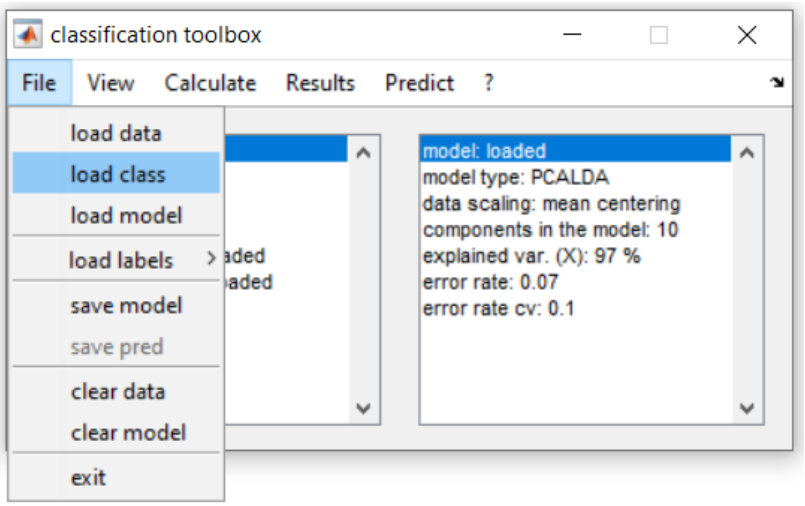
Step H14: Go to File > load data.



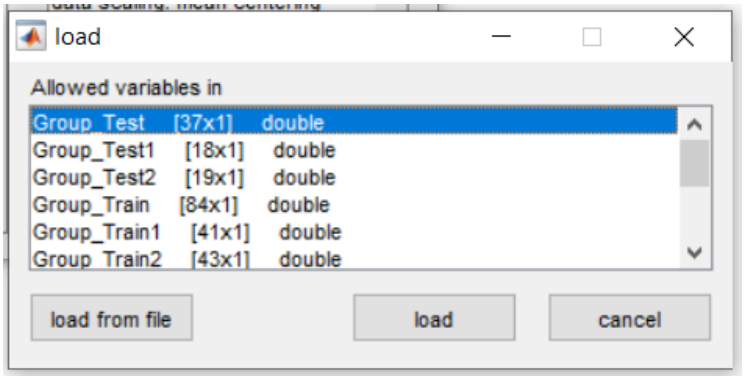
Step H15: Select the Test dataset and click on 'load'.



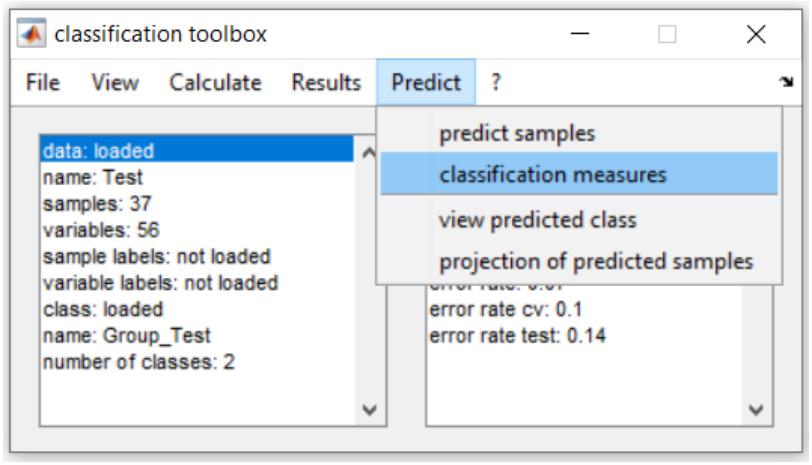
Step H16: Go to File > load class.

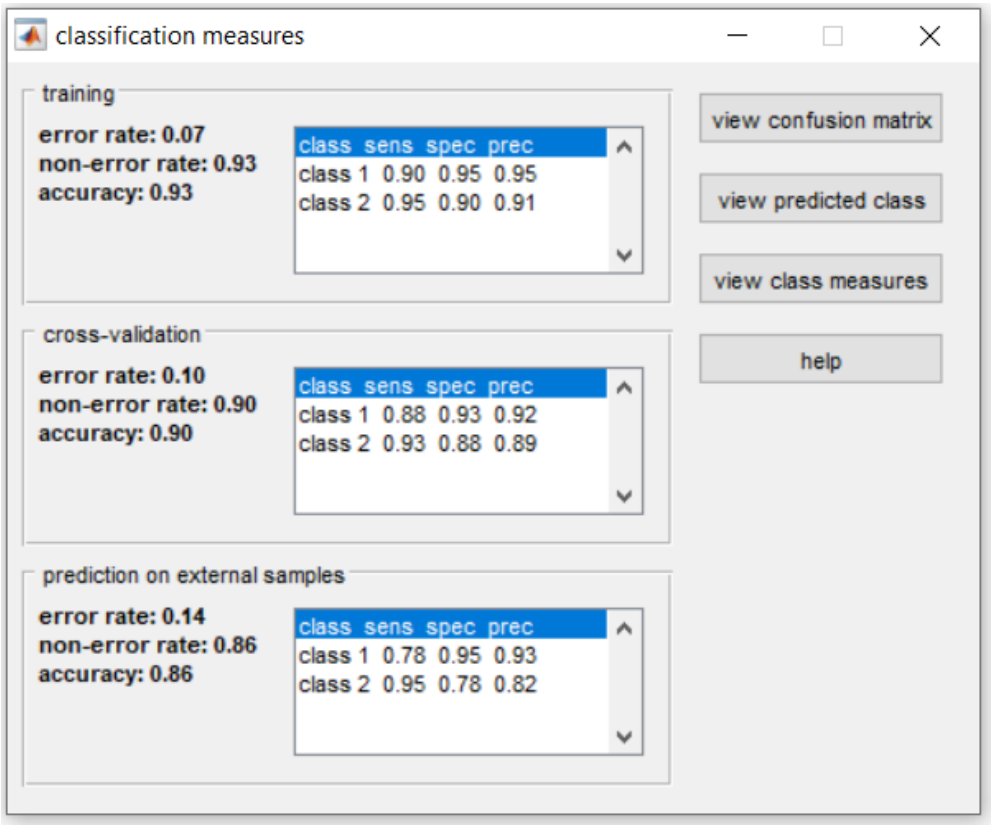


Step H17: Select the reference class labels for the Test set (this is used for calculation of the figures of merit only; the test samples are blind to the model).

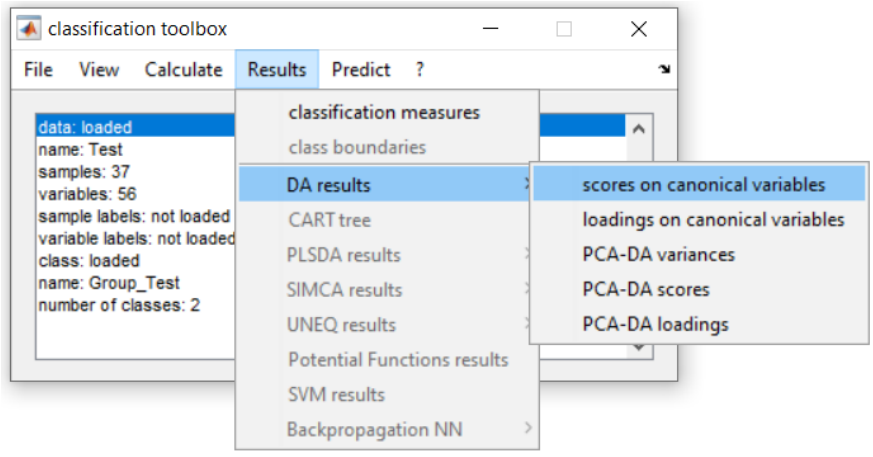


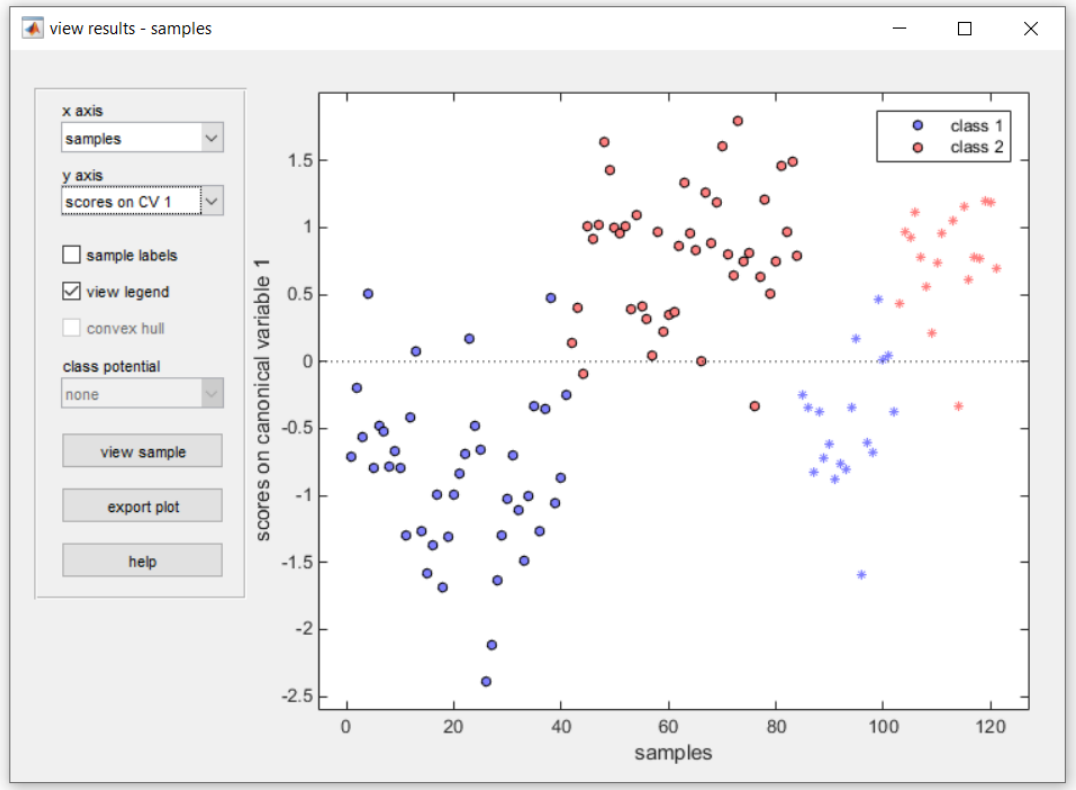
Step H18: Go to Predict > predict samples. Then, go to Predict > classification measures to see the classification performance for the test set.



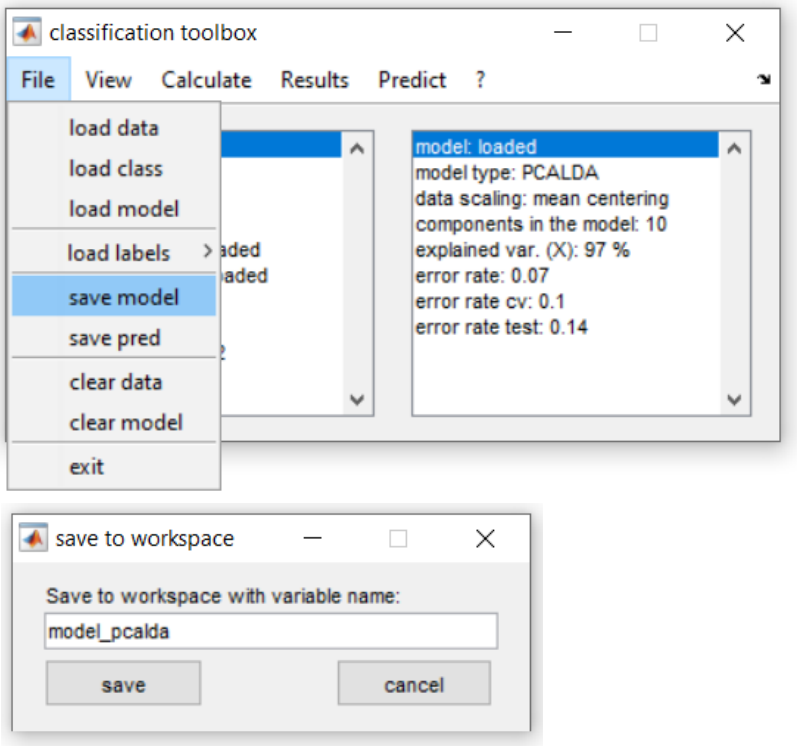


Step H19: Go to Results > DA results > scores on canonical variables to see the discriminant function (DF) plot for PCA-LDA, where o = training samples and * = test samples.



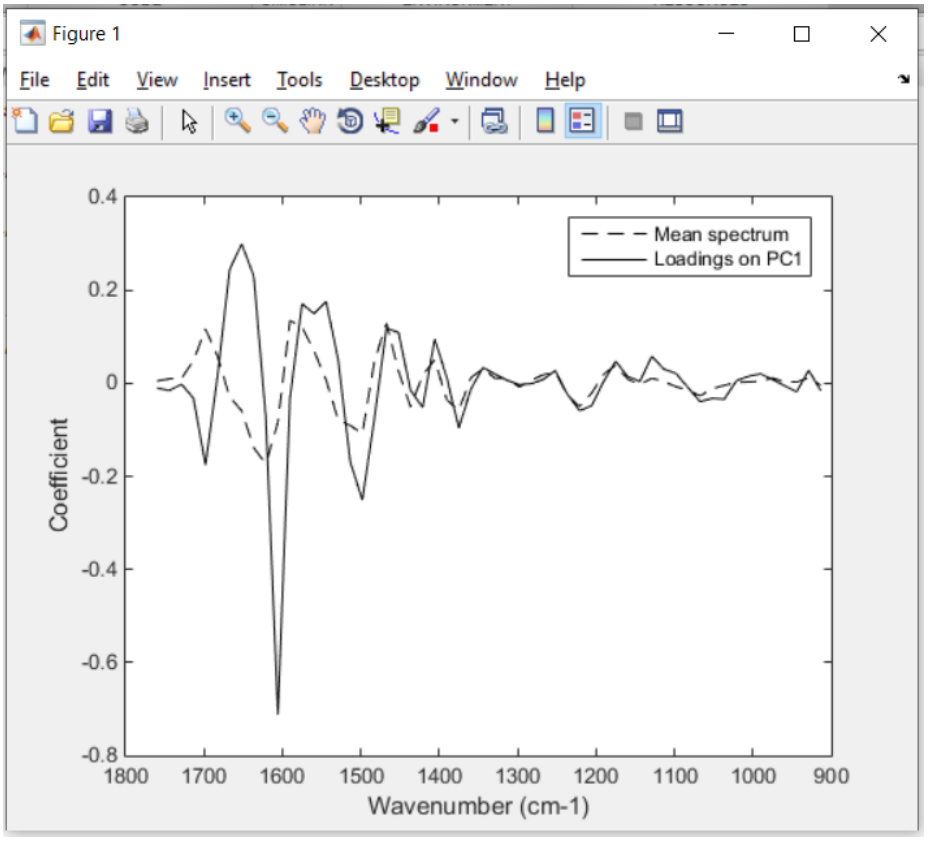


Step H20: Save the model.



Step H21: Extract the PCA loadings and plot the loadings on PC1 (main discriminant feature).

```
Command Window
>> loadings = model_pcalda.settings.modelpca.L;
>> figure,
>> plot(cm,mean(Train),'--k');
>> hold on
>> plot(cm,loadings(:,1),'-k');
>> xlabel('Wavenumber (cm-1)');
>> ylabel('Coefficient');
>> legend({'Mean spectrum','Loadings on PC1'});
>> set(gca,'Xdir','reverse');
>>
```



APPENDIX B – SUPPLEMENTARY MATERIAL FOR CHAPTER 5

Table B1. Correct classification rate for distinguishing Grade I and Grade II meningiomas.

Algorithm	Class	Training	Test
PCA-LDA	Grade I	80.0	31.6
	Grade II	66.7	85.7
PCA-QDA	Grade I	97.8	100
	Grade II	73.3	85.7
PCA-SVM	Grade I	100	73.7
	Grade II	100	28.6
SPA-LDA	Grade I	75.6	42.1
	Grade II	66.7	100
SPA-QDA	Grade I	95.6	100
	Grade II	46.7	85.7
SPA-SVM	Grade I	77.8	21.1
	Grade II	100	71.4
GA-LDA	Grade I	100	63.2
	Grade II	93.3	57.1
GA-QDA	Grade I	100	100
	Grade II	86.7	0
GA-SVM	Grade I	91.1	42.1
	Grade II	100	42.9

Figure B1. Outliers identified by a Hotelling T^2 *versus* Q residuals test (PCA with 8 PCs). (a) Meningioma Grade I samples (outliers: 58, 66); (b) meningioma Grade II samples (outliers: 11, 18); (c) meningioma Grade I outlier spectra in red; (d) meningioma Grade II outlier spectra in red.

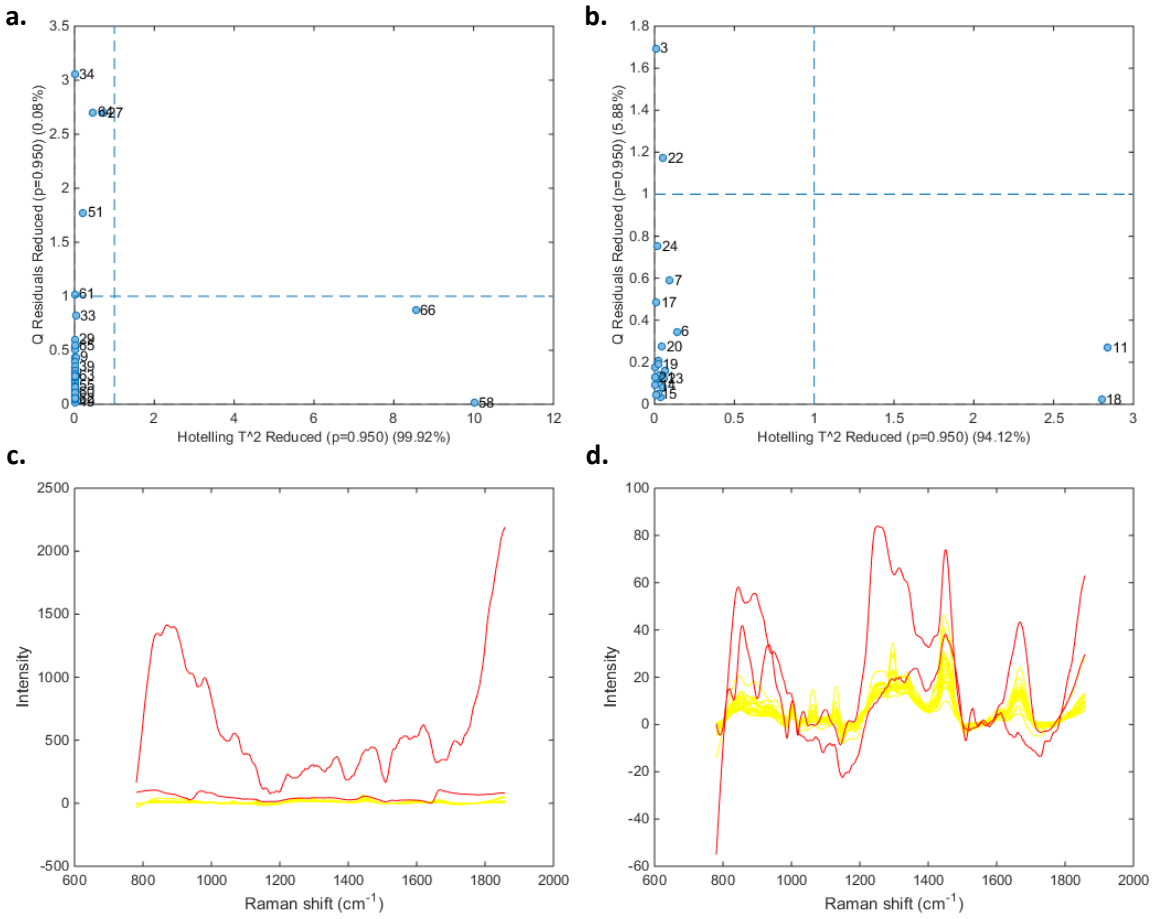


Figure B2. Singular value varying the number of principal components (PCs) of PCA.

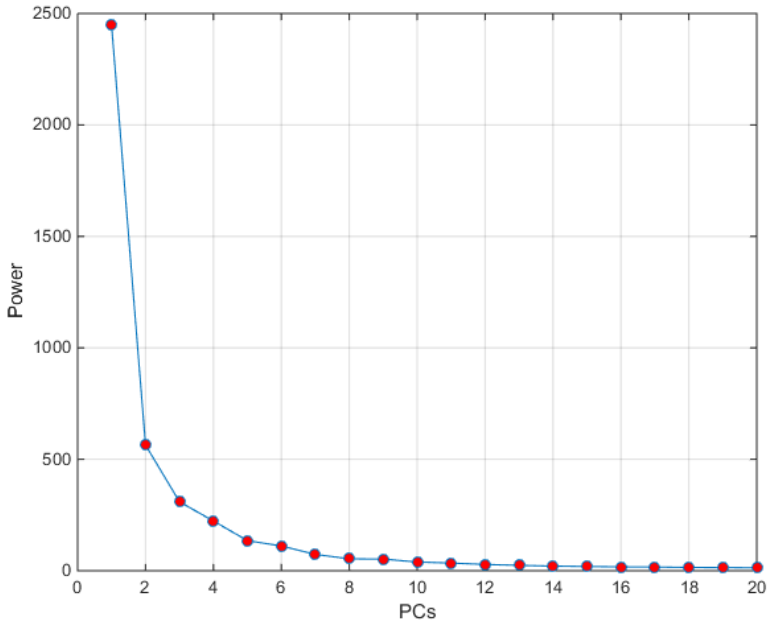
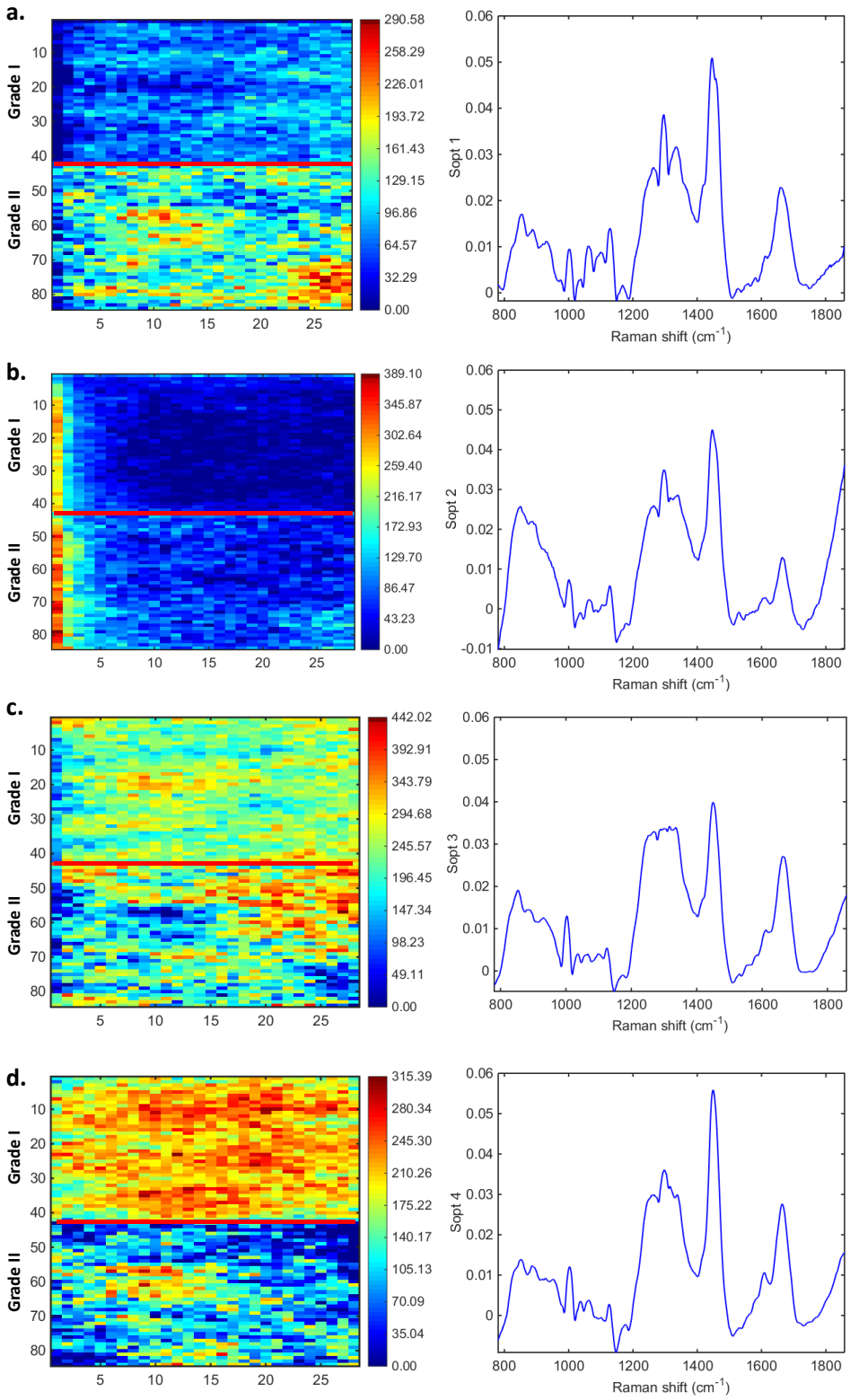


Figure B3. Concentration distribution maps and recovered spectral profiles by MCR-ALS for the 1st (a), 2nd (b), 3rd (c), and 4th (d) components. Colour bar: relative concentration.



APPENDIX C – SUPPLEMENTARY MATERIAL FOR CHAPTER 9

C1. Supplementary Material: Additional Results from Pilot Study

A. Effect of different instruments

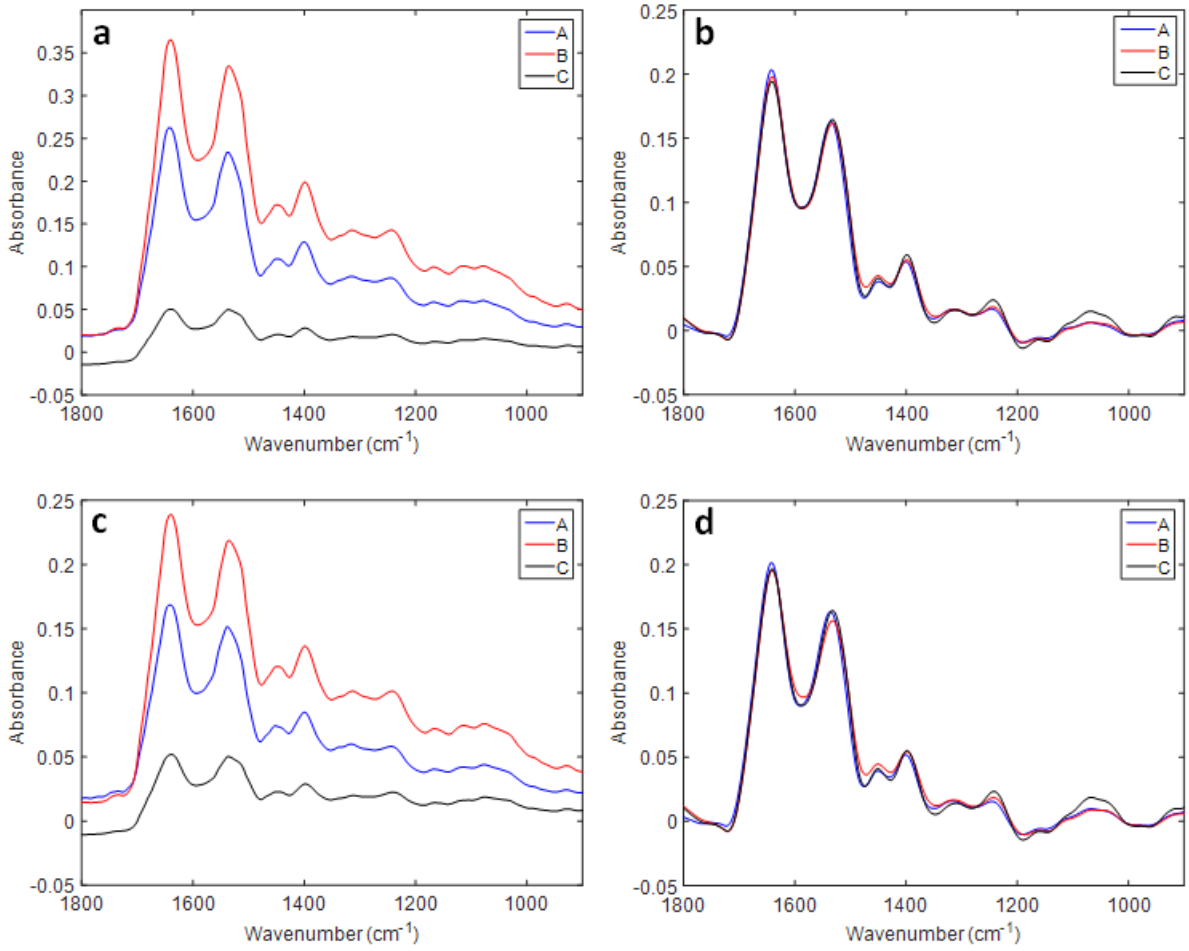


Figure C1.1. IR spectra of same type of samples measured by different ATR-FIR spectrometers at the same institution. Average (a) raw and (b) pre-processed spectra for healthy controls samples; average (c) raw and (d) pre-processed spectra for cancer samples across three different instruments (A, B and C).

B. Effect of different instruments

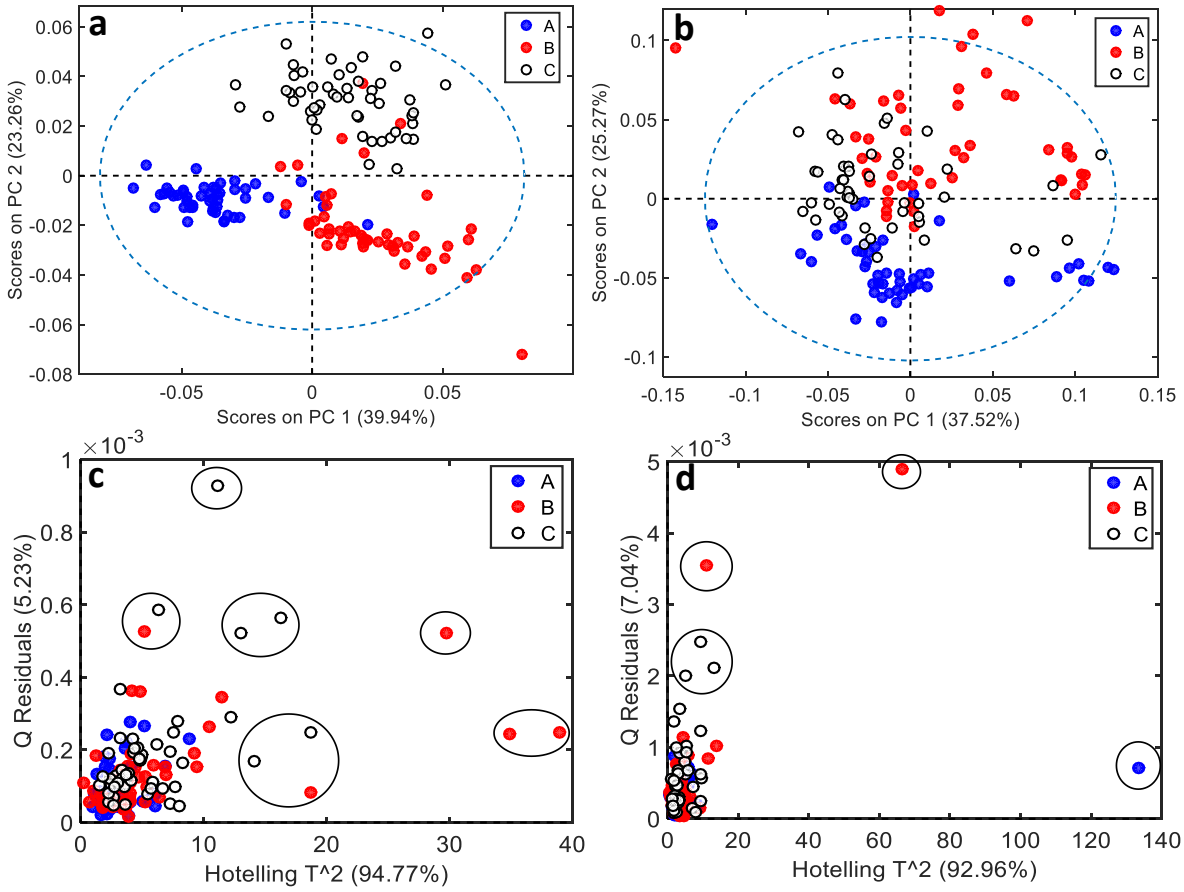


Figure C1.2. PCA scores for pre-processed spectra acquired by different ATR-FIR spectrometers at the same institution and outlier detection test. (a) PCA scores for healthy control samples according to the instrument used for spectra acquisition (A, B and C); (b) PCA scores for cancer samples according to the instrument used for spectra acquisition (A, B and C); (c) Hotelling T^2 *versus* Q residual test for healthy control samples according to the instrument used for spectra acquisition (A, B and C) based on a PCA using 5 PCs (94.77% cumulative variance); (d) Hotelling T^2 *versus* Q residual test for cancer samples according to the instrument used for spectra acquisition (A, B and C) based on a PCA using 5 PCs (92.96% cumulative variance). Circled samples in (c) and (d) indicate outliers removed. Confidence ellipse was 95%, depicted in blue in (a) and (b).

C. Effect of different instruments

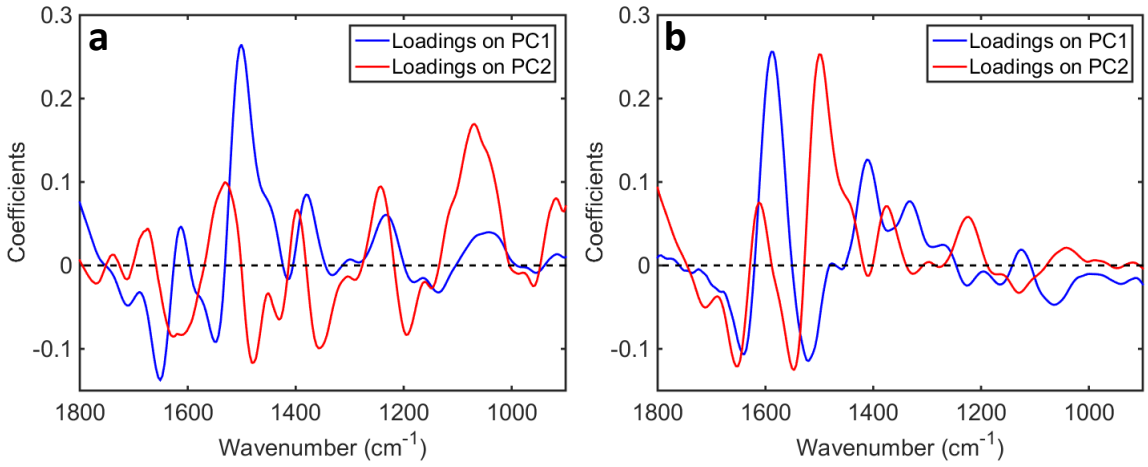


Figure C1.3. PCA loadings for pre-processed spectra acquired by different ATR-FIR spectrometers at the same institution. (a) PCA loadings for healthy control samples measured in different instruments (A, B and C); (b) PCA loadings for cancer samples measured in different instruments (A, B and C).

D. Effect of different operators

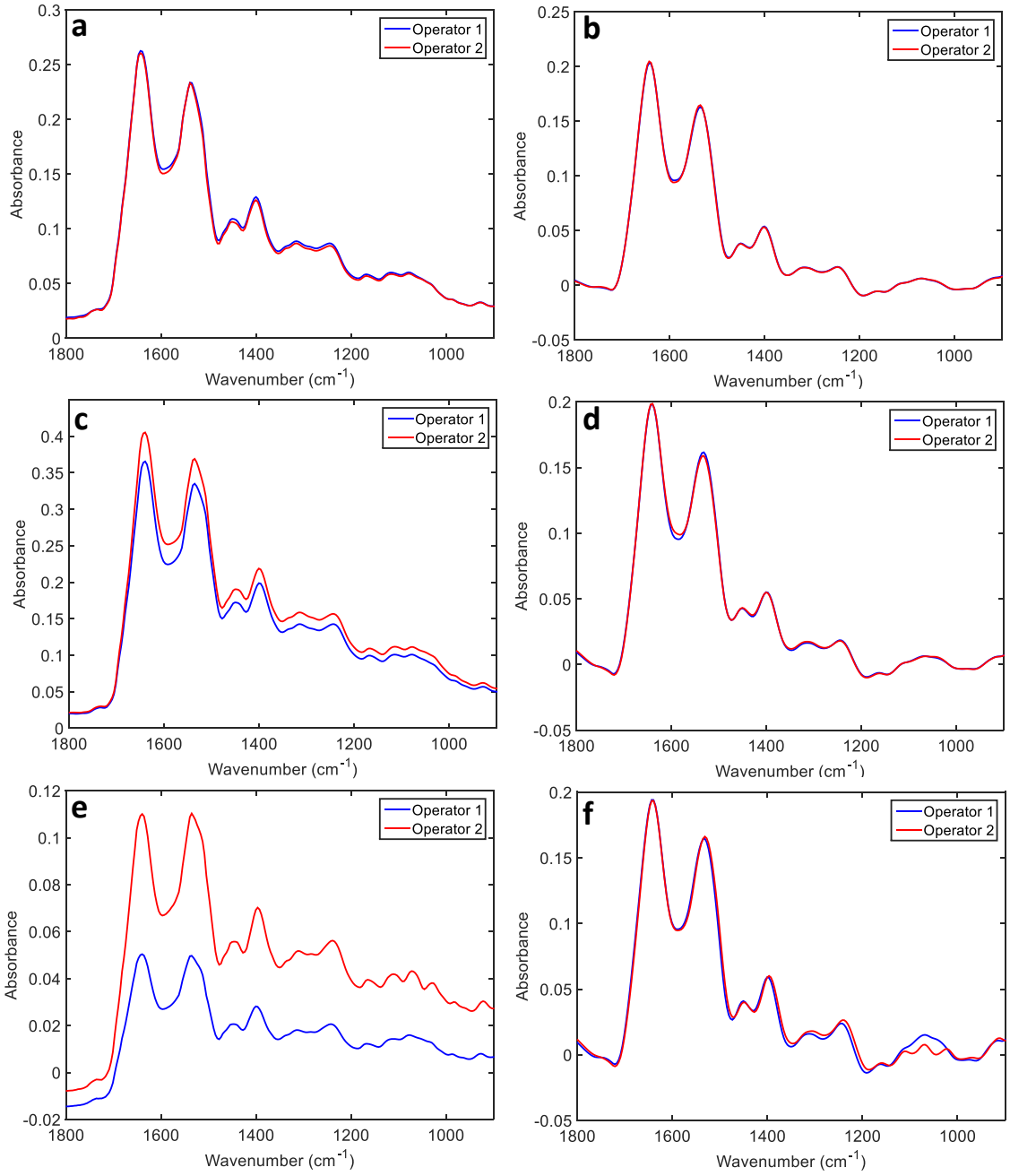


Figure C1.4. IR spectra of healthy control samples measured by different operators at the same institution. Average (a) raw and (b) pre-processed spectra for healthy control samples acquired with instrument A depending on the operator; average (c) raw and (d) pre-processed spectra for healthy control samples acquired with instrument B depending on the operator; average (e) raw and (f) pre-processed spectra for healthy control samples acquired with instrument C varying the operator.

E. Effect of different operators

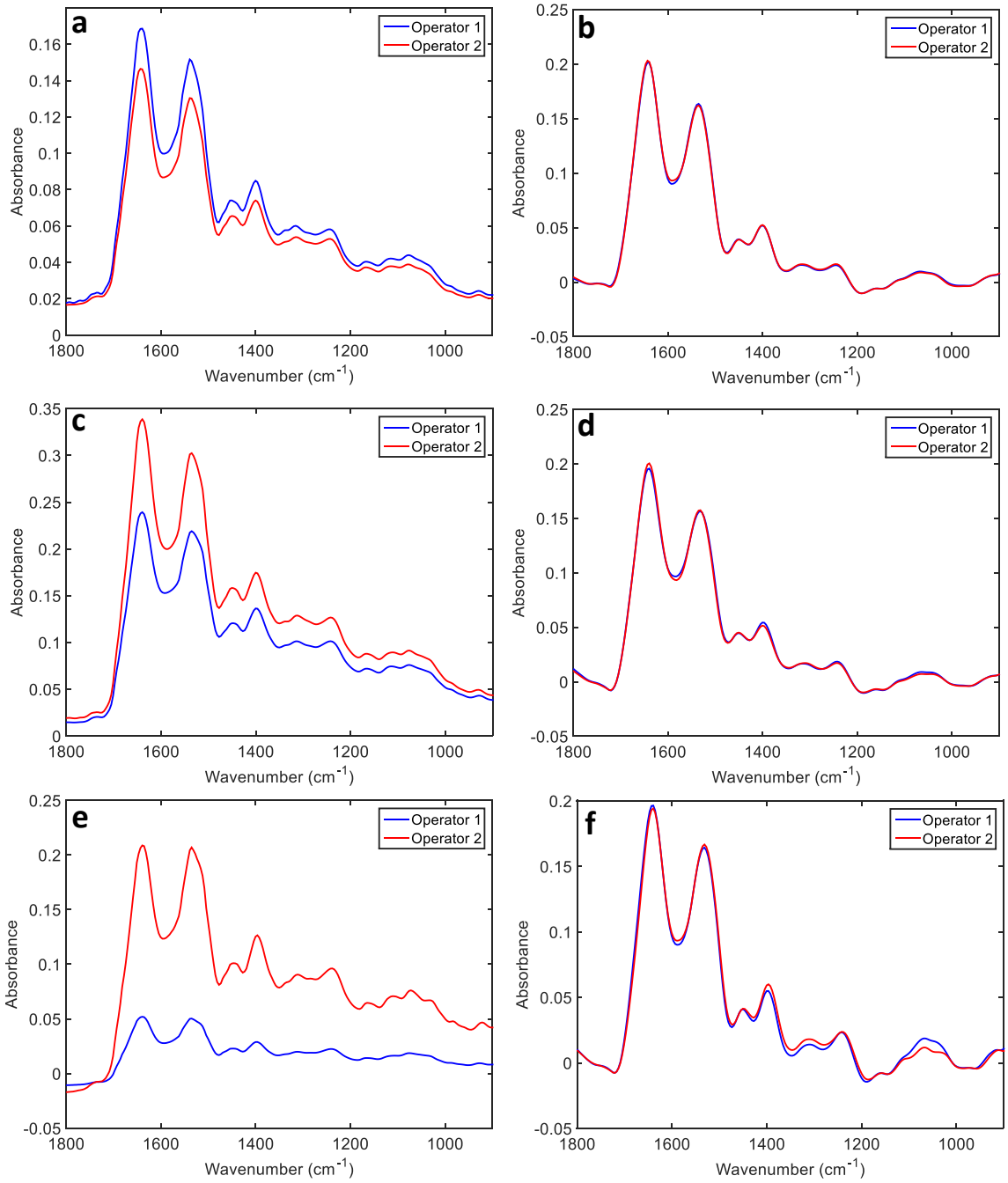


Figure C1.5. IR spectra of ovarian cancer samples measured by different operators at the same institution. Average (a) raw and (b) pre-processed spectra for cancer samples acquired with instrument A depending on the operator; average (c) raw and (d) pre-processed spectra for cancer samples acquired with instrument B depending on the operator; average (e) raw and (f) pre-processed spectra for cancer samples acquired with instrument C depending on the operator.

F. Effect of different operators

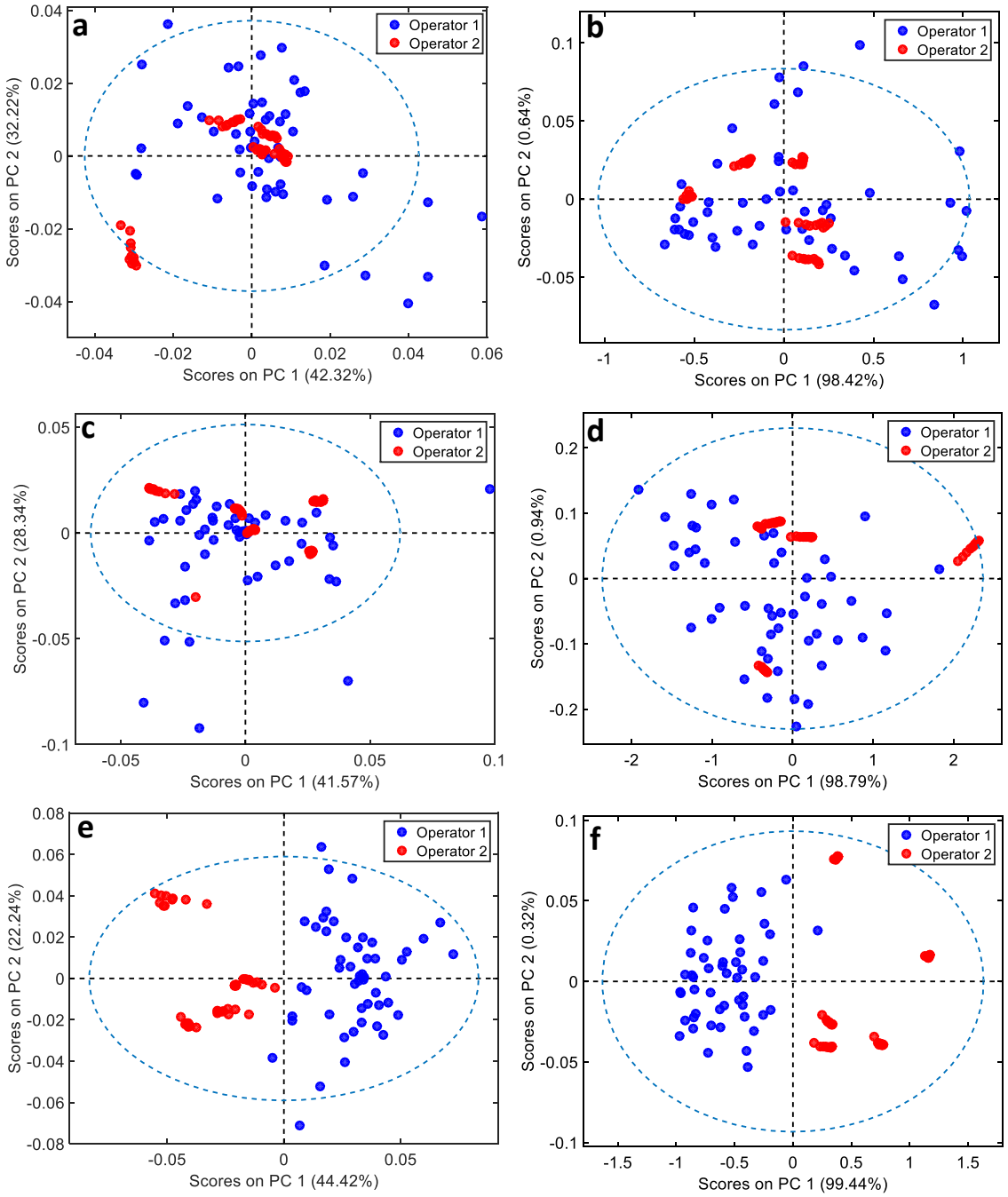


Figure C1.6. PCA scores for pre-processed spectra acquired by different operators at the same institution. PCA scores for (a) healthy control and (b) cancer samples acquired with instrument A depending on the operator; PCA scores for (c) healthy control and (d) cancer samples acquired with instrument B depending on the operator; PCA scores for (e) healthy control and (f) cancer samples acquired with instrument C depending on the operator. Confidence ellipse was 95%, depicted in blue

G. Effect of different instruments and operators

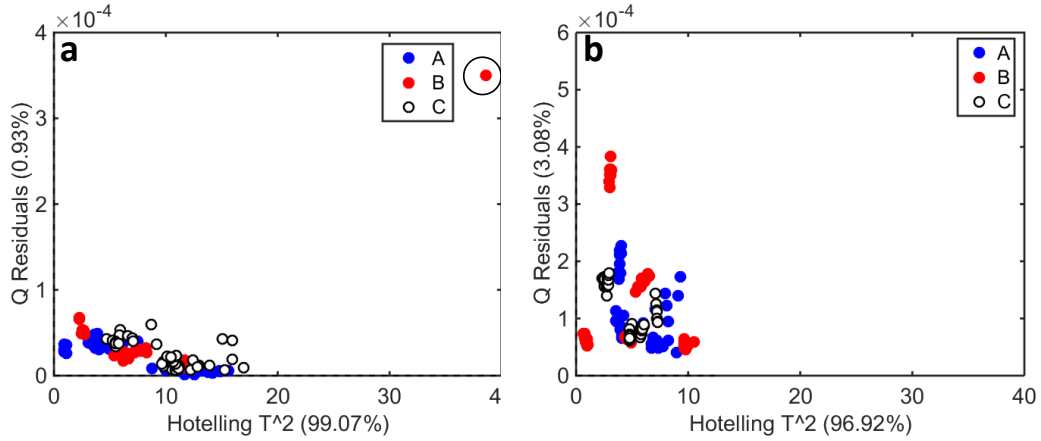


Figure C1.7. Outlier detection test for healthy controls and ovarian cancer samples. (a) Hotelling T^2 *versus* Q residual test based on a PCA using 8 PCs (99.07% cumulative variance) for healthy control samples depending on the instrument for spectra acquisition (A, B and C) used by Operator 2; (b) Hotelling T^2 *versus* Q residual test based on a PCA using 5 PCs (96.92% cumulative variance) for cancer samples depending on the instrument for spectra acquisition (A, B and C) used by Operator 2. Circled sample in a) indicates an outlier removed. The Hotelling T^2 *versus* Q residual test for Operator 1 is depicted in Fig. S2c-d.

H. Effect of different classes

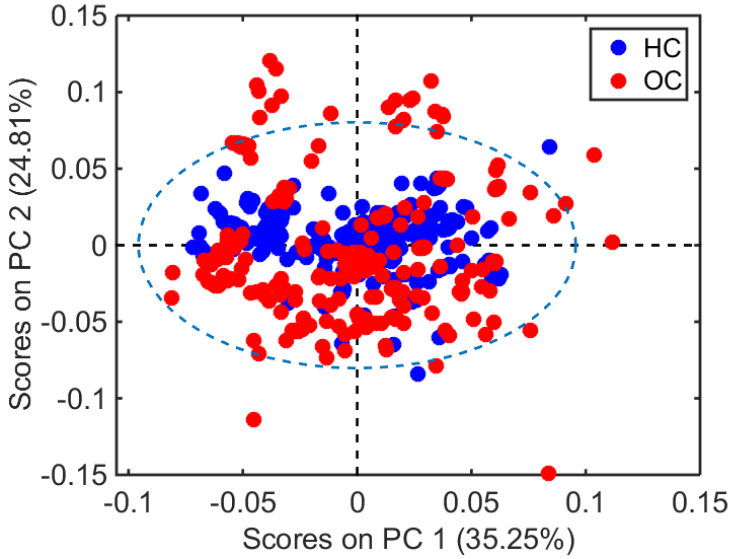


Figure C1.8. PCA scores for healthy controls (HC) and ovarian cancer (OC) samples based on the spectra acquired by both operators (1 and 2) and by all instruments (A, B and C). Confidence ellipse at a 95% confidence level is depicted in blue

C2. Supplementary Method: Protocol for Outliers Detection

A. Outlier detection using Hotelling T^2 *versus* Q residuals test

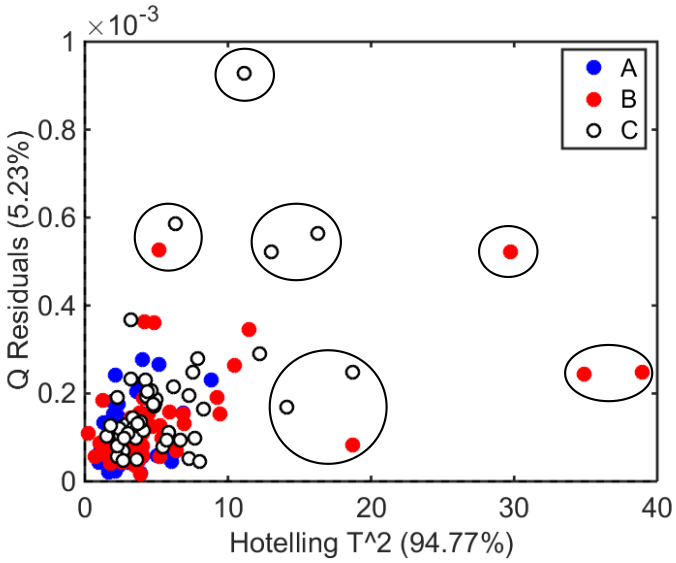
1st step: Build a PCA model.

2nd step: Calculate Hotelling T^2 and Q residuals.

3rd step: Plot Hotelling T^2 *versus* Q residuals

4th step: Select the samples which are most distant to the plot origin (0,0) and remove them one at a time from the data set. This procedure can be performed manually after visual inspection or automatically by algorithms.

Figure S1. Hotelling T^2 *versus* Q residuals for healthy control samples (blood plasma) varying the instrument for spectra acquisition (A, B and C). PCA performed with 5 PCs (94.77% cumulative variance). Circled samples indicate outliers removed.



B. Automatic outlier detection using MATLAB®

Algorithm link to download:

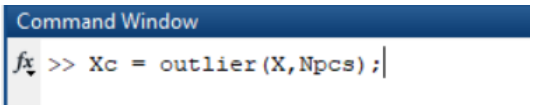
https://doi.org/10.6084/m9.figshare.7066613.v2

1st step: Add the .m files within the file downloaded to the path.

2nd step: Load the spectral data into MATLAB and organize all the spectra into a single matrix "X" containing each spectrum as a row.

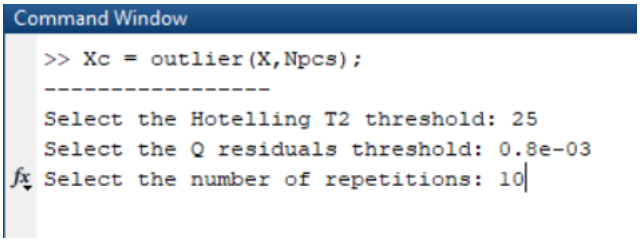
3rd step: Perform an initial PCA model to determine the number of principal components (PCs) to work with.

4th step: Run the algorithm as follows:

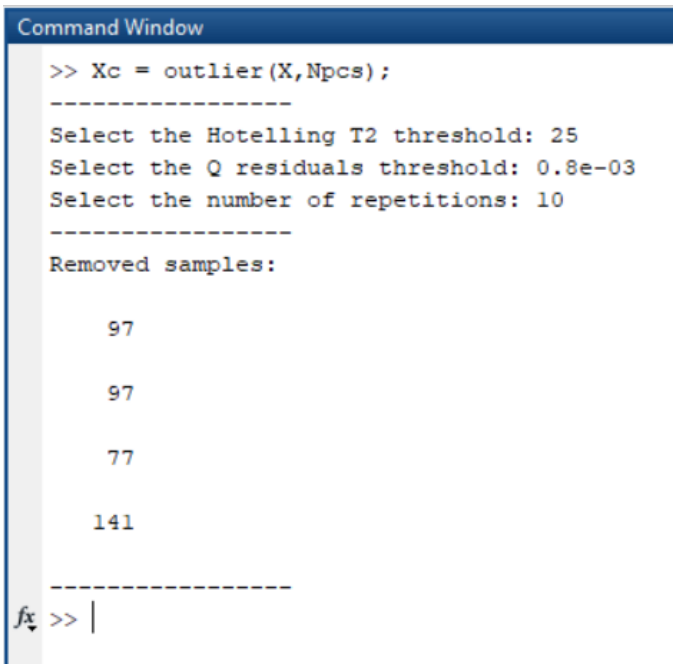


where "Xc" is the spectral matrix without outliers, "X" is the input spectral data, and "Npcs" the number of PCs for PCA.

5th step: Input optimization parameters:



In this case, the algorithm will perform a PCA model 10 times removing one sample at a time that follows one of these criteria: Hotteling $T^2 > 25$ or Q residuals > $0.8x10^{-3}$. Then, these samples are automatic excluded from the new dataset (Xc). The list of excluded samples is also displayed in MATLAB. Example:



APPENDIX D – ETHICS APPROVAL

27 July 2018

Frank Martin and Camilo De Lelis Medeiros-de-morais

School of Pharmacy and Biomedical Sciences

University of Central Lancashire

Dear Frank and Camilo

Re: STEMH Ethics Committee Application
Unique reference Number: STEMH 917

The STEMH ethics committee has granted approval of your proposal application 'Novel chemometric approaches towards handling biospectroscopy datasets'. Approval is granted up to the end of project date*. It is your responsibility to ensure that

- the project is carried out in line with the information provided in the forms you have submitted
- you regularly re-consider the ethical issues that may be raised in generating and analysing your data
- any proposed amendments/changes - including transfer of samples to another researcher - to the project are raised with, and approved, initially by BTNW and then submitted to STEMH
- you notify EthicsInfo@uclan.ac.uk if the end date changes or the project does not start
- serious adverse events that occur from the project are reported to Committee
- a closure report is submitted to complete the ethics governance procedures (Existing paperwork can be used for this purposes e.g. funder's end of grant report; abstract for student award or NRES final report. If none of these are available use e-Ethics Closure Report Proforma).
- human tissue held under this project (which has been approved by BTNW) is stored and used in accordance with the HTA licence requirements. At the end of the project any unused human tissue samples must be returned to BTNW for further use/storage or

appropriate disposal. Samples that do not fall within the HTA's definition of 'relevant material' should be disposed of in accordance with all relevant H&S requirements including any specific BTNW disposal arrangements.

Yours sincerely

Karen Rouse

Chair

STEMH Ethics Committee

Cc UCLan HT Technician

* for research degree students this will be the final lapse date

NB - Ethical approval is contingent on any health and safety checklists having been completed, and necessary approvals as a result of gained.