Nearest clusters based partial least squares discriminant analysis for the classification of spectral data

Weiran Song[a], Hui Wang[a], Paul Maguire[b], Omar Nibouche[a]

[a]*School of Computing and Mathematics,* [b]*School of Engineering, Ulster University, BT37 0QB, Newtownabbey, Co. Antrim, UK*

ABSTRACT

Partial Least Squares Discriminant Analysis (PLS-DA) is one of the most effective multivariate analysis methods for spectral data analysis, which extracts latent variables and uses them to predict responses. In particular, it is an effective method for handling high-dimensional and collinear spectral data. However, PLS-DA does not explicitly address data multimodality, i.e., within-class multimodal distribution of data. In this paper, we present a novel method termed *nearest clusters based PLS-DA* (NCPLS-DA) for addressing the multimodality and nonlinearity issues explicitly and improving the performance of PLS-DA on spectral data classification. The new method applies hierarchical clustering to divide samples into clusters and calculates the corresponding centre of every cluster. For a given query point, only clusters whose centres are nearest to such a query point are used for PLS-DA. Such a method can provide a simple and effective tool for separating multimodal and nonlinear classes into clusters which are locally linear and unimodal. Experimental results on 17 datasets, including 12 UCI and 5 spectral datasets, show that NCPLS-DA can outperform 4 baseline methods, namely, PLS-DA, kernel PLS-DA, local PLS-DA and k-NN, achieving the highest classification accuracy most of the time.

Keywords: Partial Least Squares, Clustering, Nonlinearity, Multimodality, Spectral pattern recognition.

1. Introduction

Spectral data analysis is used in many areas of science and engineering as a mean of exploring the constituents of matter. Absorption spectroscopy is a type of spectral analysis usually used to identify and possibly quantify particular substances in a sample. Infrared (IR) spectroscopy and ultraviolet-visible (UV-Vis) spectroscopy are two specific examples. Measurements of radiation intensity at a series of fixed wavelengths result in a spectrum that consists of a series of discrete peaks. Other types include transmission, reflectance or emission spectroscopies which, along with mass spectroscopy, can provide useful information on the chemical constituents of substance. Spectral data contains a molecular fingerprint of the substance of interest, which can be used to identify and/or quantify the substance. The interpretation of the fingerprint depends critically on instrumental factors. Recently, there has been an ongoing drive towards miniaturisation and field portability of such instrumentation and the development of low cost spectral-based sensors. This may significantly degrade the fingerprint data quality, thus making accurate identification problematic; in portable applications, the nature of the sample and its environment may add to the challenge. Therefore, the use of pattern recognition techniques in spectral data analysis is becoming common practice. However, the challenges in robust extraction of meaningful fingerprint data from noisy field data cannot be underestimated. These include, for example, high dimensionality [1], collinearity [2], nonlinearity [3] and a special type of nonlinearity - multimodality [4, 5].

Among the common methods for spectral data analysis are Principal Component Analysis (PCA), Support Vector Machine (SVM) and Partial Least Squares (PLS). PLS is currently the de-facto standard [6]. It is a statistical regression method that combines the features of Canonical Correlation Analysis (CCA) and Multiple Linear Regression (MLR) to predict responses based on independent variables. It searches for linear combinations of independent variables, namely *latent variables* (LV), that maximize the covariance between the latent variable and the response. PLS has proven to be a very useful method for spectral data analysis. It efficiently handles the high dimensionality and collinearity problems that widely exist in spectral data [7, 8] by stably estimating regression coefficients from low-dimensional latent variables. However, it has been reported that the PLS algorithm will degrade in performance under nonlinear conditions [9, 10, 11], which is often present in spectral data for various reasons [3] and can be identified by a quantitative numerical tool (e.g. run test) with augmented partial residual plots (APaRP) [12, 13]. Recent attempts to modify the PLS algorithm to handle nonlinear data have focused typically on two approaches. The first is kernel approaches which transforms the original input data into a feature space by nonlinear mapping, and then constructs a linear PLS model in the feature space [14]. The second approach is to combine locally weighted regression (LWR) [15] and PLS, namely, locally weighted PLS (LW-PLS) [16]. On one hand, this approach fills the gap that LWR cannot be used to handle the problem of ill-

conditioned matrices, such as small sample size and collinearity, unless a robust variable selection is implemented. On the other hand, LW-PLS constructs a local model to enlarge the contribution of neighbouring data for a given query. As a result, the global nonlinearity can be lessened.

A particular type of nonlinearity is multimodality where the data distribution within a class is multimodal possibly due to the fact that data within a class comes from different sources or different data collection sessions [11, 17, 18]. For example, if we want to identify apples from other types of fruit, the apple as a class is very likely to have multiple modes in the data distribution each corresponding to a variety, since apple varieties may be quite distinct. Further still, differences within the same variety of apple from different regions may also lead to multiple modes in the data distribution. In general, it is possible that data instances within a class are more similar to data instances of a different class than to other members of its own class. Multimodality has been studied in pattern recognition; it has been shown [19, 20] that modelling multimodality explicitly can significantly improve classification performance. However, multimodality has not been explicitly addressed in PLS although it has been implicitly addressed in variants of PLS. Kernel PLS-DA has been studied for analysing nonlinear chemometric data and it has been shown [21, 22] to have a classification performance comparable on average to kernel Support Vector Machine (KSVM) for nonlinear chemometric data. Kernel PLS-DA uses the similarity between two data vectors as the basis to map the original data into feature space. It is not directly possibly to see the contribution of each variable with respect to the final prediction as well as to interpret the obtained kernel PLS model [14, 23]. Moreover, kernel PLS-DA selects more LVs than PLS-DA on the classification of spectral data [4]. The locally weighted PLS-DA (LW-PLS-DA) has been studied for analysing nonlinear spectral data [4, 24]. LW-PLS-DA can outperform standard and kernel PLS-DA [4], also the resulting model does not require all training samples to be involved in. Common weighting matrices of LW-PLS are based on the Euclidean distance or the Mahalanobis distance, which perform less well than covariance or sparse regression coefficient for many industrial processes [25, 26, 27]. Moreover, this instance-based learning approach produces multiple models for a set of queries which results in a sharp increase in computational complexity compared to the classical PLS.

This paper presents an extension of PLS-DA that explicitly addresses data nonlinearity and multimodal distributions, namely NCPLS-DA. By using hierarchical clustering, all training samples are grouped into clusters in which the clustering centres are calculated. Clusters that contain the nearest centres towards a given query are selected for PLS modelling. This strategy handles nonlinearity and multimodality by constructing linearly separable models in neighbourhoods. Thus, more accurate results can be expected. This PLS extension has been

tested on a wide range of datasets including twelve UCI datasets of different data types and five spectral datasets (some being simulated with multimodality and nonlinearity, and some being publicly available).

The remainder of the paper is organized as follows: Section 2 briefly reviews PLS-DA and hierarchical clustering. The proposed method, NCPLS-DA, is presented in Section 3. Section 4 presents the experiments on UCI and spectral datasets, including datasets description, parameter settings, results and discussion. Conclusions are drawn in Section 5.

2. Related Work

The standard chemometrics notation is used in this paper. Capital and lowercase letters in boldface denote matrix and vector, respectively. Table 1 lists the symbols used in this paper.

Table 1

Symbols used in this paper along with their meanings.

$\boldsymbol{X}$	Matrix of independent variables
$\boldsymbol{Y}$	Matrix of response or dummy matrix of class labels
$\boldsymbol{w}$	Weight vector of $\boldsymbol{X}$
$\boldsymbol{c}$	Weight vector of $\boldsymbol{Y}$
$\boldsymbol{t}$	Score vector of $\boldsymbol{X}$
$\boldsymbol{T}$	Score matrix of $\boldsymbol{X}$, where the columns are $\boldsymbol{t}$
$\boldsymbol{P}$	Loading matrix of $\boldsymbol{X}$
$\boldsymbol{B}_{\mathrm{PLS}}$	PLS regression coefficients
$\boldsymbol{x}_q$	Query point
$\hat{\boldsymbol{y}}_q$	Prediction of query point
n	The number of samples in $\boldsymbol{X}$
d	The number of variables in $\boldsymbol{X}$
σ	Gaussian width of the kernel function
LV	The number of latent variables
CN	Clustering numbers
NC	The number of nearest clusters towards a query

2.1 PLS-DA

PLS is a classical method in multivariate analysis that maximizes the covariance between the latent variables and the responses. It is today most widely used in chemometrics including spectral data analysis. There exist

different PLS algorithms, including Nonlinear Iterative Partial Least Squares (NIPALS) [28] and SIMPLS [29], which are different in computational complexity and numerical stability. The computation time is dependent mainly on the dimensionality of data and also the number of latent variables selected, to a lesser extent [30]. The numerical stability is dependent on the numerical calculation methods used and is a factor of model precision. Theoretically, all PLS algorithms should yield the same models but in practice there are differences due to the numerical calculation methods [30]. In this paper, the SIMPLS algorithm is used, because it is faster than NIPALS and it is stable when the number of latent variables is not high [30, 31].

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are mean-centred, the SIMPLS algorithm is to find a linear combination of $\boldsymbol{X}$, $\boldsymbol{t} = \boldsymbol{Xw}$, that maximizes data covariance as follows,

$$\max \boldsymbol{w}^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{Yc}, \tag{1}$$

$$\text{s.t. } \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{c} = 1.$$

It yields the following eigen problem:

$$\boldsymbol{w}^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{Yc} = \lambda \boldsymbol{w}, \tag{2}$$

which follows a unified framework for latent modelling [32, 33]. For a given number of latent variables (LV), the singular value decomposition (SVD) of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ and its successive deflation are calculated.

SIMPLS can solve univariate responses (PLS1) as well as multivariate responses (PLS2) problems. When responses contain multi-categorical information, the PLS2 regression becomes a discriminant analysis method which can be used to classify data with high-dimensionality and collinearity. This is accomplished by transforming the categorical responses into numerical responses using dummy matrix coding [34]. One common criterion for deciding which class a sample belongs to is the following: a sample is classified by whichever class that has the maximum value in the *Y*-matrix.

2.2 Kernel PLS-DA

Kernel PLS-DA is an extension of PLS-DA that proceeds by mapping the original data $\boldsymbol{X}$ into a feature space $\boldsymbol{F}$ (ϕ: $\boldsymbol{R}^d \rightarrow \boldsymbol{F}$):

$$\boldsymbol{K}_{ij} = k\,(\boldsymbol{x_i}, \boldsymbol{x_j}) = (\phi(\boldsymbol{x}_i),\ \phi(\boldsymbol{x}_j)) = \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j) \tag{3}$$

and then building a PLS-DA model in the feature space for prediction. The mapping procedure is performed by kernel function which calculates the similarity between two sample vectors. In this paper, we use the Gaussian kernel due to its efficiency. A Gaussian kernel is defined as:

$$\boldsymbol{K}_{ij} = e^{\frac{-||\boldsymbol{x_i}-\boldsymbol{x_j}||^2}{2\sigma^2}}. \tag{4}$$

2.3 Local PLS-DA

LW-PLS-DA is an extension of LW-PLS for classification [4]. In this method, a diagonal weighting matrix is applied on a subset of neighbouring data that contains the nearest neighbors of a query. This weighting matrix gives more weight on the neighbouring data than others with respect to the query and can be based on different distance schemes such as uniform, quadratic and Gaussian. A PLS-DA model is then built on the weighted neighbouring data and used to classify the query.

Local PLS-DA is a typical form of LW-PLS-DA in which the distance scheme is uniform, i.e. all elements in the diagonal weighting matrix are equals to 1. Local PLS has been shown to perform very well on many UCI datasets [24] and outperforms other weighting matrix of LW-PLS-DA on the classification of chemometric data [4].

2.4 Hierarchical clustering

Hierarchical clustering is one common approach to clustering in pattern recognition. It builds a hierarchy of clusters based on a greedy strategy by merging clusters (agglomerative) or splitting samples (divisive). This paper adopts an agglomerative strategy which merges clusters in a bottom-up way. Initially, every sample is an individual cluster in the bottom of the tree. Once the Euclidean distance between every pair of clusters has been computed, two nearest clusters are merged to create a new cluster. This procedure continues until the tree is complete, resulting in a treelike *dendrogram*. The dendrogram returned by hierarchical clustering is informative and potentially useful so hierarchical clustering has been widely applied in chemometrics [35, 36, 37]

3. Proposed Method

A nearest clusters based strategy is introduced into PLS-DA in order to address multimodality and nonlinearity of multivariate data such as spectral data. This strategy globally divides samples into clusters and then calculates the centre of each cluster. For a query sample, certain numbers of nearest cluster centres are selected and a PLS-DA model is built for all of the samples in the selected clusters.

Here, we use some artificial, multimodal and nonlinear examples to illustrate this method. In a multimodal case, samples of the first two principal components are randomly generated and distributed in three ellipse-like areas which have the same orientation (see Fig. 1a). The central ellipse contains samples of class 1 while the outer ellipses include samples labelled as class 2. Despite the fact the three ellipses are linearly separable; a query (black cross) which belongs to class 1 is however wrongly attributed by PLS-DA to class 2. To help PLS-DA achieve correct classification for this type of query, we divide all samples into five clusters by hierarchical clustering (see Fig. 1b) and calculate the mean of each cluster as clustering centre. By comparing the distances between the query and the five clustering centres, two nearest clusters are selected for local modelling which produce more accurate predictions than global PLS-DA (Fig. 1c).

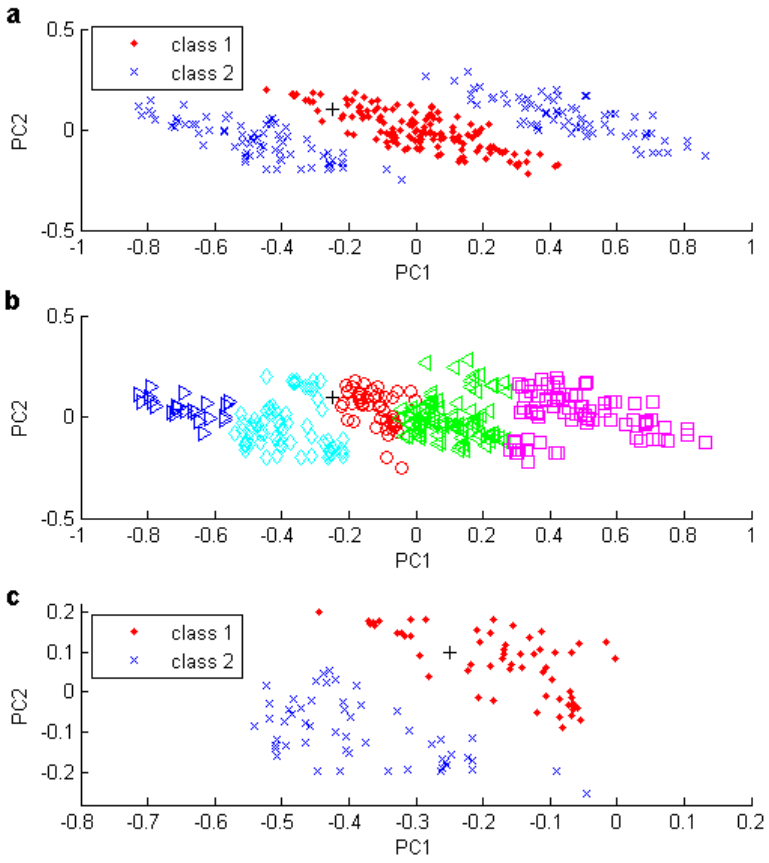

Fig. 1. (a) Projection of multimodal samples onto the space spanned by the two significant principal components; (b) hierarchical clustering of multimodal samples; (c) two nearest clusters of samples selected for modelling.

In a nonlinear example, samples of two classes are distributed in inner and outer rings, respectively of the first two principal components (see Fig. 2a). A query (black cross) located in outer ring is misclassified by PLS-DA. As in the multimodal case, if we first separate samples to five clusters (see Fig. 2b) and model with only two clusters which contain the clustering centres nearest to the query, an approximate linear cut will ensure the correct prediction of PLS-DA (Fig.2c).

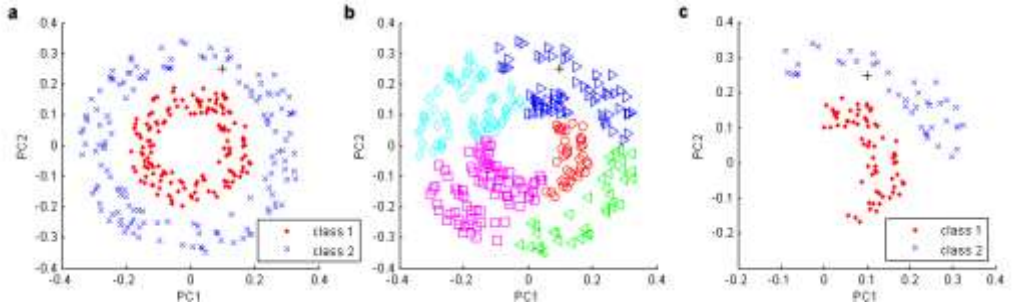

Fig. 2. (a) Projection of nonlinear samples onto the space spanned by the two significant principal components; (b) hierarchical clustering of nonlinear samples; (c) two nearest clusters of samples selected for modelling.

The proposed NCPLS-DA firstly separates training samples into CN clusters by hierarchical clustering. The clustering centre is estimated by calculating the mean of the cluster as:

$$\boldsymbol{\mu} = \frac{1}{c}\sum_{i=1}^{c} \boldsymbol{x}_{(i)} \tag{5}$$

where c is the number of samples in the cluster and $\boldsymbol{x}_{(i)}$ is the i-th sample of the cluster. The distances between a given query and each clustering centre are calculated and then sorted in order to find the nearest clusters being used for PLS modelling. If the labels of these clusters are the same, the query will be directly attributed to this single class. Otherwise, the PLS regression coefficient is calculated based on the nearest clusters of samples and further applied to predict the query. The NCPLS-DA algorithm is summarized as the following steps:

a. Cluster $\boldsymbol{X}$ into CN clusters, $\boldsymbol{X}_1$, $\boldsymbol{X}_2$,…, $\boldsymbol{X}_{\mathrm{CN}}$ with their corresponding $\boldsymbol{Y}$ values collected as $\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$,…, $\boldsymbol{Y}_{\mathrm{CN}}$.

b. Calculate each clustering centre $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$,…, $\boldsymbol{\mu}_{\mathrm{CN}}$ by (5)

c. Calculate the distance between $\boldsymbol{x}_{\mathrm{q}}$ and each $\boldsymbol{\mu}$ to find nearest centres.

d. Reconstruct local samples and corresponding labels: $\boldsymbol{X}^* = [\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}, \ldots, \boldsymbol{X}_{(NC)}]$, $\boldsymbol{Y}^* = [\boldsymbol{Y}_{(1)}, \boldsymbol{Y}_{(2)}, \ldots, \boldsymbol{Y}_{(NC)}]$.

e. Prediction:

e.1 If $\boldsymbol{Y}^*$ is an individual class, $\hat{\boldsymbol{y}}_q$ belongs to this class.

e.2 Otherwise, PLS modelling: $[\boldsymbol{T}, \boldsymbol{P}, \boldsymbol{B}_{PLS}]$ = SIMPLS ($\boldsymbol{X}^*$, $\boldsymbol{Y}^*$, LV),

then $\hat{\boldsymbol{y}}_q = \boldsymbol{x}_q \cdot \boldsymbol{B}_{PLS}$.

Two more parameters are introduced in NCPLS-DA, namely, clustering number (CN) – the number of clusters, and nearest cluster (NC) – the number of nearest clusters. CN decides to what extent the training samples are clustered. Too few clusters will result in some samples not contributing to robust classification, while too many clusters will unnecessarily divide analogous samples and ineffectively add computational cost. To find the optimal CN, we apply the *separability criterion* [20] to evaluate the performance of clustering within a range of CN. Separability criterion calculates the average Euclidean distance (AED) between the mean of samples and the means of CN clusters as:

$$AED_{\mathrm{CN}} = \frac{1}{\mathrm{CN}} \sum_{i=1}^{\mathrm{CN}} ||\bar{\boldsymbol{X}} - \boldsymbol{\mu}_i||_2. \tag{6}$$

The larger the AED value the more separated the clusters are from each other. Therefore, the maximum value of AED is used to determine the optimal CN. As the Euclidean distance tends to decrease its performance when the dimensionality increases [38], the original data is kernel-transformed before using separability criterion in this paper.

The other parameter NC determines how many nearest clusters are used to predict a query. As with CN, too few nearest clusters are less distinctive in prediction while too many nearest clusters will make NCPLS-DA tend to global PLS-DA. Given different structures and distributions of various datasets, the NC is empirically set to the nearest integer of CN/3 initially and then validated in a proper range. It is noted that this way of setting parameters may not achieve the best results in all cases but can outperform PLS-DA in almost all the time. More details about the parameter settings of NCPLS-DA will be demonstrated in the experiments section.

4. Experiments

We conducted extensive experimentation to compare k-NN, PLS-DA, kernel PLS-DA, local PLS-DA and NCPLS-DA in terms of classification on seventeen datasets -- twelve UCI and five spectral datasets. UCI is a machine learning data repository hosted at University of California, Irvine (UCI) [39]. These UCI datasets cover a wide range of data generated in the scientific community and they are selected to demonstrate the diversity of the proposed method. Further, we created simulated spectral data that are multimodal and nonlinear, as discussed in the last section, to test the performance of NCPLS-DA. Additionally, proprietary data from infrared spectral measurements on apple, olive oil and fruit puree are used to demonstrate the effectiveness of the proposed method in the prime application of PLS.

4.1 UCI datasets description

The selected twelve UCI datasets are high-dimensional, multiclass and imbalanced. Some are related to analytical chemistry, including Arcene & Forest Types (spectroscopy) and QSAR-biodeg (chemometrices). The Arcene dataset originally contains thousands of irrelevant variables and feature selection prior to classification is suggested [40]. Therefore, ReliefF algorithm [41] is applied only on this dataset to remove 9/10 of the variables. The basic information about 12 UCI datasets and the optimal parameters are shown in Table 2.

Table 2

Information on 12 UCI datasets and the optimal parameters of k-NN, PLS-DA, kernel PLS-DA, local PLS-DA and NCPLS-DA.

Dataset	Samples	Variables	Categories	k-NN	PLS-DA	KPLS-DA		LPLS-DA		NCPLS-DA		
				NN	LV	LV	log(σ)	LV	NN	LV	CN	NC
Arcene	200	1000	2	1	8	5	3	8	50	9	18	6
Breast Tissue	106	9	4	3	9	10	5	4	39	3	50	16
Ecoli	336	7	8	7	5	10	0	6	98	3	40	13
Forest Types	523	27	4	3	9	10	2	4	52	9	37	10
Glass	214	9	6	1	9	10	0	4	31	7	30	13
Ionosphere	351	34	2	1	5	10	0	2	75	3	149	31
Iris	150	4	3	9	3	5	1	2	41	2	31	10
Parkinsons	195	22	2	7	6	10	1	8	53	2	23	5
QSAR-biodeg	1055	41	2	3	8	10	2	3	65	10	37	9
Sonar	208	60	2	1	6	10	0	4	70	4	46	15
SPECTF heart	267	44	2	35	5	5	2	2	86	1	44	11
Zoo	101	16	7	1	9	9	0	6	57	8	32	10

4.2 Parameter settings and results on UCI datasets

The optimal parameter(s) of each algorithm on every dataset, as shown in Table 2, are set by 10-fold cross-validation using the metric of average classification accuracy. The number of nearest neighbours (NN) in k-NN is set from 1 to 50. A grid search approach is used in kernel PLS-DA (LV*σ), local PLS-DA (LV*NN) and NCPLS-DA (LV*NC). To prevent overfitting, the range of LVs is selected from 1 to 10 if the minimum number between n and d is above 10, otherwise, from 1 to min (n, d). The width of the radical basis functions σ in kernel PLS-DA is adjusting from 10^{-3} to 10^{5} on a logarithmic scale, while the nearest neighbours in local PLS-DA is varying from 25 to 100 or 50 to 100 depending on the sample scale. In NCPLS-DA, two extra parameters need to be determined compared to PLS-DA: the clustering numbers (CN) and the nearest clusters (NC). For CN, separability criterion is used to select the optimal CN within a range, for example, from 10 to 50 for UCI datasets (except Ionosphere) shown in Fig. 3a and Fig. 3b. From the figures, the optimal CN for Arcene and Parkinsons datasets are 18 and 23, respectively. Another parameter NC is validated within an empirical range [r-4, r+4] in most cases, where r is the nearest integer of CN/3. Next, we demonstrate the grid search of the optimal number of LV and NC, which is depicted in Fig. 4a and Fig. 4c as a mesh plot. As it can be seen in both datasets, the average accuracy increases smoothly with increasing number of LVs until a saturation level is reached. By tracing the position of NC where the highest average accuracy achieves, for example 6 and 5, respectively for Arcene and Parkinsons dataset, the optimal parameters can be easily identified.

For a fair comparison between algorithms, we use the same training and test sets for all algorithms. To see the performance of all PLS-based algorithms graphically, we chart in Fig. 4b and Fig 4d the average 10-fold cross validation accuracy over the LV parameter, fixing other parameters to their optimal values, on two datasets Arcene and Parkinsons. It is clear that the average accuracy of NCPLS-DA is usually above that of PLS-DA for each LV and gradually exceeds that of kernel PLS-DA after the specific number of LVs. NCPLS-DA can also outperform local PLS-DA in most cases, achieving the highest average accuracy.

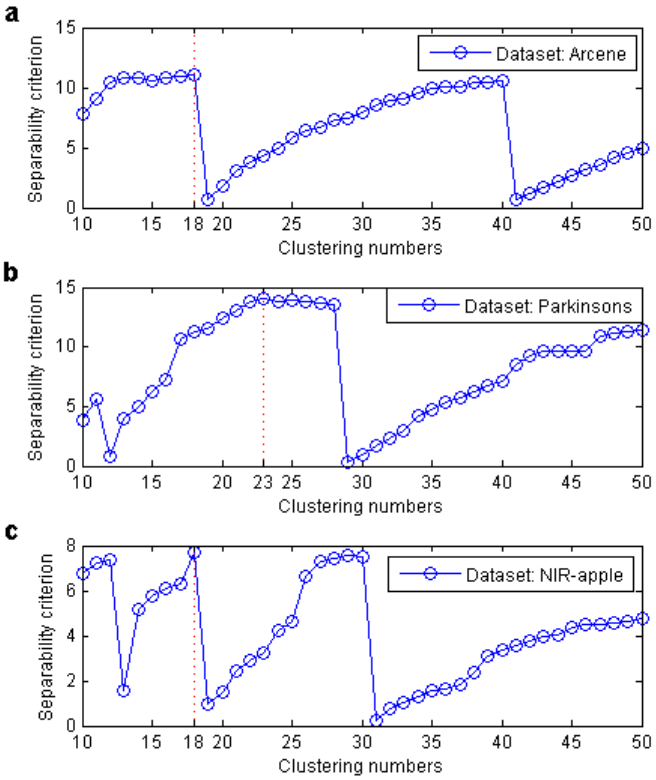

Fig. 3. Using separability criterion to determine the optimal clustering numbers of 3 datasets: (a) Arcene; (b) Parkinsons; (c) NIR-apple.

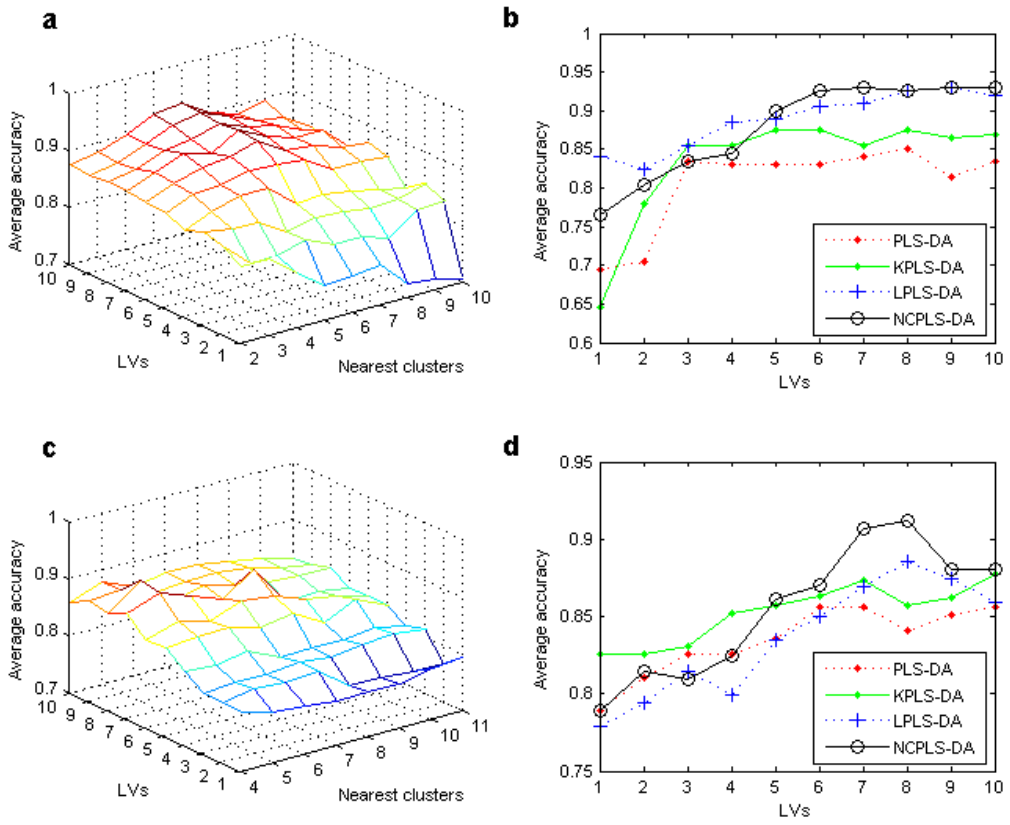

Fig. 4. Selection of the optimal number of latent variables and nearest clusters in NCPLS-DA and the average performance of 4 PLS-based algorithms on Arcene (a & b) and Parkinsons (c & d) datasets.

Using the optimal parameter(s) of the algorithms, we further repeat the 10-fold cross-validation for 10 times to obtain the average classification accuracy and its standard deviation, as shown in Table 3. Among the five algorithms, the proposed NCPLS-DA yields the best results in seven out of the 12 datasets. The last row shows the classification accuracies averaged over all datasets. This row indicates that NCPLS-DA performs better on average than four baseline algorithms, specifically, NCPLS-DA drastically outperforms PLS-DA by 5.96%. Furthermore, the optimal number of LVs in NCPLS-DA and local PLS-DA is usually smaller than that of PLS-DA (Table 2).

Table 3

Average Classification Accuracy (%) and Standard Deviation (%) of the Different Algorithms

Dataset	k-NN	PLS-DA	KPLS-DA	LPLS-DA	NCPLS-DA
Arcene	87.25 (0.86)	84.65 (0.97)	86.75 (1.18)	90.00 (1.35)	**91.95 (1.09)**
Breast Tissue	82.55 (1.66)	84.62 (0.44)	85.61 (1.86)	87.85 (1.33)	**89.60 (1.54)**
Ecoli	87.32 (0.66)	84.52 (0.97)	**87.97 (0.41)**	86.51 (0.47)	86.83(0.68)
Forest Types	88.97 (0.35)	86.07 (0.32)	88.97 (0.27)	90.14 (0.53)	**90.56 (0.46)**

Glass	**73.13 (1.05)**	59.19 (2.01)	66.68 (1.10)	71.80 (1.47)	68.67(1.16)
Ionosphere	86.43 (0.71)	86.77 (0.24)	93.24 (0.44)	94.04 (0.65)	**95.41 (0.51)**
Iris	96.60 (0.38)	82.40 (0.47)	94.80 (0.61)	**98.33 (0.35)**	98.20 (0.45)
Parkinsons	83.86 (0.73)	85.66 (0.70)	86.36 (0.84)	**89.00 (1.28)**	88.14 (1.07)
QSAR-biodeg	82.21 (0.49)	83.33 (0.37)	82.78 (0.44)	**85.57 (0.41)**	85.05 (0.39)
Sonar	82.12 (0.77)	76.33 (1.42)	83.49 (1.60)	82.63 (0.89)	**84.12 (1.16)**
SPECTF heart	79.71 (0.59)	78.28 (0.43)	80.91 (0.74)	81.13 (0.59)	**81.26 (0.61)**
Zoo	97.52 (0.68)	93.93 (0.88)	96.43 (1.30)	95.42 (0.83)	**97.53 (0.70)**
Average	85.64	82.15	86.17	87.70	**88.11**

Standard deviations are presented in parentheses.

4.3 Spectral datasets description

- '*Multimodal'*: The simulated multimodal dataset contains 300 samples belonging to two classes. In the space spanned by the first two principal components, the samples (a 300 by 2 score matrix) are distributed in three ellipses which are linearly separable along a single orientation (Fig. 1a). The first class are located in the central ellipse, and the second class are distributed in the two outer ellipses. A loading matrix (2 by 200) was simulated as orthogonal combination of two or three Gaussian peaks (see Fig. 5a). The score and loading matrices are multiplied to construct a simulated data matrix. This matrix is further processed by inverting zero-mean normalization: the data matrix is multiplied by a sequence of randomly generated values (<0.05) as standard deviations for every 'wavelengths' and added by a mean spectrum to generate one sample vector as a spectrum. Finally, 40 dB white Gaussian noises were added to each spectrum. The spectral-like representation of the multimodal dataset is shown in Fig.5b.
- '*Nonlinear*': The nonlinear dataset contains 300 samples divided into two classes. Data are distributed in co-centric rings in the space of the first two principal components, as shown in Fig.2a. The loading matrix (2 by 200) was built by orthogonalization of linear combinations of two or three Gaussian peaks (see Fig. 5c). After the multiplication of the score and loading matrices, the obtained data matrix is processed by inverting zero-mean normalization and adding Gaussian noise (40 dB). The spectral-like representation of the nonlinear dataset is shown in Fig. 5d.

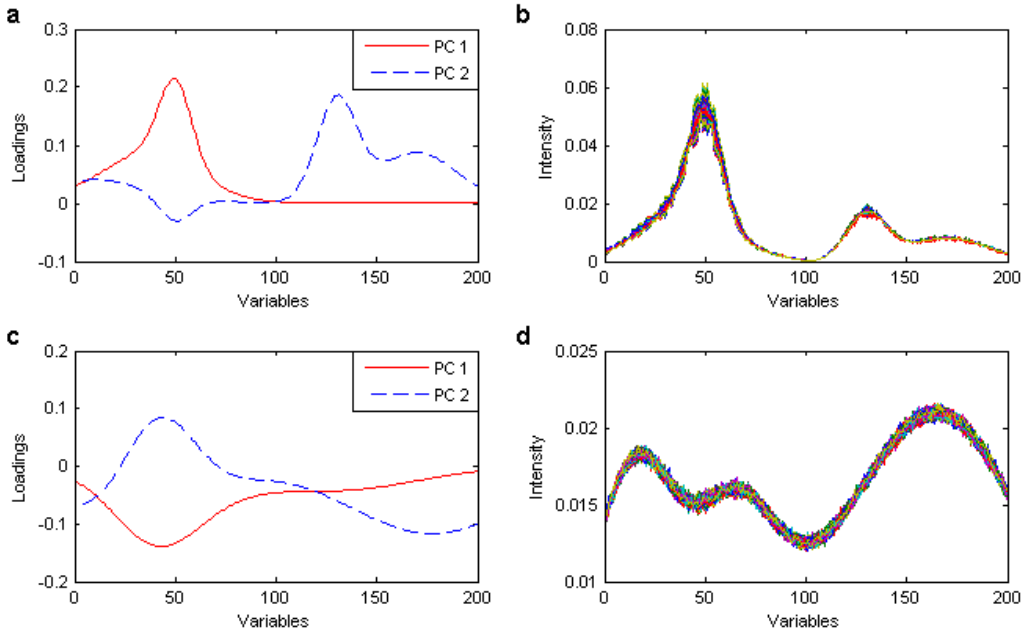

Fig.5. (a) Variable loadings for the two principal components of simulated 'multimodal' data; (b) spectral-like representation of 'multimodal' data; (c) variable loadings for the two principal components of simulated 'nonlinear' data; (d) spectral-like representation of 'nonlinear' data.

- *NIR-apple* [42]: Fresh apples were scanned in reflectance mode using a portable NIR spectrometer (NIRQuest512 spectrometer, Ocean Optics, Inc., United States) equipped with an InGaAs detector and having a wavelength range of 901.06-1721.242 nm with a 1.65 nm interval. The experiments were conducted under ambient light conditions with added illumination using a 45° diffuse reflectance probe (DR-Probe) with integrated tungsten halogen light source. NIR spectra, each with a dimensionality of 512, were collected with the OceanView software. This dataset contains 182 apples of two species (Gala and Pink Lady). The task is to differentiate organic (86 samples) and non-organic (96 samples) apples.
- *FTIR-olive oil* [43]: 120 Mid-infrared spectra (including duplicates), collected from 60 different authenticated extra virgin olive oils are used to distinguish the country of their origins: Greece, Italy, Portugal and Spain (respectively 20, 34, 16 and 50 samples of each). The wavelength of the spectra ranges from 798.892 to 1896.8085 nm with an interval of 1.9295 nm.
- *FTIR-fruit puree* [44]: using Fourier transform infrared (FTIR) spectroscopy with attenuated total reflectance (ATR) sampling, 983 spectra are collected in two classes: 'strawberry' and 'non-strawberry' purees, respectively 351 and 632 of each class. The wavelength of the spectra ranges from 899.327 to 1802.564 nm with an interval of 3.86 nm.

The raw spectra of the aforementioned three datasets are shown in Fig. 6a, c and e.

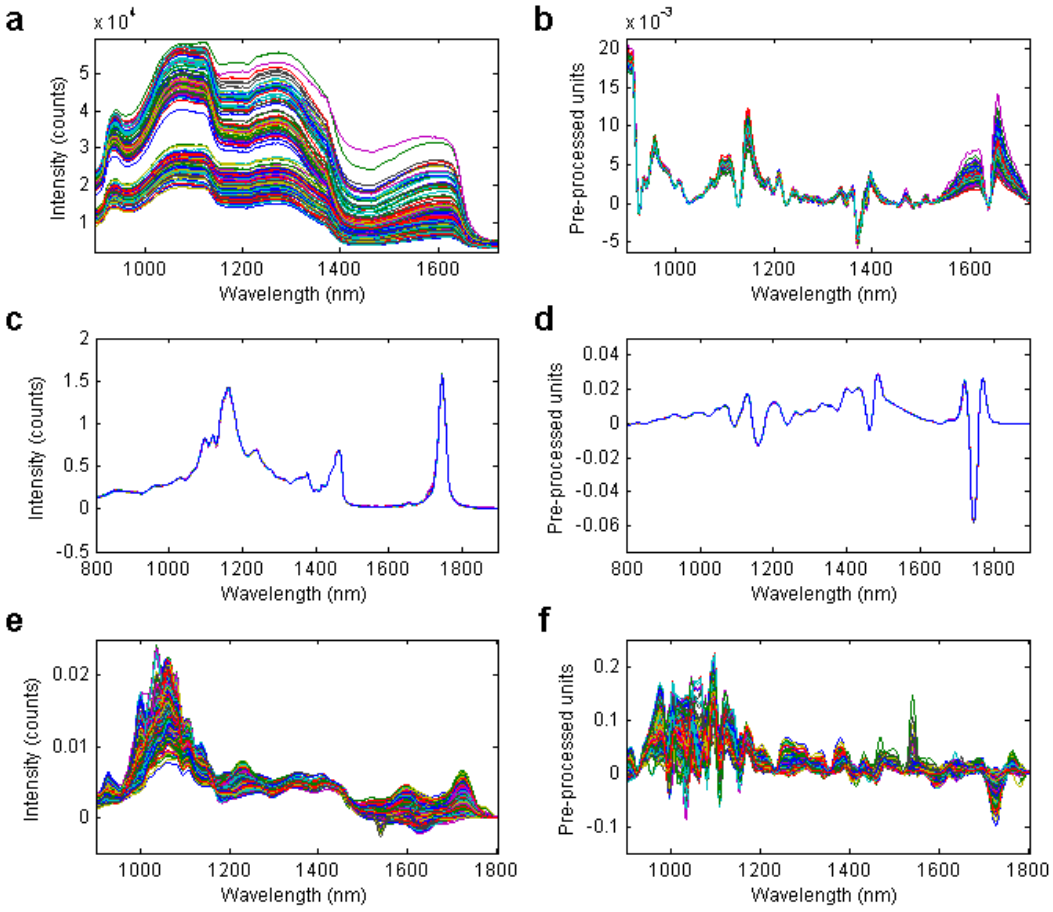

Fig. 6. Raw and pre-processed spectral data: NIR-apple (a & b), FTIR-oil (c & d), FTIR-fruit (e & f).

4.4 Parameter settings and results on spectral datasets

The DUPLEX algorithm [45] was applied to each class separately to split each spectral dataset into training and test sets with the ratio of 2:1. To improve the training and testing performance of the classification models, the real-world spectral datasets (NIR-apple, FTIR-oil and FTIR-fruit) were pre-processed by smoothing, normalization, second derivatives and baseline removal in this paper. The pre-processed spectra are shown in Fig. 6b, d and f. The optimal parameters of five algorithms are set by 10-fold cross-validation on the training set using the metric of average accuracy. The search ranges for the optimal parameters of the algorithms on spectral datasets are the same as that on UCI datasets. Fig. 7a shows the average accuracy of NCPLS-DA over LV and NC for the NIR-apple dataset. Over 85% accuracy can always be achieved after 4 LVs for any NC. As with the optimal NC

for the Arcene dataset, the optimal NC for NIR-apple dataset is set to CN/3, where CN is selected by separability criterion within the range from 10 to 50 shown in Fig. 3c.

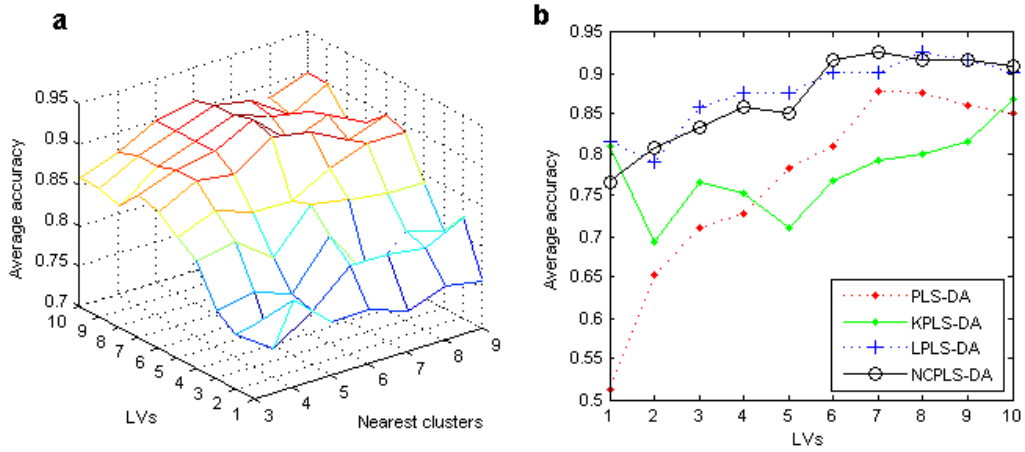

Fig. 7. Selection of the optimal LV and NC for NCPLS-DA on NIR-apple (a) and the average performance of 4 PLS-based algorithms on NIR-apple (b).

Cross-validation results of four PLS-based algorithms on NIR-apple training set, over LV, are presented in Fig. 7b. The other parameters, except LV, of kernel, local and NC PLS-DA have been set to their optimal values. Among the four PLS-based algorithms, PLS-DA performs badly with first few LVs and requires more than 6 LVs to reach 80%. Local PLS-DA and NCPLS-DA not only obtain comparably high accuracies using small LVs but also achieve the best result of 92.5%. Table 4 shows the optimal parameters of different algorithms for five spectral datasets.

Table 4

The optimal parameters of the different algorithms for spectral datasets

Dataset	**k−NN**	**PLS-DA**	**KPLS-DA**		**LPLS-DA**		**NCPLS-DA**		
	NN	LV	LV	log(σ)	LV	NN	LV	CN	NC
Multimodal	1	4	10	-1	2	23	2	46	8
Nonlinear	15	1	2	1	1	22	1	40	7
NIR-apple	7	7	10	-1	8	36	7	18	6
FTIR-oil	1	6	9	1	8	22	7	18	7
FTIR-fruit	1	10	10	-1	10	88	7	26	5

Table 5

Classification accuracy (%) of the different algorithms for test set and each class

Data set		**k-NN**	**PLS-DA**	**KPLS-DA**	**LPLS-DA**	**NCPLS-DA**
Multimodal	Test set	**100.00**	52.00	99.00	**100.00**	**100.00**
	Class 1	100.00	52.17	98.04	100.00	100.00

	Class 2	100.00	51.85	100.00	100.00	100.00
Nonlinear	Test set	**94.00**	57.00	93.00	**94.00**	**94.00**
	Class 1	89.29	56.60	95.74	92.31	94.00
	Class 2	100.00	57.45	90.57	95.83	94.00
NIR-apple	Test set	80.33	81.97	80.33	85.25	**90.16**
	Class 1	81.25	80.00	81.25	84.85	90.63
	Class 2	79.31	84.62	79.31	85.71	89.66
FTIR-oil	Test set	**100.00**	92.50	92.50	**100.00**	**100.00**
	Class 1	100.00	87.50	87.50	100.00	100.00
	Class 2	100.00	84.62	84.62	100.00	100.00
	Class 3	100.00	100.00	100.00	100.00	100.00
	Class 4	100.00	100.00	100.00	100.00	100.00
FTIR-fruit	Test set	96.34	96.65	95.43	96.95	**98.17**
	Class 1	95.85	98.54	96.23	97.63	99.04
	Class 2	97.30	93.44	93.97	95.73	96.64

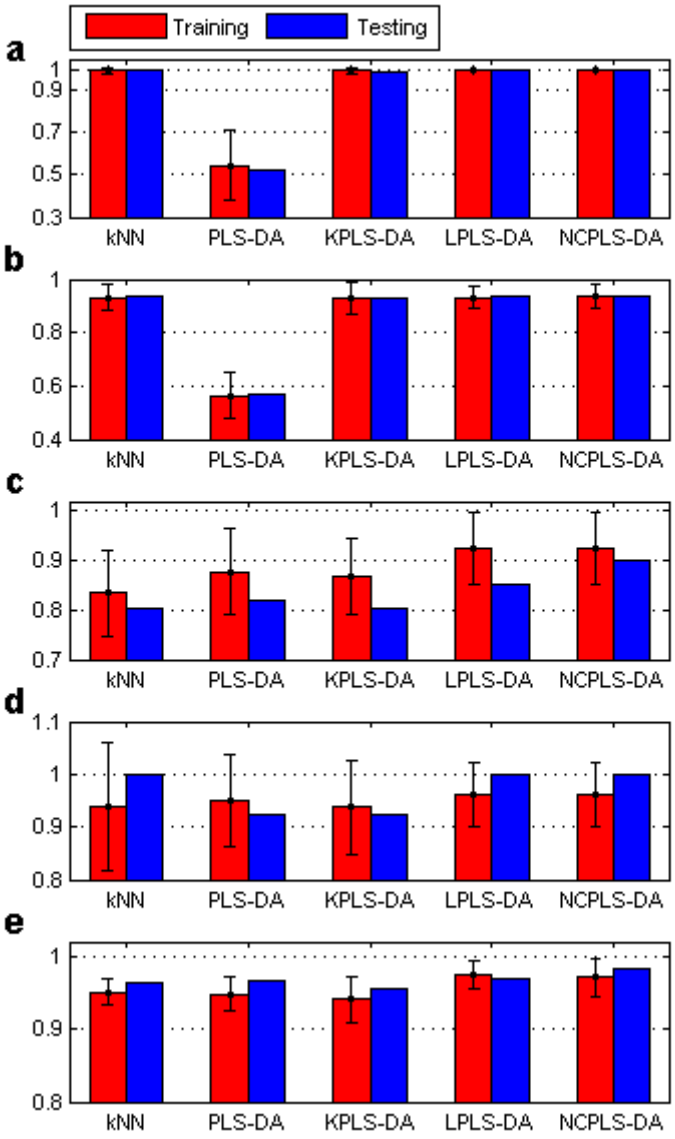

Fig. 8. Performance of five algorithms on spectral datasets: (a) Multimodal, (b) Nonlinear, (c) NIR-apple, (d) FTIR-oil and (e) FTIR-fruit. The bar in red colour is the average accuracy of 10-fold cross-validation within training set, while the bar in blue colour is the overall classification accuracy on test set.

For spectral datasets, the overall performance of different algorithms on training and test set are shown in Fig. 8 as bar plots. The red and blue bars represent the average accuracy of cross-validation on training set and the overall classification accuracy on test set respectively. An error bar is also used to represent the standard deviation in cross-validation. In the simulated spectral datasets (Fig. 8a and b), the cross-validation and classification accuracies of PLS-DA are around 50% due to the symmetric distribution of multimodal and nonlinear data. K-NN and three modified PLS-DA algorithms, by contrast, show better performance in handling multimodal and nonlinear problem. In the real world spectral datasets (Fig. 8c, d and e), the proposed NCPLS-DA always achieves

the best results in classifications. Local PLS-DA also performs very well compared to PLS-DA and kernel PLS-DA. By looking at the classification accuracy of each class demonstrated in Table 5, NCPLS-DA obtains the highest accuracy in identifying specific class, for example 89.66 % in organic apples (Class 2 of NIR-apple dataset) and 99.04% in non-strawberry purees (Class 1 of FTIR-fruit dataset).

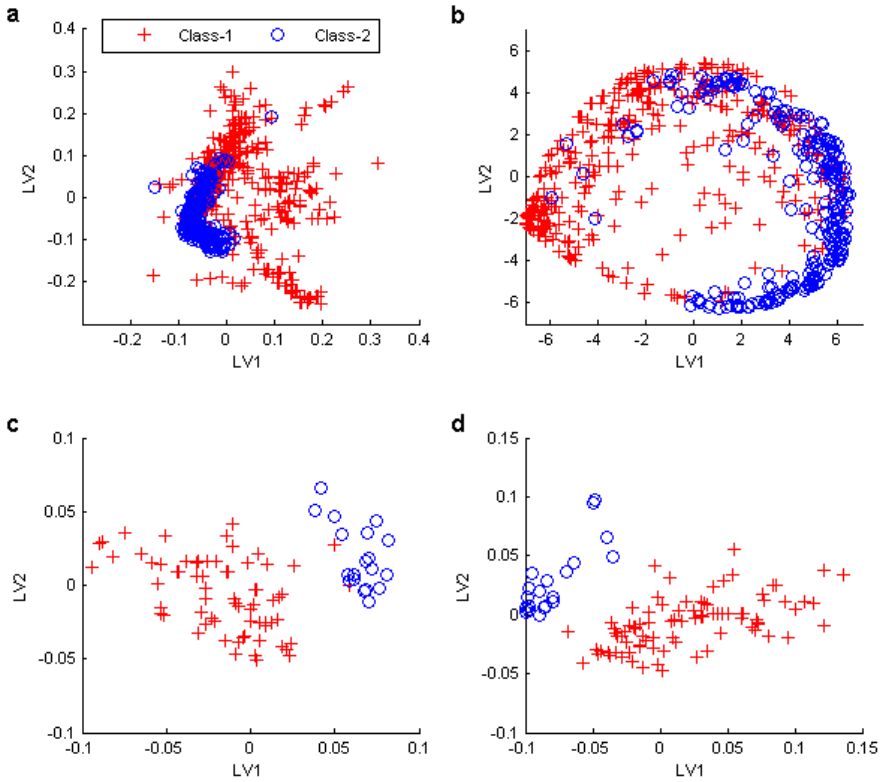

Fig. 9. Training sample (FTIR-fruit) projections in the latent variable space of the PLS-DA and kernel PLS-DA model: (a & b). Local training sample (FTIR-fruit) projections in the latent variable space of the local PLS-DA and NCPLS-DA model: (c & d).

Furthermore, the scores of FTIR-fruit samples used for modelling a query in four PLS methods are visualized in latent variables space, as shown in Fig. 9. This query is wrongly attributed by PLS-DA but correctly identified by kernel, local and NC PLS-DA. As it can be seen, global PLS modelling is unable to provide a distinctive separation between two classes in projected space, while the scores obtained from local PLS-DA and NCPLS-DA are clearly separated using 2 LVs.

4.5 Discussion

From the experiments on 12 UCI and 5 spectral datasets, it is clear that the proposed method outperforms PLS-DA almost all the time; especially when using small LVs, the outperformance is significant. It also achieves

the highest classification accuracy in most cases, 12 out of 17 datasets and all 5 spectral datasets, respectively among 5 methods. A main reason for the outperformance is the fact that NCPLS-DA simplifies the modelling procedure by removing samples which is irrelevant to a query. Meanwhile, it effectively maintains a local structure which is approximately linearly separable to handle the problems such as multimodality and nonlinearity.

We provide a simple and general way to define the parameters (CN and NC) in NCPLS-DA which merely aims to demonstrate the superiority of our method compared to standard and kernel PLS-DA. Thus, this method still has potential of further improvement. For example, we select a larger value of CN and a smaller value of NC in Ionosphere and the simulated nonlinear dataset, respectively, which achieves better classification results than the default searching range of CN and NC. This indicates that widening the validation range of parameters sometimes is required to capture the specific information in diverse data, e.g. sample size and distribution.

5. Conclusion

The PLS algorithm and its extensions have been widely applied to the analysis of multivariate data. In this paper, we present an extension of PLS-DA in order to improve its performance in the classification of multimodal and nonlinear data by embedding a nearest cluster strategy. Termed NCPLS-DA, the extended PLS-DA applies hierarchical clustering to group the data globally and applies PLS-DA on the nearest clusters of samples of a query. A simple and effective way is provided to set the cluster numbers as well as the nearest clusters. The NCPLS-DA algorithm can not only handle high dimensional and collinearity data effectively as PLS-DA, but it can also locally handle multimodal and nonlinear distributions.

The experimental results clearly demonstrate that NCPLS-DA is superior to similar methods in the literature in terms of classification accuracy. It outperforms PLS-DA for small LVs and keeps such outperformance even when LVs increase. Moreover, the sample distribution in the projected space composed by nearest clusters is more distinctive and approximately linear-separable compared to global PLS model.

NCPLS-DA is a flexible approach in multivariate analysis. The components of hieratical clustering and PLS classification can be adjusted to fit analysing tasks in different fields. Distance functions in clustering and finding nearest clustering centres can be changed. On one hand, Euclidean distance can be replaced by Manhattan or fractional distance to further improve the classification accuracy on high dimensional data. On the other hand, dimensionality reduction is suggested prior to clustering because the removal of irrelevant variables usually

provides more accurate results. Our future work will improve the clustering procedure that can efficiently capture the distribution of data.

References

[1] Z. Xiaobo, Z. Jiewen, M. J. W. Povey, M. Holmes, and M. Hanpin, "Variables selection methods in near-infrared spectroscopy," *Analytica Chimica Acta*, vol. 667, no. 1–2. pp. 14–32, 2010.

[2] L. Nørgaard, R. Bro, F. Westad, and S. B. Engelsen, "A modification of canonical variates analysis to handle highly collinear multivariate data," *J. Chemom.*, vol. 20, no. 8–10, pp. 425–435, 2006.

[3] V. Centner, O. E. De Noord, and D. L. Massart, "Detection of nonlinearity in multivariate calibration," *Anal. Chim. Acta*, vol. 376, no. 2, pp. 153–168, 1998.

[4] M. Bevilacqua and F. Marini, "Local classification: Locally weighted-partial least squares-discriminant analysis (LW-PLS-DA)," *Anal. Chim. Acta*, vol. 838, pp. 20–30, 2014.

[5] Q. P. He and J. Wang, "Large-scale semiconductor process fault detection using a fast pattern recognition-based method," in *IEEE Transactions on Semiconductor Manufacturing*, 2010, vol. 23, no. 2, pp. 194–200.

[6] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø, "A review of variable selection methods in Partial Least Squares Regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 118. pp. 62–69, 2012.

[7] E. K. Kemsley, "Discriminant analysis of high-dimensional data: A comparison of principal components analysis and partial least squares data reduction methods," *Chemom. Intell. Lab. Syst.*, vol. 33, no. 1, pp. 47–61, 1996.

[8] Q. S. Xu, Y. Z. Liang, and H. L. Shen, "Generalized PLS regression," *J. Chemom.*, vol. 15, no. 3, pp. 135–148, 2001.

[9] F. Despagne, D. Luc Massart, and P. Chabot, "Development of a robust calibration model for nonlinear in-line process data," *Anal. Chem.*, vol. 72, no. 7, pp. 1657–1665, 2000.

[10] U. Thissen, M. Pepers, B. Üstün, W. J. Melssen, and L. M. C. Buydens, "Comparing support vector machines to PLS for spectral regression applications," *Chemom. Intell. Lab. Syst.*, vol. 73, no. 2, pp. 169–179, 2004.

[11] S. R. Araújo, J. Wetterlind, J. A. M. Demattê, and B. Stenberg, "Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques," *Eur. J. Soil Sci.*, vol. 65, no. 5, pp. 718–729, 2014.

[12] H. Lin, J. Zhao, L. Sun, Q. Chen, and F. Zhou, "Freshness measurement of eggs using near infrared (NIR) spectroscopy and multivariate data analysis," *Innov. Food Sci. Emerg. Technol.*, vol. 12, no. 2, pp. 182–186, 2011.

[13] W. Pan, J. Zhao, Q. Chen, and D. Zhang, "Simultaneous and Rapid Measurement of Main Compositions in Black Tea Infusion Using a Developed Spectroscopy System Combined with Multivariate Calibration," *Food Anal. Methods*, vol. 8, no. 3, pp. 749–757, 2015.

[14] G. J. Postma, P. W. T. Krooshof, and L. M. C. Buydens, "Opening the kernel of kernel partial least squares and support vector machines," *Anal. Chim. Acta*, vol. 705, no. 1–2, pp. 123–134, 2011.

[15] W. S. Cleveland and S. J. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting," *J. Am. Stat. Assoc.*, vol. 83, no. 403, pp. 596–610, 1988.

[16] S. Kim, R. Okajima, M. Kano, and S. Hasebe, "Development of soft-sensor using locally weighted PLS with adaptive similarity measure," *Chemom. Intell. Lab. Syst.*, vol. 124, pp. 43–49, 2013.

[17] V. Giovenzana, R. Beghi, S. Buratti, R. Civelli, and R. Guidetti, "Monitoring of fresh-cut Valerianella locusta Laterr. shelf life by electronic nose and VIS-NIR spectroscopy," *Talanta*, vol. 120, pp. 368–375, 2014.

[18] M. L. McDowell, G. L. Bruland, J. L. Deenik, and S. Grunwald, "Effects of subsetting by carbon content, soil order, and spectral classification on prediction of soil total carbon with diffuse reflectance spectroscopy," *Appl. Environ. Soil Sci.*, vol. 2012, 2012.

[19] M. Zhu and A. M. Martinez, "Subclass discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1274–1286, 2006.

[20] H. Wan, H. Wang, G. Guo, and W. Xin, "Separability-Oriented Subclass Discriminant Analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.

[21] L. Zhang and F. C. Tian, "A new kernel discriminant analysis framework for electronic nose recognition," *Anal. Chim. Acta*, vol. 816, pp. 8–17, 2014.

[22] T. Czekaj, W. Wu, and B. Walczak, "About kernel latent variable approaches and SVM," *J. Chemom.*, vol. 19, no. 5–7, pp. 341–354, 2005.

[23] R. Rosipal, "Nonlinear partial least squares: An overview," *Chemoinformatics Adv. Mach. Learn. Perspect. Complex Comput. Methods Collab. Tech.*, pp. 169–189, 2011.

[24] W. Song, H. Wang, P. Maguire, and O. Nibouche, "Local Partial Least Square classifier in high dimensionality classification," *Neurocomputing*, vol. 234, pp. 126–136, 2017.

[25] K. Hazama and M. Kano, "Covariance-based locally weighted partial least squares for high-performance adaptive modeling," *Chemom. Intell. Lab. Syst.*, vol. 146, pp. 55–62, 2015.

[26] T. Uchimaru and M. Kano, "Sparse Sample Regression Based Just-In-Time Modeling (SSR-JIT): Beyond Locally Weighted Approach," *IFAC-PapersOnLine*, vol. 49, no. 7, pp. 502–507, 2016.

[27] X. Zhang, M. Kano, and Y. Li, "Locally weighted kernel partial least squares regression based on sparse nonlinear features for virtual sensing of nonlinear time-varying processes," *Comput. Chem. Eng.*, vol. 104, pp. 164–171, 2017.

[28] H. Wold, "Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments A2 - KRISHNAIAH, PARUCHURI R," in *Multivariate Analysis–III*, 1973, pp. 383–407.

[29] S. de Jong, "SIMPLS: An alternative approach to partial least squares regression," *Chemom. Intell. Lab. Syst.*, vol. 18, no. 3, pp. 251–263, 1993.

[30] J. P. A. Martins, R. F. Teófilo, and M. M. C. Ferreira, "Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets," *J. Chemom.*, p. n/a-n/a, 2010.

[31] A. Alin, "Comparison of PLS algorithms when number of objects is much larger than number of variables," *Stat. Pap.*, vol. 50, no. 4, pp. 711–720, 2009.

[32] T. N. Tran, L. Blanchet, N. L. Afanador, and L. M. C. Buydens, "Novel unified framework for latent modeling and its interpretation," *Chemom. Intell. Lab. Syst.*, vol. 149, pp. 127–139, 2015.

[33] Z. He, H. Zhou, J. Wang, and S. Zhai, "A unified framework for contrast research of the latent variable multivariate regression methods," *Chemom. Intell. Lab. Syst.*, vol. 143, pp. 136–145, 2015.

[34] M. Barker and W. Rayens, "Partial least squares for discrimination," *J. Chemom.*, vol. 17, no. 3, pp. 166–173, 2003.

[35] M. T. Bona and J. M. Andrés, "Coal analysis by diffuse reflectance near-infrared spectroscopy: Hierarchical cluster and linear discriminant analysis," *Talanta*, vol. 72, no. 4, pp. 1423–1431, 2007.

[36] Y. Gong, L. Guan, X. Feng, L. Wang, and X. Yu, "In-situ lubricating oil condition sensoring method based on two-channel and differential dielectric spectroscopy combined with supervised hierarchical clustering analysis," *Chemom. Intell. Lab. Syst.*, vol. 158, pp. 155–164, 2016.

[37] K. H. Liland, A. Kohler, and V. Shapaval, "Hot PLS-a framework for hierarchically ordered taxonomic classification by partial least squares," *Chemom. Intell. Lab. Syst.*, vol. 138, pp. 41–47, 2014.

[38] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, “On the Surprising Behavior of Distance Metrics in High Dimensional Space BT - Database theory,” in *Database theory*, vol. 1973, no. Chapter 27, 2001, pp. 420–434.

[39] K. Bache and M. Lichman, “UCI Machine Learning Repository,” *University of California Irvine School of Information*, vol. 2008, no. 14/8. p. 0, 2013.

[40] I. Guyon, J. Li, T. Mader, P. A. Pletscher, G. Schneider, and M. Uhr, “Competitive baseline methods set new standards for the NIPS 2003 feature selection benchmark,” *Pattern Recognit. Lett.*, vol. 28, no. 12, pp. 1438–1444, 2007.

[41] M. Robnik-Šikonja and I. Kononenko, “Theoretical and Empirical Analysis of ReliefF and RReliefF,” *Mach. Learn.*, vol. 53, no. 1–2, pp. 23–69, 2003.

[42] W. Song, H. Wang, P. Maguire, and O. Nibouche, “Differentiation of organic and non-organic apples using near infrared reflectance spectroscopy — A pattern recognition approach,” *2016 IEEE Sensors*, pp. 1–3, 2016.

[43] H. S. Tapp, M. Defernez, and E. K. Kemsley, “FTIR Spectroscopy and Multivariate Analysis Can Distinguish the Geographic Origin of Extra Virgin Olive Oils,” *J. Agric. Food Chem.*, vol. 51, no. 21, pp. 6110–6115, 2003.

[44] J. K. Holland, E. K. Kemsley, and R. H. Wilson, “Use of fourier transform infrared spectroscopy and partial least squares regression for the detection of adulteration of strawberry purees,” *J. Sci. Food Agric.*, vol. 76, no. 2, pp. 263–269, 1998.

[45] R. D. Snee, “Validation of Regression Models: Methods and Examples,” *Technometrics*, vol. 19, no. 4, pp. 415–428, 1977.