NWU Spectra Analysis (NWUSA) Toolbox

User manual

by

Dongliang Song, Shuang Wang*

Institute of Photonics and Photon-Technology, Northwest University,
Xi’an, Shaanxi, 710127, China

*Corresponding Author:
Dr. Shuang Wang
Associate Professor
State Key Laboratory of Photon-Technology in Western China Energy
Institute of Photonics and Photon-technology
Northwest University
#1 Xuefu Avenue, Guodu Education and Technology Industrial Zone
Chang’an District, Xi’an, 710127
Shaanxi, China
Email: swang@nwu.edu.cn
Tel: +86-29-8830-3281
Fax: +86-29-8830-3281

Contents

1. Declaration

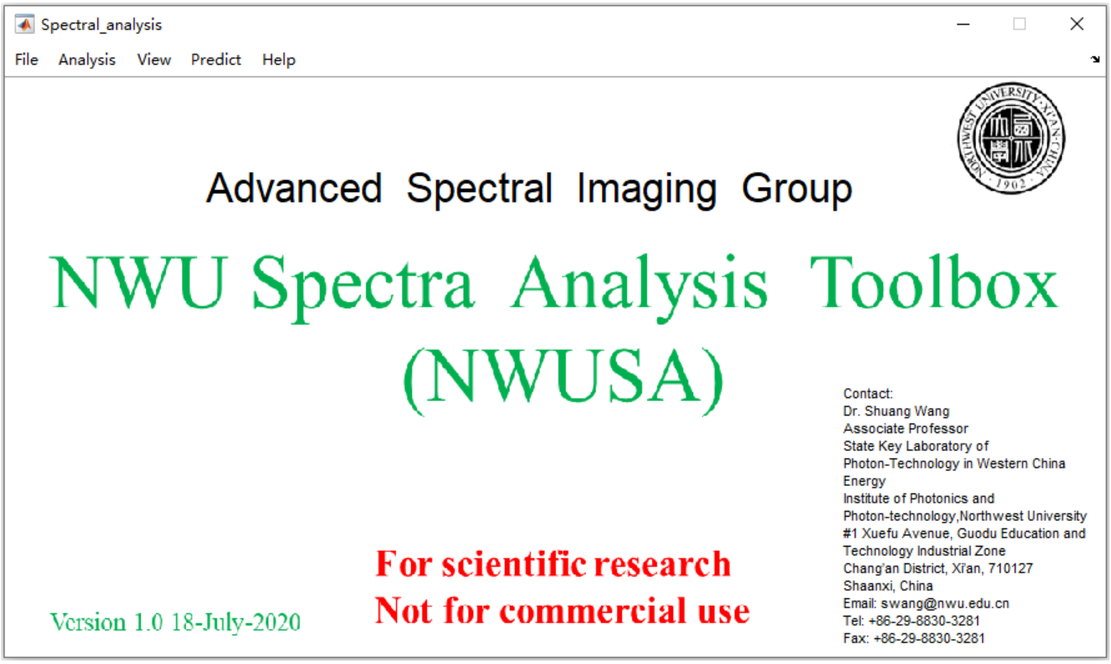
NWUSA toolbox is developed by Advanced Spectral Imaging Group, Institute of Photonics and Photonics Technology, Northwest University, Xi'an 710127, Shaanxi, China. If you need more detailed information about the software usage and operation or have certain suggestions for software improvement, please feel free to contact Dr. Shuang Wang, Institute of Photonics and Photonics Technology (Email: swang@nwu.edu.cn; wsnwuphy@163.com). The NWUSA toolbox can be downloaded at https://github.com/wsnwuphy/NWU-Spectra-Analysis-Toolbox.git.

This software is ONLY for academic research but NOT for commercial usage.

The references where this software has been used are listed in the following.

1. Song D, Yu F, Chen S, et al. Raman spectroscopy combined with multivariate analysis to study the biochemical mechanism of lung cancer microwave ablation. *Biomed Opt Express*. 2020;11(2):1061-1072. https://www.osapublishing.org/boe/abstract.cfm?URI=boe-11-2-1061
2. Li J, Wang R, Qin J, et al. Confocal Raman Spectral Imaging Study of DAPT, a γ-secretase Inhibitor, Induced Physiological and Biochemical Reponses in Osteosarcoma Cells. *Int J Med Sci*. 2020, 17(5):577-590. https://pubmed.ncbi.nlm.nih.gov/32210707/
3. Li J, Li J, Qin J, et al. Confocal Raman microspectroscopic analysis on the time-dependent impact of DAPT, a γ-secretase inhibitor, to osteosarcoma cells. *Spectroc. Acta Pt. A-Molec. Biomolec. Spectr*, 2020, 239:118372. https://www.sciencedirect.com/science/article/pii/S1386142520303504
4. Li J, Qin J, Zeng H, et al. Unveiling dose- and time-dependent osteosarcoma cell responses to the γ-secretase inhibitor, DAPT, by confocal Raman microscopy. *J. Biophotonics*, 2020, doi:10.1002/jbio.202000238. https://www.x-mol.com/paper/1286002611886669824?adv
5. Song D, Chen T, Wang S, et al. Study on the biochemical mechanisms of the micro-wave ablation treatment of lung cancer by ex vivo confocal Raman microspectral imaging. *Analyst*, 2020, 145:626-635. https://pubs.rsc.org/en/content/articlelanding/2020/an/c9an01524h
6. Li J, Qin J, Zhang X, et al. Label-free Raman imaging of live osteosarcoma cells with multivariate analysis. *Appl Microbiol Biotechnol*, 2019:6759-6769. https://pubmed.ncbi.nlm.nih.gov/31230100/
7. 王爽, Haishan Zeng. 实时拉曼光谱分析技术及其在临床早期癌症检测中的应用[J]. 中国激光, 2018, 45(002):35-49. http://www.cnki.com.cn/Article/CJFDTotal-JJZZ201802003.htm

2. Introduction



Advanced Spectral Imaging Group

NWU Spectra Analysis Toolbox
(NWUSA)

Contact:
Dr. Shuang Wang
Associate Professor
State Key Laboratory of
Photon-Technology in Western China
Energy
Institute of Photonics and
Photon-technology,Northwest University
#1 Xuefu Avenue, Guodu Education and
Technology Industrial Zone
Chang'an District, Xi'an, 710127
Shaanxi, China
Email: swang@nwu.edu.cn
Tel: +86-29-8830-3281
Fax: +86-29-8830-3281

For scientific research
Not for commercial use

Version 1.0 18-July-2020

Figure 1. The main interface of the NWUSA toolbox.

NWUSA toolbox is designed for Raman spectral preprocessing and multivariate analysis. Its main functions include the spectra preprocessing techniques (wavenumber range selection, cosmic ray removal, background noise subtraction based on polynomial fitting and extended multiplicative signal correction (EMSC) algorithm, and Savitzky-Golay smoothing), additional spectral analysis functions (area normalization, mean-centered, vector normalization, the first and second derivatives, and spectral normalization under the specific peak), principal component analysis (PCA), linear discriminant analysis (LDA), partial least squares-discriminant analysis (PLS-DA), support vector machine (SVM), and the PCA-SVM.

2.1 Principal Component Analysis

The PCA represents a multivariable analysis technique. It simplifies a complex dataset by reducing the data dimension and extracts significant features from the dataset. Therefore, the PCA can identify subtle differences between different types of samples. Usually, the first principal component (PC) can explain the largest variance in a dataset, and variances explained by subsequent PCs decrease gradually. In addition, the PC scores reflects a difference between different categories, and each PC loading is related to the characteristic spectrum through the PC scores.

2.2 Linear Discriminant Analysis

As a classical supervised classification algorithm, LDA aims to identify one or more optimal projection directions that maximize inter-group differences while minimizing intra-group differences. The number of discriminant functions generated by the LDA is equal to the number of sample categories decreased by one, and the projection data can be calculated by the discriminant functions. In this software, the PC scores obtained by the PCA dimensionality reduction process can be used as the input of the LDA algorithm to generate an effective discrimination model. In the LDA classification, it is feasible to calculate the probability of a sample belonging to a particular category using the Gaussian probability density function. In addition, in order to verify the performance of the classification model based on the PCA-LDA algorithm, it is necessary to perform the cross-validation.

2.3 Partial Least Squares–Discriminant Analysis

PLS-DA is a supervised technique that employs the fundamental principle of PCA, but further rotates the latent variables (LVs) by maximizing the covariance between spectral changes and group affinity, thus enabling the LVs to explain changes related to the diagnosis rather than the most significant variations in the spectral dataset. Before the PLS-DA modeling, it is necessary to select the number of LVs with the minimum sum square error as the optimal number of components using the leave-one-out cross-validation (LOOCV) method. Since the predicted response value obtained by the PLS-DA model is not strictly equal to zero or one, it is necessary to calculate the posterior probabilities of samples belonging to each category using the probability density function and Bayesian formula, and the category with the highest probability is used as the prediction label.

2.4 Support Vector Machine

The SVM is also a supervised classification algorithm based on the VC dimension theory in the statistical learning theory and the principle of structural risk minimization.

One of the basic ideas of the SVM is nonlinear mapping. When data are nonlinearly separable, the SVM maps a low-dimensional input space to the high-dimensional feature space by introducing kernel functions so that data samples may become linearly separable in the high-dimensional space. Another basic idea of the SVM is to construct an optimal hyperplane with the largest classification margin for separating different data types. The SVM was originally applied to solve the binary classification problems. As for multi-classification problems, it is feasible to construct multiple classifiers using the “one-versus-one” and “one-versus-all” strategies and then use these classifiers to conduct voting; next, the category with the largest number of votes should be selected as a prediction label. The SVM analysis process consists of model training and testing. First, the initial dataset is divided into training and test sets according to a certain proportion. Random selection and the Kennard-Stone algorithm are used to split the dataset. The training set is used for model construction and optimization of classification parameters. In the training process, the grid search and cross-validation are combined to obtain optimal parameters of each kernel, which are then used to build the final model. The performance of the final model is validated on the test set. The software uses three types of kernels: linear, polynomial, and radial basis function (RBF).

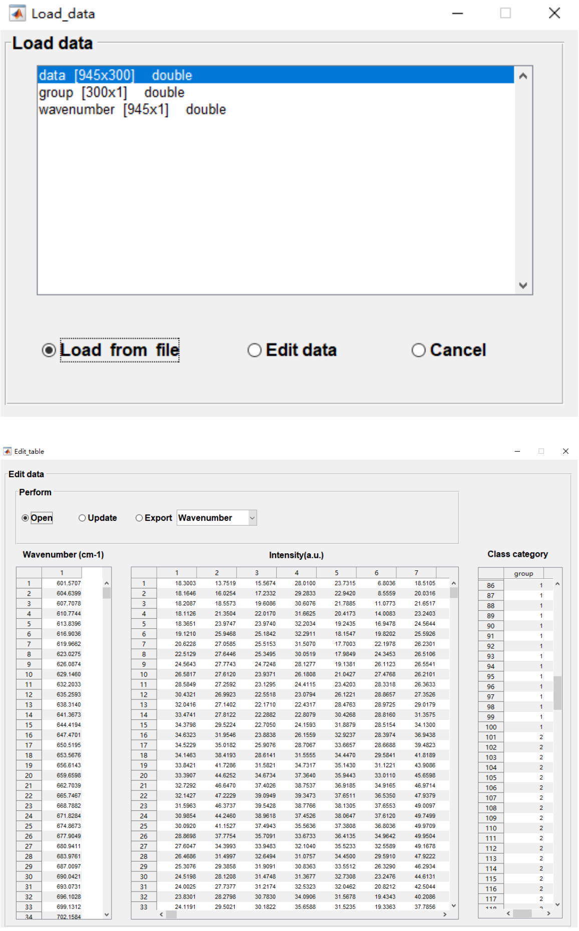
2.5 Principal Component Analysis–Support Vector Machine

In order to realize the visualization of the SVM classification and simplification of calculation without sacrificing the classification performance of the model, this software uses the PCA to reduce the dimensions of the spectral dataset. Then, the first two most significant PC scores are used as the input variables of the SVM. The training and testing process of the PCA-SVM model are the same as those described in Section 2.4. PCA-SVM can not only extract the significant features of the spectral dataset and realize visual classification, but also reduce the computational complexity by reducing the input data characteristics, thus shortening the running time of the program.

3. Operation Guide

3.1 Data Loading

At present, the NWUSA toolbox accepts spectra data in .mat and .txt file formats. If the spectral data are given in a .mat file format, it needs to include the spectra intensity matrix (N×M), a wavenumber variable (N×1), and category label (M×1) to which each spectrum belongs, where N denotes the number of spectral features, and M represents the total number of spectra; it is recommended to use numerical vectors instead of category labels, such as [1, 1... 2, 2... 3, 3…]. If the spectral data are given a .txt file format, it is allowed to load multiple .txt files in batches, and a user only needs to create a .txt file containing the category labels. In addition, the software contains an editable tool that allows users to edit the spectra category and intensity information in a table manually. In order to keep the data balanced, all categories should contain the same number of spectra.



Load_data

Load data

data [945x300] double
group [300x1] double
wavenumber [945x1] double

● Load from file ○ Edit data ○ Cancel

Edit_table

Edit data

Perform

● Open ○ Update ○ Export Wavenumber

Wavenumber (cm-1)

	1
1	601.5707
2	604.6399
3	607.7078
4	610.7744
5	613.8396
6	616.9036
7	619.9662
8	623.0275
9	626.0874
10	629.1460
11	632.2033
12	635.2593
13	638.3140
14	641.3673
15	644.4194
16	647.4701
17	650.5195
18	653.5676
19	656.6143
20	659.6598
21	662.7039
22	665.7467
23	668.7882
24	671.8284
25	674.8673
26	677.9049
27	680.9411
28	683.9761
29	687.0097
30	690.0421
31	693.0731
32	696.1028
33	699.1312
34	702.1584

Intensity(a.u.)

	1	2	3	4	5	6	7
1	18.3003	13.7519	15.5674	28.0100	23.7315	6.8036	18.5105
2	18.1646	16.0254	17.2332	29.2833	22.9420	8.5559	20.0316
3	18.2087	18.5573	19.6086	30.6076	21.7885	11.0773	21.6517
4	18.1126	21.3504	22.0170	31.6625	20.4173	14.0083	23.2403
5	18.3651	23.9747	23.9740	32.2034	19.2435	16.9478	24.5644
6	19.1210	25.9468	25.1842	32.2911	18.1547	19.8202	25.5926
7	20.6228	27.0585	25.5153	31.5070	17.7003	22.1978	26.2301
8	22.5129	27.6446	25.3495	30.0519	17.9849	24.3453	26.5106
9	24.5643	27.7743	24.7248	28.1277	19.1381	26.1123	26.5541
10	26.5817	27.6120	23.9371	26.1808	21.0427	27.4768	26.2101
11	28.5849	27.2592	23.1295	24.4115	23.4203	28.3318	26.3633
12	30.4321	26.9923	22.5518	23.0794	26.1221	28.8657	27.3526
13	32.0416	27.1402	22.1710	22.4317	28.4763	28.9725	29.0179
14	33.4741	27.8122	22.2882	22.8079	30.4268	28.8160	31.3575
15	34.3798	29.5224	22.7050	24.1593	31.8879	28.5154	34.1300
16	34.6323	31.9546	23.8838	26.1559	32.9237	28.3974	36.9438
17	34.5229	35.0182	25.9076	28.7067	33.6657	28.6688	39.4823
18	34.1463	38.4193	28.6141	31.5555	34.4470	29.5841	41.8189
19	33.8421	41.7286	31.5821	34.7317	35.1430	31.1221	43.9086
20	33.3907	44.6252	34.6734	37.3640	35.9443	33.0110	45.6598
21	32.7292	46.6470	37.4026	38.7537	36.9185	34.9165	46.9714
22	32.1427	47.2229	39.0949	39.3473	37.6511	36.5350	47.9379
23	31.5963	46.3737	39.5428	38.7766	38.1305	37.6553	49.0097
24	30.9854	44.2460	38.9618	37.4526	38.0647	37.6120	49.7499
25	30.0920	41.1527	37.4943	35.5636	37.3808	36.8036	49.9709
26	28.8698	37.7754	35.7091	33.6733	36.4135	34.9642	49.9504
27	27.6047	34.3993	33.9483	32.1040	35.5233	32.5589	49.1678
28	26.4686	31.4997	32.6494	31.0757	34.4500	29.5910	47.9222
29	25.3076	29.3858	31.9091	30.8363	33.5512	26.3290	46.2934
30	24.5198	28.1208	31.4748	31.3677	32.7308	23.2476	44.6131
31	24.0025	27.7377	31.2174	32.5323	32.0462	20.8212	42.5044
32	23.8301	28.2798	30.7830	34.0906	31.5678	19.4343	40.2086
33	24.1191	29.5021	30.1822	35.6588	31.5235	19.3363	37.7856

Class category

	group
86	1
87	1
88	1
89	1
90	1
91	1
92	1
93	1
94	1
95	1
96	1
97	1
98	1
99	1
100	1
101	2
102	2
103	2
104	2
105	2
106	2
107	2
108	2
109	2
110	2
111	2
112	2
113	2
114	2
115	2
116	2
117	2

Figure 2. Data loading and editing interface.

3.2 Spectra Preprocessing

When performing the Raman spectra preprocessing process, a user first needs to choose the following options to enter the spectral preprocessing interface: “Analysis > Preprocessing > Spectral Preprocessing,” and then to click on the button “Load” on the right side of the preprocessing interface to load the spectral data in the form of a .mat file or multiple .txt files. The spectral data in .txt format are selected to be processed in the list box, and then the spectral selection is completed by clicking on the “Select” button. The user can view the processing results in the drawing window in real time by clicking different buttons and parameters in the “Methods” panel on the left of the preprocessing interface. For the spectral data in the format of a .mat file, the user only needs to enter the spectra serial number to be viewed in the “Spectrum sequence number” on the right side, and the rest of the operation is the same as in the previous case. After that, the user can query peak information by clicking on the “Find peaks” button and save the preprocessed spectral data by clicking on the “Output data” button.

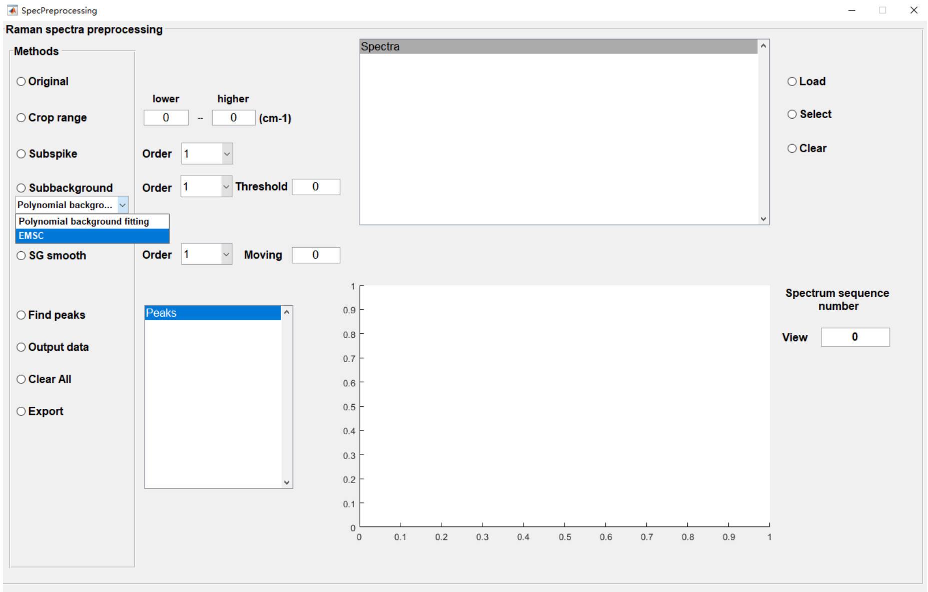


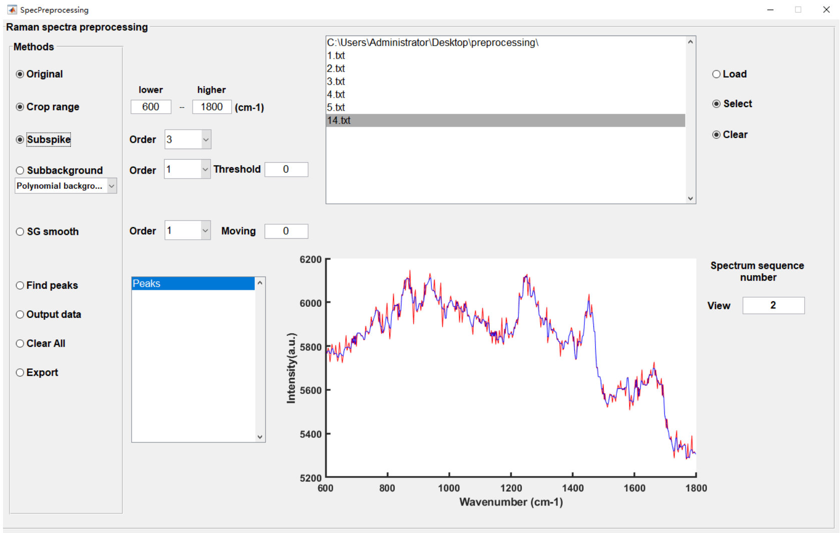
Figure 3. The preprocessing interface in the NWUSA toolbox.

3.2.1 Selecting Spectral Range (Crop Range)

The user needs to input the spectral range of interest in the box and clicks the “Crop range” button.

3.2.2 Removing Cosmic Rays (Subspike)

The user can select appropriate parameter (order) for removing cosmic rays. As shown in the figure below, the red lines represent the spectrum after the spectral range is cut, and the blue lines represent the spectrum after the cosmic rays are removed.



SpecPreprocessing

Raman spectra preprocessing

Methods

● Original

● Crop range   lower 600 -- higher 1800 (cm-1)

● Subspike   Order 3

○ Subbackground   Polynomial backgro...   Order 1   Threshold 0

○ SG smooth   Order 1   Moving 0

○ Find peaks

○ Output data

○ Clear All

○ Export

C:\Users\Administrator\Desktop\preprocessing\
1.txt
2.txt
3.txt
4.txt
5.txt
14.txt

○ Load

● Select

● Clear

Peaks

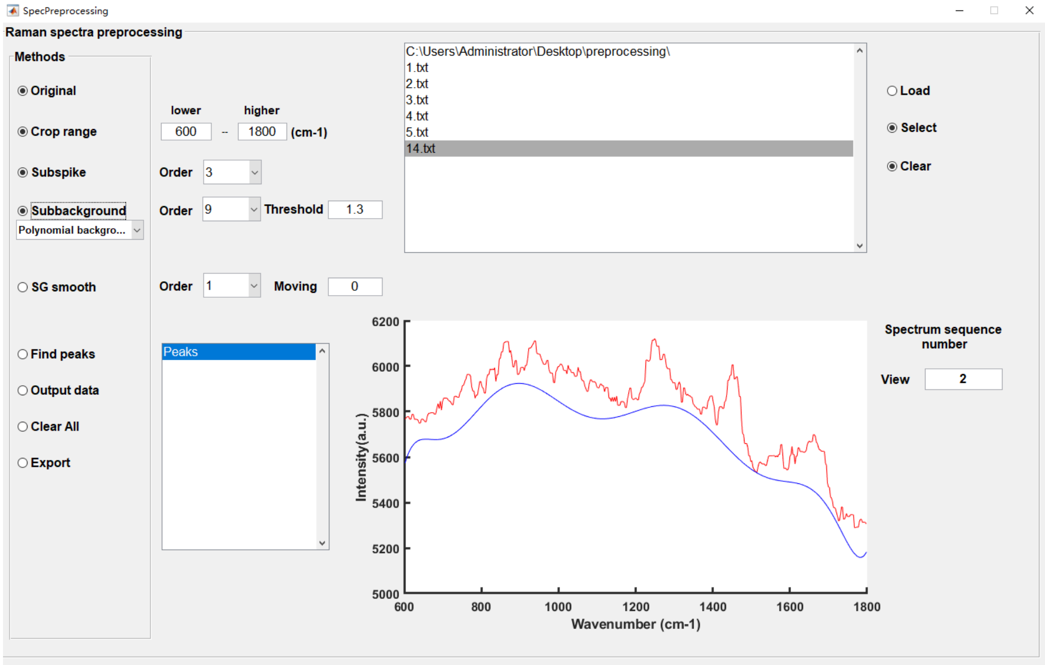
Spectrum sequence number

View 2

Figure 4. Spectrum with cosmic rays removed.

3.2.3 Background Noise Subtraction (Subbackground)

Before using the background subtraction function, the user needs to select the method for background subtraction from the pop-up menu, such as “Polynomial background fitting” or “EMSC”. In this example, the red lines represent the spectrum after removing cosmic rays, and the blue lines represent the baseline of the polynomial fitting. Subtracting these two will yield the spectral data after background noise subtraction. “Order” stands for polynomial order, and “Threshold” can be used to adjust the position of the baseline. In addition, the open-source EMSC toolbox developed by Kristian *et al*. is included in the software, and when the EMSC is used as the light scattering correction method, the baseline correction is not required. The EMSC toolbox has been approved for use in the NWUSA toolbox by Prof. Kristian Hovde Liland in Norwegian University of Life Sciences. If the user chooses to use “EMSC” for baseline correction, he/she needs to click “Subbackground” to enter the EMSC menu, click “Select data” button, select “rawdata” variable in the list box of the subinterface, and click “OK” button to complete data selection. After that, the user clicks the “EMSC” button to enter the EMSC processing interface, in which the user is free to set parameters and reference spectrum, and clicks the “Apply EMSC” button to complete the baseline correction.



C:\Users\Administrator\Desktop\preprocessing\
1.txt
2.txt
3.txt
4.txt
5.txt
14.txt

● Original

● Crop range: lower 600 – higher 1800 (cm-1)

● Subspike: Order 3

● Subbackground: Polynomial backgro... Order 9 Threshold 1.3

○ SG smooth: Order 1 Moving 0

○ Find peaks

○ Output data

○ Clear All

○ Export

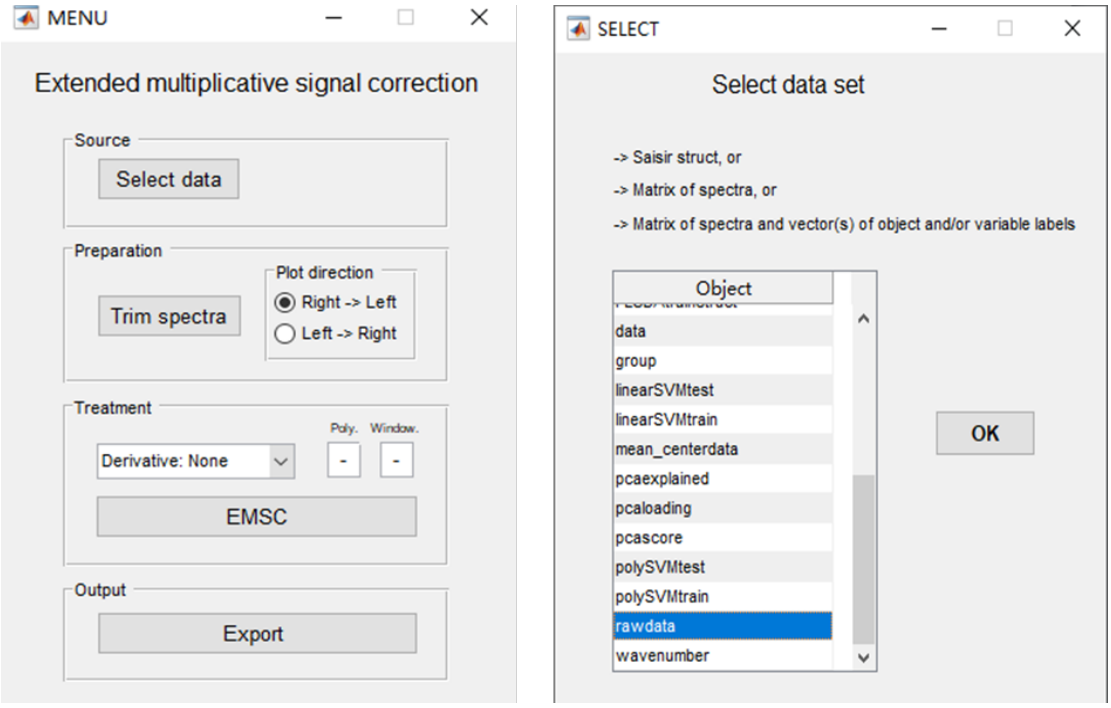
○ Load

● Select

● Clear

View 2

Figure 5. The baseline of the polynomial fitting.



● Right -> Left

○ Left -> Right

Derivative: None

-> Saisir struct, or

-> Matrix of spectra, or

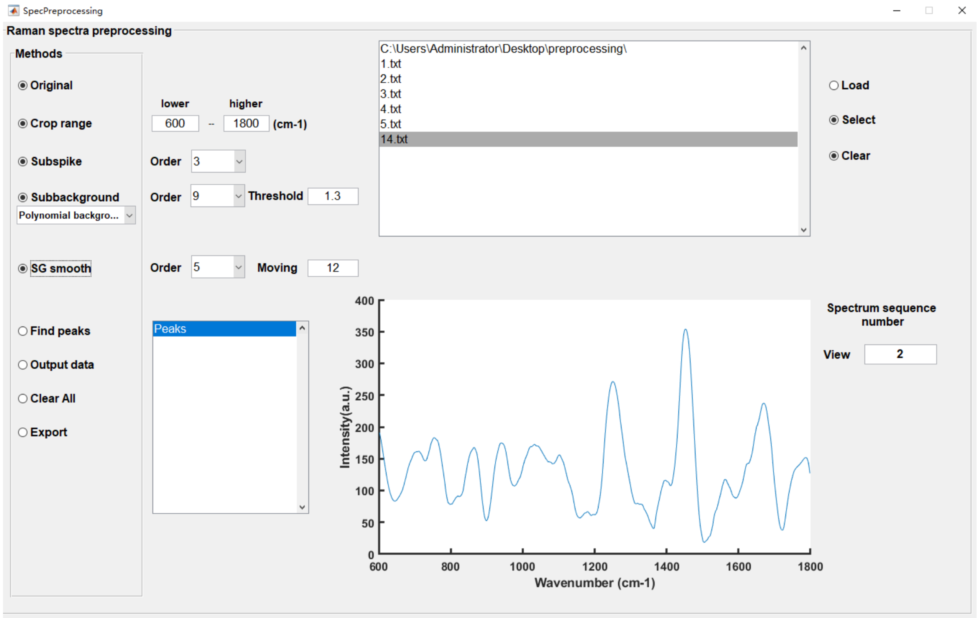
-> Matrix of spectra and vector(s) of object and/or variable labels

data
group
linearSVMtest
linearSVMtrain
mean_centerdata
pcaexplained
pcaloading
pcascore
polySVMtest
polySVMtrain
rawdata
wavenumber

Figure 6. EMSC menu and data selection interface.

3.2.4 Savitzky-Golay Smoothing (SG Smoothing)

Finally, a smoother spectrum can be obtained by performing Savitzky-Golay smoothing based on the least square principle, as shown in the figure below.



lower 600 – higher 1800 (cm-1)

Order 3

Order 9 Threshold 1.3

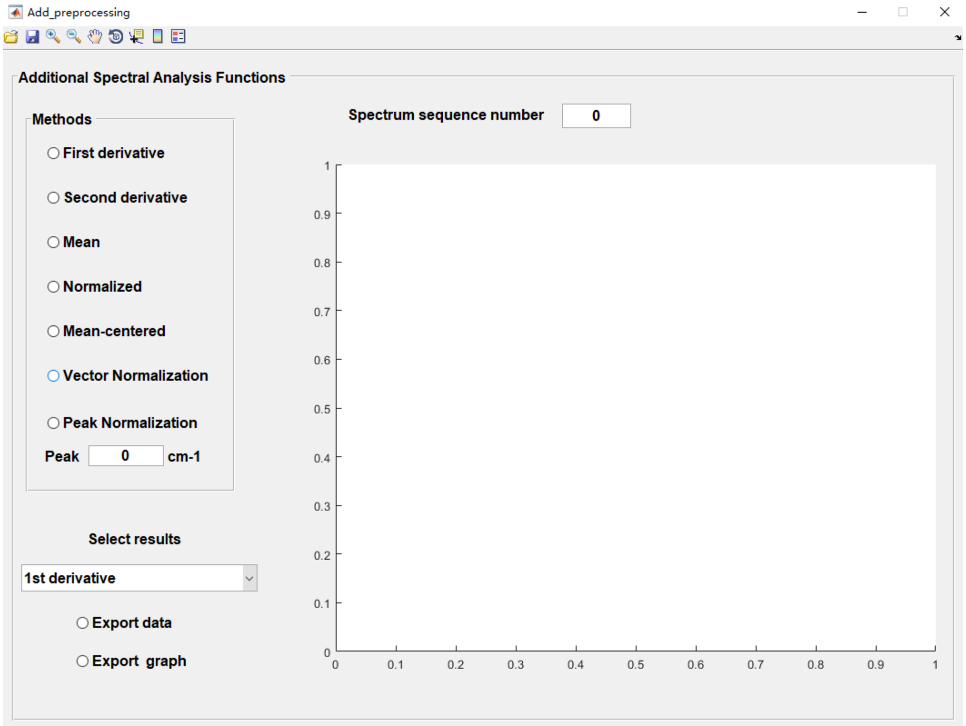
Order 5 Moving 12

View 2

C:\Users\Administrator\Desktop\preprocessing\
1.txt
2.txt
3.txt
4.txt
5.txt
14.txt

Figure 7. The smoothed spectral by Savitzky-Golay algorithm.

3.3 Additional Spectral Analysis Functions



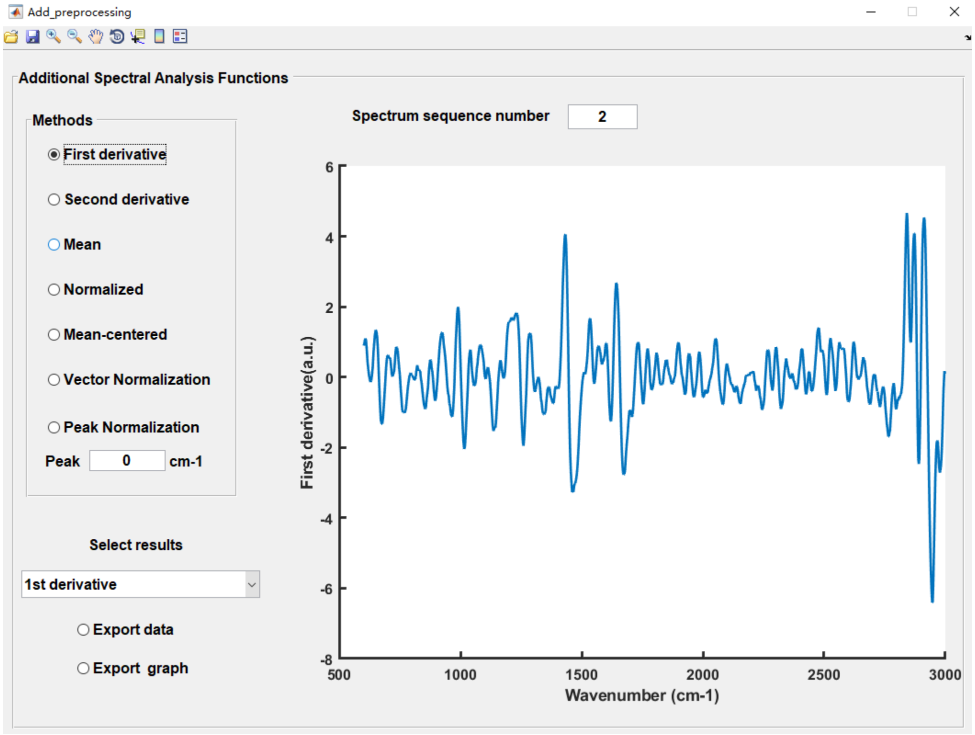
Spectrum sequence number 0

Peak 0 cm-1

1st derivative

Figure 8. Additional spectral analysis functions in the NWUSA toolbox.

3.3.1 First and Second Derivatives



Spectrum sequence number 2

Peak 0 cm-1

Select results: 1st derivative

Figure 9. The first derivative of an example spectrum.

3.3.2 Mean spectral intensity (Mean)

This function can be used in two modes. One is to obtain the mean intensity of the preprocessed spectral data, and another is to obtain the mean intensity of the spectral data after area normalization. Users can choose the mode according to their needs.

3.3.3 Area Normalization

Spectrum normalization through the integration of the area under the spectral curve can reduce the influence of certain factors, such as laser power disturbance. This function can also be used in two modes. One is to normalize the area under the preprocessed spectral curve, and the other is to perform normalization for the mean spectral. Users can choose the mode according to their needs.

3.3.4 Mean-centering

The software subtracts the mean value from the spectral intensity matrix after area normalization to obtain the mean-centered matrix, and used for subsequent multivariate analysis. The mean-centered approach can eliminate the influence of intra-subject and/or inter-subject spectral variability on multivariate analysis.

3.3.5 Vector normalization

During vector normalization, the software first calculates the norm of the spectrum, which is defined as the square root of the sum of squares of spectrum intensities. Then, the intensity corresponding to each Raman shift is divided by the norm to obtain the normalized spectrum.

3.3.6 Spectral Normalization Under Specific Peak (Peak Normalization)

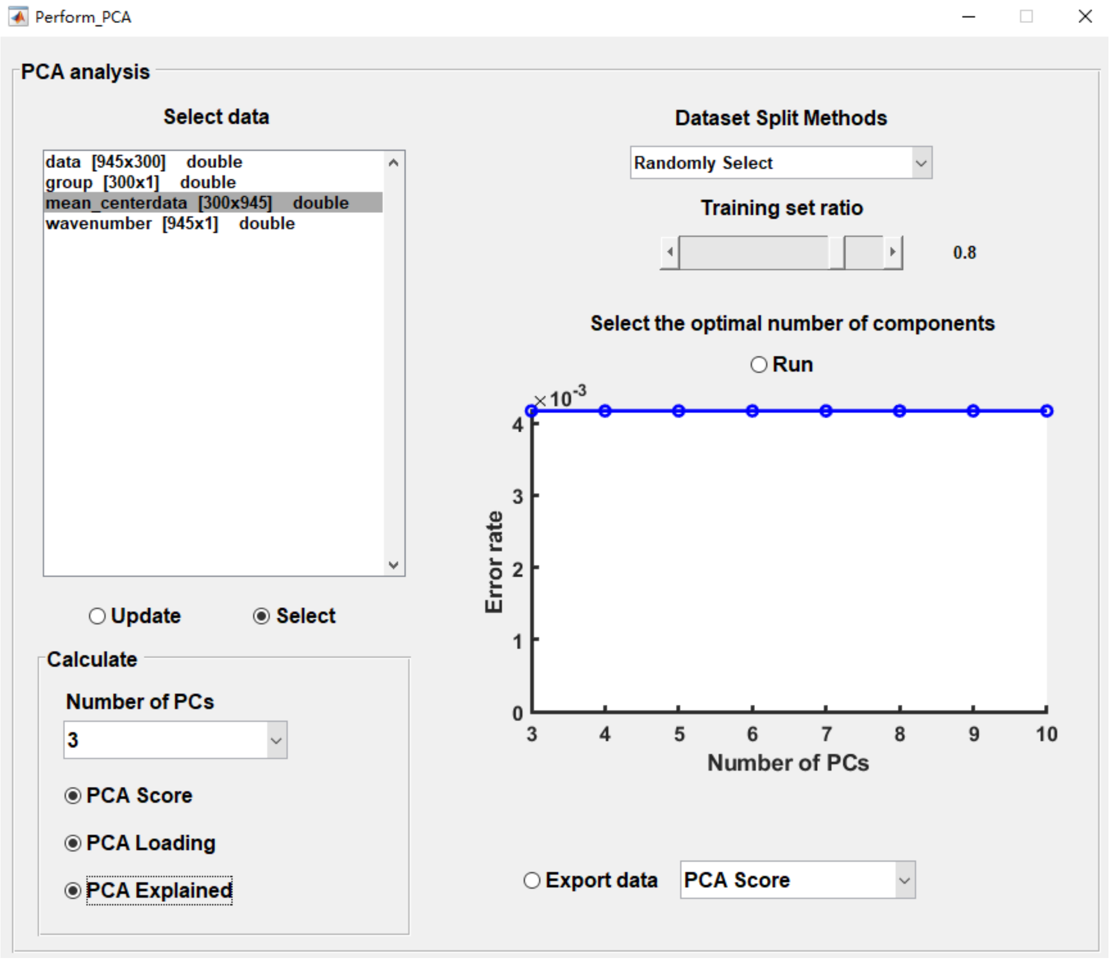
During the normalization under specific peak, the intensity value corresponding to the center frequency of a specific Raman shift is used as a reference and defined as I_P. Next, the Raman intensity of the whole spectrum is divided by I_P to obtain the normalized spectrum. The user can set the position of the reference peak freely.

Users have an option to save results using the pop-up menu on the “Add preprocessing” interface, and the data and graph can be saved in a specified file by clicking on the “Export data” and “Export graph” buttons. The software supports exporting data in several formats, including .txt, .mat, and .xls formats.

3.4 Multivariate Analysis

3.4.1 Principal Component Analysis

The user can enter the PCA analysis interface by clicking “Analysis > PCA,” where he can update the variable name in the list box by clicking on the “Update” button and select the variable used for the PCA analysis by clicking on the “Select” button; it is recommended to select “mean_centerdata” in the list box. Then, the user can customize the dataset splitting method and training set ratio to divide the dataset into training and test sets. Click the “Run” button to calculate and select the number of components with the minimum error rate as the optimal number of components. After that, the user can select the optimal number of PCs to be retained and click “PCA Score / Loading / Explained” in the calculation panel to complete the PCA analysis. Finally, the user selects the result to be exported from the pop-up menu at the bottom of the interface and completes the exporting process by clicking on the “Export data” button.



Perform_PCA

PCA analysis

Select data

```
data  [945x300]    double
group  [300x1]    double
mean_centerdata  [300x945]    double
wavenumber  [945x1]    double
```

○ Update   ● Select

Calculate

Number of PCs: 3

● PCA Score

● PCA Loading

● PCA Explained

Dataset Split Methods: Randomly Select

Training set ratio: 0.8

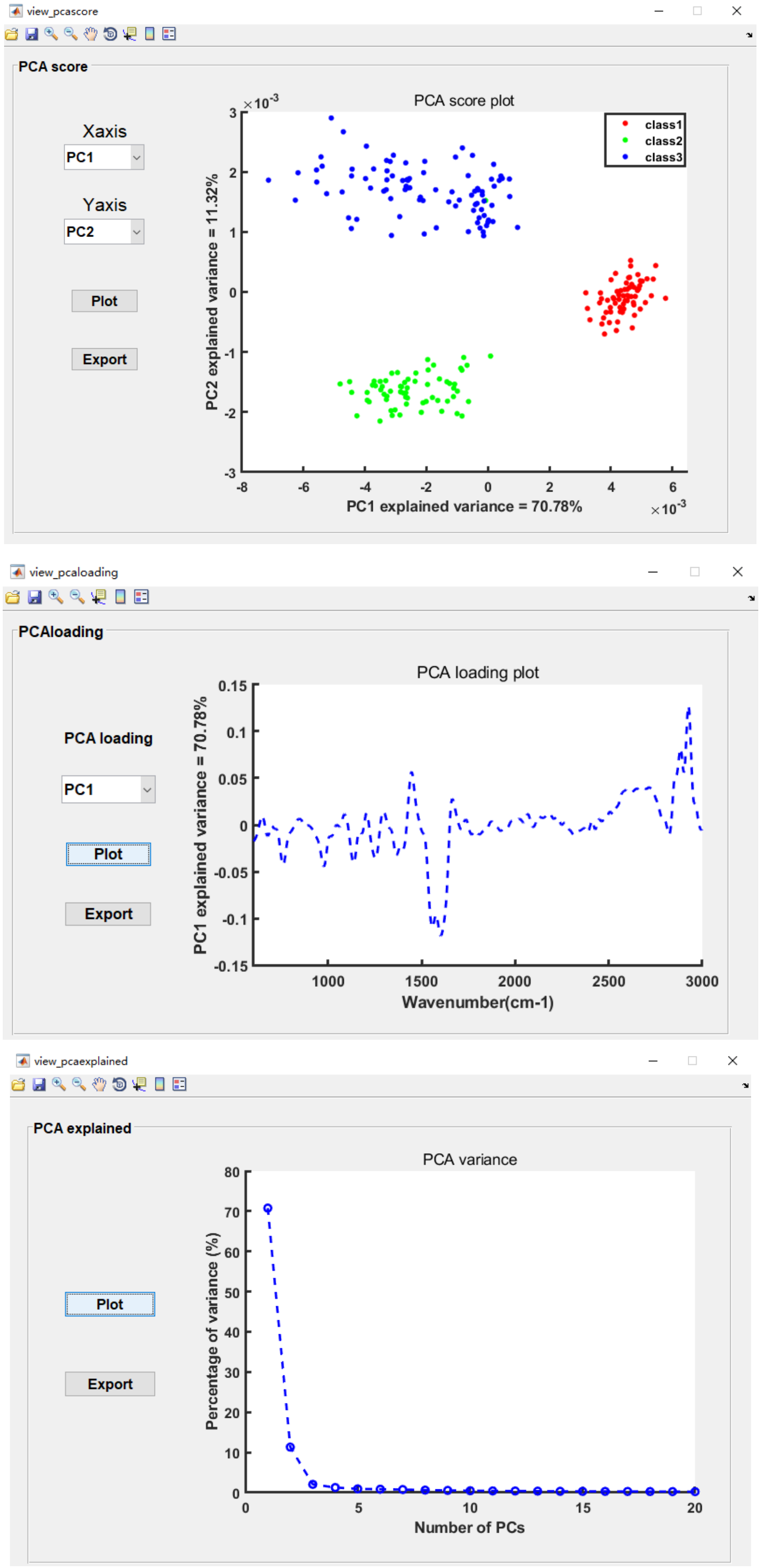
Select the optimal number of components

○ Run

○ Export data   PCA Score

Figure 10. PCA analysis interface in the NWUSA toolbox.

The user can view the PCA results, including the projection data (score), loading, and variance percentage, by clicking “View > viewPCA > view PCAscore / loading / explained” to open the corresponding drawing interface and then can freely select different PCs for drawing. By clicking “Export” in each drawing window, the graph is output.



Xaxis: PC1

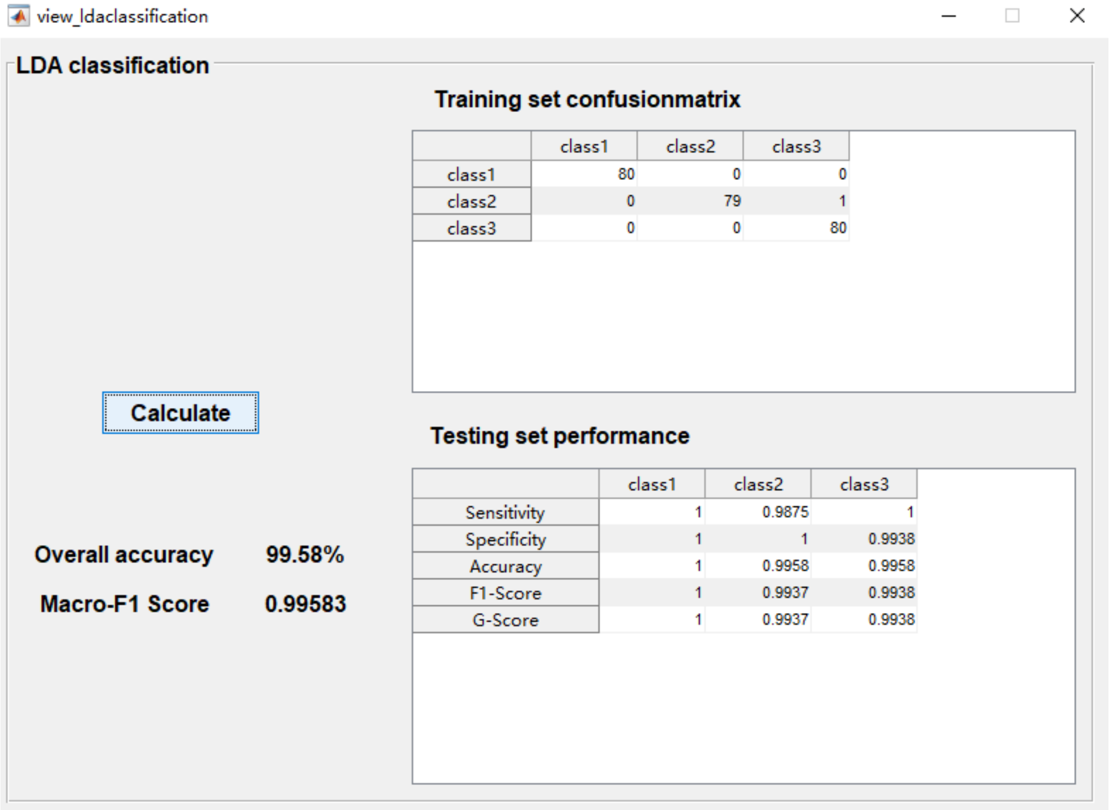
Yaxis: PC2

PCA loading: PC1

Figure 11. PCA scores scatter plot, loading of PC1, and variance percentage of the first 20 PCs.

3.4.2 Linear Discriminant Analysis

To execute the LDA, the user needs to click “Analysis > LDA > ProjData.” After entering the interface, the user has to select the number of components (the number of selected components should be at least equal to the number of sample categories) and then to click on the “Calculate” button to obtain the data after the LDA projection and the posterior probability of samples belonging to each category. Similarly, the user can select the result to be exported from the pop-up menu and click on the “Export data” button to complete the export. Then, the user can click “View > view LDA > view LDAprojData” to view the scatter plot of linear discriminant scores. Also, the user can click “View > view LDA > view LDAprobability” to view the posterior probability of samples. By clicking on “Analysis > LDA > LDA classification” view the classification results and performance evaluation indicators of the PCA-LDA model on the training set, including sensitivity, specificity, accuracy, F1-score, G-score, Macro-F1 score, overall accuracy and the confusion matrices.



Training set confusionmatrix

	class1	class2	class3
class1	80	0	0
class2	0	79	1
class3	0	0	80

Testing set performance

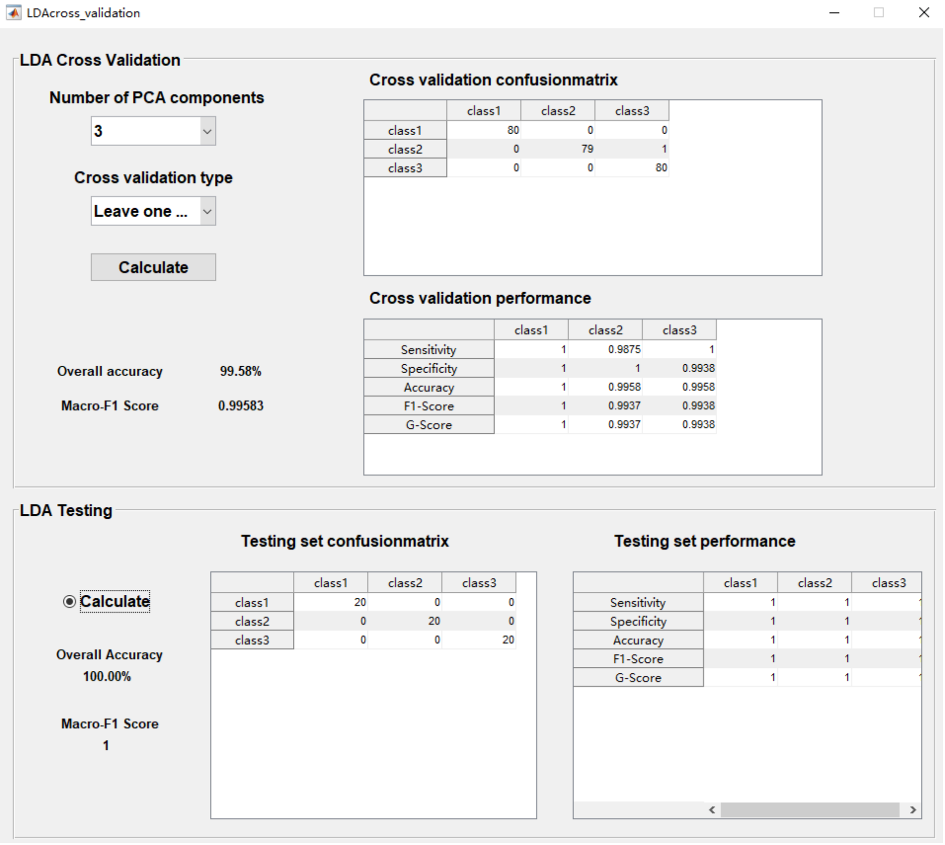
	class1	class2	class3
Sensitivity	1	0.9875	1
Specificity	1	1	0.9938
Accuracy	1	0.9958	0.9958
F1-Score	1	0.9937	0.9938
G-Score	1	0.9937	0.9938

Overall accuracy 99.58%

Macro-F1 Score 0.99583

Figure 12. The classification results and performance evaluation of the PCA-LDA model on the training set.

The next step is to assess the classification performance of the model by performing cross-validation. This is done by clicking “Analysis > LDA > Cross validation” to enter the cross-validation interface and then selecting the number of PCs and groups for cross-validation. The number of PCs is the same as the number of components selected previously. The type of cross-validation can be set to five-fold, ten-fold, or leave-one-out cross-validation. The user can click on the “Calculate” button to view the classification results of the model for different categories of samples represented by the confusion matrices, as well as the corresponding performance evaluation indicators, including sensitivity, specificity, accuracy, F1-score, G-score, Macro-F1 score, and the overall accuracy. In the “LDA Testing” panel, the user can view the classification result and performance evaluation indicators of the PCA-LDA model on the test set by clicking the “Calculate” button. In addition, the user can obtain the ROC curve of each category by clicking “Analysis > LDA > ROC.”



Number of PCA components: 3

Cross validation type: Leave one ...

Overall accuracy 99.58%

Macro-F1 Score 0.99583

Cross validation confusionmatrix

	class1	class2	class3
class1	80	0	0
class2	0	79	1
class3	0	0	80

Cross validation performance

	class1	class2	class3
Sensitivity	1	0.9875	1
Specificity	1	1	0.9938
Accuracy	1	0.9958	0.9958
F1-Score	1	0.9937	0.9938
G-Score	1	0.9937	0.9938

◉ Calculate

Overall Accuracy
100.00%

Macro-F1 Score
1

Testing set confusionmatrix

	class1	class2	class3
class1	20	0	0
class2	0	20	0
class3	0	0	20

Testing set performance

	class1	class2	class3
Sensitivity	1	1	1
Specificity	1	1	1
Accuracy	1	1	1
F1-Score	1	1	1
G-Score	1	1	1

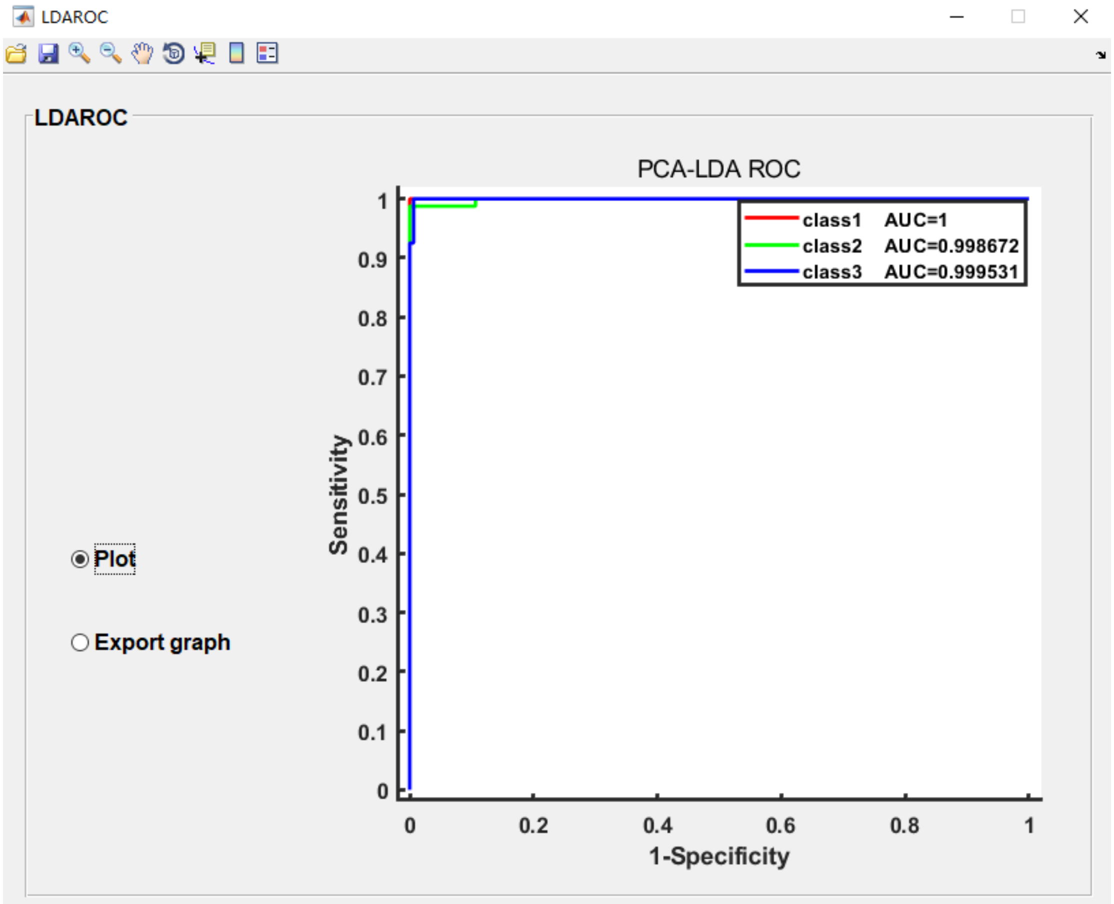
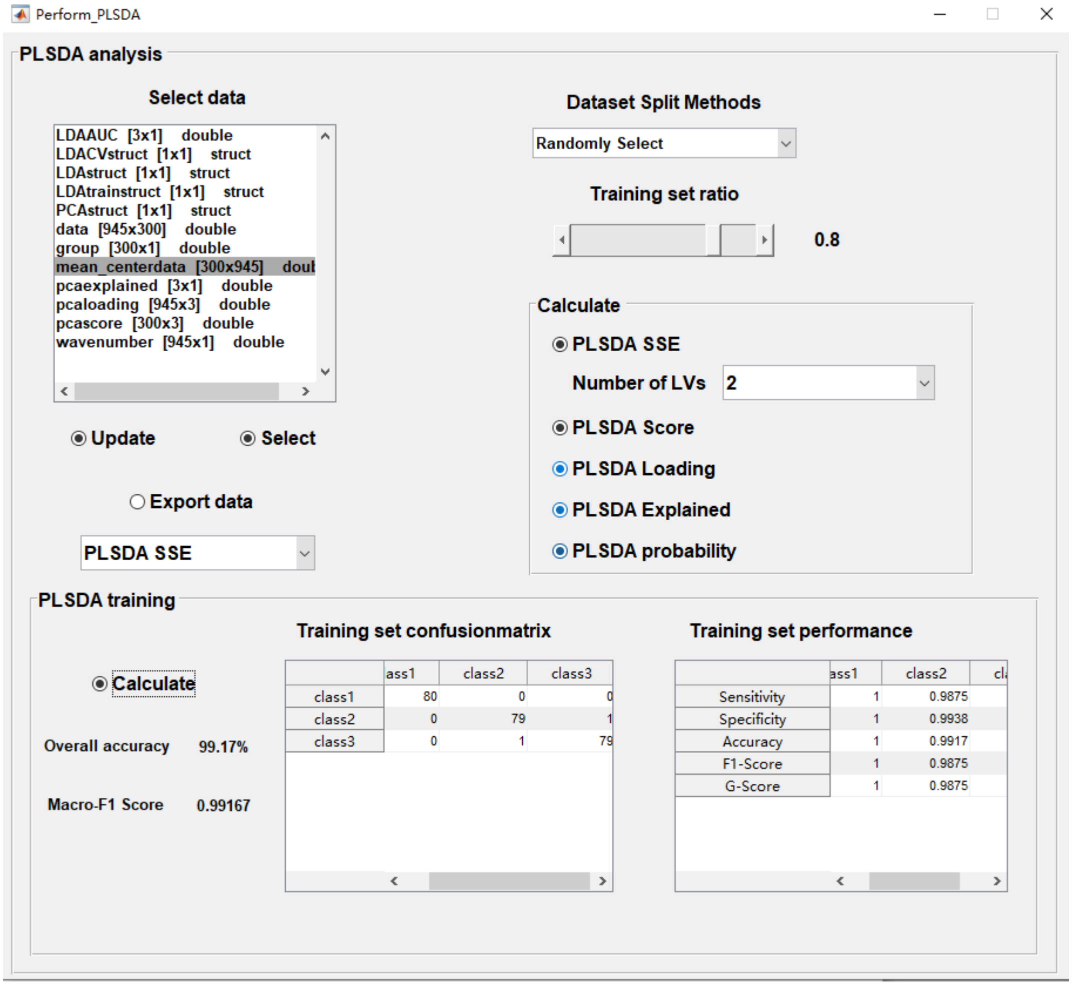


Figure 13. The cross validation results, classification results and performance evaluation of the model on the test set, and the ROC curve of each class.

3.4.3 Partial Least Squares-Discriminant Analysis

First, the user can enter the PLS-DA analysis interface by clicking "Analysis > PLSDA > Perform PLSDA," and then the "Update" button should be clicked to update the variable name in the list box; then, the variable for the PLS-DA analysis should be selected, and lastly, the "Select" button should be clicked to complete data selection; it is recommended to select "mean_centerdata" in the list box. Then, the user can customize the dataset splitting method and training set ratio to divide the dataset into training and test sets. Click the "PLSDA SSE" button in the calculation panel to get the sum of squared errors (SSE) of the first 20 latent variables (LVs), and the SSE curve can be drawn by clicking "View > viewPLSDA > view PLSDAsumsqerr"; the number of LVs with the minimum SSE is chosen as the optimal number of components required for modeling. Next, the corresponding PLSDA execution results can be obtained by selecting the number of LVs and clicking "PLSDA Score / Loading / Explained / probability". The drawing interface of PLSDA analysis results can be opened by clicking "View > viewPLSDA > view PLSDAscore / loading," and different LVs can be chosen for drawing. To output the graph in each drawing interface, it is needed to click on the "Export" button.



PLSDA analysis

Select data

LDAAUC [3x1] double
LDACVstruct [1x1] struct
LDAstruct [1x1] struct
LDAtrainstruct [1x1] struct
PCAstruct [1x1] struct
data [945x300] double
group [300x1] double
mean_centerdata [300x945] doul
pcaexplained [3x1] double
pcaloading [945x3] double
pcascore [300x3] double
wavenumber [945x1] double

◉ Update ◉ Select

○ Export data

PLSDA SSE

Dataset Split Methods

Randomly Select

Training set ratio 0.8

Calculate

◉ PLSDA SSE

Number of LVs 2

◉ PLSDA Score

◉ PLSDA Loading

◉ PLSDA Explained

◉ PLSDA probability

PLSDA training

◉ Calculate

Overall accuracy 99.17%

Macro-F1 Score 0.99167

Training set confusionmatrix

	ass1	class2	class3
class1	80	0	0
class2	0	79	1
class3	0	1	79

Training set performance

	ass1	class2	cl
Sensitivity	1	0.9875	
Specificity	1	0.9938	
Accuracy	1	0.9917	
F1-Score	1	0.9875	
G-Score	1	0.9875	

Figure 14. PLSDA analysis interface in the NWUSA toolbox.

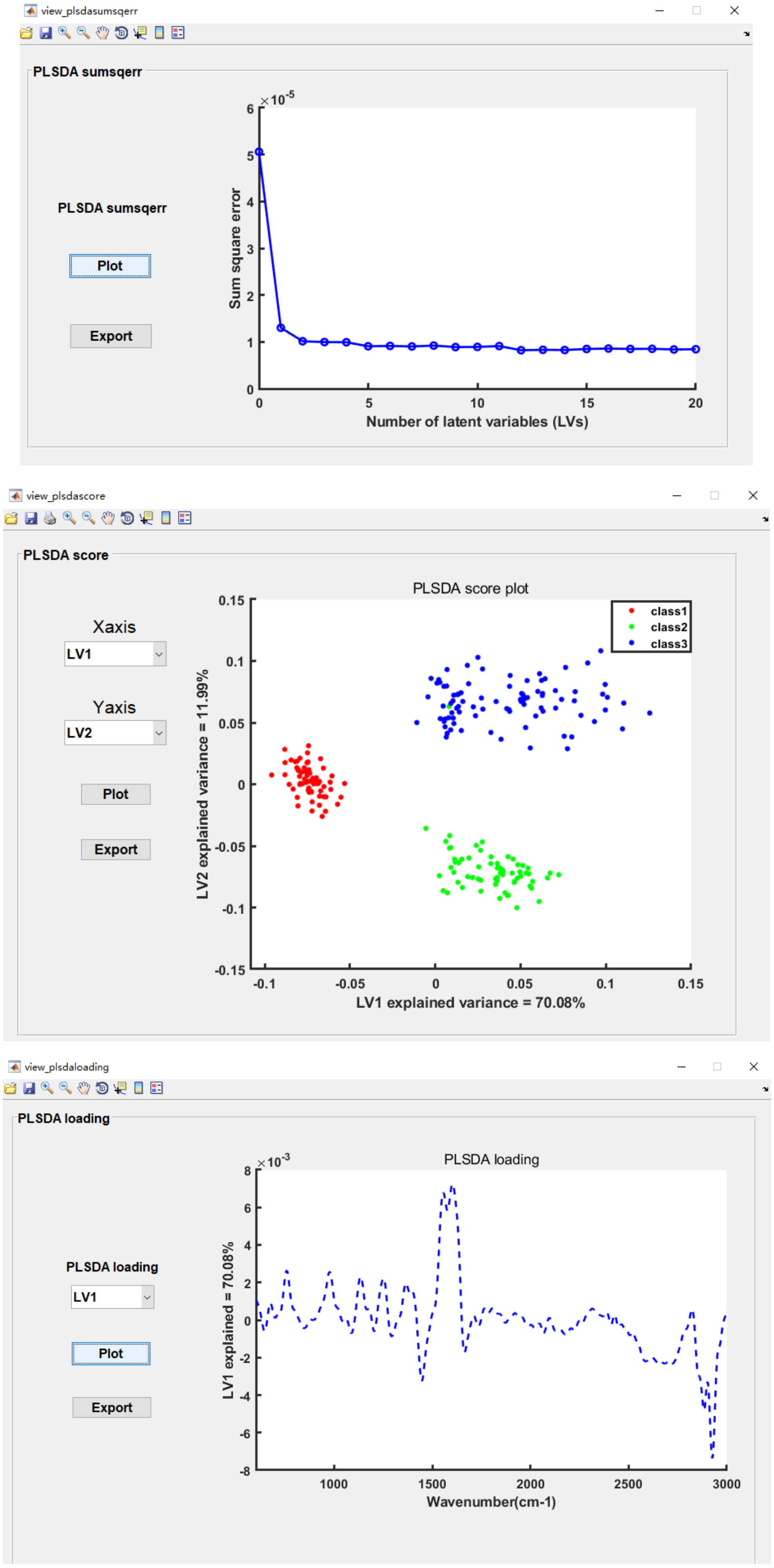
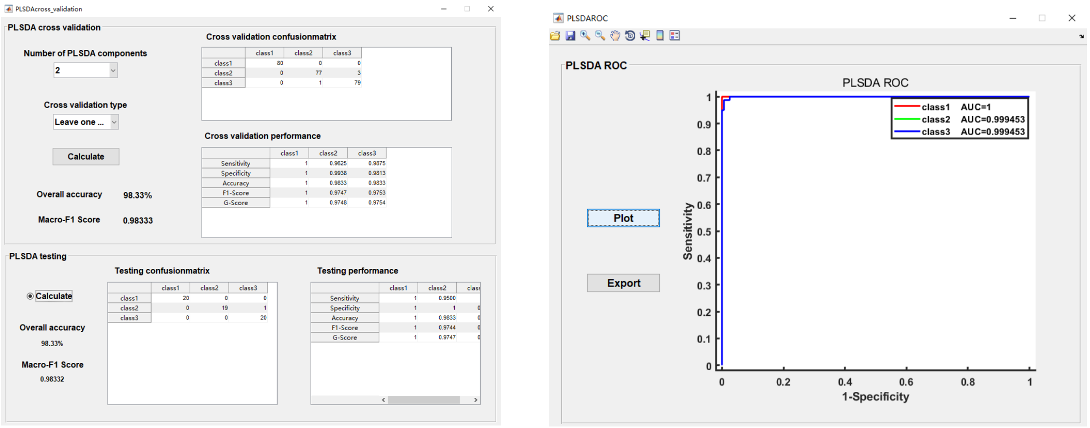


Figure 15. The variance percentage of the first 20 LVs, scores scatter plot, and loading of LV1.

In the "PLSDA training" panel, the user can click on the "Calculation" button to view the classification results and performance evaluation indicators of the PLSDA model on the training set, including sensitivity, specificity, accuracy, F1-score, G-score, Macro-F1 score, overall accuracy and the confusion matrices. After that, the cross-validation function is performed by clicking "Analysis > PLSDA > PLSDA Cross validation." The number of PLSDA components is the optimal number of LVs selected in the first step, and the cross-validation types can be set to five-fold, ten-fold, or leave-one-out cross-validation. The user can click on the "Calculate" button to view the classification results of the PLSDA model for training set samples represented by the confusion matrices, as well as the corresponding performance evaluation indicators, including sensitivity, specificity, accuracy, F1-score, G-score, Macro-F1 score, and the overall accuracy. In the "PLSDA testing" panel, the user can obtain the classification results and performance evaluation indicators of the PLS-DA model on the test set by clicking the "Calculate" button. In addition, the user can obtain the ROC curve by clicking "Analysis > PLSDA > PLSDA ROC."



Number of PLSDA components: 2

Cross validation type: Leave one ...

Overall accuracy 98.33%

Macro-F1 Score 0.98333

Cross validation confusionmatrix

	class1	class2	class3
class1	80	0	0
class2	0	77	3
class3	0	1	79

Cross validation performance

	class1	class2	class3
Sensitivity	1	0.9625	0.9875
Specificity	1	0.9938	0.9813
Accuracy	1	0.9833	0.9833
F1-Score	1	0.9747	0.9753
G-Score	1	0.9748	0.9754

PLSDA testing

Overall accuracy 98.33%

Macro-F1 Score 0.98332

Testing confusionmatrix

	class1	class2	class3
class1	20	0	0
class2	0	19	1
class3	0	0	20

Testing performance

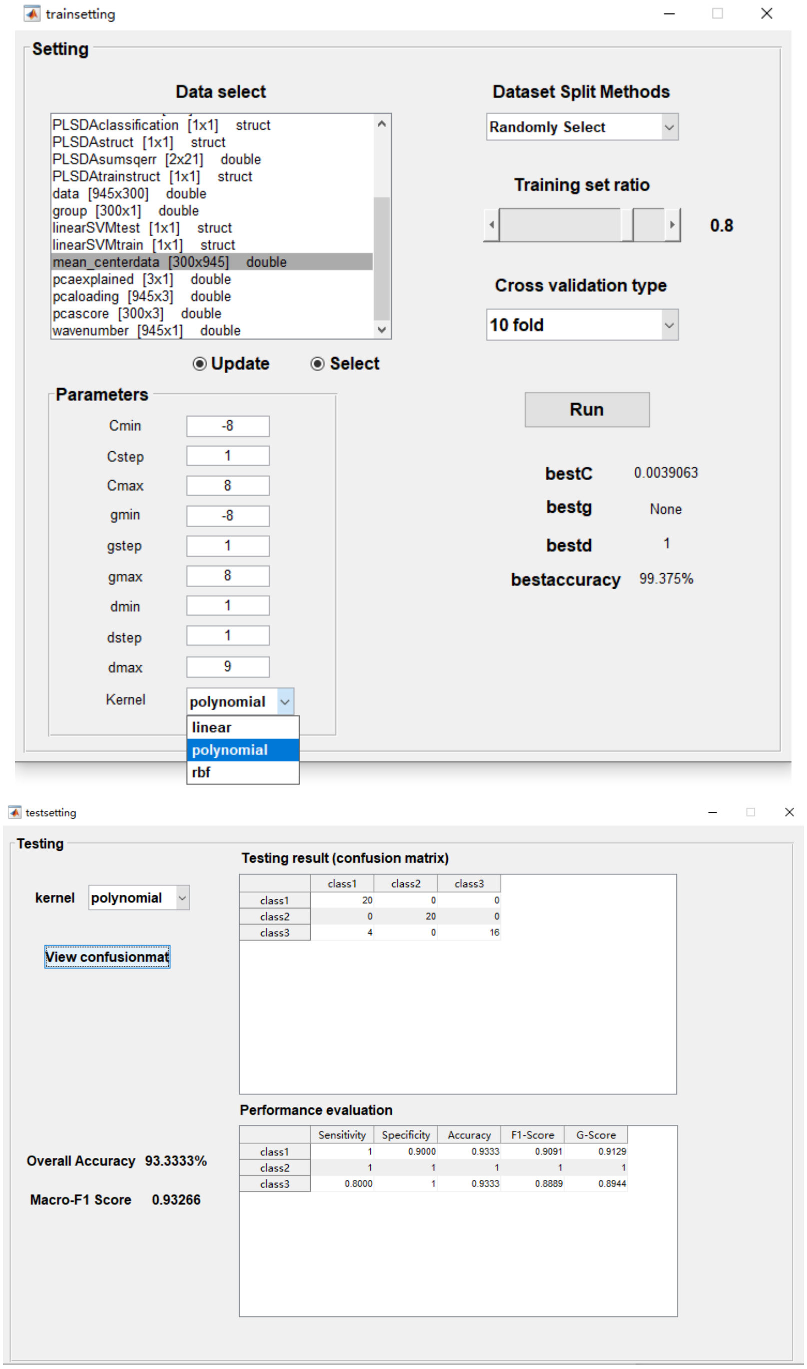
	class1	class2	clas
Sensitivity	1	0.9500	
Specificity	1	1	0
Accuracy	1	0.9833	0
F1-Score	1	0.9744	0
G-Score	1	0.9747	0

Figure 16. The cross validation results, classification results and performance evaluation of the model on the test set, and the ROC curve of each class.

3.4.4 Support Vector Machine

In the execution process of the SVM, the selection of kernel type and parameters is crucial. User can enter the interface for setting the SVM training parameters by clicking "Analysis > SVM > TrainSVM," and then click on the "Update" button to update the variable name in the list box, select the variable used for the SVM training, and click on the “Select” button to complete data selection; it is recommended to select “mean_centerdata” in the list box. Then, the user can customize the dataset splitting method, the proportion of the training set, the cross-validation types, kernel type, and the corresponding parameter ranges. It is recommended to use the range of 0.6–0.8 for the proportion of the training set, that is, the training set accounts for 60%–80% of the overall number of data samples.

For a linear kernel, it is only necessary to set the range of parameter C (C_{min}, C_{step}, and C_{max} represent the minimum value, step size, and maximum value, respectively). For a polynomial kernel, it is necessary to specify the ranges of parameter C and polynomial order d (d_{min}, d_{step}, and d_{max} represent the minimum value, step size, and maximum value of polynomial order, respectively, and they are all positive integers). For an RBF kernel, the ranges of parameters C and *gamma* (g_{min}, g_{step}, and g_{max} represent the minimum value, step size, and maximum value of *gamma*, respectively) should be specified. It is worth noting that for the convenience of use, the parameters C and g (*gamma*) in the setting interface can be expressed in the form of rational numbers (e.g., C_{min} = -10, and g_{min} = 10), but they are expressed by default in the form of logarithm with the base number of two (e.g., $C_{min} = 2^{-10}$, $g_{min} = 2^{10}$) in the process of program execution, so the parameter ranges should be properly selected. Finally, the user can click on the “Run” button to view the combination of optimal parameters (*bestC* / *bestg* / *bestd*) and the corresponding best accuracy. To verify the SVM performance on the test set, it is necessary to click “Analysis > SVM > TestSVM,” select the kernel type, and then click the “View confusionmat” button to view the classification results of each kernel on the test set, as well as performance evaluation indicators, including sensitivity, specificity, accuracy, F1-score, G-score, Macro-F1 score, and the overall accuracy.



Setting

Data select

```
PLSDAclassification  [1x1]    struct
PLSDAstruct  [1x1]    struct
PLSDAsumsqerr  [2x21]    double
PLSDAtrainstruct  [1x1]    struct
data  [945x300]    double
group  [300x1]    double
linearSVMtest  [1x1]    struct
linearSVMtrain  [1x1]    struct
mean_centerdata  [300x945]    double
pcaexplained  [3x1]    double
pcaloading  [945x3]    double
pcascore  [300x3]    double
wavenumber  [945x1]    double
```

● Update ● Select

Dataset Split Methods: Randomly Select

Training set ratio: 0.8

Cross validation type: 10 fold

Parameters

Cmin	-8
Cstep	1
Cmax	8
gmin	-8
gstep	1
gmax	8
dmin	1
dstep	1
dmax	9
Kernel	polynomial

bestC	0.0039063
bestg	None
bestd	1
bestaccuracy	99.375%

Testing

kernel: polynomial

Overall Accuracy 93.3333%

Macro-F1 Score 0.93266

Testing result (confusion matrix)

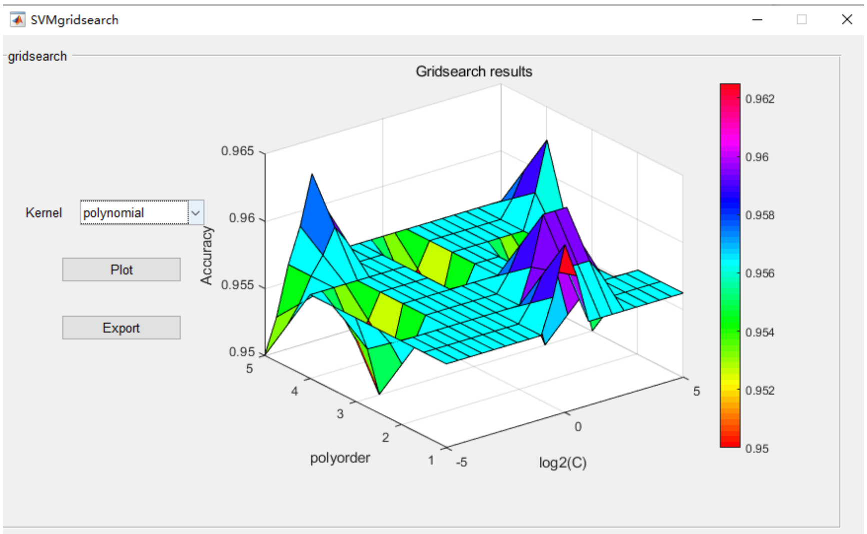
	class1	class2	class3
class1	20	0	0
class2	0	20	0
class3	4	0	16

Performance evaluation

	Sensitivity	Specificity	Accuracy	F1-Score	G-Score
class1	1	0.9000	0.9333	0.9091	0.9129
class2	1	1	1	1	1
class3	0.8000	1	0.9333	0.8889	0.8944

Figure 17. Training and testing interfaces of SVM and PCA-SVM models.

The influence of each parameter on the accuracy of the SVM model when a combination of optimal grid search parameters is used can be depicted by a 2D graph or a 3D surface. To perform this function, it is needed to click “View > viewSVM > SVMgridsearch,” select the kernel type, click on the “Plot” button to draw the diagram, and then click on the “Export” button to export the graph. To make the 3D surface more uniform, parameters *C* and *g* are expressed in the exponential form, i.e., $\log_2 C$ and $\log_2 g$, respectively.

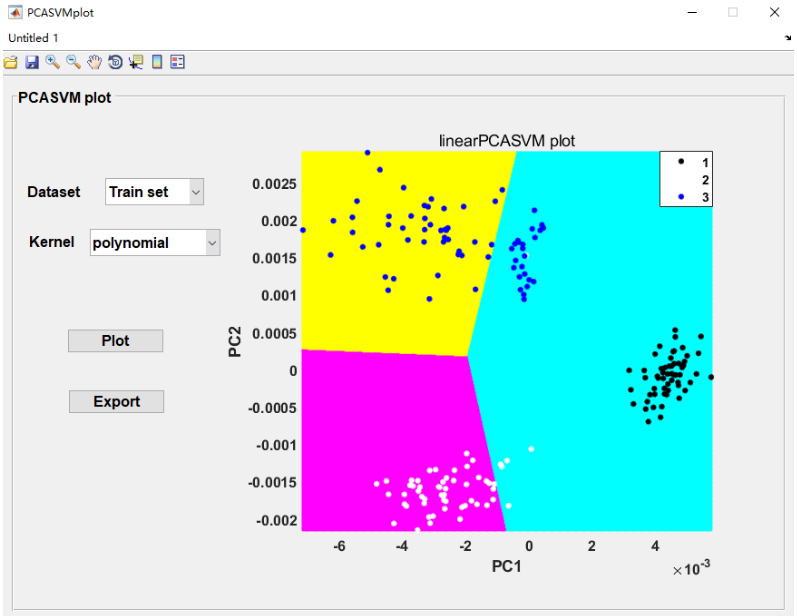


Kernel polynomial

Figure 18. The 3D surface plot of classification accuracy as the function of the parameter C and polynomial order d in the polynomial kernel.

3.4.5 PCA-SVM

The implementation process of the PCA-SVM is basically the same as that of the SVM, with the exception of the form of input data. In the PCA-SVM, the data to be selected in the list box is "pcascore," and only the first two significant PCs are used for the PCA-SVM model. To view the visualized classification result, the user can click "View > viewPCASVM > PCASVMplot," select the dataset and kernel type, and then click on the "Plot" button to draw a diagram. Finally, by clicking on the "Export" button, the graph is exported.



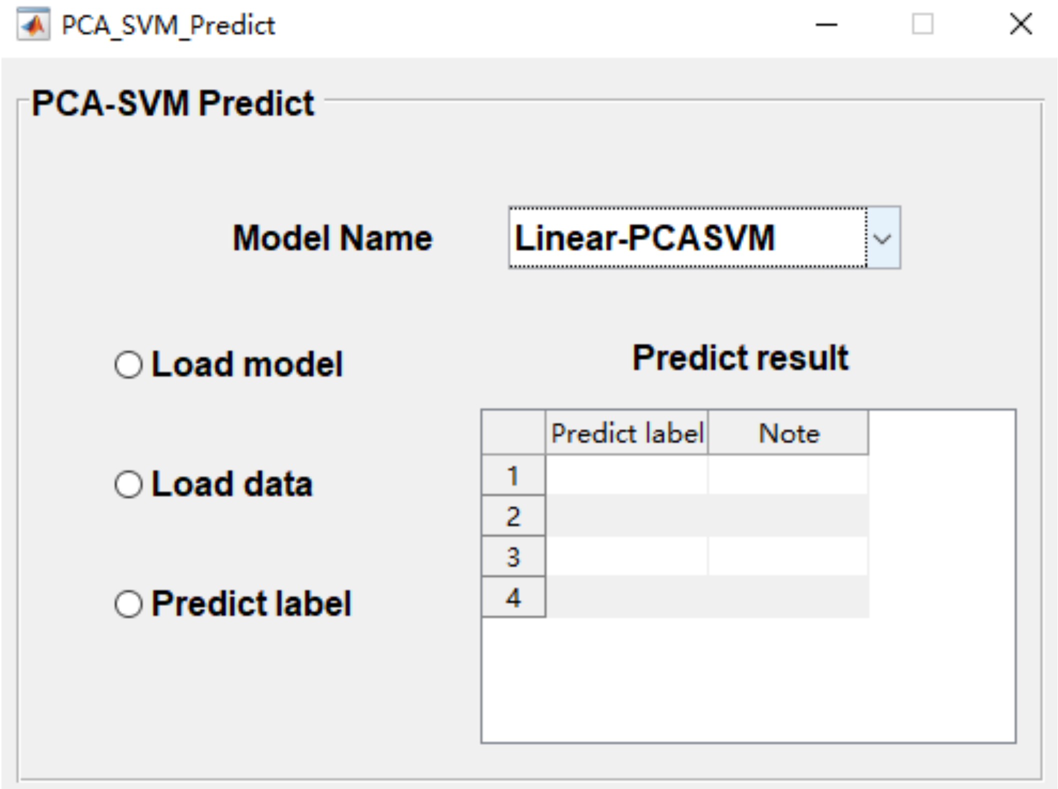
Dataset Train set

Kernel polynomial

Figure 19. The polynomial kernel PCA-SVM classification scatter plots.

4. Model saving and loading

After each model training, the user can save all the information (e.g., training parameters, training methods, and other parameters) on the trained model by clicking "Save model" in the "File" in the main menu. The user selects the name of the model to be saved in the pop-up menu of the "Save Model" interface, and clicking on the "Save" button to save the model to the specified location. After the information on the trained model is saved, there is no need to perform data preprocessing and establish classifiers again when the model is used to predict a new dataset. When the user wants to predict an unknown spectral dataset, he should click "Predict" in the main menu, select the required model, and then click "Load model" and "Load data" in the subinterface to load the model and new dataset, respectively. By clicking on the "Predict label" button in the subinterface, the new dataset is predicted, and prediction results are displayed in the table.



Model Name: Linear-PCASVM

	Predict label	Note
1		
2		
3		
4		

Figure 20. The saved model's prediction interface for unknown data. The example is the PCA-SVM model.

5. References

1. Wang X, Tang X. Dual-space linear discriminant analysis for face recognition. IEEE Computer Society Conference on Computer Vision & Pattern Recognition. IEEE, 2004.
2. Ballabio D, Consonni V. Classification tools in chemistry. Part 1: linear models. PLS-DA. Analytical Methods. 2013, 5(16):3790-3798.
3. Néstor F. Pérez, Joan Ferré, Ricard Boqué. Calculation of the reliability of classification in discriminant partial least-squares binary classification. Chemometrics & Intelligent Laboratory Systems. 2009, 95(2):122-128.
4. De Almeida M R, Correa D N, Rocha W F C, et al. Discrimination between authentic and counterfeit banknotes using Raman spectroscopy and PLS-DA with uncertainty estimation. Microchemical Journal. 2013, 109:170-177.
5. Chih-Chung Chang, Chih-Jen Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011.
6. Cortes C, Vapnik V N. Support-Vector Networks. Machine Learning. 1995, 20(3):273-297.
7. K. H. Liland, A. Kohler and N. K. Afseth. Model-based pre-processing in Raman spectroscopy of biological samples. Journal of Raman Spectroscopy. 2016, 47: 643-650.