University of Central Lancashire
UCLan

Central Lancashire Online Knowledge (CLoK)

Title	TTWD-DA: A MATLAB toolbox for discriminant analysis based on trilinear three-way data
Type	Article
URL	https://clok.uclan.ac.uk/27711/
DOI	https://doi.org/10.1016/j.chemolab.2019.03.007
Date	2019
Citation	Medeiros-De-morais, Camilo De lelis orcid iconORCID: 0000-0003-2573-787X, Lima, Kassio M.G. and Martin, Francis L (2019) TTWD-DA: A MATLAB toolbox for discriminant analysis based on trilinear three-way data. Chemometrics and Intelligent Laboratory Systems, 188. pp. 46-53. ISSN 0169-7439
Creators	Medeiros-De-morais, Camilo De lelis, Lima, Kassio M.G. and Martin, Francis L

It is advisable to refer to the publisher's version if you intend to cite from the work.
https://doi.org/10.1016/j.chemolab.2019.03.007

For information about Research at UCLan please go to http://www.uclan.ac.uk/research/

All outputs in CLoK are protected by Intellectual Property Rights law, including Copyright law. Copyright, IPR and Moral Rights for the works on this site are retained by the individual authors and/or other copyright owners. Terms and conditions for use of this material are defined in the http://clok.uclan.ac.uk/policies/

TTWD-DA: A MATLAB toolbox for discriminant analysis based on trilinear three-way data

Camilo L. M. Morais[1]*, Kássio M. G. Lima[2], Francis L. Martin[1]*

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

[2]*Biological Chemistry and Chemometrics, Federal University of Rio Grande do Norte, Natal 59072-970, Brazil*

***Corresponding authors:** Camilo L. M. Morais / Prof. Francis L. Martin, School of Pharmacy and Biomedical Sciences, Maudland Building, University of Central Lancashire, Preston PR1 2HE, UK; Email: cdlmedeiros-de-morai@uclan.ac.uk / flmartin@uclan.ac.uk

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 free available online.

Keywords: Discriminant analysis; three-way data; EEM; MATLAB; software; GUI

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 [1,2]. The advantages of molecular fluorescence spectroscopy are its high sensitivity and relatively low-cost instrumentation [2]. In addition, the EEM data generated is contemplated by the "second-order advantage" [3], 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 [4,5].

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) [6] and Tucker3 algorithm [7]. 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 [6,8]. 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 [7-9]. 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 [10,11]. 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 [11]. 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 [12]. There are many applications for which chemometric techniques are employed for analysing three-way data, such as for assessing food quality [13-15], detection of substances in the atmosphere [16], and differentiation of fungi [17] using EEM spectroscopy; analysis of heavy metal ions using spectral-kinematic responses [18]; and evaluation of different juices colorants via spectral-pH responses [19]. 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 [17, 20-22]. 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.

2. Software

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) [23] 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.

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 [6]. It decomposes the three-way data $\underline{\mathbf{X}}$ as follows [17]:

$$\underline{\mathbf{X}} = \mathbf{A}(\mathbf{C}|\otimes|\mathbf{B})^{\mathrm{T}} + \underline{\mathbf{E}} \quad (01)$$

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 [24].

Tucker3 is another multivariate deconvolution method for higher-order data also known as "3-way principal component analysis (PCA)" [25]. It decomposes the three-way data $\underline{\mathbf{X}}$ as follows [9]:

$$\underline{\mathbf{X}} = \mathbf{A}\mathbf{G}(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}} + \underline{\mathbf{E}} \quad (02)$$

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 [26].

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 [10,11]:

$$L_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}}\mathbf{C}_{pooled}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_k) \quad (03)$$

$$Q_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}}\mathbf{C}_k^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_k) \quad (04)$$

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:

$$C_{pooled} = \frac{1}{n}\sum_{k=1}^{K} n_k \mathbf{C}_k \quad (05)$$

$$\mathbf{C}_k = \frac{1}{n_k - 1}\sum_{i=1}^{n_k}(\mathbf{x}_i - \bar{\mathbf{x}}_k)\,(\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}} \quad (06)$$

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 [12]. PLS decomposition takes the form [12]:

$$\mathbf{X} = \mathbf{TP} + \mathbf{E} \quad (07)$$

$$\mathbf{y} = \mathbf{Tq} + \mathbf{f} \quad (08)$$

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 [12]. 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 [12]:

$$\mathbf{b} = \mathbf{W}(\mathbf{PW})^{-1}\mathbf{q} \quad (09)$$

$$\hat{\mathbf{y}} = \mathbf{Xb} \quad (10)$$

where **b** are PLS coefficients; **W** is a weight matrix; and $\hat{\mathbf{y}}$ is the predicted response vector.

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 [20]. These parameters are calculated as follows:

$$\mathrm{CC\%} = 100 - \frac{(\varepsilon_1 - \varepsilon_2)}{N} \times 100 \tag{11}$$

$$\mathrm{AC(\%)} = \left(\frac{\mathrm{TP+TN}}{\mathrm{TP+FP+TN+FN}}\right) \times 100 \tag{12}$$

$$\mathrm{SENS(\%)} = \left(\frac{\mathrm{TP}}{\mathrm{TP+FN}}\right) \times 100 \tag{13}$$

$$\mathrm{SPEC(\%)} = \left(\frac{\mathrm{TN}}{\mathrm{TN+FP}}\right) \times 100 \tag{14}$$

$$\mathrm{F-score} = \frac{2 \times \mathrm{SENS} \times \mathrm{SPEC}}{\mathrm{SENS+SPEC}} \tag{15}$$

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.

2.4 Software overview

The main GUI features of TTWD-DA are depicted in Figure 1.

[Insert Figure 1 here]

The main classification interface (Figure 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) [20] 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.

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.* [27]. 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 2. More details about the experimental procedure for data acquisition can be found at Andersen *et al.* [27].

[Insert Figure 2 here]

4. Software Application

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 [28], 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 [29]. 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 3 depicts these type of data in MATLAB.

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 3). Only previously saved .mat files can be used as input.

[Insert Figure 3 here]

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 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".

4.3 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 4). This window also can be accessed by clicking on View > Figures of Merit.

[Insert Figure 4 here]

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 [20,30]. 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 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 [11], however it has an inferior performance compared to LDA for datasets with small number of samples [31]. Comparing the variance among the three classes in dataset 3 (Figure 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 [31].

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 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.

[Insert Figure 5 here]

[Insert Figure 6 here]

After the model is built, the measured and predicted sample indexes for each class can be viewed in the Menu: View > Classification Index (Figure 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.

[Insert Figure 6 here]

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.

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".

Acknowledgments

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.

References

[1] L. Bachmann, D.M. Zezell, A.C. Ribeiro, L. Gomes, A.S. Ito, Fluorescence Spectroscopy of Biological Tissues – A Review. Appl. Spectrosc. Rev. 41 (2006) 575–590.

[2] M.C.D. Santos, C.L.M. Morais, Y.M. Nascimento, J.M.G. Araujo, K.M.G. Lima, Spectroscopy with computational analysis in virological studies: A decade (2006–2016). Trends Analyt. Chem. 97 (2017) 244–256.

[3] K.S. Booksh, B.R. Kowalski, Theory of Analytical Chemistry. Anal. Chem. 66 (1994) 782A–791A.

[4] Y.-N. Li, H.-L. Wu, X.-D. Qing, C.-C. Nie, S.-F. Li, Y.-J. Yu, S.-R. Zhang, R.-Q. Yu, 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. 85 (2011) 325–332.

[5] K. Calimag-Williams, G. Knobel, H.C. Goicoechea, A.D. Campiglia, 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. 811 (2014):60–69.

[6] R. Bro, PARAFAC. Tutorial and applications. Chemometr. Intell. Lab. Syst. 38 (1997) 149–171.

[7] L.R. Tucker, Some mathematical notes on three-mode factor analysis. Psychometrika. 31 (1966) 279–311.

[8] R. Bro, R.A. Harshman, N.D. Sidiropoulos, M.E. Lundy, Modeling multi-way data with linearly dependent loadings. J. Chemometrics. 23 (2009) 324–340.

[9] M. Gallo, Tucker3 Model for Compositional Data. Communications in Statistics – Theory and Methods. 44 (2015) 4441–4453.

[10] S.J. Dixon, R.G. Brereton, 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. 95 (2009) 1–17.

[11] C.L.M. Morais, K.M.G. Lima, Principal Component Analysis with Linear and Quadratic Discriminant Analysis for Identification of Cancer Samples Based on Mass Spectrometry. J. Braz. Chem. Soc. 29 (2018) 472–481.

[12] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis: taking the magic away, J. Chemometrics 28 (2014) 213–225.

[13] J. Sádecká, M. Jakubíková, P. Májek, Fluorescence spectroscopy for discrimination of botrytized wines. Food Control. 88 (2018) 75–84.

[14] I.D. Merás, J.D. Manzano, D.A. Rodríguez, A.M. Peña, Detection and quantification of extra virgin olive oil adulteration by means of autofluorescence excitation-emission profiles combined with multi-way classification. Talanta. 178 (2018) 751–762.

[15] S.M. Azcarate, A.A. Gomes, M.R. Alcaraz, M.C.U. Araújo, J.M. Camiña, H.C. Goicoechea, Modeling excitation–emission fluorescence matrices with pattern recognition algorithms for classification of Argentine white wines according grape variety. Food Chem. 184 (2015) 214–219.

[16] Y.-L. Pan, Detection and characterization of biological and other organic-carbon aerosol particles in atmosphere using fluorescence. J. Quant. Spectrosc. Radiat. Transfer. 150 (2015) 12–35.

[17] F.S.L. Costa, P.P. Silva, C.L.M. Morais, R.C. Theodoro, T.D. Arantes, K.M.G. Lima, Comparison of multivariate classification algorithms using EEM fluorescence data to distinguish Cryptococcus neoformans and Cryptococcus gattii pathogenic fungi. Anal. Methods. 9 (2017) 3968–3976.

[18] J.C.G. Esteves da Silva, C.J.S. Oliveira, Parafac decomposition of three-way kinetic-spectrophotometric spectral matrices corresponding to mixtures of heavy metal ions, Talanta 49 (1999) 889–897.

[19] N.R. Marsili, A. Lista, B.S. Fernandez Band, H.C. Goicoechea, A.C. Olivieri, Evaluation of complex spectral-pH three-way arrays by modified bilinear least-squares: determination of four different dyes in interfering systems, Analyst 130 (2005) 1291–1298.

[20] C.L.M. Morais, K.M.G. Lima, Comparing unfolded and two-dimensional discriminant analysis and support vector machines for classification of EEM data. Chemometr. Intell. Lab. Syst. 170 (2017) 1–12.

[21] F. Stelzle, C. Knipfer, W. Adler, M. Rohde, N. Oetter, E. Nkenke, M. Schmidt, K. Tangermann-Gerk, Tissue Discrimination by Uncorrected Autofluorescence Spectra: A Proof-of-Principle Study for Tissue-Specific Laser Surgery. Sensors. 13 (2013) 13717–13731.

[22] F. Stelzle, M. Rohde, M. Riemann, N. Oetter, W. Alder, K. Tangermann-Gerk, M. Schmidt, C. Knipfer, Autofluorescence spectroscopy for nerve-sparing laser surgery of the head and neck—the influence of laser-tissue interaction. Lasers Med. Sci. 32 (2017) 1289–1300.

[23] C.A. Andersson, R. Bro, The N-way Toolbox for MATLAB. Chemometr. Intell. Lab. Syst. 52 (2000) 1–4.

[24] S. Liu, Matrix results on the Khatri-Rao and Tracy-Singh products. Linear Algebra Appl. 289 (1999) 267–277.

[25] R. Henrion, N-way principal component analysis theory, algorithms and applications. Chemometr. Intell. Lab. Syst. 25 (1994) 1–23.

[26] C.F. Van Loan, The ubiquitous Kronecker product. J. Comput. Appl. Math. 123 (2000) 85–100.

[27] C.M. Andersen, R. Bro, Practical aspects of PARAFAC modeling of fluorescence excitation-emission data. J. Chemom. 17 (2003) 200–215.

[28] M. Bahram, R. Bro, C. Stedmon, A. Afkhami, Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation. J. Chemom. 20 (2006) 99–105.

[29] R.W. Kennard, L.A. Stone, Computer Aided Design of Experiments. Technometrics. 11 (1969) 137–148.

[30] H. Nilsen, M. Esaiassen, K. Heia, F. Sigernes, Visible/Near-Infrared Spectroscopy: A New Tool for the Evaluation of Fish Freshness? J. Food Sci. 67 (2002) 1821–1826.

[31] W. Wu, Y. Mallet, B. Walczak, W. Penninckx, D.L. Massart, S. Heuerding, F. Erni, Comparison of regularized discriminant analysis, linear discriminant analysis and quadratic discriminant analysis, applied to NIR data. Anal. Chim. Acta. 329 (1996) 257–265.

Captions for Figures

Figure 1: EEM-DA main interface overview. Insets (A)-(N) refer to the text.

Figure 2: Average EEM for the test dataset.

Figure 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.

Figure 4: Figures of merit for A) PARAFAC-LDA, B) PARAFAC-QDA, C) Tucker3-LDA, D) Tucker3-QDA, E) PLS-DA.

Figure 5: Variance calcualted for the test dataset.

Figure 6: PLS-DA canonical scores (left) and predicted class (right).

Figure 7: Classification indexes predicted by the toolbox for the PARAFAC-LDA model.

Figure 1

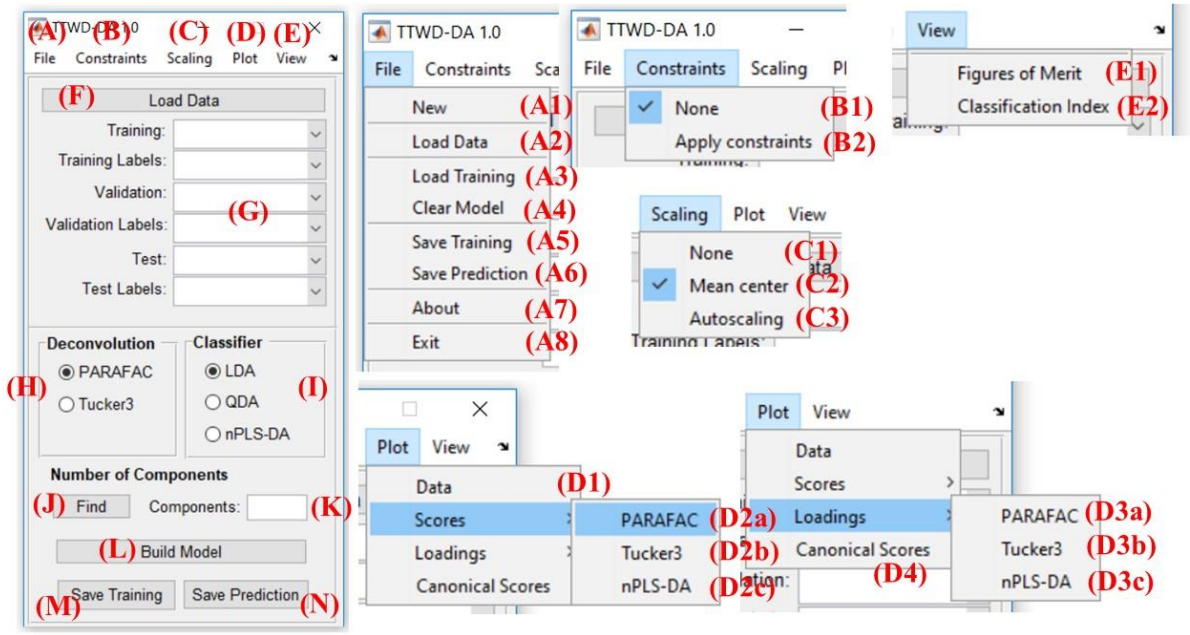

(A)
(B)
(C)
(D)
(E)
File
Constraints
Scaling
Plot
View
(F)
Load Data
Training:
Training Labels:
Validation:
Validation Labels:
Test:
Test Labels:
(G)
Deconvolution
PARAFAC
Tucker3
(H)
Classifier
LDA
QDA
nPLS-DA
(I)
Number of Components
(J)
Find
Components:
(K)
(L)
Build Model
Save Training
Save Prediction
(M)
(N)
TTWD-DA 1.0
New
(A1)
Load Data
(A2)
Load Training
(A3)
Clear Model
(A4)
Save Training
(A5)
Save Prediction
(A6)
About
(A7)
Exit
(A8)
None
(B1)
Apply constraints
(B2)
Figures of Merit
(E1)
Classification Index
(E2)
None
(C1)
Mean center
(C2)
Autoscaling
(C3)
Data
(D1)
Scores
Loadings
Canonical Scores
PARAFAC
(D2a)
Tucker3
(D2b)
nPLS-DA
(D2c)
(D3a)
(D3b)
(D3c)
(D4)

Figure 2

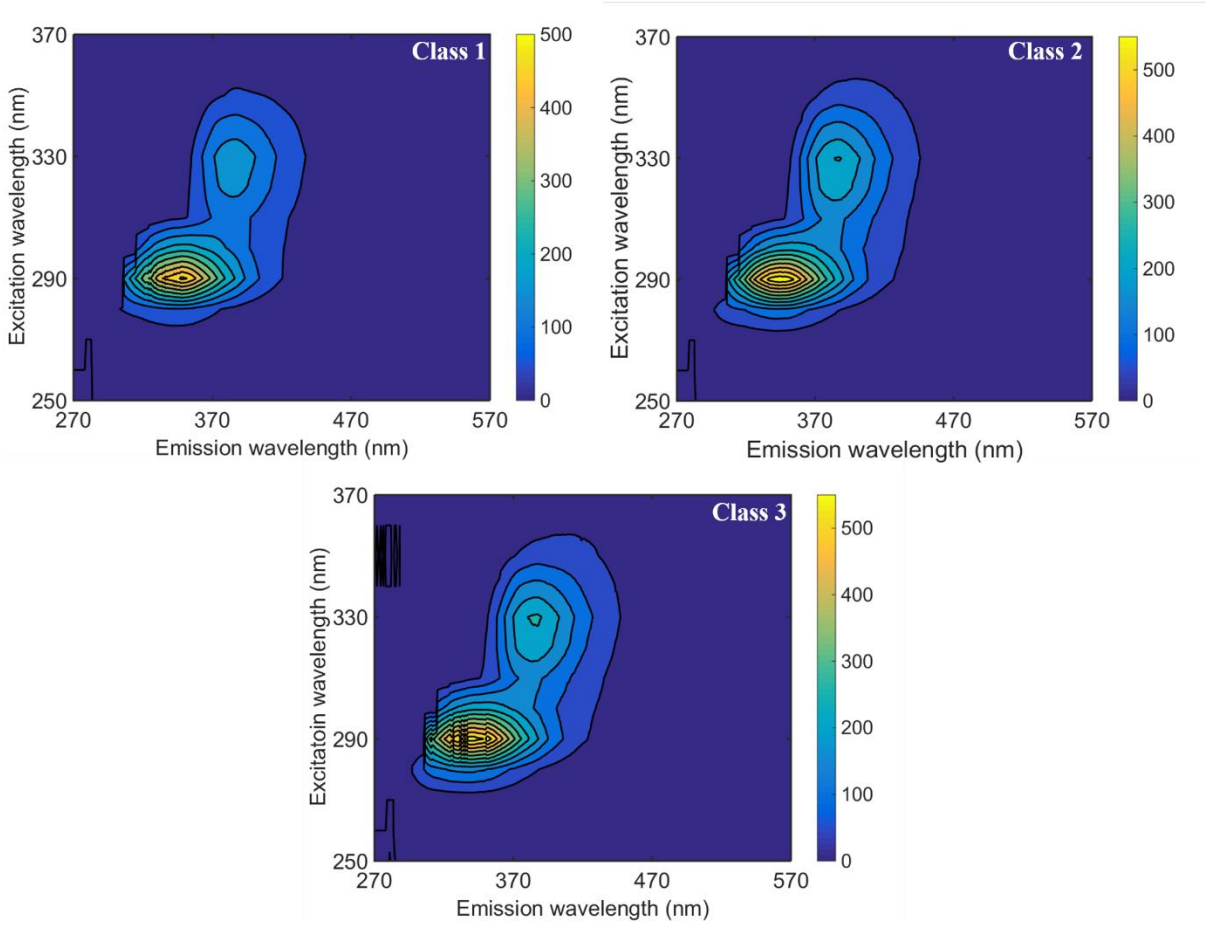

Class 1
Class 2
Class 3
Excitation wavelength (nm)
Excitatoin wavelength (nm)
Emission wavelength (nm)
370
330
290
250
270
370
470
570
500
400
300
200
100
0

Figure 3

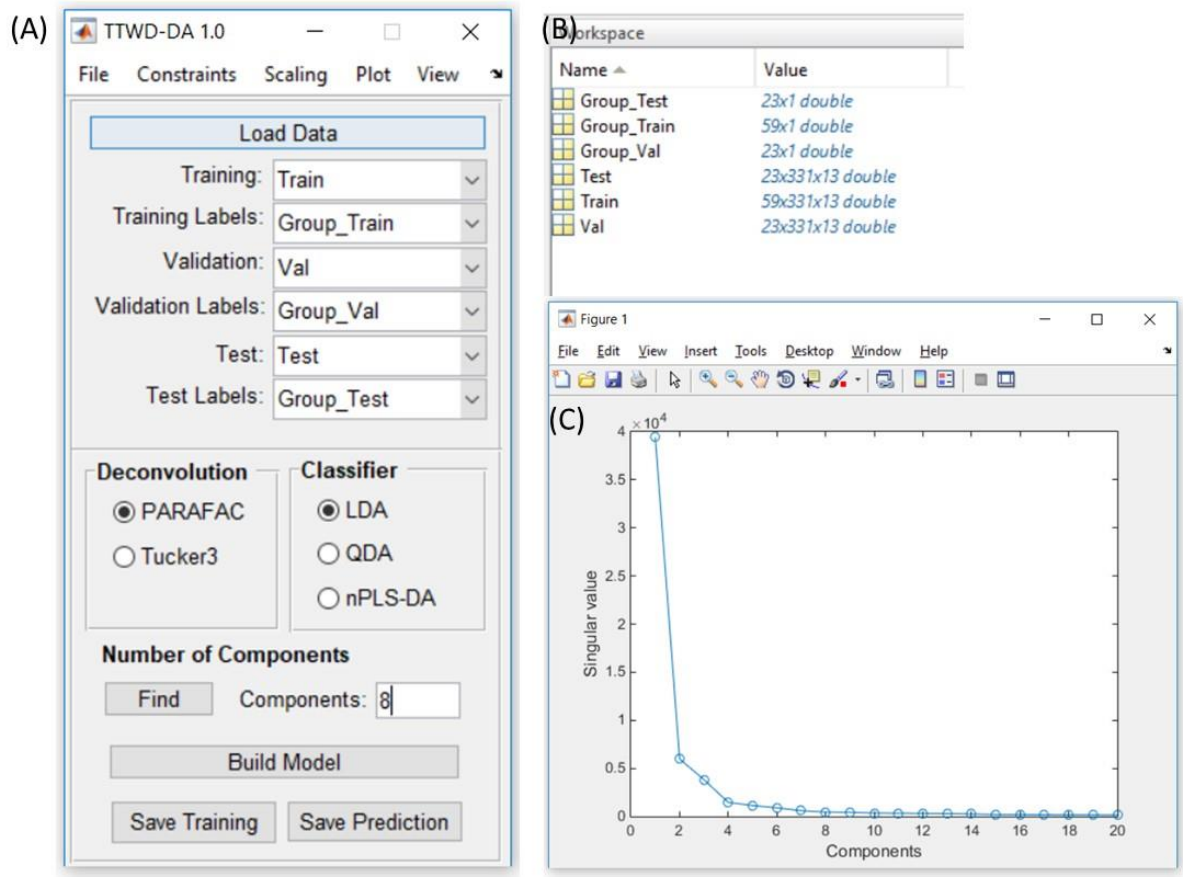

(A)
TTWD-DA 1.0
File Constraints Scaling Plot View
Load Data
Training: Train
Training Labels: Group_Train
Validation: Val
Validation Labels: Group_Val
Test: Test
Test Labels: Group_Test
Deconvolution
PARAFAC
Tucker3
Classifier
LDA
QDA
nPLS-DA
Number of Components
Find
Components: 8
Build Model
Save Training
Save Prediction
(B)
Workspace
Name
Value
Group_Test 23x1 double
Group_Train 59x1 double
Group_Val 23x1 double
Test 23x331x13 double
Train 59x331x13 double
Val 23x331x13 double
Figure 1
File Edit View Insert Tools Desktop Window Help
(C)
×10^4
Singular value
Components

Figure 4

(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 5

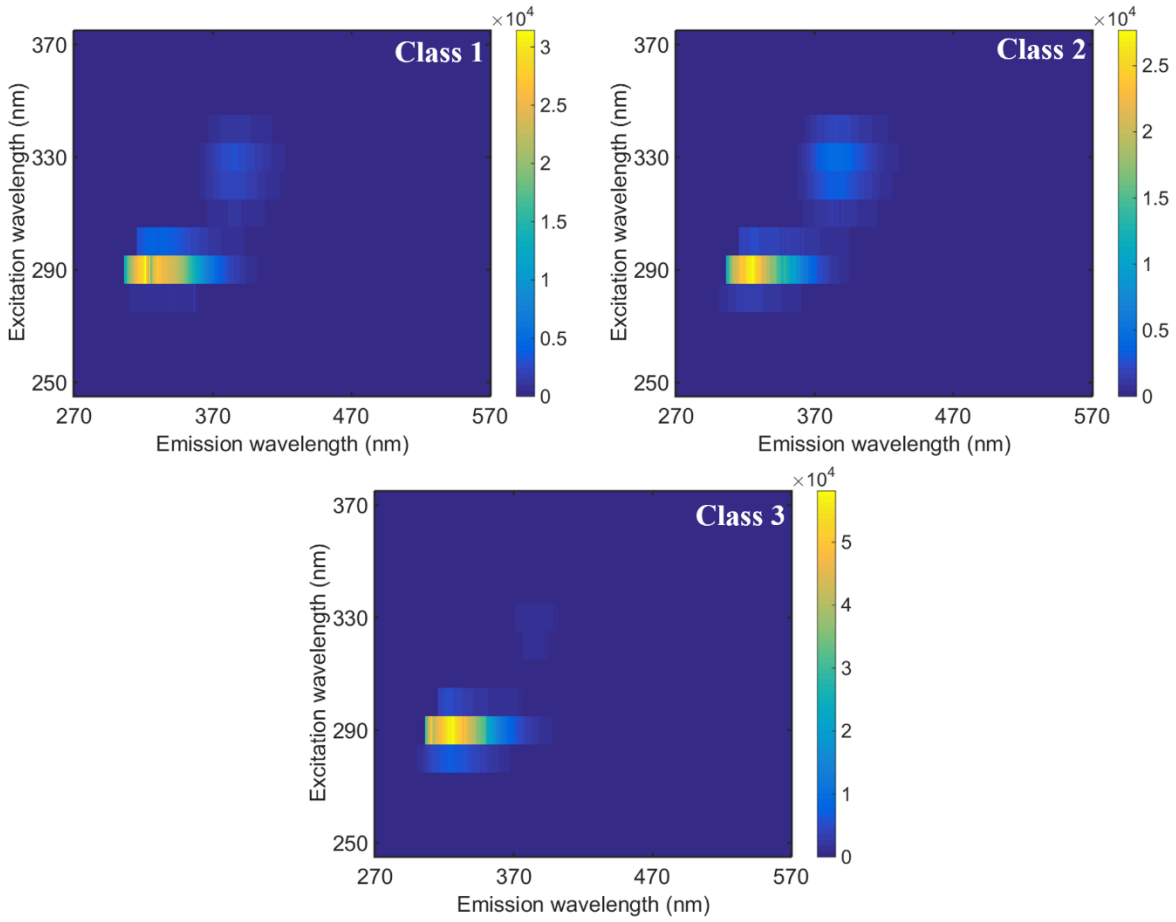

Class 1
Class 2
Class 3
Excitation wavelength (nm)
Emission wavelength (nm)
370
330
290
250
270
370
470
570
×10^4

Figure 6

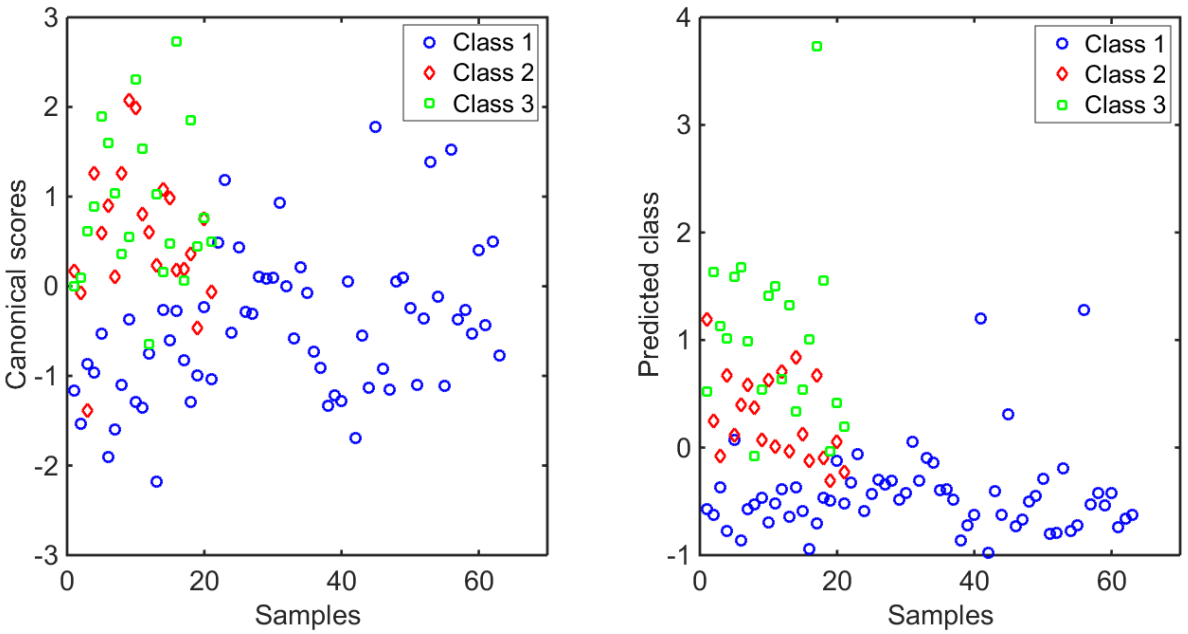
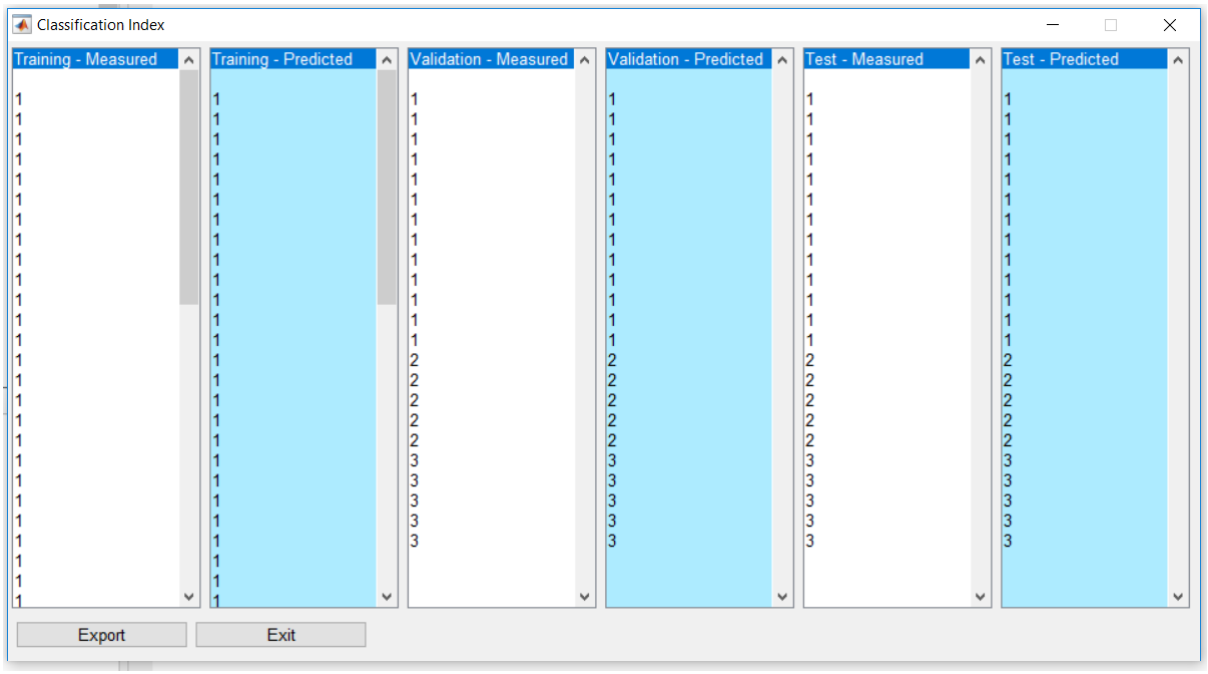

Class 1
Class 2
Class 3
Canonical scores
Samples
Predicted class
Samples

Figure 7


Classification Index
Training - Measured
Training - Predicted
Validation - Measured
Validation - Predicted
Test - Measured
Test - Predicted
Export
Exit