A three-dimensional discriminant analysis approach for hyperspectral images

Camilo L. M. Morais[1]*, Panagiotis Giamougiannis[2], Rita Grabowska[1], Nicholas J. Wood[2], Pierre L. Martin-Hirsch[2], Francis L. Martin[3]*

[1]School of Pharmacy and Biomedical Sciences, University of Central Lancashire, Preston PR1 2HE, United Kingdom

[2]Department of Obstetrics and Gynaecology, Lancashire Teaching Hospitals NHS Foundation Trust, Preston PR2 9HT, United Kingdom

[3]Biocel Ltd, Hull 10 7TS, United Kingdom

*cdlmedeiros-de-morai@uclan.ac.uk / flm13@biocel.uk

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 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 discriminate simulated and real-world data, comprising benign controls and ovarian cancer samples based on Raman hyperspectral imaging, in which 3D-PCA-LDA and 3D-PCA-QDA achieved far superior performance than classical algorithms using unfolding procedures (PCA-LDA, PCA-QDA, partial lest squares discriminant analysis [PLS-DA], and support vector machines [SVM]), where the classification accuracies improved from 66% to 83% (simulated data) and from 50% to 100% (real-world dataset) 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.

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".[1]

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.[2-5] One of these imaging techniques that has found increasing application is Raman hyperspectral imaging,[6] which is a generally non-destructive technique where a spectral response is obtained based on molecular polarizability changes.[7] Raman hyperspectral imaging has been used in a wide range of applications, such as, pharmaceutical analysis,[8] food quality control,[9] forensic studies,[10] and to investigate biological materials.[11] 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 successfully applied to identify brain,[12] breast,[13] cervical,[14] lung,[14] and skin cancer.[15]

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.[16] 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;[17] in which as series of algorithms are employed to perform these tasks, *e.g.*, principal component analysis (PCA) for feature extraction and exploratory analysis,[18] partial least squares (PLS) for feature extraction and quantification,[19] partial least squares discriminant analysis (PLS-DA) for feature extraction and classification,[20] and multivariate curve resolution alternating least squares (MCR-ALS) for feature extraction, exploratory analysis, calibration and construction of concentration distribution maps.[16,21]

Since most algorithms used to process hyperspectral images are first-order-based, *i.e.*, applied to a 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.*[22] reported a 2D linear discriminant analysis (2D-LDA) algorithm to classify three-way chemical data; Morais and Lima[23] 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.*[1] have reported a 3D-PCA approach to perform exploratory analysis in hyperspectral images.

In this paper, 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 simulated and real-world data comprised of benign controls and ovarian cancer patients based on the Raman hyperspectral imaging.

2. Methods

2.1 Samples

Two datasets were used in this study. Dataset 1 is a simulated dataset composed of 60 hyperspectral images with dimensions 20 × 20 × 301 (301 spectral wavenumbers randomly generated using a normal distribution function) divided into two classes. Class 1 contains 30 hyperspectral images with mean ranging from 0.0722–0.2146 and standard-deviation ranging from 0.1988–0.5815 intensity units, and class 2 contains 30 hyperspectral images with mean ranging from 0.1421–0.2544 and standard-deviation ranging from 0.2592–0.7134 intensity units.

Dataset 2 is composed of 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 (numerical aperture [NA] = 0.45) 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 (1.20 cm^{-1} data spacing, 725–1813 cm^{-1}).

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,[24] and 3D discriminant analysis (3D-PCA-LDA, 3D-PCA-QDA) were performed using lab-made algorithms. The simulated dataset was not pre-processed. The real-world pre-processed hyperspectral images were split into training (70%) and test (30%) sets using the Kennard-Stone (KS) uniform sample selection algorithm.[25]

2.3 Computational analysis

Unfolded PCA-LDA and PCA-QDA where 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.[1,18,26]

In this 3D-PCA approach,[1] a local bilinear PCA model is performed for each pixel position across the hyperspectral image dataset as follows:

$$\mathbf{X}_{ij}^{*} = \mathbf{T}_{ij}\mathbf{P}_{ij}^{\mathrm{T}} + \mathbf{E}_{ij} \quad (1)$$

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 ($\underline{\mathbf{T}}$), loadings ($\underline{\mathbf{P}}$), and residuals ($\underline{\mathbf{E}}$). 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;[27,28] and also different of Tucker3 model with orthogonal factors, known as "three-way PCA".[29,30]

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:[31]

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

$$Q_{sk} = (\mathbf{x}_s - \overline{\mathbf{x}}_k)^{\mathrm{T}}\mathbf{C}_k^{-1}(\mathbf{x}_s - \overline{\mathbf{x}}_k) + \log_e|\mathbf{C}_k| - 2\log_e \pi_k \quad (3)$$

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; $\mathbf{C}_{\mathrm{pooled}}$ 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*. $\mathbf{C}_{\mathrm{pooled}}$, $\mathbf{C}_k$ and π_k are calculated as follows:

$$\mathbf{C}_{\text{pooled}} = \frac{1}{n}\sum_{k=1}^{K} n_k \mathbf{C}_k \quad (4)$$

$$\mathbf{C}_k = \frac{1}{n_k - 1}\sum_{s=1}^{n_k}(\mathbf{x}_s - \bar{\mathbf{x}}_k)(\mathbf{x}_s - \bar{\mathbf{x}}_k)^{\mathrm{T}} \quad (5)$$

$$\pi_k = \frac{n_k}{n} \quad (6)$$

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 2–6 are performed with the PCA scores of the unfolded 3D array. The unfolded data were also tested using partial least squares discriminant analysis (PLS-DA) and support vector machines (SVM) with a radial basis function (RBF) kernel as comparative classification methods. All unfolded and 3D models were built using cross-validation leave-one-out for optimization of the number of factors (number of latent variables (LVs) in PLS-DA and kernel parameters in SVM), and evaluated using an external test set. The external test set contained 30% of the total number of samples (whole hyperspectral images) in the simulated and real-world datasets that did not participate in the model construction phase.

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).[23] These parameters are calculated as follows:

$$\text{Accuracy (\%)} = \left(\frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}\right) \times 100 \quad (7)$$

$$\text{Sensitivity (\%)} = \left(\frac{\text{TP}}{\text{TP+FN}}\right) \times 100 \quad (8)$$

$$\text{Specificity (\%)} = \left(\frac{\text{TN}}{\text{TN+FP}}\right) \times 100 \quad (9)$$

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.

3. Results and discussion

The mean simulated hyperspectral images for classes 1 (n = 30 hyperspectral images) and 2 (n = 30 hyperspectral images) of dataset 1 along with their mean spectral profiles are shown in Figure 1. This dataset was randomly generated using a normal distribution function. Distinguishing spectral features between classes 1 and 2 are clearly shown between the entire variable range. For dataset 2, 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 2a and 2b, 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.[32] The raw and pre-processed (SG smoothing and AWLS baseline correction) mean spectra for both groups of samples are depicted in Figures 2c and 2d, respectively.

Unfolded PCA-LDA, PCA-QDA, PLS-DA and SVM were applied to the pre-processed data of both datasets using cross-validation, and tested using an external test set containing 18 samples (hyperspectral images) for the simulated dataset and 11 samples (hyperspectral images) for the real-world dataset. Table 1 contains confusion matrices representing the training, cross-validation and test performances for the simulated dataset. In the simulated and real-world datasets, the minimum cross-validation error was used to

estimate the number of LVs in PLS-DA (simulated dataset: 6 LVs (84% cumulative explained variance); real-world dataset: 1 LV (11% cumulative explained variance)) and the RBF kernel parameters of SVM (simulated dataset: kernel parameter = 9, cost = 10, support vectors = 31; real-world dataset: kernel parameter = 0.14, cost = 1000 cost, support vectors = 27).

The unfolded and 3D-PCA-LDA/QDA models were built using 2 PCs. Figures 3a and 3c show the unfolded PCA-LDA and PCA-QDA calculated classification boundaries between class 1 and 2 of the simulated dataset, and Figures 3b and 3d shows the 3D-PCA-LDA and 3D-PCA-QDA calculated classification boundaries between these two classes. For the real-world dataset, the unfolded and 3D-PCA-LDA/QDA calculated classification boundaries between benign controls and ovarian cancer samples are shown in Figure 4. The PCA scores response were average per sample, so each point in Figures 3 and 4 represents a sample (image). The superposition pattern observed in Figures 4a and 4c reflects the poor classification of unfolded PCA-LDA and PCA-QDA as demonstrated in Table 2, where ovarian cancer samples are being highly misclassified in the test set (80% misclassification), and in Table 3, where accuracies (64%) and sensitivities (20%) for PCA-LDA and PCA-QDA are substantially low.

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 4b and 4d 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 4b), 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 2). The classification performance in the test set also substantially improved for the simulated dataset using the 3D-based algorithms (Table 3), where both 3D-PCA-LDA and 3D-PCA-QDA had a classification accuracy of 83%, in comparison with PCA-LDA (66%), PCA-QDA (67%), PLS-DA (72%) and SVM (72%). In the real-world dataset, there was a perfect discrimination in the test set using the 3D algorithms (accuracy, sensitivity and specificity equal to 100%) (Table 3), while the unfolded methods provided a maximum of 64% accuracy using PCA-LDA or PCA-QDA. These findings indicate the potential of 3D discriminant analysis compared to the unfolding procedure.

The difference-between-mean (DBM) spectrum and 3D-PCA loadings are show in Figures 5 (simulated dataset) and 6 (real-world dataset). The simulated data show distinguishing features particularly between variables 0 to 200 (Figure 5a and 5b). Both loadings on PC1 (Figure 5c) and PC2 (Figure 5d) indicate higher coefficients along variables 0 to 100, indicating this region was the most important for the separation pattern observed in the PCA scores. The ovarian cancer samples spectra appear to have overall higher intensity values than benign controls, as demonstrated in Figures 6a and 6b, 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 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 4d) (29). 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.[33] Vibrations around 820 cm^{-1} and 1400 cm^{-1} (PC1) have been previously reported as protein markers for cervical tumors[34] and ovarian cancer.[1]

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 unfolded methods (PCA-LDA, PCA-QDA, PLS-DA and SVM), where a much superior performance was obtained with the 3D-based techniques to discriminate simulated and real-world data composed of benign controls and ovarian cancer patients based on the Raman hyperspectral imaging. An improvement in the accuracy (66% to 83% (simulated data), 50% to 100% (real-world data)) and sensitivity (33% to 89% (simulated data), 0% to 100% (real-world data)) in the test set was observed when the 3D discriminant algorithms were applied. For the real-world dataset, 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.

Acknowledgment

CLMM would like to thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Brazil (grant 88881.128982/2016-01) for financial support.

Conflicts of interest

F.L.M. has recently joined Biocel Ltd, a company developing analytical techniques similar to those used herein.

References

1. C. L. M. Morais, P. L. Martin-Hirsch and F. L. Martin, *Analyst*, 2019, **144**, 2312–2319.
2. K. J. Zuzak, M. D. Schaeberle, E. Neil Lewis and I. W. Levin, *Anal. Chem.*, 2002, **74**, 2021–2028.
3. S. Türker-Kaya and C. W. Huck, *Molecules*, 2017, **22**, 68.
4. M. Pilling and P. Gardner, *Chem. Soc. Rev.*, 2016, **45**, 1935–1957.
5. A. R. Buchberger, K. DeLaney, J. Johnson and L. Li, *Anal. Chem.*, 2018, **90**, 240–265.
6. S. Lohumi, M. S. Kim, J. Qin and B. -K. Cho, *Trends Analyt. Chem.*, 2017, **93**, 183–198.
7. M. C. D. Santos, C. L. M. Morais, Y. M. Nascimento, J. M. G. Araujo and K. M. G. Lima, *Trends Analyt. Chem.*, 2017, **97**, 244–256.
8. L. M. Kandpal, B. -K. Cho, J. Tewari and N. Gopinathan, *Sens. Actuators B Chem.*, 2018, **260**, 213–222.
9. T. Yaseen, D. -W. Sun and J. -H. Cheng, *Trends Food Sci. Technol.*, 2017, **62**, 177–189.
10. M. R. Almeida, L. P. L. Logrado, J. J. Zacca, D. N. Correa and R.J. Poppi, *Talanta*, 2017, **174**, 628–632.
11. H. J. Butler, L. Ashton, B. Bird, G. Cinque, K. Curtis, J. Dorney, K. Esmonde-White, N. J. Fullwood, B. Gardner, P. L. Martin-Hirsch, M. J. Walsh, M. R. McAinsh, N. Stone and F. L. Martin, *Nat. Protoc.*, 2016, **11**, 664–687.
12. M. Kirsch, G. Schackert, R. Salzer and C. Krafft, *Anal. Bioanal. Chem.*, 2010, **398**, 1707–1713.

13. H. Abramczyk and B. Brozek-Pluska, *Chem. Rev.*, 2013, **113**, 5766–5781.

14. M. Diem, A. Mazur, K. Lenau, J. Schubert, B. Bird, M. Miljković, C. Krafft and J. Popp, *J. Biophotonics*, 2013, **6**, 855–886.

15. H. Lui, J. Zhao, D. McLean and H. Zeng, *Cancer Res.*, 2012, **72**, 2491–2500.

16. J. M. Prats-Montalbán, A. de Juan and A. Ferrer, *Chemometr. Intell. Lab. Syst.*, 2011, **107**, 1–23.

17. C. Duchesne, J. J. Liu and J. F. MacGregor, *Chemometr. Intell. Lab. Syst.*, 2012, **117**, 116–128.

18. R. Bro and A. K. Smilde, *Anal. Methods*, 2014, **6**, 2812–2831.

19. S. Wold, M. Sjöström and L. Eriksson, *Chemometr. Intell. Lab. Syst.*, 2001, **58**, 109–130.

20. R. G. Brereton and G. R. Lloyd, *J. Chemom.*, 2014, **28**, 213–225.

21. J. Jaumot, A. de Juan and R. Tauler, *Chemometr. Intell. Lab. Syst.*, 2015, **140**, 1–12.

22. A. C. da Silva, S. F. C. Soares, M. Insausti, R. K. H. Galvão, B. S. F. Band and M. C. U. de Araújo, *Anal. Chim. Acta*, 2016, **938**, 53–62.

23. C. L. M. Morais and K. M. G. Lima, *Chemometr. Intell. Lab. Syst.*, 2017, **170**, 1–12.

24. D. Ballabio and V. Consonni, *Anal. Methods*, 2013, **5**, 3790–3798.

25. R. W. Kennard and L. A. Stone, *Technometrics*, 1969, **11**, 137–148.

26. P. Geladi and B.R. Kowalski, *Anal. Chim. Acta*, 1986, **185**, 1–17.

27. L. R. Tucker, *Psychometrika*, 1966, **31**, 279–311.

28. C. L. M. Morais, K. M. G. Lima and F. L. Martin, *Chemometr. Intell. Lab. Syst.*, 2019, **188**, 46–53.

29. P. J. Gemperline, K. H. Miller, T. L. West, J. E. Weinstein, J. C. Hamilton and J. T. Bray, *Anal. Chem.*, 1992, **64**, 523A–532A.

30. P. M. Kroonenberg, K. E. Basford and P. J. Gemperline, *J. Chemom.*, 2004, **18**, 508–518.

31. C. L. M. Morais and K. M. G. Lima, *J. Braz. Chem. Soc.*, 2018, **29**, 472–481.

32. J. G. Kelly, J. Trevisan, A. D. Scott, P. L. Carmichael, H. M. Pollock, P. L. Martin-Hirsch and F. L. Martin, *J. Proteome Res.*, 2011, **10**, 1437–1448.

33. Z. Movasaghi, S. Rehman and I. U. Rehman, *Appl. Spectrosc. Rev.*, 2007, **42**, 493–541.

34. U. Utzinger, D. L. Heintzelman, A. Mahadevan-Jansen, A. Malpica, M. Follen and R. Richards-Kortum, *Appl. Spectrosc.*, 2001, **55**, 955–959.

Figure captions

Figure 1. Simulated hyperspectral data. (a) Mean hyperspectral image for class 1; (b) mean hyperspectral image for class 2; (c) mean spectra of class 1 and class 2. Colour bar: mean image intensity.

Figure 2. Raw Raman hyperspectral images for the real-world dataset. (a) Benign controls; (b) ovarian cancer patients; (c) mean raw Raman spectra for benign controls and ovarian cancer samples; (d) mean pre-processed (Savitzky-Golay (SG) smoothing (window of 15 points, 2^{nd} order polynomial fitting) and automatic weighted least squares (AWLS) baseline correction) Raman spectra for benign controls and ovarian cancer samples. False-colour images represented by the mean of the spectral dimension (725–1813 cm^{-1}).

Figure 3. Calculated class boundaries on the PCA scores for the training set of the simulated dataset. (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).

Figure 4. Calculated class boundaries on the PCA scores for the training set of the real-world dataset. (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).

Figure 5. 3D-PCA loadings for the simulated dataset. (a) Average pre-processed spectra for class 1 (continuous line) and class 2 (dashed line) samples; (b) difference-between-mean spectrum for class 1 and class 2 (negative signal indicates higher intensity in class 2); (c) 3D-PCA loadings on PC1; (d) 3D-PCA loadings on PC2.

Figure 6. 3D-PCA loadings for the real-world dataset. (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.

Tables

Table 1. Confusion matrices for the training, cross-validation and test sets using the unfolded and 3D hyperspectral images of the simulated dataset. PCA-LDA: principal component analysis linear discriminant analysis; PCA-QDA: principal component analysis quadratic discriminant analysis; PLS-DA: partial least squares discriminant analysis; SVM: support vector machines; 3D-PCA-LDA: three-dimensional principal component analysis linear discriminant analysis; 3D-PCA-QDA: three-dimensional principal component analysis quadratic discriminant analysis.

Unfolded		**Training**		**Cross-validation**		**Test**	
PCA-LDA		Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
	Class 1	**67%**	33%	**52%**	48%	**100%**	0%
	Class 2	52%	**48%**	48%	**52%**	67%	**33%**
PCA-QDA							
	Class 1	**71%**	29%	**71%**	29%	**78%**	22%
	Class 2	43%	**57%**	48%	**52%**	44%	**56%**
PLS-DA							
	Class 1	**86%**	14%	**81%**	19%	**78%**	22%
	Class 2	14%	**86%**	24%	**76%**	33%	**67%**
SVM							
	Class 1	**91%**	9%	**67%**	33%	**67%**	33%
	Class 2	0%	**100%**	33%	**67%**	22%	**78%**
3D							
3D-PCA-LDA							
	Class 1	**71%**	29%	**71%**	29%	**78%**	22%
	Class 2	24%	**76%**	24%	**76%**	11%	**89%**
3D-PCA-QDA							
	Class 1	**81%**	19%	**76%**	24%	**78%**	22%
	Class 2	24%	**76%**	29%	**71%**	11%	**89%**

Table 2. Confusion matrices for the training, cross-validation and test sets using the unfolded and 3D hypespectral images of real-world dataset. PCA-LDA: principal component analysis linear discriminant analysis; PCA-QDA: principal component analysis quadratic discriminant analysis; PLS-DA: partial least squares discriminant analysis; SVM: support vector machines; 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.

Unfolded		**Training**		**Cross-validation**		**Test**	
PCA-LDA		Control	Cancer	Control	Cancer	Control	Cancer
	Control	**79%**	21%	**70%**	30%	**100%**	0%
	Cancer	46%	**54%**	51%	**49%**	80%	**20%**
PCA-QDA							
	Control	**86%**	14%	**72%**	28%	**100%**	0%
	Cancer	54%	**46%**	57%	**43%**	80%	**20%**
PLS-DA							
	Control	**86%**	14%	**64%**	36%	**100%**	0%
	Cancer	23%	**77%**	46%	**54%**	100%	**0%**
SVM							
	Control	**100%**	0%	**71%**	29%	**100%**	0%
	Cancer	0%	**100%**	46%	**54%**	100%	**0%**
3D							
3D-PCA-LDA							
	Control	**100%**	0%	**99%**	1%	**100%**	0%
	Cancer	8%	**92%**	7%	**93%**	0%	**100%**
3D-PCA-QDA							
	Control	**100%**	0%	**93%**	7%	**100%**	0%
	Cancer	8%	**92%**	8%	**92%**	0%	**100%**

Table 3. Quality parameters for the models using the unfolded hyperspectral images and the full three-dimensional arrays in the test set. 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.

Dataset	Method	Accuracy	Sensitivity	Specificity
Simulated	**Unfolded**			
	PCA-LDA	66%	33%	100%
	PCA-QDA	67%	56%	78%
	PLS-DA	72%	67%	78%
	SVM	72%	78%	67%
	3D			
	3D-PCA-LDA	83%	89%	78%
	3D-PCA-QDA	83%	89%	78%
Real	**Unfolded**			
	PCA-LDA	64%	20%	100%
	PCA-QDA	64%	20%	100%
	PLS-DA	50%	0%	100%
	SVM	50%	0%	100%
	3D			
	3D-PCA-LDA	100%	100%	100%
	3D-PCA-QDA	100%	100%	100%

Figure 1

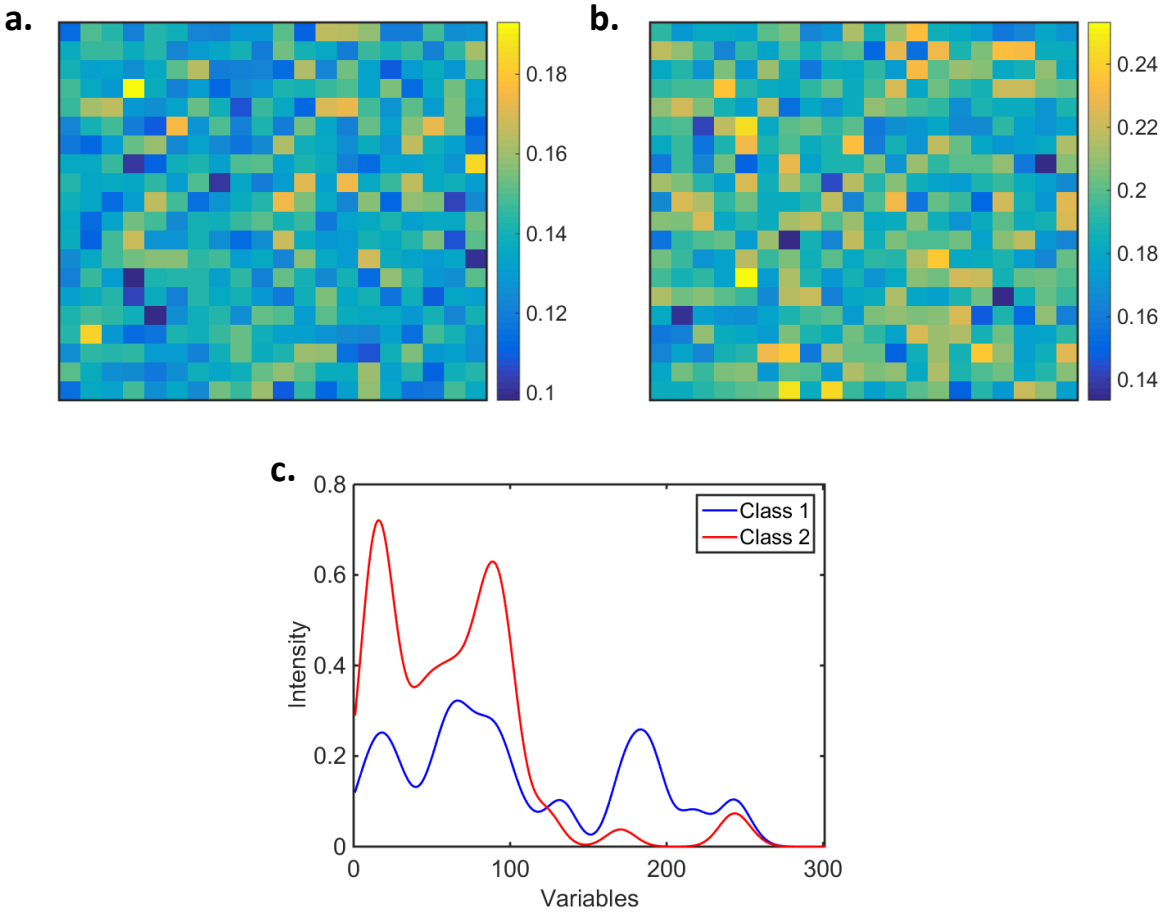

a.
0.18
0.16
0.14
0.12
0.1
b.
0.24
0.22
0.2
0.18
0.16
0.14
c.
0.8
0.6
0.4
0.2
0
Intensity
0
100
200
300
Variables
Class 1
Class 2

Figure 2

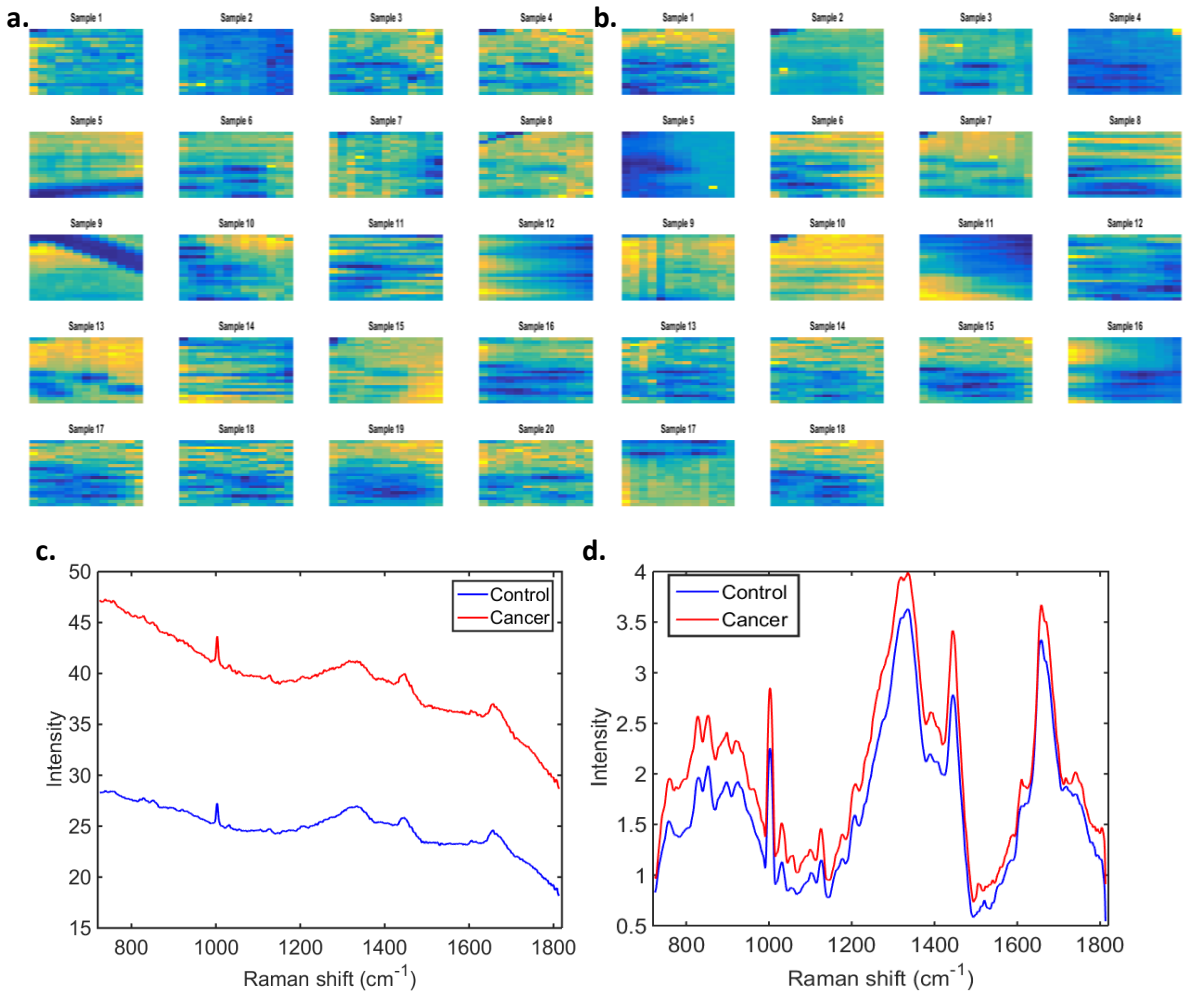


Figure 3

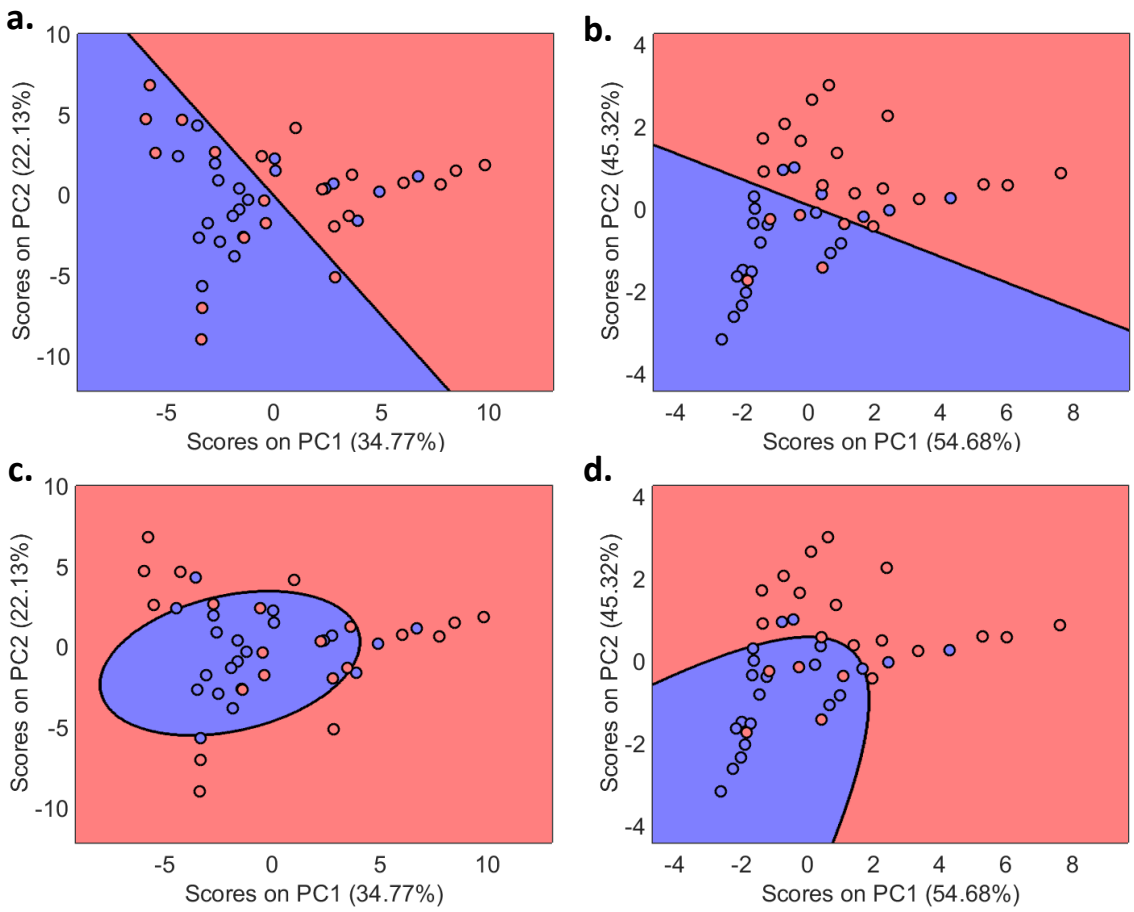

a.
b.
c.
d.
Scores on PC2 (22.13%)
Scores on PC1 (34.77%)
Scores on PC2 (45.32%)
Scores on PC1 (54.68%)

Figure 4

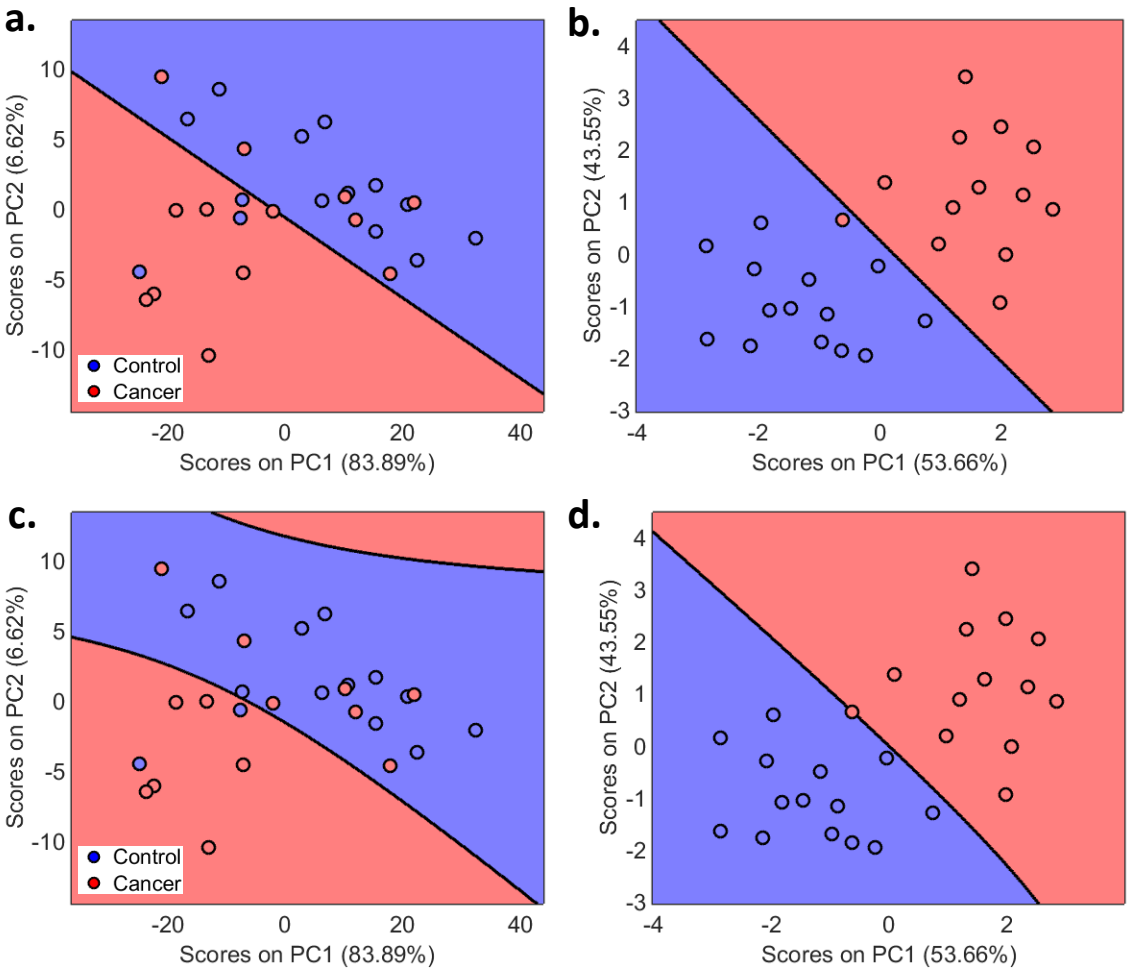

a.
b.
c.
d.
Scores on PC2 (6.62%)
Scores on PC1 (83.89%)
Scores on PC2 (43.55%)
Scores on PC1 (53.66%)
Control
Cancer

Figure 5

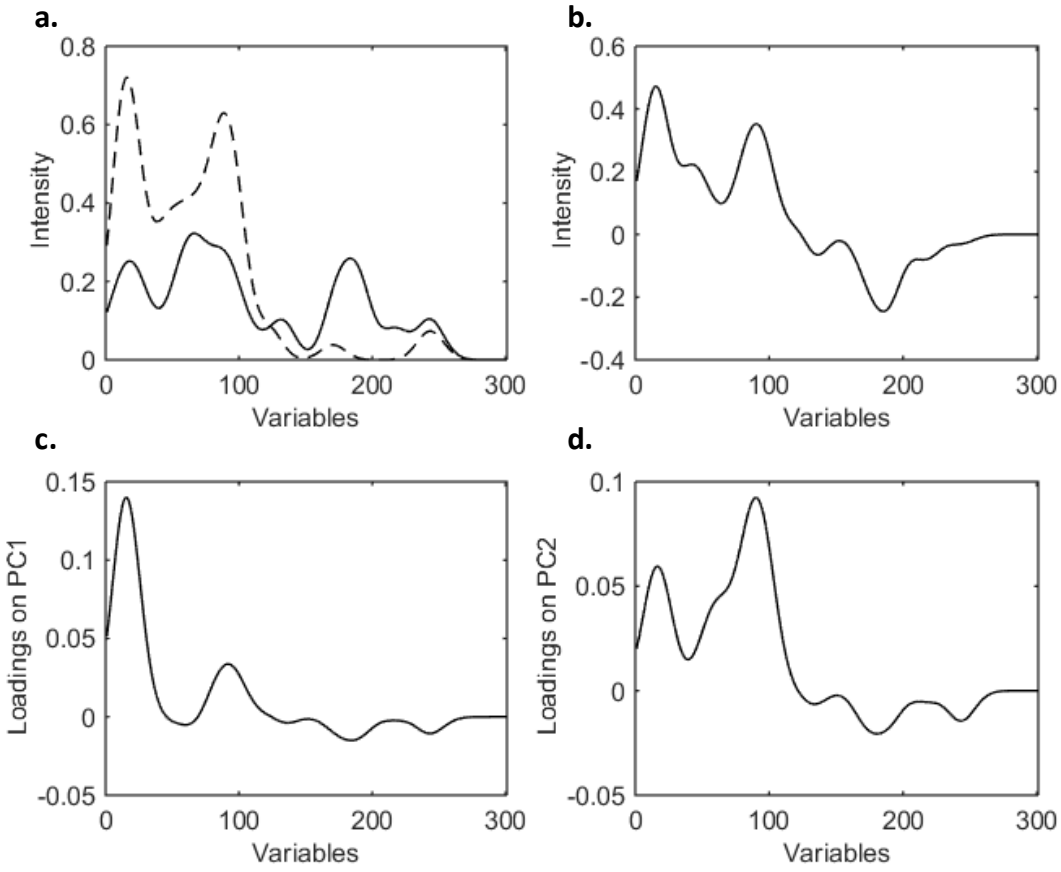
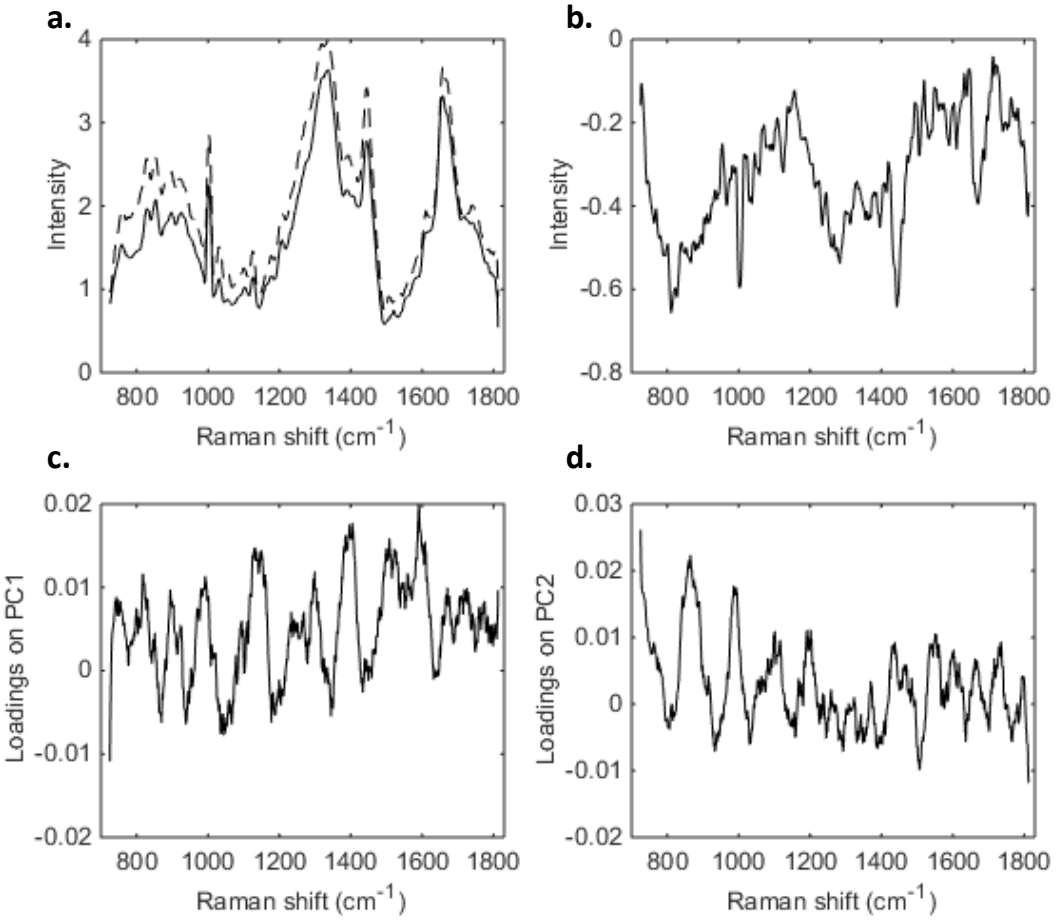
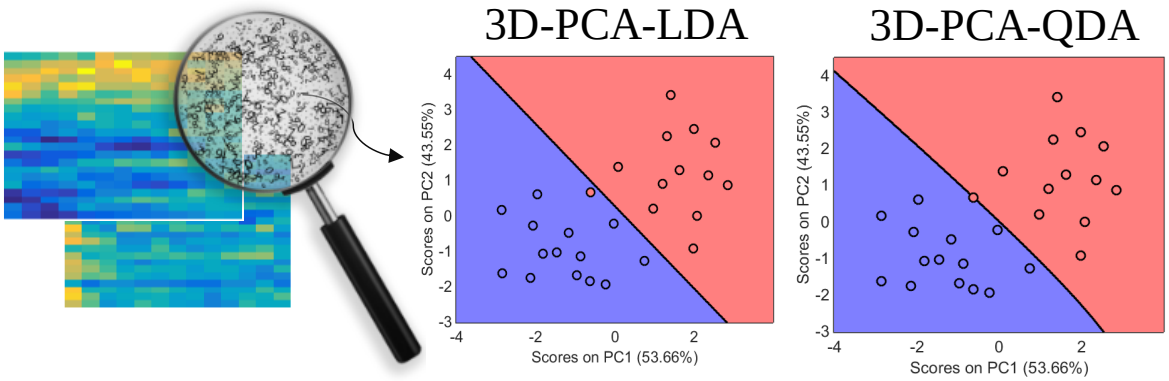

a.
b.
c.
d.
Intensity
Variables
Loadings on PC1
Loadings on PC2
0
100
200
300

Figure 6


a.
b.
c.
d.
Intensity
Raman shift (cm-1)
Loadings on PC1
Loadings on PC2
800
1000
1200
1400
1600
1800

ToC


3D-PCA-LDA
3D-PCA-QDA
Scores on PC2 (43.55%)
Scores on PC1 (53.66%)
Scores on PC2 (43.55%)
Scores on PC1 (53.66%)