A computational protocol for sample selection in biological-derived infrared spectroscopy datasets using Morais-Lima-Martin (MLM) algorithm

Camilo LM Morais (cdlmedeiros-de-morai@uclan.ac.uk)
University of Central Lancashire
Francis L Martin (flmartin@uclan.ac.uk)
University of Central Lancashire
Kassio MG Lima
Federal University of Rio Grande do Norte

Method Article

Keywords: infrared spectroscopy, sample selection, data splitting, pre-processing, computational analysis, chemometrics, algorithm

Posted Date: December 20th, 2018

DOI: 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.

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 \(1), since it allows building protocols for analysing tissues, cells and biofluids in a non-destructive, fast and low-cost fashion \(1,2). 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 \(3), identification of neurodegenerative diseases \(4), cervical cancer screening \(5), endometrial and ovarian cancer identification \(6), identification of prostate cancer tissue samples \(7), differentiation of endometrial tissue regions \(8), toxicology screening \(9,10), and microbiologic studies involving fungi and virus identification \(11-13). 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 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 \(14). 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 1b.

Equipment

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). 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 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\(1).

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 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). 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 2b in the MATLAB Command Window. In this figure, the following steps are performed: \(1) 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).

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.

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').

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 3a). Further details about this dataset can be found in Gajjar et al. \(15). Samples were divided into training \(70%) and test \(30%) sets as depicted in Figure 2b. Two classification algorithms were applied: principal component analysis linear discriminant analysis \(PCA-LDA) \(16) and partial least squares discriminant analysis \(PLS-DA) \(17). 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/) \(18) 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 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 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 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 3d). Gliobastoma 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 \(15). 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.

References

1. Baker, M.J. et al. Using Fourier transform IR spectroscopy to analyze biological materials. Nat. Protoc. 9, 1771–1791 \(2014). 2. Martin, F.L. et al. Distinguishing cell types or populations based on the computational analysis of their infrared spectra. Nat. Protoc. 5, 1748–1760 \(2010). 3. Bury, D. et al. Spectral classification for diagnosis involving numerous pathologies in a complex clinical setting: A neuro-oncology example. Spectrochim. Acta A Mol. Biomol. Spectrosc. 206, 89–96 \(2019). 4. Paraskevaidi, M. et al. Differential diagnosis of Alzheimer's disease using spectrochemical analysis of blood. Proc. Natl. Acad. Sci. U.S.A. 114, E7929–E7938 \(2017). 5. Neves, A.C.O. et al. ATR-FTIR and multivariate analysis as a screening tool for cervical cancer in women from northeast Brazil: a biospectroscopic approach. RSC Adv. 6, 99648–99655 \(2016). 6. Paraskevaidi, M. et al. Potential of mid-infrared spectroscopy as a non-invasive diagnostic test in urine for endometrial or ovarian cancer. Analyst 143, 3156–3163 \(2018). 7. Siqueira, L.F.S. et al. LDA vs. QDA for FT-MIR prostate cancer tissue classification. Chemometr. Intell. Lab. Syst. 162, 123–129 \(2017). 8. Theophilou, G. et al. 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. 410, 4541–4554 \(2018). 9. Duan, P. et al. 4-Nonylphenol effects on rat testis and sertoli cells determined by spectrochemical techniques coupled with chemometric analysis. Chemosphere 218, 64–75 \(2019). 10. Morais, C.L.M. et al. Assessing Binary Mixture Effects from Genotoxic and Endocrine Disrupting Environmental Contaminants Using Infrared Spectroscopy. ACS Omega 3, 13399–13412 \(2018). 11. Costa, F.S.L. et al. Attenuated total reflection Fourier transform infrared \(ATR-FTIR) spectroscopy as a new technology for discrimination between Cryptococcus neoformans and Cryptococcus gattii. Anal. Methods 8, 7107–7115 \(2016). 12. Morais, C.L.M.; Costa, F.S.L. & Lima, K.M.G. Variable selection with a support vector machine for discriminating Cryptococcus fungal species based on ATR-FTIR spectroscopy. Anal. Methods 9, 2964–2970 \(2017). 13. Santos, M.C.D. et al. Spectroscopy with computational analysis in virological studies: A decade \(2006–2016). Trends Analyt. Chem. 97, 244–256 \(2017). 14. Kennard, R.W. & Stone, L.A. Computer Aided Design of Experiments. Technometrics 11, 137–148 \(1969). 15. Gajjar, K. et al. Diagnostic segregation of human brain tumours using Fourier-transform infrared and/or Raman spectroscopy coupled with discriminant analysis. Anal. Methods 5, 89–102 \(2013). 16. Morais, C.L.M. & Lima, K.M.G. Principal Component Analysis with Linear and Quadratic Discriminant Analysis for Identification of Cancer Samples Based on Mass Spectrometry. J. Braz. Chem. Soc. 29, 472–481 \(2018). 17. Brereton, R.G. & Lloyd, G.R. Partial least squares discriminant analysis: taking the magic away. J. Chemom. 28, 213–225 \(2014). 18. Ballabio, D. & Consonni, V. Classification tools in chemistry. Part 1: linear models. PLS-DA. Anal. Methods 5, 3790–3798 \(2013).

Acknowledgements

Camilo L. M. Morais would like to thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior \(CAPES) - Brazil \(grant 88881.128982/2016-01) for financial support.

Figures

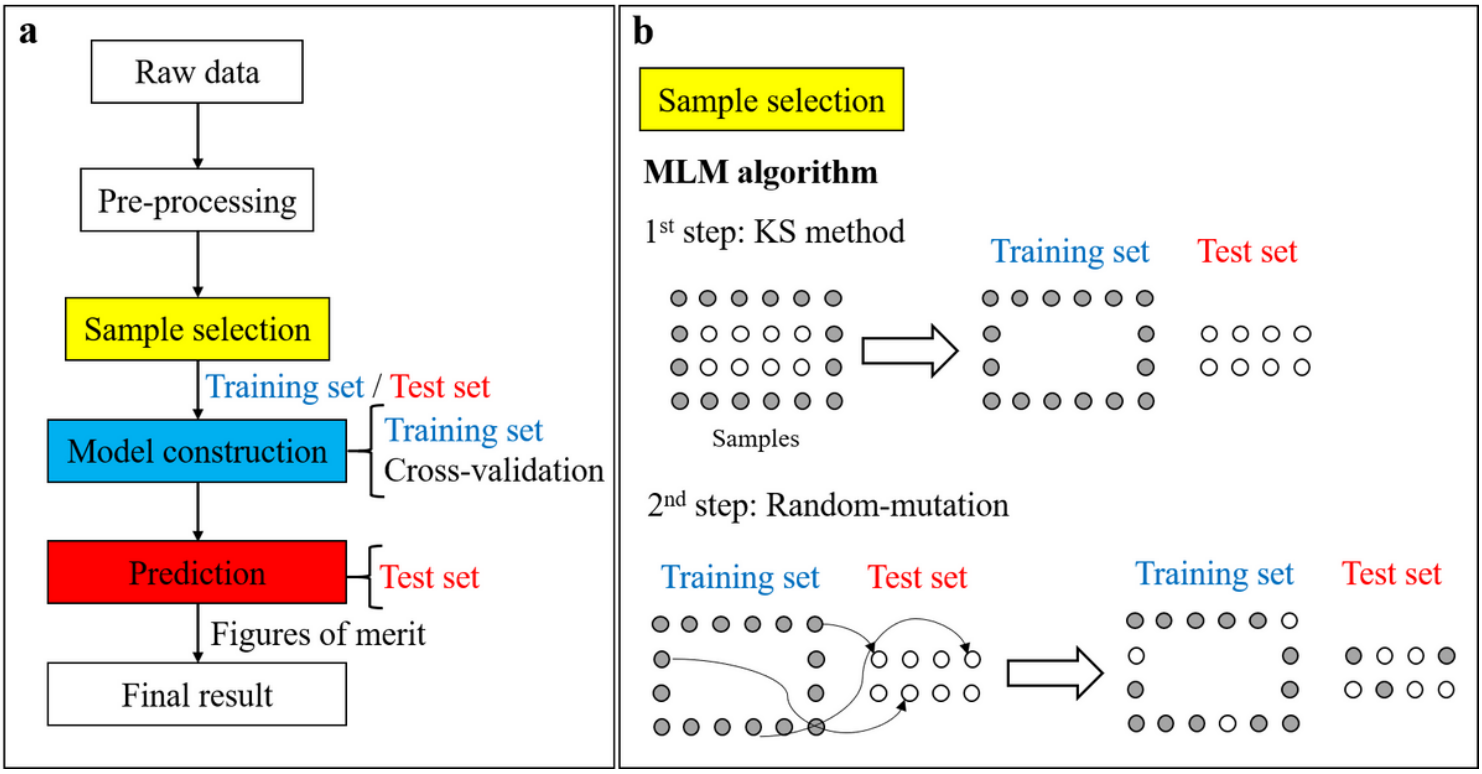

Figure 1

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.

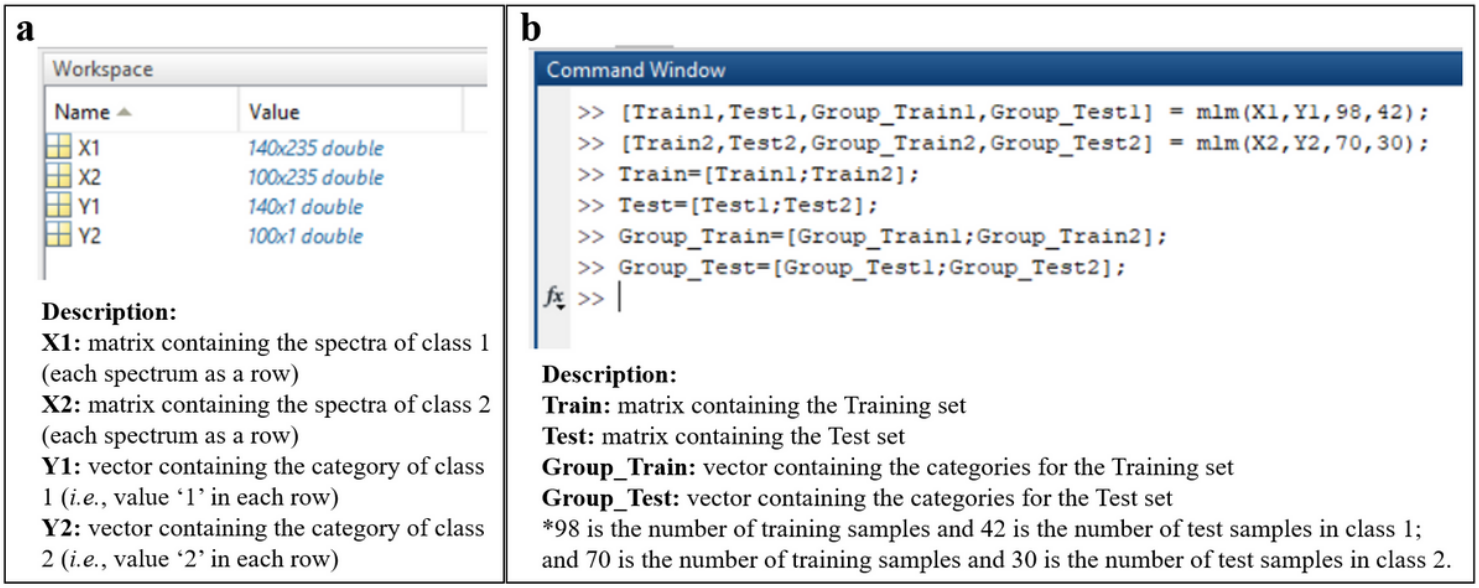

Figure 2

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.

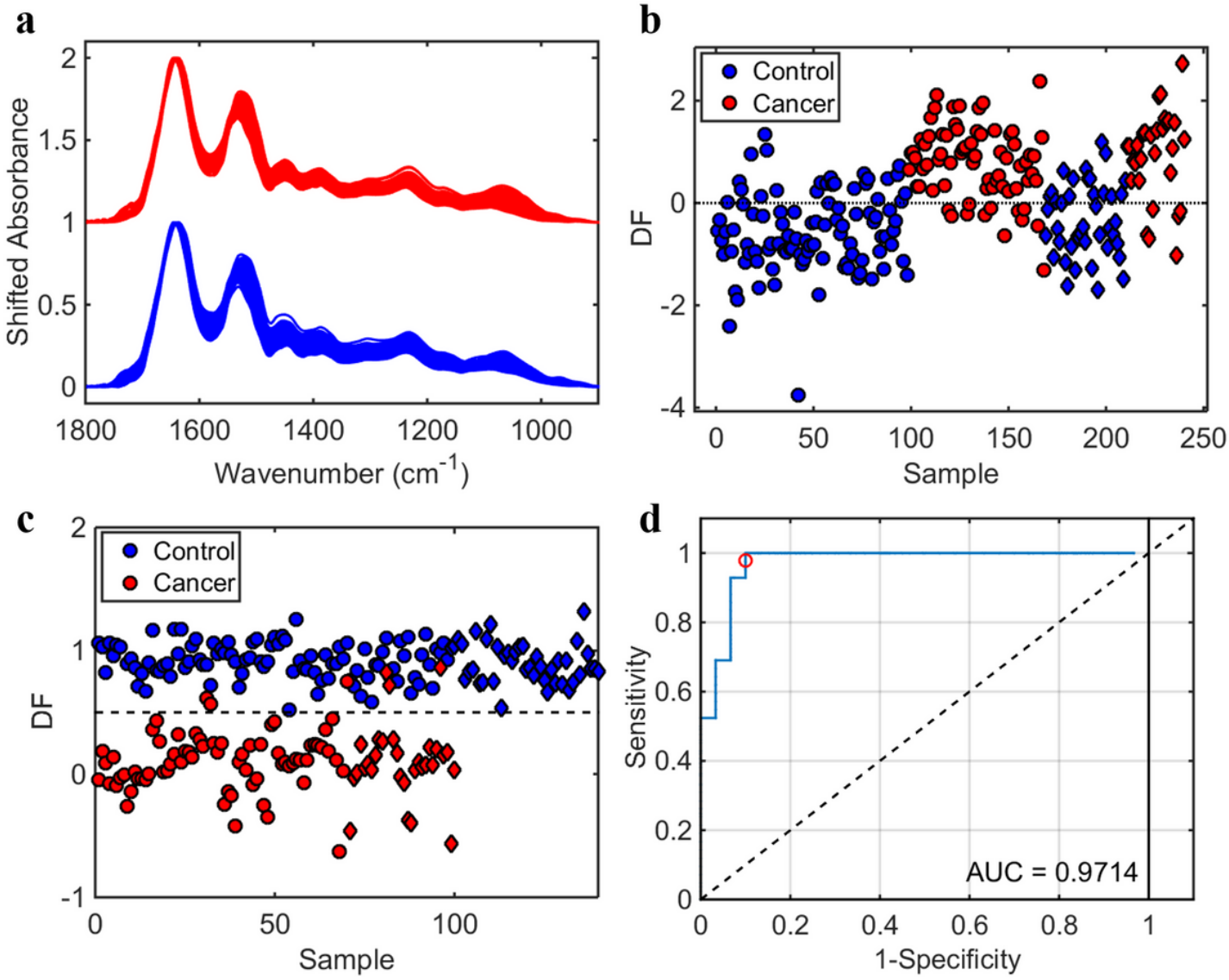

Figure 3

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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement0.docx