Analytica Chimica Acta 902 (2016) 70–81

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Variable selection in multivariate calibration based on clustering of variable concept

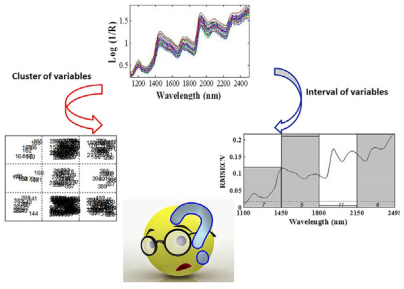
Maryam Farrokhnia[a], Sadegh Karimi[b,*]

[a] The Persian Gulf Marine Biotechnology Research Center, Bushehr University of Medical Sciences, Bushehr, Iran
[b] Department of Chemistry, College of Sciences, Persian Gulf University, Bushehr, Iran

HIGHLIGHTS

- A new and efficient variable selection based on clustering of variable concept has been suggested for PLS.
- Selection the most useful variable is simple and straightforward.
- CLoVA concept can be used as alternative instead of using interval based variable selections for PLS.
- Analyses of different data sets indicate the superiority of CLoVA respect to available variable selection algorithms.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 3 September 2015
Received in revised form 31 October 2015
Accepted 4 November 2015
Available online 17 November 2015

Keywords:
Variable selection
Partial least square
Clustering of variable – partial least square
Self organization map
Interval based partial least square

ABSTRACT

Recently we have proposed a new variable selection algorithm, based on clustering of variable concept (CLoVA) in classification problem. With the same idea, this new concept has been applied to a regression problem and then the obtained results have been compared with conventional variable selection strategies for PLS. The basic idea behind the clustering of variable is that, the instrument channels are clustered into different clusters via clustering algorithms. Then, the spectral data of each cluster are subjected to PLS regression. Different real data sets (Cargill corn, Biscuit dough, ACE QSAR, Soy, and Tablet) have been used to evaluate the influence of the clustering of variables on the prediction performances of PLS. Almost in the all cases, the statistical parameter especially in prediction error shows the superiority of CLoVA-PLS respect to other variable selection strategies. Finally the synergy clustering of variable (sCLoVA-PLS), which is used the combination of cluster, has been proposed as an efficient and modification of CLoVA algorithm. The obtained statistical parameter indicates that variable clustering can split useful part from redundant ones, and then based on informative cluster; stable model can be reached.

1. Introduction

Multivariate calibration/pattern recognition models, such as partial least squares regression/discrimination, are usually applied when the number of variables is much higher than the number of samples [1]. Nevertheless, certainly all the variables are not related to the response vector and some of them are redundant and their responses do not possess useful information about the studied parameter(s). So that, the input variables of multivariate calibration contain the mix information from both useful and redundant parts [2]. In other words, both regions, useful and uninformative variables, participate in model building. This subject can be

* Corresponding author.
E-mail address: karimi.sadegh@gmail.com, sakarimi@pgu.ac.ir (S. Karimi).

http://dx.doi.org/10.1016/j.aca.2015.11.002
0003-2670/

considered as the main weak point of multivariate calibration. Usually this problem can be solved using variable selection concept. Huge numbers of publications regards to this subject [3–7] highlight its importance.

Variable selection in multivariate calibration can be used for several reasons. One of the main advantages is to reduce the number of variables. With this aim, the model prediction can be improved. Another benefit of variable reduction is the better interpretation of obtained model [1]. Various mathematical strategies for variable selection have been proposed in the literature which have been tried to achieve the aforementioned purposes. These algorithm has own strengths and weaknesses which Anderson et al [1] has been addressed some of them in his article. On the other hand, variable selection might be complicated and fails to obtain the promising result when the number of predictor variables respect to the number of samples substantially exceeds [8,9]. This study has the aim to propose a simple and efficient variable selection in multivariate calibration. Moreover a comparative study has been done to evaluate the prediction performance of this approach respect to other variable selection algorithms for all studied data sets.

2. Theory

2.1. Notations

The standard chemometrics notations will be used. Capital and lowercase letters in boldface demonstrate matrix and vector, respectively. Matrix dimensions are shown as ($I \times J$), where I and J are the number of rows and columns, respectively. The data matrix is denoted by **X**, in which the rows are absorbance spectrum of each sample. The data matrix is divided to different sub-matrices (clusters), shown by $\mathbf{X_i}$ so that $\mathbf{X} = \mathbf{[[X_1]\ [X_2]\ \ldots\ [X_q]]}$, where q is the number of clusters and the dependent variable (measured property) is denoted Y.

2.2. Partial least square regression (PLS)

Different algorithms have been proposed for PLS in literature [10]. The goal of all of them is to finds components that compromise between fitting of X and predicting Y. The central idea of partial least square regression is to approximate X by a few, say R, specifically constructed component (the partial least squares regression components) and to regress Y on the R components. Hence, partial least squares regression tries to model X and Y using the common score component T:

$$X = TP^T + E_X \quad (1)$$

$$Y = TQ^T + E_Y \quad (2)$$

Where T is an $I \times R$ matrix of scores; P is a $J \times R$ matrix of X-loadings; Q is a matrix of Y loadings; E_X and E_Y are residual matrix [11]. The R components are constructed such that they have maximum covariance with Y. The algorithm starts using normalized weight vector (W) which has been calculated according to Eq. (3).

$$W = X'Y \quad (3)$$

The PLS was first applied to evaluate near infrared (NIR) spectra by Martens and Jensen in 1983, and is now used routinely in academic institutions and industry to correlate spectroscopic measurements with related chemical/physical data.

2.3. Interval partial least square regression (iPLS)

As the Anderson et al mentioned [1] "if data are highly correlated, such as spectral data, windows of variables should be used instead of doing variable selection on each variable individually". Hence, interval-PLS has been introduced and now is one of more generally used variable selection methods. In *i*PLS, the whole spectral data is divided into some intervals (equal or unequal length) and then PLS models are applied on each of these intervals separately [12]. The main idea behind the *i*PLS algorithms is to find the interval's, which gives the better prediction respect to situation when the full spectrum is used. It should be mentioned that the comparison between interval performances is usually based on root mean square error for cross validation (RMSECV). This is an attractive and simple approach to wavelength selection, but improper selection of interval size in *i*PLS can corrupt the predictive performance in regression model [13].

2.4. Synergy interval partial least square regression (siPLS)

Synergy-*i*PLS is the modification of *i*PLS using different interval combinations and it selects the lowest RMSECV combination. Although, investigation of all combination of variables seems perfect and simple in theory, it is impossible in practice. Remarkably according to Bro comment "Even with the most advanced computers the number of variables combinations to investigate, becomes prohibitive even for, say, 50–100 variables" [1].

2.5. Backward interval partial least squares (biPLS)

In this efficient variable selection algorithm, the concept of intervals (spectral regions) has been reserved. The main idea of this algorithm is that, *i*PLS is applied to the data and then followed by backward elimination. The algorithm is continued such that in each time, the interval whose removal causes the lowest RMSECV is eliminated. According to Leardi comment, a key point which is to be optimized in *bi*PLS is the number of intervals. Small and large number of intervals has their own problem which should be considered [14].

2.6. GA-PLS

Theory of evolution is a fundamental concept in Genetic Algorithm (GA) technique. This technique consists of several steps [15,16]. First, a vector with the size corresponding to the number of variables is created. This vector is called chromosome. The zeros and ones are randomly defined for a vector. These zeros and ones resemble genes and a PLS model with selected genes is defined as an individual. Number of different individuals can make a start population (typically in the range between 20 and 500). The quality of each PLS model can be given in the term of RMSECV. The recombination of initial chromosome produces offspring. In this step, those chromosomes with higher prediction ability have more chance to be copied. Two phenomena, cross-over and mutations are performed on the chromosome. Finally when the predefined numbers of iterations has been met, the variable evaluation by GA is stopped.

2.7. Competitive adaptive reweighted sampling (CARS)

CARS is an efficient regression coefficients-based variable selection method which has been proposed by Li et al [17]. Briefly in this method, the regression coefficients are first computed on full spectra. The exponentially decreasing function (EDF) is then employed to put in force feature selection which led to removing variables with small absolute regression coefficients. Consecutively,

adaptive reweighted sampling (ARS) is performed to realize a competitive feature selection based on the regression coefficients.

2.8. Moving window-PLS

Moving window-PLS uses a fixed-size window that moves through the entire spectra and establishes PLS models of different latent variables (LVs) for each window [18]. As a result, a series of PLS models together with the sums of squared residues (SSR) are calculated. Consequently, the wavelength interval with smaller SSR and fewer LVs is selected to build the final calibration model.

2.9. Interval random frog (IRF)

Since the identification of important genes is a challenge in microarray based disease diagnosis, Li et al [19] used the reversible jump Markov Chain Monte Carlo (RJMCMC)-like strategy for variable selection. This method which is called random frog starts with a randomly selected variable subset. Then new variable subset is generated based on the previous one and is accepted with certain probability. This step iteratively continued until predefined iterations are finished. As the author [19] indicated proposed algorithm possesses the advantages of RJMCMC methods and is much easier to implement.

2.10. Iteratively retaining informative variables (IRIV)

To obtain the optimal combination of variables in a high dimensional data set, an efficient variable selection namely iterative retaining informative variable (IRIV) has been proposed by Yun et al [20]. The basic idea behind the strategy is that considers the possible interaction effect among variables through random combinations. Besides, with this algorithm the variables are classified into four classes as strongly informative, weakly informative, uninformative and interfering variables. Strongly and weak informative variables have been retains in every iteration until no uninformative and interfering variables exist.

2.11. Segmented principal component regression (SPCR)

In SPCR algorithm, which has been applied firstly in QSAR analysis, a segmentation approach is combined with PCR. Briefly the descriptors are segmented into different segments and then principal component analysis (PCA) is applied to each segment separately to extract significant principal components (PCs). Consequently, with this strategy the PCs having useful and redundant information are separated. Finally, a linear regression analysis based on stepwise selection of variables is then employed to connect a relationship between the informative extracted PCs and biological activity of compound. It's abilities to regression analysis also have been investigated by Hemmateenejad et al research group [21].

2.12. Stacked PLS

The stacking concept has been introduced by Wolpert in 1990 [22] and then generalized in 1996 by Breiman [23]. The central idea of stacking regressions is to exploit the information in the entire spectral response. The basic idea of stack PLS is to apply a linear combination of different predictors (wavelengths) to improved prediction accuracy. Stacked PLS is based on cross validation criteria, so that aims to have a set of weights from whole region that minimize the cross-validated error in the stacked regression model.

2.13. Interval variable iterative space shrinkage approach (iVISSA)

This new variable selection algorithm which has been proposed by Deng et al [24] combines the global and local searches toward iteratively. It is also intelligently optimize the locations, widths and combinations of the spectral intervals. For global search procedure, uses the advantage of soft shrinkage from VISSA to search the locations and combinations of informative wavelengths, whilst for the local search procedure, it utilizes the information of continuity in spectroscopic data to determine the widths of wavelength intervals. Both global and local search procedures are carried out alternatively to realize wavelength interval selection.

2.14. Clustering of variable using Kohonen self- organizing map (SOM)

A Kohonen self- organizing map (SOM) is a two dimensional array of neurons, which each neuron containing a weight vector that has the same dimension as the experimental variable data set. A SOM is trained to reflect as much as possible the relationship between individual pieces of data. That is able to map multidimensional information into a surface (the 2D array). Likewise principal component analysis, SOM reduces multidimensional information to two dimensions with maintaining the topology of information. However, in contrast to PCA, SOM has advantages to use the nonlinear relationship between the variables in data matrix. Fig. 1 shows the structural design of a Kohonen network. Each column in the grid represents a neuron and each box in such a column represents a weight (a number). In our case, the objects are the samples and the variables are wavelengths, wavenumbers, descriptors and etc. First of all, before the starting training, the weights take the random values. It should be noted that the learning is a competitive process. This step includes the adjustment of the weight during the training phase. The procedure is as follow. (1) A variable from training set is introduced to the network. (2) The neuron that its weight vector is the most similar (determined using the Euclidean distance) to the input variable is called the wining neuron or the best matching unit (BMU). (3) The network modifies this winner neuron weights to become much more similar to the input descriptors. (4) With the same aim, neighborhood neurons are also corrected. However the amounts of these corrections depend on their distance from the winning neuron. (5) All these steps repeat iteratively to reach a predefined number of cycles (epoch) and then the process stops. Finally, when the entire variables are entered in the Kohonen network and the process is completed, similar input (in our case similar spectral information) vectors are clustered based on their similarities [25].

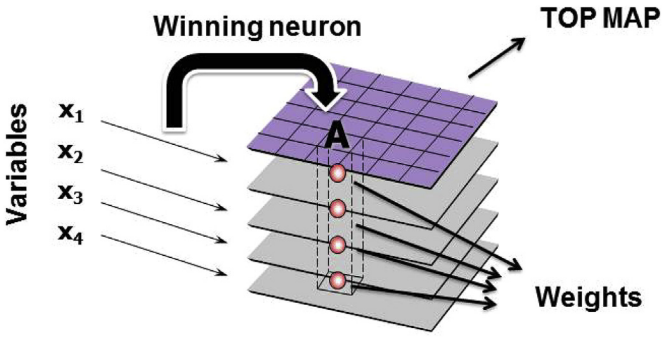


Fig. 1. Architecture of a Kohonen self-organizing map or Kohonen network.

2.15. Clustering of variable based PLS (CLoVA-PLS)

Recently we have proposed an efficient variable selection algorithm for classification problem [26] based on clustering of variable concept. The strategy of our algorithms (like the other variable selection methods) is that, all the variables cannot be informative for modeling. In other words, some of them are related to Y-variable (dependent variable) and the others not. Variables contain useful information about y-variable, as well correlate with each other, have similar information and can be considered as collinear variables. The same rule also is true for non-informative ones. On the other hand, the input variable of PLS multivariate calibration is the latent variables (LVs) which are constructed from whole region of spectral data. This is the main subject of this study; because the all variables have been contribute to construction of latent variables (LVs). The objective function of clustering of variable concept is to find the cluster of variables (instead of interval) based on their similarities, which has high correlation respect to dependent variable (Y). Hence, the variables of each cluster are used as input of the multivariate calibration/QSAR analysis separately. However it should be noted that, one of the common features between using clustering of variable strategy and other variable selection algorithms is that, the dimension of huge data sets (LC-MS, GC-MS, gene expression) is reduced. In these data sets, the ratio of variables to samples are high and this leads to be obtained the over fitted model which is known as small sample size problem in literatures [9,27]. In the following section we will show the potential ability of variable clustering for building the stable models in regression.

The algorithm of CLoVA-PLS can be described in three stages:

1. Suppose we have a data (**X**) which is arranged in the matrix with ($I_{sample} \times J_{variable}$) dimension. The variable dimension can be wavelength (uv–vis), wavenumber (IR spectra), chemical shift (NMR), descriptors (QSAR), mass value (GC/LS-MS) and so on. In general, clustering strategy can be applied in both dimensions (sample and variable space). However, since the CLoVA-PLS algorithm has been proposed to clustering of variables instead of samples, the data matrix should be transposed before applying the Kohonen SOM.
   (a) Clustering of variables into q cluster (which is defined by user), using Kohonen self- organizing map:

$$X = [X_1|X_2|\dots|X_q] \tag{6}$$

When the variables are clusterised, each ith cluster set of variables constitutes a reduced set of data $\mathbf{X}_i$ (with i = 1,2,…,q). For example for $q = 2$, Kohonen SOM produced 4 clusters. Since the number of cluster size (q) is very important parameter for building the stable multivariate calibration, its value should be optimized. So that, in order to find the optimum network size following steps should be considered:

2. For each cluster size, in order to find the most useful cluster of variable, all of the produced sub-matrix (clusters) has been investigated using PLS regression separately. The statistical parameters (RMSECV, RMSEP) of constructed model from each cluster, are used to judgment for selecting the informative one(s). It is worthy to mention that, calibration samples have been used to train and select variables, while test samples have never used during the optimization stage and there subsequently predicted by means of the models optimized in the training samples.

Generally, suppose PLS is applied in the selected cluster X_{ic} of calibration data and the latent variables which have high covariance with Y variable is calculated:

$$X_{ic} = T_{ic}P_{ic}^{T} \tag{7}$$

T and P matrices are the latent variable and loading matrices of the selected cluster (X_i) respectively. The superscript "T" indicates the matrix transpose notation. This equation suggests that the latent variable of CLoVA-PLS is calculated from specific parts of X-variables while those of conventional PLS (without variable selection) are calculated from all variables in the X matrix. In other words, these latent variables constructed just form useful region of spectral data.

Besides according to equation (2), the score and loadings of Y matrix also obtained. These matrices are used to calculate the PLS regression coefficient vectors (b_i) for specific cluster [10]:

$$b_i = W_i\left(P_i'W_i\right)Q_i \tag{8}$$

Like conventional PLS procedure, this regression coefficient (obtained from specific part of data) with Y variable has been used to construction the model in calibration step.

Finally, the prediction of Y-value of unknown sample (Y_{iup}) is calculated as follows:

$$Y_{iup} = X_{iu}b_i = X_{iup}W_i\left(P_i'W_i\right)Q_i \tag{9}$$

3. The variables of selected cluster(s) have spectral information that is more correlated with chemical property of studied samples. Consequently in order to know which subsets of variables are more useful for model building, PLS regression coefficient vector of constructed model for informative cluster(s) have been searched. The MATLAB codes for CLoVA-PLS are freely available upon request. Also in the parallel study the combination of clusters has been investigated. In this strategy, which we would like to call it synergy CLoVA-PLS, several combinations (two, three and four) have been examined and the best ones (based on statistical parameter) are selected. By this approach you can use the useful information in the other clusters.

As it was mentioned, in the CLoVA based PLS, a clustering algorithm is followed by a regression method. Therefor the key question is that which clustering algorithms can be used for clustering of variables. Recently [21], we have shown that nonlinear clustering algorithm like Kohonen self- organizing map (SOM), has superiority respect to PCA (loading plot), K-means, Fuzzy – c-means, and hierarchical for clustering of variables. Consequently, in the present study, Kohonen self-organizing map (SOM) has been used as clustering method.

3. Experimental

3.1. Real data sets

In order to investigation the efficiency of proposed variable selection algorithm, five experimental data sets, including near-infrared (NIR), and QSAR of different samples have been analyzed. The descriptions of data sets are summarized in Table SI from supplementary materials. The first benchmark data is Cargill corn data set. This data can be obtained from Eigenvector Research (available from http://www.eigenvector.com/Data/Data_sets.html; accessed on 14 September 2014). NIR spectra of Cargill corn data set involve the estimation of four properties of corn samples including moisture, oil, protein, and starch. In accordance with Eigenvector research incorporated, these properties have been

measured on three separate instruments (m5, mp5, andmp6) over a wavelength range of 1100–2498 nm at 2-nm intervals (700 wavelengths). In the present study, moisture and starch contents of 80 corn samples from the m5 (first data set) instrument have been analyzed. As proposed by Brown et al [28] two samples, (75 and 77) have been removed as outliers. The remaining samples have been divided into model building and prediction test sets (39/39) [28].

The second data set comes from an experiment designed to check the possibility of NIR spectroscopy to obtain accurate measurements of four properties (dependent variables) of biscuit dough pieces. These characteristics include percentages of four ingredients fat, sucrose, flour and water from unbaked biscuits. The training set consist of 40 samples with 700 wavelengths (1100–2498 nm in 2-nm intervals) and a further 32 samples were used as a separate validation set. In this study the fat content of biscuit dough has been analyzed. A further description of this data set can be found in Brown's article [29].

One of the important research areas which illustrate the capabilities of variable clustering is QSAR data analysis. For this purpose, a dipeptide data set has been evaluated which contains a set of 58 angiotensin-converting enzyme (ACE) inhibitors. The structures and description of the dipeptide with their experimental activity are found in Ref. [30]. Because the calculated descriptors have different scales, they were subjected to scale unit variance before analysis [30].

The fourth data set consists of 54 soy flour samples which measured on NIR spectrometers. The spectra were recorded from 1104 to 2496 nm in 8 nm intervals (175 wavelengths) [31]. The moisture values were used as the responses. According to the Leardi comments the samples have been divided into 40 samples calibration set, and the other 14 samples have been used as the independent test.

Fifth data set is NIR transmittance spectra of pharmaceutical tablets which are available on the website: www.models.kvl.dk/datasets [24]. The major property of this data set is high number of studied samples. The spectra were measured for 310 tablets having different dosage of active substances (4.3–22.8 mg) and manufactured by various production scales (full, pilot and laboratory scales). The spectral data acquisition range is 7400–10,507 cm^{-1}, led to a total of 404 variables per each sample. The response variable is the relative content of active substance in the tablets (% w/w) which is measured using high performance liquid chromatography (HPLC). According to Liang comment Kennard-Stone algorithm has been applied to divide the 310 samples into calibration and test set group with 210 and 100 samples, respectively. The more description of data set can be found in Ref. [24].

In the present study, except for QSAR data set, the maximum latent variable was set to 20, and the optimum number of latent variables obtained by five-fold cross validation. Also all the data sets were mean- centered before model building.

3.2. Computational details

Computational processing was performed in the framework of MATLAB software (Math works, Inc., Natick, MA, USA, version 7.2). PLS calibrations was based on the-PLS Toolbox version 4 from Eigenvector Research. The genetic algorithm provided by Leardi, has been downloaded from the website of (http://www.models.kvl.dk/GAPLS). The Kohonen self- organizing map Toolbox, provided by Todeschini and Ballabio has been downloaded from the website of Milano Chemometrics and QSAR research group (http://michem.disat.unimib.it/chm/download/kohoneninfo.htm). *i*PLS, *si*PLS and *bi*PLS regression have been calculated using iPLS Toolbox which is available at www.models.life.ku.dk.

4. Results and discussion

4.1. Real data sets

4.1.1. Data set 1 (Cargill Corn)

The NIR spectra of the Cargill corn samples are depicted in Fig. 2a. As we mentioned in the previous section two properties (moisture and starch) have been considered for our study. Since all the wavelengths are not informative related to studied parameters, the useful ones need to be extracted. Some of them have the similar information and it helps us to collect them in one cluster. Therefore, as the first step in CLoVA based PLS regression, Kohonen self-organizing map (SOM) is used to cluster the wavelengths based on their similarities. The SOM network projects the similar wavelengths in the $(q \times q)$ array of neurons. Hence, the cluster of variables numbers which is created by each Kohonen SOM is q^2. Fig. 1b shows the distribution pattern of variables in the (4×4) SOM network size for moisture content. The numbers from 1 to 700 in this figure refer to wavelengths of 1100–2498 nm (in 2 nm intervals) respectively. Each cluster is identified as $S_{i,j}$, where i and j are the coordinate of the rows and columns of the clusters in Kohonen SOM map. The first feature which is evident in Fig. 2b is the non- homogenous distribution pattern of NIR wavelengths. Clusters $S_{1,1}$, $S_{1,3}$, $S_{2,3}$, and $S_{3,3}$ have a high population of variables whereas in $S_{4,1}$ and $S_{2,4}$, a small number of variables are observed. One of the important parameters which should be optimized in all clustering algorithms is the cluster size. The number of clusters can be varied from 1 to the number of variables. For example if one set the number of cluster size (q) to 1, all the variables contribute in model building and can be considered as the conventional PLS. In practice, the number of clusters size can be optimized by gradually increasing the cluster size (q) and followed the statistical parameter to find a model with the satisfied result (usually the least prediction error). Here, seven SOM network sizes (2–8) have been examined. In order to find the most informative wavelength(s) related to our studied parameter, each cluster (X_i) has been separately subjected to PLS regression. Some statistical parameters of the PLS models, which have been obtained for moisture property from different clusters of network size $q = 2$ are listed in Table 1. By studying the obtained result, cluster $S_{2,1}$ has been selected as a most informative ones, which their variables leads to more appropriate regression model than the full spectral data. This cluster possesses root mean square errors of 0.0083, 0.0109 and 0.0090 for calibration, cross-validation and prediction, respectively. The important subject is that the selected cluster ($S_{2,1}$) does not have high number of variable. The cluster(s) are not necessarily selected in accordance with their population but they are chosen based on their correlations with the studied properties. Finally in order to know that which subset of wavelengths are more useful for model building of the moisture content, the corresponding regression coefficient of the best constructed model ($S_{2,1}$) has been investigated. The selected variables (wavelengths) are shown in Fig. 2c.

Once the ability of CLoVA based PLS has been confirmed, its efficiency over the other variable selection should be investigated. This part has been divided to two sections. The first ones (section A of each table) involve the results of three interval based variable selection methods (*i*PLS, *si*PLS and *bi*PLS). Moreover the results of GA-PLS, as a conventional variable selection, and PLS have been also reported in this part. In the second section (section B of each table) the results of available methods (in the literature) have been presented for each data set. In the case of CARS result, to the best of our knowledge there were no available results in the literatures for Corn, Biscuit and QSAR data sets, therefore we have also calculated this algorithm and the obtained result are reported in the section A of each table. But for those with the reports of CARS's results in literature, the section B of each table (available method)

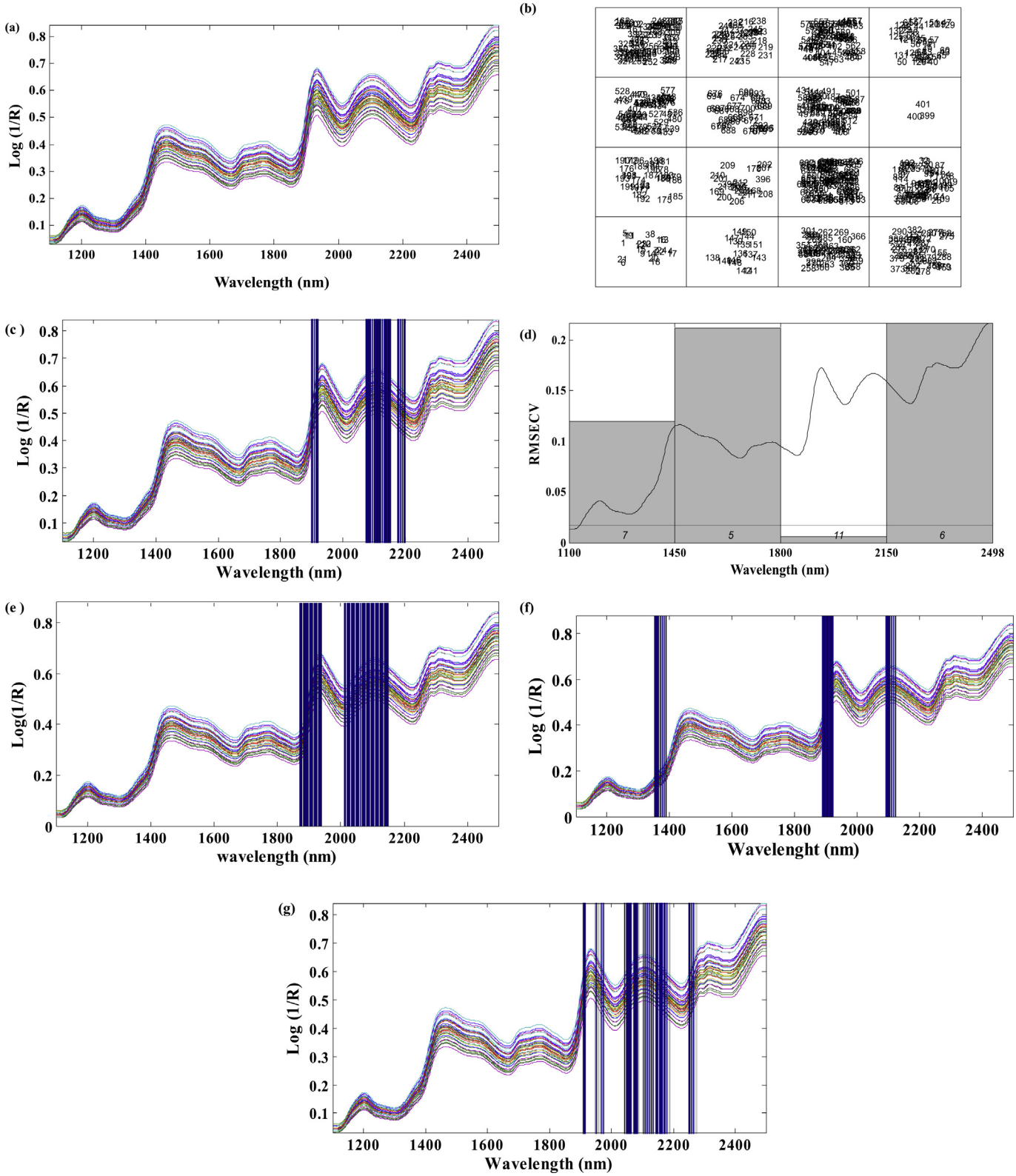


Fig. 2. (**a**) NIR spectra of Cargill corn data set. (**b**) Distribution pattern for the wavelengths of the Cargill corn data set obtained by 4 × 4 Kohonen network size. The numbers from 1 to 700 refer to the wavelengths of 1000–2498 nm in 2 nm intervals, respectively. (**c**) selected variable based on CLoVA PLS of $S_{2,1}$ form network size $q = 2$ (**d**) Results of *i*PLS analysis of the NIR spectral data. Bars are the number of RMSECV for each interval and the numbers are the PLS latent variables used. The line is error of the full-spectrum model. (**e**) selected regions by *bi*PLS algorithm (**f**) selected regions by *si*PLS algorithm (**g**) selected region synergy clustering of variable strategy (sCLoVA).

Table 1
Statistical parameters of the CLoVA-PLS models obtained from different cluster of network size $q = 2$: Cargill corn data (moisture content).

Clusters in (2 × 2) Kohonen map	N_w[a]	R^2_C	RMSC	RMSECV	R^2_P	RMSEP
CLoVA-PLS ($S_{1,1}$)	250	0.941	0.0817	0.1094	0.907	0.1127
CLoVA-PLS($S_{2,1}$)	176	0.991	0.0083	0.0109	0.993	0.0090
CLoVA-PLS($S_{1,2}$)	149	0.841	0.1607	0.2132	0.678	0.2033
CLoVA-PLS($S_{2,2}$)	125	0.957	0.0378	0.0707	0.957	0.0665

[a] Number of wavelength in each cluster.

includes their CARS's results. Since CARS is selective and predictive algorithm therefore has been selected for our comparison. At first, CLoVA-PLS has been compared with conventional PLS regression without variable selection. Obviously by clustering of variable, the prediction ability of PLS model is improved. This is very common issue because the PLS uses all parts of the spectral data and thus the irrelevant regions reduce its prediction ability. Furthermore obtained result has been compared with interval based-PLS. Since the number of interval in *i*PLS is critical parameter to obtain the appropriate model, its values should be optimized. Different interval numbers have been examined and *i*PLS model of 4 equidistant sub-intervals (resulted in lower prediction error) has been selected (Fig. 2d). This figure shows the prediction error of cross-validation (RMSECV) for each interval (bars) using optimized number of latent variables. Apparently, the interval number 3, which is related to the wavelength range of 1800–2150 nm, produced better results than other intervals. The statistical results of selected interval (*i*PLS) have been reported in Table 2. It is evident that the CLoVA based PLS model presents much better results than *i*PLS for both calibration and prediction. Although the advent of *i*PLS caused the improvement of PLS statistical parameter, it has its own limitation. The main disadvantage of this algorithm is that *i*PLS just selects the small part of data for modeling and discards the information of other spectral region. The selection of one interval is similar to lose the information of other intervals for regression problem. On the other hand, although CLoVA gives better results than *i*PLS, but *i*PLS appears simpler because of less optimization required and faster.

Two efficient variable selection algorithms namely backward *i*PLS (*bi*PLS) and synergy *i*PLS (*si*PLS) have been applied in this data and the results are shown in Table 2. From the obtained result, though both algorithms produced promising result, but CLoVA-PLS has lower prediction than these methods. The selected regions by these algorithms have been depicted in Fig. 2(e) and (f) for moisture content.

Furthermore the result of genetic algorithm (GA-PLS) as a conventional single variable selection is also given in Table 2. One can observe the superiority of the proposed methodology for discovering more predictive models.

Besides, for the investigating of clustering combinations on PLS performance, the results of synergy CLoVA-PLS have been calculated for moisture and starch parameters. It is evident from this table, synergy CLoVA has been improved the prediction power of PLS regression even respect to interval based methods and also CARS algorithms. When the prediction ability of *s*CLoVA is compared with the *bi*PLS and CARS results, the RMSEP values decrease 35.2 and 31.3% respectively for moisture content. This improvement for starch content is 65.9 and 52.4% respectively. The selected wavelengths using *s*CLoVA algorithm for moisture content has been depicted in Fig. 2(g). Based on Fig. 2, we can observe that the selected regions by interval based methods are selected by clustering based methods (*s*CLoVA) too; nevertheless *s*CLoVA not only identifies these wavelengths but the other wavelengths which their information is correlated with these wavelengths, are selected as well. The informative spectral regions were 1900–1924 nm and 2100–2124 nm, which correspond to the water absorption and the combination of O–H bond. Our result is in agreement with previous report for this data set [24]. According to Liang et al comments both regions are responsible for constructing the satisfied model.

Although Stacked and its moving window strategies which has been proposed by Brown [28] are powerful algorithms, these methods could not produce RMS errors lower than cluster based algorithms. It should be mentioned that, it might be that different variable selection results exist for each data set, but in order

Table 2
Comparison between root mean square errors of validation and prediction of the obtained models by clustering based PLS regression methods for Cargill corn.

Methods	Moisture		Starch		Ref
Section (A)	RMSECV	RMSEP	RMSECV	RMSEP	
PLS	0.0334	0.0406	0.2920	0.2874	
*i*PLS	0.0227 (4 int)	0.0197	0.2740(12 int)	0.2567	
*si*PLS[a]	0.0121(20int,3comb)	0.0122	0.1178(20int, 3comb)	0.1392	
*bi*PLS[b]	0.0094(40int,3comb)	0.0071	0.2478(16int,3comb)	0.2482	
GA-PLS	0.0068	0.0065	0.1102	0.1394	
CARS[c]	0.0067	0.0067	0.0749	0.1776	
CLoVA-PLS	0.0109(2×2,$S_{2,1}$)	0.009	0.1742(3×3,$S_{3,1}$)	0.1837	This work
*s*CLOVA-PLS[d]	0.0053(4×4, $S_{1,3}$&$S_{3,3}$)	0.0046	0.0832(6×6, $S_{3,1}$&$S_{4,1}$)	0.0845	This work
Section (B)					
Available method					
SPLS[e]	0.0224	0.0197	0.1778	0.2003	[28]
SMWPLS[f]	0.0234	0.0137	0.1790	0.2143	[28]
OSC-PLS	0.0238	0.0189	0.2209	0.02759	[29]

[a] Synergy interval PLS.
[b] Backward interval PLS.
[c] Competitive adaptive reweighted sampling.
[d] Synergy clustering of variable PLS.
[e] Stacked PLS.
[f] Stacked moving-window PLS.

to have reasonable comparison, those published model which have the same condition (preprocessing, validation algorithm, data splitting, etc) have been considered.

The same procedure has been done for starch property and the results have been reported in the second column of Table 2. For this property, almost all the CLoVA-based PLS models, resulted in promising prediction errors respect to other variable selection algorithms.

In order to avoid a prolonged manuscript, the results of first data set (Cargill Corn data) have been discussed in details and the remaining data sets have been described briefly.

4.1.2. Data set 2 (NIR spectra of biscuit dough)

Another real NIR data set which is used to indicate the ability of variable clustering for variable selection is biscuit dough data set. The NIR spectra of the biscuit dough samples are shown in Fig. 3a. In this study, the fat property has been selected for analysis. Table 3 summarized the statistical parameter obtained from different variable selection algorithms along with clustering based PLS in the optimum of their network sizes. In accordance with the results of Table 3, CLoVA-PLS using (4×4) Kohonen map has lower error especially for prediction than all interval based methods (*si*PLS, *bi*PLS and *i*PLS). Cluster $S_{2,1}$ of this network size has 0.186 RMSEP which improves the performance of conventional PLS (52.7%). The similar trend has been seen for prediction errors in GA-PLS (47.4%) and *si*PLS (11.8%). Moreover, we have investigated the synergy-CLoVA strategy for this property. This shows completely similar results to CLoVA which discloses that all the useful variables are located in the selected cluster. In other word, the other clusters have no useful information for our modeling. Selected region using CLoVA-PLS and interval based methods have been depicted in Fig. 3b for fat property. As it is clear from Fig. 3d and e, both *si*PLS and *bi*PLS have been selected the similar regions. Although CLoVA algorithm has some common feature with them, it also selects other region, which seems has high correlation related to fat property.

Interestingly, CLoVA – PLS has also lower prediction error (36.5%) than SPCAR as an efficient variable selection method. Yiming Bi et al [32] was also investigated the fat property of Biscuit dough using modified version of stacked PLS which was

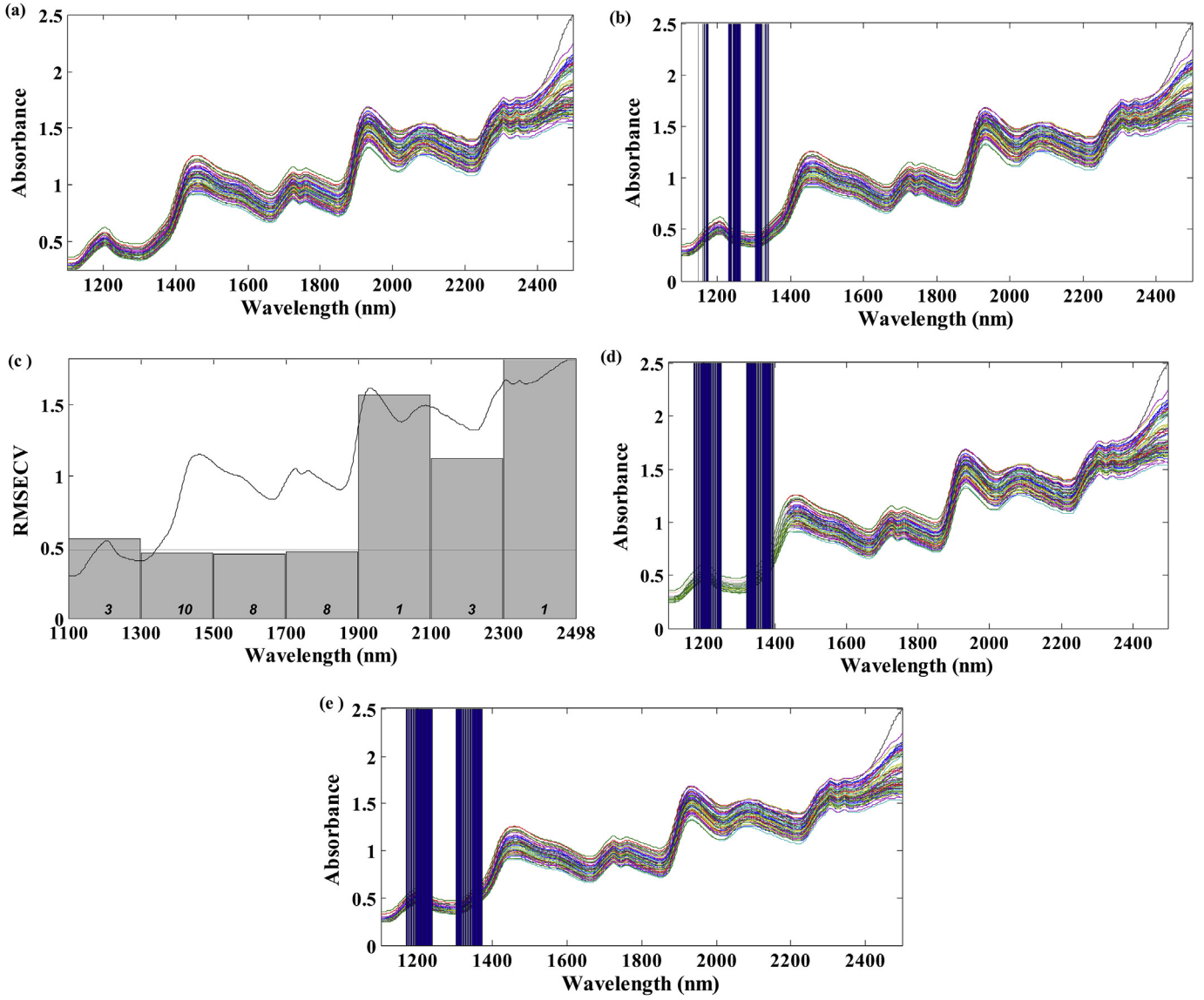


Fig. 3. **(a)** NIR spectra of Biscuit dough samples **(b)** selected wavelengths for synergy CLoVA-PLS algorithm of $S_{2,1}$ from network size $q = 4$ **(c)** Cross-validation (RMSECV) of *i*PLS model for biscuit dough data (fat). Numbers shown in each bar represent the latent variables in each interval **(d)** selected region for *si*PLS strategy **(e)** interval selection using *bi*PLS algorithm.

Table 3
Comparison between root mean square errors of validation and prediction of the obtained models for fat property of Biscuit dough data set.

Methods	Fat		Ref
Section (A)	RMSECV	RMSEP	
PLS	0.617	0.394	
*i*PLS	0.362(7 int)	0.322	
*si*PLS	0.332(21int, 2comb)	0.208	
*bi*PLS	0.387(19int,2comb)	0.191	
GA-PLS	0.236	0.354	
CARS	0.342	0.256	
CLoVA-PLS	0.359(4 × 4,$S_{2,1}$)	0.186	This work
*s*CLoVA-PLS	0.359	0.186	This work
Section (B)			
Available method			
SPCR[a]	0.327	0.293	[21]
DSPLS[b]	NR[c]	0.202	[32]
Stepwise MLR	NR	0.209	[29]

[a] Segmented principal component analysis and regression.
[b] Dual stack PLS.
[c] Not reported.

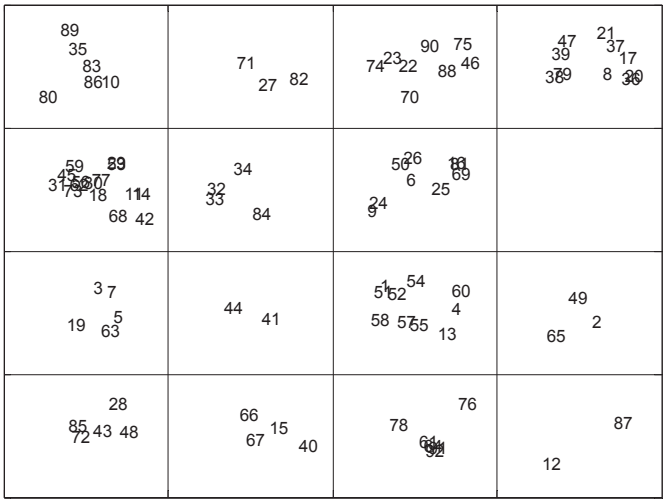


Fig. 4. Distribution pattern for the descriptors of the ACE data set obtained by (4×4) Kohonen network size clustering of variables. The numbers from 1 to 98 refer to the number of extracted descriptors.

called Dual stacked PLS (DSPLS). Although this efficient algorithm which contains two steps of stacked regression and PLS resulted in promising prediction error, CLoVA methods are more predictive than DSPLS.

Fat contents are divided into saturated and unsaturated content. the saturated ones contain no C–C double bond, so that they are solid at room temperature while the unsaturated fat which usually is referred as oil is unsaturated with at least one C=C bond. The other functional group of fat is ester, carboxylic or hydroxyl functional group. Therefor we expect that in accordance with the fat kind of studied samples, at least one of these functional groups exist in the IR spectrum region as fat indicator [33]. Our selected region may refer to C–O stretching of carboxylic acid functional group. This C–O stretching frequency usually is in the region 1000–1300 nm. We believe that this region is particularly indicator of fat (fatty acid) because other region are more probable to be in common in different kind of macronutrients such as C=C, O–H and C=O stretching in acid, aldehyde and ketone functional groups.

4.1.3. Data set 3 (QSAR of amino acid (AA) indices)

QSAR analysis is one of the most important research areas which can be studied by clustering of variable concept. Due to high important role of amino this acid (AA) in building the blocks of proteins, they are vital to life. Clustering of variable, as a new policy in quantitative structure–activity relationship (QSAR) has been applied to define new amino acid indices. The chemical structures of the amino acids (AAs) have been drawn by HyperChem software (Version 7, Hypercube Inc). Semi-empirical (AM1) method has been used for geometry optimization of their structures [30]. Different descriptors (Constitutional, topological indices, Galves charge-topological indices, charge, geometrical, functional groups and empirical) have been extracted by Dragon software (Milano Chemometrics and QSAR research group). In this way, 108 descriptors have been calculated for each amino acid. Therefore, our data matrix has 20 (naturally occur AA) rows and 108 columns (descriptors). To find the best Kohonen network size, different nodes have been examined. The statistical parameters of the CLoVA based PLS for optimum network size are listed in Table 4. Among them, the CLoVA model of (4 × 4) has been selected as the best, according to cross-validation and prediction abilities (those are highlighted in Table 4). The distributions of the original descriptors of AA indices in a (4×4) network size are shown in Fig. 4. The numbers which are located in each cluster represent the extracted descriptors for ACE data. Interestingly, descriptors which are in cluster $S_{1,3}$ have relevant information for modeling the ACE activity of the dipeptides. In conventional PLS all calculated descriptors are used to extract the AA indices (i.e.,$q = 1$ in CLoVA-PLS) and it is obvious from Table 4 (first row) that the performance of this model is significantly lower than the 4 × 4 CLoVA-PLS [34]. Thus, variable clustering concept can partition the contained information within the extracted scores into informative and redundant parts. In other words, it is the possible to get rid of redundant variable and obtains more appropriate models.

Table 4
Comparison between root mean square errors of validation and prediction of obtained models by CLoVA based PLS regression method for ACE inhibitors QSAR data set.

Methods	RMSC	RMSECV	R^2_C	R^2_{CV}	RMSEP	R^2_P	Ref
Section (A)							
PLS	0.404	0.505	0.770	0.745	0.48	0.688	
GA-PLS	0.375	0.427	0.856	0.803	0.37	0.924	
CARS	0.318	0.407	0.922	0.900	0.40	0.925	
CLoVA-PLS (4 × 4, $S_{1,3}$)	0.337	0.483	0.892	0.890	0.23	0.970	This work
*s*CLoVA-PLS(4 × 4, $S_{2,1}$, $S_{1,3}$, $S_{3,3}$, $S_{4,4}$)	0.320	0.481	0.902	0.895	0.19	0.982	This work
Section (B)							
Available method							
SPCR based loading plot[a] (3 × 3 cluster)	0.356	0.414	0.875	0.840	0.39	0.896	[30]
SPLS based loading plot[b] (4 × 4 cluster)	0.393	0.453	0.852	0.813	0.36	0.921	[30]

[a] Segmented principal component regression.
[b] Segmented partial least square.

Table 5
Selected descriptors using clustering of variable for ACE data set.

Descriptor	Description	Type
PW4	Path/walk 4 – Randic shape index	Topological
BAC	Balaban centric index	Topological
Me	Mean atomic Sanderson electronegativity (scaled on Carbon atom	Constitutional
nN	Number of nitrogen atoms	Constitutional
nC	Number of carbon atoms	Constitutional
AAC	mean information index on atomic composition	Information
BIC2	Bond information content index (neighborhood symmetry of 2-order)	Information
SIC3	Structural information content index (neighborhood symmetry of 3-order)	Information
TIC4	Total information content index (neighborhood symmetry of 4-order	Information
GGI5	Topological charge index of order 5	2D autocorrelations
JGI2	Mean topological charge index of order 2	2D autocorrelations
X5v	valence connectivity index of order 5	Connectivity
nCp	Number of terminal primary C(sp^3)	Functional group counts
G(N … O)	Sum of geometrical distances between N … O	3D Atom pairs

As it is presented in Table 4, the efficiency of CLoVA-PLS is much better than obtained results by Hemmateenejad et al [35] (QTMS indices). Variable clustering possessed higher R^2_{cv} and lower prediction error than that model. This research group also has analyzed the AA data set based on newly methods called Segmented PCR and Segmented PLS based on loading plot. The description of this algorithm can be found in Ref. [30]. Interestingly our proposed algorithm also shows the lower validation and prediction errors than segmented-PCR and segmented- PLS algorithms [30]. Because these methods used the scores from all clusters instead of few cluster.

Table 5 shows the selected descriptors by our proposed approach. It is evident from table that different groups of descriptors are important in building a QSAR model. The selected groups include constitutional, information, topological, connectivity, 2D autocorrelation, 3D atoms pairs and functional group count indices. Constitutional indices reflect the chemical composition of samples but they give no information about sample geometry or its atoms connectivity. Information, topological and connectivity indices generally characterize structures according to neighborhood symmetry, size and degree of branching and overall shape. 2D autocorrelation indices are independent of the original atom numbering and can be applied as descriptors which reveal physico-chemical properties of compound. The 3D atom pairs indices show the importance of all pairs of atoms in molecules, number of π bonding electrons, the length of the shortest bond by bond path between atoms. Finally, the functional group count indices reveal the count of compound functional groups.

Table 6
Results of soy data set: nLVs: the number of latent variables; RMSEC: root mean square error of calibration; RMSECV: root mean square error of cross-validation; RMSEP: root mean square error of prediction.

Method	nLV	RMSC	RMSECV	RMSEP	Ref
Section (A)					
PLS	4	0.7230	0.8702	1.1090	
*i*PLS(20 int)	3	0.7026	0.7842	1.0311	
*si*PLS	3	0.6965	0.7353	1.0250	
*bi*PLS (15int, 3comb)	6	0.6894	0.7314	0.8870	
GA-PLS	5	0.7188	0.7375	0.9853	
CLoVA-PLS (4×4 , $S_{3,4}$)	6	0.6620	0.8216	0.8900	This work
*s*CLoVA-PLS (4×4, $S_{3,3}$ and $S_{3,4}$)	7	0.6234	0.8211	0.8498	This work
Section (B)					
Available method					
MW-PLS[a]	2	0.7165	0.7375	1.0122	[24]
CARS	3	0.7091	0.7351	1.0062	[24]
*i*RF[b]	2	0.7083	0.7319	0.9967	[24]
*i*VISSA[c]	2	0.7067	0.7273	0.9950	[24]
IRIV	4	0.7789	NR[d]	1.0578	[24]

[a] Moving window partial least squares.
[b] Interval random frog.
[c] Interval variable iterative space shrinkage approach.
[d] Not reported.

4.1.4. Data set 4 (NIR spectra of soy data)

Related to soy data set the optimum network size was set to $q = 4$ for both CLoVA and synergy CLoVA PLS algorithms. The maximum number of latent variables for both strategies was set to 6 and 7 using 5 fold cross-validations respectively [24]. The results of CLoVA and *s*CLoVA besides other variable selection (*si*PLS ,*bi*PLS, MW-PLS, CARS, GA-PLS, *i*-RF and *i*-VISSA) are reported in Table 6. Respect to full spectrum model (PLS) all the variable selection methods indicate the improved prediction ability. However synergy CLoVA showed the lowest RMSEP (0.8498) than *bi*PLS (0.8870), CLoVA (0.8900), GA-PLS (0.9853) and even *i*VISSA (0.9950). Compared to the result of full spectra, the RMSEC, RMSECV and RMSEP of *s*CLoVA decreased by 13.8%, 4.1% and 23.4% respectively, which

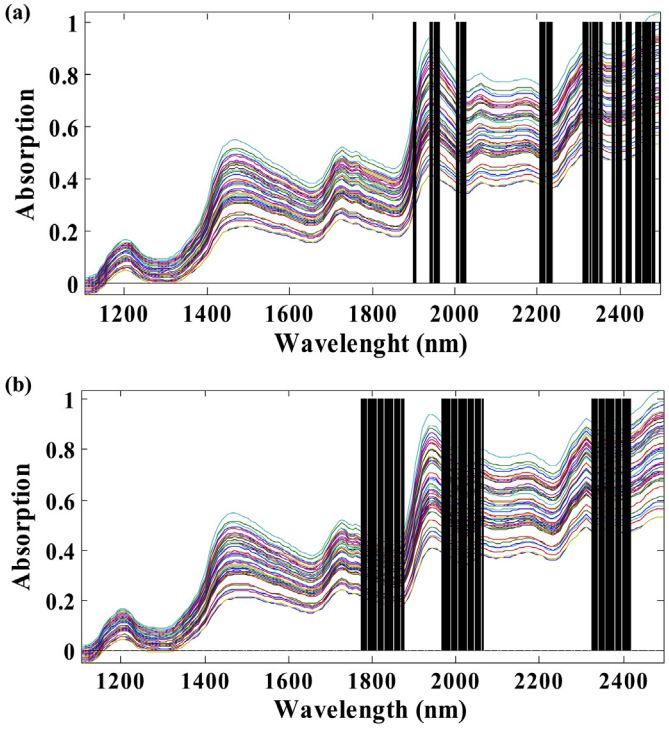


Fig. 5. Wavelengths selection using different methods on Soy data set. (a) *s*CLoVA-PLS (b) *bi*PLS.

Table 7
Results of Tablet data set: nLVs: the number of latent variables; RMSEC: root mean square error of calibration; RMSECV: root mean square error of cross-validation; RMSEP: root mean square error of prediction.

Method	nLV	RMSC	RMSECV	RMSEP	Ref
Section(A)					
PLS	6	0.3200	0.3497	0.3655	
*i*PLS (20 int)	7	0.3326	0.3454	0.3650	
*si*PLS	6	0.3140	0.3413	0.3635	
*bi*PLS (20int, 3comb)	6	0.3312	0.3469	0.3438	
GA-PLS	6	0.3127	0.3264	0.3559	
CLoVA-PLS (3×3, $S_{1,1}$)	8	0.2984	0.3472	0.3485	This work
*s*CLoVA-PLS (3×3, $S_{1,1}$ $S_{3,1}$, and $S_{2,3}$)	7	0.3157	0.3450	0.3345	This work
Section (B)					
Available method					
MW-PLS[a]	6	0.3168	0.3443	0.3620	[24]
CARS	6	0.3152	0.3243	0.3577	[24]
*i*RF	6	0.3135	0.3497	0.3594	[24]
*i*VISSA	6	0.3075	0.3259	0.3552	[24]

[a] Moving window partial least squares.

can be consider remarkable. The selected wavelength for *s*CLoVA and *bi*PLS are displayed in Fig. 5. The spectral region around 1944–2024 nm is due to water absorption which both algorithms are succeed to select this informative region [24]. On the other hand, wavelength region correspond to 2032–2500 nm is rich of information but is also complex, because it includes various combination of OH stretching with different CH, COH and OCO bending or stretching . Therefor it is reasonable to consider this region because of its valuable information [33].

4.1.5. Data set 5 (NIR transmittance spectra of tablet data)

Finally tablet data set with highly number of samples have been considered for analyzing using variable clustering strategy. Table 7 and Fig. 6 show the results of the tablet data set. For reasonable comparison with other variable selection algorithm, the maximum number of latent variables has been selected using 5-fold cross-validation on full spectra [24]. Although the combination of variable selection with PLS led to improvement of prediction results, *s*CLoVA-PLS showed the lowest RMSEP respect to all of them (0.3345). By Comparing the prediction ability of intervals based methods with variable clustering, the superiority of second ones is revealed. *si*PLS, *bi*PLS, *i*PLS, *i*RF and even *i*VISSA have RMSEP of 0.3635, 0.3438, 0.3650, 0.3594 and 0.3552 respectively. Single variable selection algorithms, GA (0.3559) and CARS (0.3577), do not produce the model with lower prediction than variable clustering algorithm.

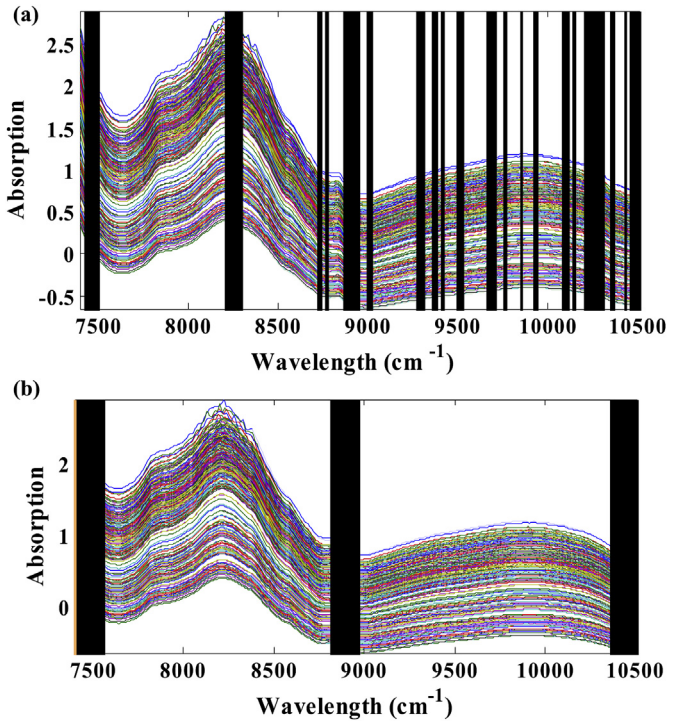


Fig. 6. Wavelengths selection using different methods on Tablet data set. (a) *s*CLoVA-PLS (b) *bi*PLS.

Fig. 6 shows the selected informative regions using clustering of variable algorithm and *bi*PLS method. Liang and coworkers [24], shows that genetic algorithm and their new methods, interval variable iterative space shrinkage approach (*i*VISSA), have been selected the similar regions. This is due to the fact that both algorithms use the RMSECV as an objective function. As it is evident from Fig. 6, CLoVA algorithm has been selected several regions for construction the stable model. One of the useful regions which have been selected using synergy CLoVA is 8800–9000 cm^{-1}. According to Dyerby et al [36]comments, the mentioned region can be used as finger print for estimation of active substance tablet data set. Additional two finger print regions (7400–7500 and 8200–8350 cm^{-1}) which have been previously reported also have been selected using clustering of variables strategy. Based on obtained result of Linag's group [24], these regions also were selected using *i*VISSA and GA-PLS algorithms. The selection of other spectral region(s) *e.g* 10,000–10,200 cm^{-1} is not clear; however it led to improve the prediction ability of obtained model. This is in agreement with Liang group's results [24].

5. Conclusion

In the present study a simple and efficient variable selection based on variable clustering concept has been proposed. In the CLoVA-PLS, the variable is divided into some clusters using unsupervised pattern recognition based on similarities. Besides the effect of clustering combinations has been investigated (synergy-CLoVA) on PLS regression. Informative scores and corresponding loadings are simply selected by applying PLS in each cluster separately. Selection the important variable is very straightforward which can be done by analyzing the regression vector of selected clusters. Analyzing of different data set indicates that variable clustering and its modifications (*s*CLoVA) combined with PLS has potential to use in model building and also classification problems. Therefore clustering of variable PLS has been suggested as an al-

ternative candidate instead of interval based and individual variable selections algorithms.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.aca.2015.11.002.

References

[1] C.M. Andersen, R. Bro, Variable selection in regression—a tutorial, J. Chemom. 24 (2010) 728–737.
[2] B. Hemmateenejad, S. Karimi, Construction of stable multivariate calibration models using unsupervised segmented principal component regression, J. Chemom. 25 (2011) 139–150.
[3] F. Allegrini, A.C. Olivieri, A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least-squares analysis, Anal. Chim. acta 699 (2011) 18–25.
[4] B.r.K. Alsberg, D.B. Kell, R. Goodacre, Variable selection in discriminant partial least-squares analysis, Anal. Chem. 70 (1998) 4126–4133.
[5] R.M. Balabin, S.V. Smirnov, Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data, Anal. Chim. acta 692 (2011) 63–72.
[6] T. Chen, E. Martin, Bayesian linear regression and variable selection for spectroscopic calibration, Anal. Chim. acta 631 (2009) 13–21.
[7] H.W. Lee, A. Bawn, S. Yoon, Reproducibility, complementary measure of predictability for robustness improvement of multivariate calibration models via variable selections, Anal. Chim. acta 757 (2012) 11–18.
[8] D. Ballabio, T. Skov, R. Leardi, R. Bro, Classification of GC-MS measurements of wines by combining data dimension reduction and variable selection techniques, J. Chemom. 22 (2008) 457–463.
[9] S. Karimi, M. Farrokhnia, Leukemia and small round blue-cell tumor cancer detection using microarray gene expression data set: Combining data dimension reduction and variable selection technique, Chemom. Intell. Lab. Syst. 139 (2014) 6–14.
[10] M. Andersson, A comparison of nine PLS1 algorithms, J. Chemom. 23 (2009) 518–529.
[11] U.G. Indahl, The geometry of PLS1 explained properly: 10 key notes on mathematical properties of and some alternative algorithmic approaches to PLS1 modelling, J. Chemom. 28 (2014) 168–180.
[12] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, Appl. Spectrosc. 54 (2000) 413–419.
[13] K. Javidnia, M. Parish, S. Karimi, B. Hemmateenejad, Discrimination of edible oils and fats by combination of multivariate pattern recognition and FT-IR spectroscopy: A comparative study between different modeling methods, Spectrochimica Acta Part A Mol. Biomol. Spectrosc. 104 (2013) 175–181.
[14] R. Leardi, L. Norgaard, Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, J. Chemom. 18 (2004) 486–497.
[15] A.S. Bangalore, R.E. Shaffer, G.W. Small, M.A. Arnold, Genetic algorithm-based method for selecting wavelengths and model size for use with partial least-squares regression: application to near-infrared spectroscopy, Anal. Chem. 68 (1996) 4200–4212.
[16] Q. Ding, G.W. Small, M.A. Arnold, Genetic algorithm-based wavelength selection for the near-infrared determination of glucose in biological matrixes: initialization strategies and effects of spectral resolution, Anal. Chem. 70 (1998) 4472–4479.
[17] G. Tang, Y. Huang, K. Tian, X. Song, H. Yan, J. Hu, Y. Xiong, S. Min, A new spectral variable selection pattern using competitive adaptive reweighted sampling combined with successive projections algorithm, Analyst 139 (2014) 4894–4902.
[18] J.-H. Jiang, R.J. Berry, H.W. Siesler, Y. Ozaki, Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data, Anal. Chem. 74 (2002) 3555–3565.
[19] H.-D. Li, Q.-S. Xu, Y.-Z. Liang, Random frog: an efficient reversible jump Markov Chain Monte Carlo-like approach for variable selection with applications to gene selection and disease classification, Anal. Chim. acta 740 (2012) 20–26.
[20] Y.-H. Yun, W.-T. Wang, M.-L. Tan, Y.-Z. Liang, H.-D. Li, D.-S. Cao, H.-M. Lu, Q.-S. Xu, A strategy that iteratively retains informative variables for selecting optimal variable subset in multivariate calibration, Anal. Chim. Acta 807 (2014) 36–43.
[21] B. Hemmateenejad, S. Karimi, N. Mobaraki, Clustering of variables in regression analysis: a comparative study between different algorithms, J. Chemom. 27 (2013) 306–317.
[22] D.H. Wolpert, Stacked generalization, Neural Netw. 5 (1992) 241–259.
[23] L. Breiman, stacked regression, Mach. Learn. 24 (1996) 49–64.
[24] B.-C. Deng, Y.-H. Yun, P. Ma, C.-C. Lin, D.-B. Ren, Y.-Z. Liang, A new method for wavelength interval selection that intelligently optimizes the locations, widths and combinations of the intervals, Analyst 140 (2015) 1876–1885.
[25] A.M. Fonseca, J.L. Biscaya, J.o. Aires-de-Sousa, A.M. Lobo, Geographical classification of crude oils by Kohonen self-organizing maps, Anal. Chim. Acta 556 (2006) 374–382.
[26] S. Karimi, B. Hemmateenejad, Identification of discriminatory variables in proteomics data analysis by clustering of variables, Anal. Chim. Acta 767 (2013) 35–43.
[27] L. Kanal, B. Chandrasekaran, On dimensionality and sample size in statistical pattern classification, Pattern Recognit. 3 (1971) 225–234.
[28] W. Ni, S.D. Brown, R. Man, Stacked partial least squares regression analysis for spectral calibration and prediction, J. Chemom. 23 (2009) 505–517.
[29] P.J. Brown, T. Fearn, M. Vannucci, Bayesian wavelet regression on curves with application to a spectroscopic calibration problem, J. Am. Stat. Assoc. 96 (2001) 398–408.
[30] B. Hemmateenejad, R. Miri, M. Elyasi, A segmented principal component analysis—regression approach to QSAR study of peptides, J. Theor. Biol. 305 (2012) 37–44.
[31] R. Leardi, A.L. Gonzalez, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemom. Intell. Lab. Syst. 41 (1998) 195–207.
[32] Y. Bi, Q. Xie, S. Peng, L. Tang, Y. Hu, J. Tan, Y. Zhao, C. Li, Dual stacked partial least squares for analysis of near-infrared spectra, Anal. Chim. Acta 792 (2013) 19–27.
[33] J. Workman Jr., L. Weyer, Practical Guide and Spectral Atlas for Interpretive Near-infrared Spectroscopy, CRC Press, 2012.
[34] H.U. Mei, Z.H. Liao, Y. Zhou, S.Z. Li, A new set of amino acid descriptors and its application in peptide QSARs, Peptide Sci. 80 (2005) 775–786.
[35] B. Hemmateenejad, S. Yousefinejad, A.R. Mehdipour, Novel amino acids indices based on quantum topological molecular similarity and their application to QSAR study of peptides, Amino Acids 40 (2011) 1169–1183.
[36] M. Dyrby, S.B. Engelsen, L. Norgaard, M. Bruhn, L. Lundsberg-Nielsen, Chemometric quantitation of the active substance (containing C≡N) in a pharmaceutical tablet using near-infrared (NIR) transmittance and NIR FT-Raman spectra, Appl. Spectrosc. 56 (2002) 579–585.