Prediction of a new object location in a case of splitting the data set

Nasser A. M. Barakat, Taha E. Farrage, Reda. A. Abu Beah
and Emad A. Mahmoud
Chemical Eng. Dept., Faculty of Eng., Minia University, Minia, Egypt

When a single linear calibration model fails to represent a data set properly, splitting the data into linear subsets is an interesting solution for the problem. In such a case, the data set is represented by a pack of different linear calibration models; that makes estimation of the sought-for predicted variable " $\hat{y}$ " for a new object being a dilemma. The present paper introduces three different nonparametric strategies to be used as discrimination tools for the new objects in a case of splitting the data set. The proposed strategies are not affected by the methodology used to split the data; so, each one can be appended any splitting algorithm. The introduced discrimination strategies have been applied on simulated and real data sets divided into training subsets and real ones. The obtained results were satisfactory.

احيانا يكون خط مستقيم وحيد غير كافى لمعايرة مجموعة من البيانات وذلك عندما يكون قيمة الخطأ بين المتغير الحقيقى y والمتغير المحسوب $\hat{y}$ غير مقبول وتظهر هذه المشكلة عادة مع البيانات المتعددة الابعاد. توجد عدة طرق للتغلب على هذه المشكلة وافضلها هى طريقة تقسيم البيانات الى عدة مجموعات صغيرة كلا منها يتم معايرته بخط مستقيم منفصل يعطى قيمة خطأ قليلة جدا ومقبولة. وحيث ان هناك مجموعة من الخطوط المستقيمة تمثل البيانات تكون عملية اختيار الخط المستقيم الملائم لحساب قيمة المتغير $\hat{y}$ لنقطة جديدة مشكلة تواجه طريقة تقسيم البيانات. لذلك الهدف من هذا البحث هو حل هذه المشكلة حيث يقوم البحث بتقديم عدة طرق حسابية كلا منها يمكن ان يستخدم لاختيار الخط المستقيم المناسب. ولإختبار الطرق المقترحة بطريقة صحيحة تم تقسيم مجموعات البيانات الصغيرة (بعد التقسيم) الى مجموعتين الاولى (training subset) وتستخدم فى الحسابات، والثانية (test subset) وتستخدم لإختبار الطرق المقترحة. ودلت النتائج التى تم الحصول عليها ان الطرق المقترحة يمكن استخدامها بأطمئنان فى اختيار الخط المستقيم الملائم لحساب قيمة المتغير $\hat{y}$ لنقاط جديدة فى حالة تقسيم البيانات.

Keywords: Discrimination, Linear machine, Genetic algorithm, Data splitting, Local linear substructures

1. Introduction

In some practical calibration situations it is difficult to construct one universal calibration model representing the whole population of interest [1]. This is sometimes common case rather than an exception in chemical calibration. An important reason can be the lack of an adequate model for all objects producing the desired error. For instance, in QSAR studies, one is unavoidably involved in the case of modeling some activity of chemicals with multiple templates which may exhibit significant structural distinction between each other, in such a case there might be difficulties to use a single linear model to describe these compounds, and it is better to search an alternative model describing the compounds with consideration of the template influence. Likewise, in chemical engineering data such as corrosion diagrams, calibration curves for adsorption of toxic compounds, kinetic study diagrams …etc. sometimes need results with very small error for the new cases. To circumvent this problem, besides the non-linear approaches using sophisticated nonlinear functions, an alternative approach is to split the whole data set into subsets and treat the problem as a quasi-linear one which models each subset as linear substructures.

The idea of splitting the whole data set followed by linear modeling in each subset is not new [2-7], so it is not the interesting point of the present study. However, predicting the variable 'y' for a new object is the actual predicament facing most the published splitting methodologies since they have not introduce discrimination tools for the new objects. A special warning is given in prediction of a new object belonging to a split data set, in which correct classification is difficult and misclassification dangerous. Actually, systematic treatment of how optimal prediction of new objects is done when it is not known to which group they belong has rarely presented [8]. Therefore, splitting the data sets into linear subsets algorithm should be appended by a discrimination methodology for the new objects to medicate the problem properly.

In the present study, the authors introduce three different discrimination strategies; the proposed strategies can be invoked as an integral epilogue for any splitting methodology. The first strategy is based on the linear machine which being used to check the linear separation of subsets in *x*-space. In this strategy, to obtain a linear machine applicable for multi subsets system; an improved Minimum-Squared-of-Errors (MSE) approach has been used to estimate the linear discrimination functions. However, the second and third strategies are excerpted from the BiLinear Modeling (BLM). The popular PCR and PLSR algorithms which being used as bilinear calibration techniques have been modified to be discrimination tools in the second and third strategies respectively. The proposed strategies have been applied in minute exposition on a simple simulated data set to clarify discrimination procedures for the readers. Moreover, the strategies have been checked by two real data sets. The obtained results in a case of simulated and real data sets were acceptable.

2. Theory

2.1. *The problem*

Mathematically, the problem can be explained as selecting the best linear model form a pack of linear models describing the data set to determine the variable *y* as well as possible from an observation of a vector *x*. In other words, discriminate the new prediction object. Discriminant analysis is usually defined as the construction of a rule that can be used to allocate a vector *x* to one of *c* different groups. A training set containing several observations from each group is needed together with a mathematical model for the distribution within each class. Note that in discriminant analysis the membership of all samples in the training set to the *c* different groups must be known, as opposite to cluster analysis. Discrimination analysis can be viewed as a calibration with a discrete response variable instead of the continuous *y*.

2.2. Splitting the data

A satisfactory splitting of calibration data into subsets has to focus in the size of the residuals from linear fitting within each subset and pay attention to the closeness of the samples in each subset in the data space, but the closeness care should not have bad influence on the linearity in each subset. As aforementioned above, many splitting methodologies have been introduced in the literature. Some splitting methods were based on cluster analysis [4-6]. Another approach has been proposed for splitting a 4-dimensional data set into linear substructures via a high-breakdown-point robust regression method [9]. However, in this study an approach being introduced by the present first author [10] has been invoked as splitting procedure. In this approach, the data set is split into a sequence of subsets which are described by linear models with desired error level. As the linear models describing the data subsets can be transformed into hyperplanes in an augmented data space, then, the proposed approach reduces the splitting of data to the search for a series of hyperplanes which successively maximize the number of data points near these hyperplanes within desired error and simultaneously producing linearly separable subsets in *x*-space. Genetic Algorithms (GAs) have an interest growing as an optimization technique for chemical problems [11-15], so a modified genetic algorithm is invoked to determine sequentially the optimum hyperplanes representing linearly separable subsets. The modification in the used genetic algorithm came from using asexual crossover process instead of using sexual one in the conventional ones since only one parent has been used to create a new generation. Moreover, a multi-parturition operation has been applying to create the individuals in the used modified GA. The reader can find detailed information about the splitting algorithm in the original paper.

2.3. The discrimination strategies

The proposed strategies are nonparametric ones to avoid the difficulty of the statistical methods.

2.3.1. Linear machine with improved MSE strategy

Linear machine is a powerful method can be used in discrimination of multicategory data sets [16]. It defines *c* linear discriminant functions *g_i*(x) for a data set containing *c* subsets, where

$$g_i(x) = xw_i^T + w_{i0} \quad i = 1,2,\ldots,c. \tag{1}$$

Where w_i is called the weight vector of the subset *i* and w_{i0} is the threshold. Assigning *x* to the subset *i* if $g_i(x) > g_j(x)$ for all $j \neq i$. Actually, the linear machine divides the feature space into *c* decision regions, with $g_i(x)$ being largest discriminant if *x* is in the region corresponding to the subset *i*. Minimum squared errors procedure (MSE) is a common method used to estimate the weight vectors and the thresholds for binary problems [17.a]. In this work, the MSE methodology has been improved to be applicable in multicategory problems, in other words, it has been modified to be adequate for estimation the discrimination functions in a case of multicategory problems.

The constraint used for estimating the discriminant surface in case of two-class problems will be invoked, this constraint is:

$$\begin{aligned} xw^T + w_0 &= 1 && \text{for all } \boldsymbol{x} \in \text{class 1} \\ xw^T + w_0 &= -1\,. && \text{for all } \boldsymbol{x} \in \text{class 2.} \end{aligned} \tag{2}$$

However, in a case of multicategory problems; the constraint is [18]:

$$\begin{aligned} xw_i^T + w_{i0} &= 1 && \text{for all } \boldsymbol{x} \in \text{subset } i \\ xw_i^T + w_{i0} &= 0 && \text{for all } \boldsymbol{x} \notin \text{subset } i. \end{aligned} \tag{3}$$

The present study introduces an estimation methodology of the discrimination function corresponding to a subset the subset X_i (i.e. estimating w_i and w_{i0}) as follow:

Let b_i is a column vector with all entries being ones and $b_1, b_2, .. b_{i-1}, b_{i+1}, \ldots b_c$ are *c*-1 column vectors with all entries being zeros. Also, let $Z_i = [u_i \quad X_i]$ and $a_i = \begin{bmatrix} w_{i0} \\ w_i \end{bmatrix}$ where u_i is a column vector with all entries being ones. Consequently, according to the constraint in eq. (3), one can write:

$$\mathrm{Z_i a_i = b_i} \qquad \mathrm{Z_i^* a_i} = \begin{bmatrix} \mathrm{b_1} \\ \vdots \\ \mathrm{b_{i-1}} \\ \mathrm{b_{i+1}} \\ \vdots \\ \mathrm{b_c} \end{bmatrix}, \tag{4}$$

Where Z_i^* is a matrix containing the remaining subsets, viz. $Z_i^* = [Z_1, \ldots Z_{i-1}, Z_{i+1}, \ldots Z_c]^T$. Eqs. (4) can be merged to be one equation as follow

$$Za_i = b\ . \tag{5}$$

Where Z is the whole data set, therefore, Z and *b* can be defined as follow

$$Z = [Z_1, \ldots Z_i, \ldots Z_c]^T$$

and

$$b = [b_1, \ldots b_i, \ldots b_c]^T\ . \tag{6}$$

If *Z* is not nonsingular, one can write $a_i = Z^{-1} b$ and obtain a formal solution at once. However, *Z* is rectangular; usually the number of samples (rows) is more than the number of variables (columns). When there are more equations than unknowns, a_i is overdetermined, and ordinarily not exact solution exists. However, one can seek a vector a_i that minimizes some functions of error between Za_i and b. If the error vector e_i is defined by

$$e_i = Za_i - b \,. \tag{7}$$

then one approach is to try to minimize the squared length of error vector. This equivalent to minimizing the sum-of-squared error criterion function *j*:

$$J(a_i) = \left\| Za_i - b \right\|^2 = \sum_{k=1}^{N} \left(a_i^t z_k - b(k)\right)^2 \,. \tag{8}$$

N is the total number of samples in the data set. The problem of minimizing the sum of squared errors is a classical one. It can be solved by a gradient search procedure, it can be simply found by forming the gradient as follow

$$\nabla J(a_i) = \sum_{k=1}^{N} 2\left(a_i^t z_k - b_{ik}\right) z_k = 2Z^t\left(Za_i - b_i\right), \tag{9}$$

and setting it equal to zero. This yield the necessary condition

$$Z^t Za_i = Z^t b_i \,. \tag{10}$$

Eq. (10) can be rewritten in terms of partitioned matrices:

$$\begin{bmatrix} u_1^t & \dots & u_i^t & \dots & u_c^t \\ X_1^t & \dots & X_i^t & \dots & X_c^t \end{bmatrix} \begin{bmatrix} u_1 & X_1 \\ \vdots & \vdots \\ u_i & X_i \\ \vdots & \vdots \\ u_c & X_c \end{bmatrix} \begin{bmatrix} w_{i0} \\ w_i \end{bmatrix} = \begin{bmatrix} u_1^t & \dots & u_i^t & \dots & u_c^t \\ X_1^t & \dots & X_i^t & \dots & X_c^t \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ u_i \\ \vdots \\ 0 \end{bmatrix} . \tag{11}$$

By defining the sample means m_i and the pooled sample scatter matrix S_W as

$$\boldsymbol{m}_i = \frac{1}{n_i} \sum_{x \in subset\ i} \boldsymbol{x} \qquad i = 1,2,\dots c \,. \tag{12}$$

$$S_W = \sum_{k=1}^{c} \sum_{x \in subset\, i} (x - m_i)(x - m_i)^t \,. \tag{13}$$

One can multiply the matrices in eq. (11) and obtain

$$\begin{bmatrix} N & (n_1 m_1 + \dots n_i m_i + \dots n_c m_c)^t \\ (n_1 m_1 + \dots n_i m_i + \dots n_c m_c) & S_W + n_1 m_1 m_1^t + \dots + n_i m_i m_i^t + \dots + n_c m_c m_c^t \end{bmatrix} \begin{bmatrix} w_{i0} \\ w_i \end{bmatrix} = \begin{bmatrix} n_i \\ n_i m_i \end{bmatrix}, \tag{14}$$

This can be viewed as a pair of equations, the first of which can be solved for w_{i0} in terms of w_i

$$w_{i0} = \frac{n_i}{N} - m^t w_i \,. \tag{15}$$

Where *m* is the mean of all of the samples, it equals

$$m = \frac{n_1}{N} m_1 + \dots \frac{n_i}{N} m_i + \dots \frac{n_c}{N} m_c \,. \tag{16}$$

Substituting the value of w_{i0} form eq. (15) in the second equation obtained from the solution of eq. (14) and performing a few algebraic manipulations, we obtain

$$\left[S_W - Nmm^t + \sum_{j=1}^{c} n_j m_j m_j^t \right] w_i = n_i(m_i - m) \,. \tag{17}$$

By knowing the value w_i; the value of w_{i0} can be calculated from eq. (15). By the aforementioned way the discriminant function of the *i*-th subset may be obtained. The procedure should be repeated with all the subsets to estimate all the discriminant functions to apply the linear machine as a discriminating tool.

2.3.2. Modification of bilinear models to be discrimination methods.

2.3.2.1. A brief introduction to the bilinear models. BLM is a significant multivariate calibration methodology [19-a]. The basic structure of BLM is that the information in the many observed variables $x' = \left(x_1, x_2, ..., x_n\right)$ is concentrated onto a few underlying variables, called components, scores, regression factors or just factors $t_1, t_2, ... t_A$ i.e.

$$\left(t_1, t_2, .. t_A\right)' = h_1\left[\left(x_1, x_2, ... x_n\right)'\right] A \le n \,. \tag{18}$$

Here h_1 is the transformation function and A indicates the number of scores can simulate the original variables. These scores are used as regressors in the regression equation with y. i.e.

$$y = h_2\left[\left(t_1, t_2, ... t_A\right)'\right] + f \,. \tag{19}$$

Here f represents those contributions to y which can not be explained by t he scores t, (t = (t_1,t_2,....t_A)). Let X and y represent the centered input data i.e.

$$X = X_{input} - 1\,\bar{X} \quad , \tag{20}$$

$$y = y_{input} - 1\,\bar{y} \,. \tag{21}$$

In terms of the centered X and y the full bilinear calibration model can be written as:

$$X = TP' + E \,, \tag{22}$$

$$y = Tq + f \,. \tag{23}$$

A loading matrix P represents the regression coefficients of X on the scores T, in the same way as q represents the regression coefficient of y on T. The residuals E and f represent the unique variation in X and y that being not explained by the A factors (T) in the bilinear structure. Really the name *'bilinear modeling"* comes from the way X itself is approximated by the model in eq. (22) that is the product of two sets of linear parameters to be estimated, termed the scores (T) and loadings (P) plus noise E. Estimating of the scores T and the loadings P is the main problem in any bilinear calibration methodology. Simply, the scores T can be represented according to eq. (18) in the following form

$$T = XV \,. \tag{24}$$

Where V is unknown weight matrix. Principal Component Regression (PCR) and Partial Least Squares (PLS) are interested bilinear calibration methods eq. [19-b]. They introduce two different ways to calculate P and T which can be used to estimate a linear model describing the data set. In this work, these methods have been modified to be discriminating techniques rather than multivariate calibration methods.

2.3.2.2. PCR bilinear modeling discrimination strategy. When PCR is used as a bilinear calibration modeling method V and P are identical, and representing the chemically meaningful eigenvectors extracted from the data set. Below, the authors report the procedure of developing the principal component regression to be a discrimination tool. The following procedure should be established on all subsets individually to elect the chemically meaningful eigenvectors in each subset, i.e. the loading matrix P for each subset. X_{input} and y_{input} denote a subset and the corresponding y values, respectively.

1. Centered the data points for the first subset, eqs. (20 and 21), they will be named here as X_0 and y_0.
2. Estimate the eigenvectors for $X_0'X_0$ (in this study the singular value decomposition method was used).
3. Use the first eigenvector (name it as p_1) to estimate the scores t_1 (t_1 = $X_0\,p_1$) and the residuals E_1 (E_1 = X_0 – $t_1\,p_1'$).

4. If the entries in the residuals matrix E_1 are very small (i.e. it can be considered as noise) that means one eigenvector embeds most of the information in the data set and the loading matrix P_1 corresponding to the first subset consists of only one column (the first eigenvector). Otherwise, the second eigenvector is invoked as p_2. Also, the residuals E_1 is used instead of X_0, it is named X_1. Estimate t_2 and E_2 via p_2 (i.e. $t_2 = X_1 p_2$ and $E_2 = X_1 - t_2 p'_2$).

5. Likewise, if the entries in E_2 still not so small this means two eigenvectors are not enough and the third one should be used as p_3, then; t_3 and E_3 are estimated, and so on and so forth. The process should be repeated until getting residuals (E_A) very small and can be considered as noise. In such a case the loading matrix P_1 consists of the exploited eigenvectors.

Discrimination of a new object x can be performed by calculating the residual obtained with each subset as follow:

$$\begin{aligned} e_{i1} &= (x - \bar{x}_i) - (x - \bar{x})p_{i1}p'_{i1} \quad \Rightarrow \\ e_{i2} &= e_{i1} - e_{i1}p_{i2}p'_{i2}\ldots\ldots \Rightarrow \\ e_{iA} &= e_{i(A-1)} - e_{i(A_i-1)}p_{iA_i}p_{iA_i} \\ i &= 1,2,\ldots c. \end{aligned} \tag{25}$$

Where $p_{i1}, p_{i2}, \ldots. p_{i(A_i-1)}, p_{iA_i}$. are the chemically meaningful eigenvectors of the i-the subset, $\bar{x}_i$ is the mean of this subset and A_i is the number of the valuable eigenvectors in the loading matrix P_i. Actually, the number of the valuable eigenvectors may change from subset to another one. The object x belongs to the subset i if

$$e_{iA_i}(x) \times e'_{iA_i}(x) < e_{jA_j}(x) \times e^{t}_{jA_j}(x) \text{ for all } j \neq i\,. \tag{26}$$

2.3.2.3. PLSR bilinear modeling discrimination strategy. As shown when PCR is used as bilinear model, allocation of a new object depends mainly on the residuals corresponding to predictor variable x, in other words, in the above strategy discrimination is achieved without taken in consideration the effect of y value.

To do discrimination depends on both of X and y this strategy is introduced; it utilizes the linear model corresponding to each subset. The strategy modifies the algorithm introduced by Partial Least Squares Regression (PLSR) as bilinear multivariate calibration model. In the literature, there are two famous algorithms based on PLSR have been introduced as bilinear models, namely; the orthogonalized algorithm [20] and non-orthogonalized one [21], in this work the orthogonal algorithm will be used because it is more simple and also there is no big difference between the final results obtained from them. Generally, PSLR approach needs an additional set of loadings called loading weights 'W" where

$$V = W(P'W)^{-1}\,. \tag{27}$$

To modify the orthogonalized PLSR algorithm to be used as a discriminant technique, the following steps should be achieved for each subset individually to calculate the loading weights *Wi* corresponding the subset i. Therefore, *Xinput* and *yinput* denote a subset and the corresponding y values.

1. Center the input data points eqs. (20 and 21).

2. Use the variability in $y0$ to find the first loading weights w1 for the subset in study using Least Squares (LS) and the local model

$$X_0 = y_0 w'_1 + E\,, \tag{28}$$

and scale the vector to length 1. The solution is

$$w_1 = (y'_0 X_0 X'_0 y_0)^{-0.5} X'_0 y_0\,. \tag{29}$$

3. Estimate the scores $t1$ using the local model

$$X_0 = t_1 w'_1 + E\,. \tag{30}$$

The LS solution (since $w'_1 w_1 = 1$)

$$t_1 = X_0 w_1 \ . \tag{31}$$

4. Estimate the spectral loadings $p1$ using the local model

$$X_0 = t_1 p_1' + E, \tag{32}$$

which gives the LS solution

$$p_1 = X_0' t_1 / t_1' t_1 \ . \tag{33}$$

5. Estimate the chemical loading $q1$ using Least Squares (LS), as follow

$$q_1 = y_0' t_1 / t_1' t_1 . \tag{34}$$

6. Estimate the residuals $f1$ and $E1$ as follow

$$f_1 = y_0 - t_1 q_1 \quad . \tag{35}$$

$$E_1 = X_0 - t_1 p_1' \quad . \tag{36}$$

7. If the residuals $f1$ it is so small this means one loading weights is enough otherwise another loading weights should be estimated, in such a case $E1$ and $f1$ will be used instead of $X0$ and $y0$ respectively, they are named $X1$ and $y1$.
8. The process is repeated until having very small fA where A is the optimum number of loading weights giving very small residuals. The required loading weight matrix Wi consists of all the exploited loading weight vectors.

To allocate an object x**,** the final residuals $f_{iA_i}(x)$ should be estimated with all subsets. The residuals can be estimated as follow

$$f_{i1}(x) = (\hat{y}_i - \bar{y}_i) - (x - \bar{x}_i) w_{i1} q_{i1} \quad \Rightarrow$$

$$f_{i2}(x) = e_{i1} - e_{i1} w_{i2} q_{i2} ... \Rightarrow f_{iA}\ (x) = f_{i(A_i-1)}(x)$$

$$= f_{i(A_i-1)}(x) - e_i(A_{i-1}) \quad i = 1,2,...c. \tag{37}$$

Where A_i is the number of the weight loadings embedding most of the information in the subset i. While $\hat{y}_i$ is the predict value for the object calculated from the corresponding linear model of the subset i which is estimated by the splitting algorithm. However, e_{i1}, e_{i2}, ... $e_{i(A_i-1)}$ are the residuals in the x-domain, they can be calculated from eqs. (25), but the loading weights (W_i) are used instead of the eigenvectors (P_i). Now, one can say the new object x belongs to a subset i if

$$f_{iA_i}(x) \times f_{iA_i}(x) < f_{jA_j}(x) \times f_{jA_j}(x) \text{for all } j \neq i. \tag{38}$$

3. Data sets

Three different data sets have been invoked to check out the efficiency of the proposed strategies. The used data sets can not be represented by a single linear calibration model since the corresponding errors were unacceptable as shown below. Therefore, the splitting algorithm has been utilized to split the data sets into the optimum linear subsets. To check the proposed strategies properly, the obtained subsets have been further divided into training and test ones.

3.1. Simulated data

This is a simple data set for detailed explanation of the second and third strategies. It consists of two subsets X_1 and X_2. The data points in each subset validate a specific hyperplane in the x-space eq. (39 and 41). Therefore, one can say this data set can not be represented by a single calibration model especially the chosen hyperplanes are not parallel. Training sets and test one have been established by randomly choosing data points validating the hyperplanes' equations. The corresponding y values for each subset have been calculated by using a random linear model for each subset. The used hyperplanes and linear models in these simulated data are given below

The first subset

$$\textit{Hyperplane} \ x_1 + x_2 - 2x_3 = 5.5 \ . \tag{39}$$

$$\textit{Linear mode} \, y = x_1 + x_2 - 0.5 \, . \tag{40}$$

The second one

$$\textit{Hyperplane}\ x_1 + x_2 - 3x_3 = 3. \tag{41}$$

$$\textit{Linear model}\ y = 2x_1 - 3x_2 + 3. \tag{42}$$

The training subsets and the corresponding y values are

$$X_1 = \begin{pmatrix} 10 & 8.5 & 6.5 \\ 8 & 5.5 & 4 \\ 6 & 2.5 & 1.5 \\ 9 & 4.5 & 4 \\ 7 & 1 & 1.5 \end{pmatrix} \qquad y_1 = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}$$

$$X_2 = \begin{pmatrix} 10.5 & 6 & 4.5 \\ 12.5 & 7 & 5.5 \\ 7 & 3 & 2.34 \\ 9 & 4 & 3.34 \\ 5 & 1 & 1 \end{pmatrix} \qquad y_2 = \begin{pmatrix} 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{pmatrix}. \tag{43}$$

More four data points have been randomly chosen validating the aforementioned hyperplanes to be the test set. The first two points belong to the first subset, while the remaining ones belong to the second one. The chosen test set is

$$\text{The test set} = \begin{pmatrix} 6 & 3 & 1.75 \\ 5.5 & 2 & 1 \\ 6 & 1.5 & 1.5 \\ 13 & 7.16 & 5.72 \end{pmatrix}. \tag{44}$$

3.2. *The vibration frequency data*

This data set consists of 75 samples of tetrahedral halide species with tetra-coordinated central atoms, the data set is in 6 dimension space [22].

3.3. *Descriptors and retention indices of alkenes*

Many molecular descriptors have been reported to describe the relationship between molecular structure and retention behavior, such as, molecular connectivity indices series, kappa indices series and quantum chemical parameters. In order to obtain a regression model with good fitting and predicting ability, 20 topological indices were used as descriptors of molecular structure for alkenes compounds. Therefore, these data are in 21 dimension space [23]

4. Results and discussion

The simulated data have been used to explain in details the second and third strategies. Table 1 shows the result obtained when the PCR have been used as bilinear modeling technique, as shown in the table; two eigenvectors embedded most of the information in the first subset since the residuals became negligible after the second eigenvector. Table 2 shows the result obtained when the same method have been applied for the second subset, as shown in the table also two eigenvectors having most of the variance. As aforementioned, the first strategy discriminate a new object according to the residuals obtained from eq. (25) and the constraint in eq. (26), i.e. the object belongs to the subset producing a minimum residual. Table 5 indicates that the strategy has discriminated the test set successfully since the residuals ($e_{12} \times e'_{12}$) of first two data points in the training set were very small when the eigenvectors of the first subset have been utilized in eqs. (25). However, when the eigenvectors corresponding to the second subset have been utilized; the obtained residuals ($e_{22} \times e'_{22}$) were relatively higher. That means the proposed strategy has discriminated the first two points successfully. Likewise, the remaining data points, the strategy has allocated it successfully as shown in table 5.

Table 1
PCR calibration results for the first subset

	x_1	x_2	x_3	Y	
Average $\overline{x},\overline{y}$	8	4.5	3.5	3	
Object no	X_0			y_0	t_1
1	2	4	3	-2	5.4
2	0	1	0.5	-1	1
3	-2	-2	-2	0	-3.4
4	1	0	0.5	1	0.67
5	-1	-3	-2	2	-3.7
The first eigenvector, p_1	0.4	0.73	0.56	q_1 = -0.33	
Object no	X_1			y_1	t_2
1	-0.123	0.082	-0.02	-0.205	-0.15
2	-0.398	0.266	-0.06	-0.664	-0.48
3	-0.673	0.45	-0.11	-1.123	-0.82
4	0.734	-0.491	0.12	1.225	0.89
5	0.459	-0.307	0.08	0.766	0.56
The second eigenvector, p_2	0.82	-0.55	0.14	q_1 = 1.374	
Object no	X_2			y_2	
1	0	0	0	0	
2	0	0	0	0	
3	0	0	0	0	
4	0	0	0	0	
5	0	0	0	0	

Table 2
PCR calibration results for the second subset

	x_1	x_2	X_3	Y	
Average $\overline{x},\overline{y}$	8.8	4.2	3.334	3	
Object no	X_0			y_0	t_1
1	1.7	1.8	1.167	-2	-2.7
2	3.7	2.8	2.167	-1	-5.2
3	-1.8	-1.2	-1	0	-2.4
4	0.2	-0.2	0	1	-0.02
5	-3.8	-3.2	-2.34	2	5.5
The first eigenvector**,** p_1	0.7	0.57	0.43	q_1=-0.31	
Object no	X_1			y_1	t_2
1	-0.21	0.25	0.01	-1.16	-0.32
2	-0.11	0.13	0	0.59	0.16
3	-0.13	0.16	0	-0.74	-0.21
4	0.18	-0.22	0.01	1	0.28
5	0.05	0.06	0	0.3	0.08
The second eigenvector, p_2	0.64	-0.76	-0.04	q_2 = 3.58	
Object no	X_2			y_2	
1	0	0	0	0	
2	0	0	0	0	
3	0	0	0	0	
4	0	0	0	0	
5	0	0	0	0	

Tables 3 and 4 show the results obtained when PLSR have been used to estimate weight loadings having most of the variances in the first and second subsets. As shown in the tables for both subsets two weight loadings embedding most of the information since the y residuals were so small after the second loading weight. As aforementioned, in the third strategy the discrimination process depends on y values and x vector of the new object. Therefore, when it was supposing that all the data points belonging to the first subset, the linear model describing this subset eq. (40) has been used to estimate the values of y for the data points (y_1 in table 5). Then, eq. (37) has been exploited to estimate the y-residuals obtained with the test set, however, the residuals in x-domain in such a case have been estimated from equations 25 by using the loading weights of the first subset. By the same way, eqs. (42, 37 and 25) have been used to estimate y-residuals and residuals in x-domain when it was supposing that the data points belonging to the second subsets. As sown in table 5, the third strategy discriminated all the data points in the test set successfully since the residuals

Table 3
PLSR calibration results for the first subset

	x_1	x_2	X_3	Y	
Average $\bar{x},\bar{y}$	8	4.5	3.5	3	
Object no	X_0			y_0	t_1
1	2	4	3	-2	-5.3
2	0	1	0.5	-1	-1
3	-2	-2	-2	0	3.2
4	1	0	0.5	1	-0.5
5	-1	-3	-2	2	3.7
The first loading weight, w_1	-0.27	-0.8	-0.53		
P_1	-0.4	-0.74	-0.56	q_1 = 0.34	
Object no	X_1			y_1	t_2
1	-0.1	0.05	-0.03	-0.16	-0.15
2	-0.42	0.21	-0.1	-0.63	-0.48
3	-0.74	0.37	-0.18	-1.1	-0.82
4	0.79	-0.4	0.19	1.18	0.89
5	0.47	-0.24	0.12	0.71	0.56
The second loading weight,w_2	0.82	-0.55	0.14	q_2 = 1.374	
P_2	0.82	-0.55	0.14		
Object no	X_2			Y_2	
1	0	0	0	0	
2	0	0	0	0	
3	0	0	0	0	
4	0	0	0	0	
5	0	0	0	0	

Table 4
PLSR calibration results for the second subset

	x_1	x_2	X_3	Y	
Average $\bar{x},\bar{y}$	8.8	4.2	3.334	3	
Object no	X_0			y_0	t_1
1	1.7	1.8	1.167	-2	-2.73
2	3.7	2.8	2.167	-1	-5.1
3	-1.8	-1.2	-1	0	2.36
4	0.2	-0.2	0	1	-0.01
5	-3.8	-3.2	-2.34	2	5.49
The first loading weight , w_1	-0.67	-0.6	-0.42	q_1=-0.31	
P1	-0.7	-0.57	-0.42		
Object no	X_1			y_1	t_2
1	-0.22	0.24	0	-1.16	-0.32
2	0.11	-0.12	0	0.59	0.16
3	-0.14	0.15	0	-0.74	-0.21
4	0.19	-0.21	0	1	0.28
5	0.06	-0.06	0	0.3	0.1
The second loading weigh ,w_2	0.67	-0.74	-0.02		
P_2	0.67	-0.74	-0.02	q_2 = 3.56	
Object no	X_2			Y_2	
1	0	0	0	0	
2	0	0	0	0	
3	0	0	0	0	
4	0	0	0	0	
5	0	0	0	0	

Table 5
The results obtained from the second and the third strategies for the training set of the simulated data

Object	Actual subset	The second strategy PCR			The third strategy PLSR				
		The first subset	The second subset	Pr. subset	The first subset		The second subset		Pr. subset
		$e_{12}{}^*e'_{12}$	$e_{22}{}^*e'_{22}$		y_1	$f_{12_1}(x) \times f_{12_1}(x)$	y_2	$f_{22_2}(x) \times f_{22_{21}}(x)$	
1	1	1×10^{-24}	0.05	1	2.5	0.005	6	0.0764	1
2	1	4×10^{-30}	0.026	1	3	0.054	8	0.14	1
3	2	0.167	1×10^{-5}	2	4.5	1.024	10.5	0.0009	2
4	2	1.75	3×10^{-6}	2	5.3	5.47	7.5	0.0014	2

obtained with the first two data points when the loading weights of the first subset have been utilized were very small, while they were relatively high when the loading weights of the second subset were used. Likewise, the remaining data points have allocated successfully as indicated from the residuals in table 5, this means the strategy has classified the test set probably.

To check properly the discrimination ability of the proposed strategies, two QSAR real data sets were used; each one has been divided training set and test one.

The first real data set is data 2, as mentioned above; this set consists of 75 samples. Since the residuals obtained when a single linear model estimated by least squares method was used to represent this data were unacceptable; the splitting algorithm was invoked to split this data set. The splitting algorithm has split data set into two linear subsets; the first one contains 33 data points while the second containing 42. Fig. 1 shows a comparison between the error percents obtained in a case of splitting the data into two subsets (the solid line) and when only a single linear model has been used to describe the data set (the light line). As shown in the figure, splitting the data set decreases the errors of the linear calibration. As mentioned above, the obtained subsets has been furthermore divided into training set and test one to check up properly the ability proposed discrimination strategies. Actually, three are two famous techniques used to check any proposed algorithm; namely; leave-one-out technique and subtracting a portion (usually 10%) from data points to be a test set (17, b). In the present study, the second technique was utilized. Therefore, the test set in this data set consists of seven data points; three points from the first subset and four ones from the second one. The remaining data points in the two subsets have been used as a training set. Since the splitting algorithm has split this data into two subsets the problem was reduced into two-class one, so, only one weight vector and its corresponding threshold were enough to achieve the discrimination analysis in the first strategy. The obtained results were satisfactory since all the data points in the test set have been discriminated successfully. For the second and the third strategies, they gave good results also since the two strategies classified all the data points in the test set successfully. The number of the used eigenvectors and the loading weights in the second and the third strategies were the same; they were four eigenvectors in second strategy and four loading weights in the third one.

The second real data set has been previously treated [10], the obtained results indicated that if this data set is split into three subsets (the first subset contains 65 data points, the second contains 35 and the third contains 49), the corresponding errors will be a relatively smaller than if only single linear model is used to represent the whole data set, as shown in fig. 2. This data set has been divided into training containing 135 samples (58 from the first subset, 32 from the second one and 45 from the third) and test set containing 14 randomly chosen samples (6 from the first, 3 from the second and 5 from the third). Since this data set is a multicategory one (containing 3 subsets), three discrimination function have been estimated in the linear machine strategy. The obtained results were satisfactory since only one data point from the training set has been misclassified (a data point had to be discriminated as a first subset was misclassified into a third subset) while the other 13 points have been discriminated successfully. The calculation in the second and the third strategies were a bit complicated since the data set is in relatively high dimensional space, in the second strategy the residuals in the *x*-space were small after 15 eigenvectors used, (i.e. E_{15} was very small). Also, in a case of the third strategy, the used loading weights were 15 to get small *y* residuals. Although the computational times in these strategies were relatively higher than the first one but the obtained result was better since these strategies have discriminated all the data points in the test set successfully.

5. Conclusions

Splitting the data into linear subsets may be used to get better chemical calibration since it decreases the error obtained. The problem of predicting of the membership of unknown object in a case of splitting the original data set was treated by three nonparametric proposed discrimination strategies to avoid the complexities of the statistical methods. The result obtained when the proposed strategies have been applied on the real data sets indicated that the strategies might be appended any splitting algorithm to properly calibrate the chemical data.

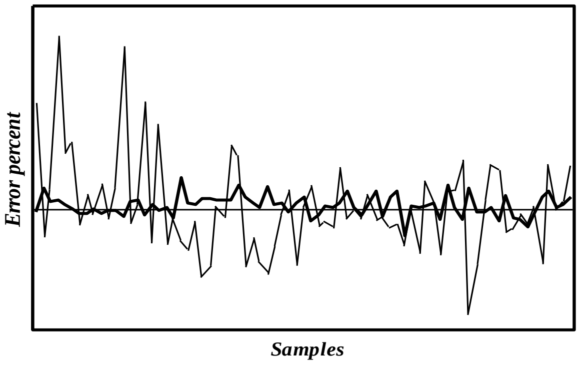

Fig. 1. The error percent obtained when one linear model represents data 2 (the light line) and when the data set was split into two subsets (the solid line).

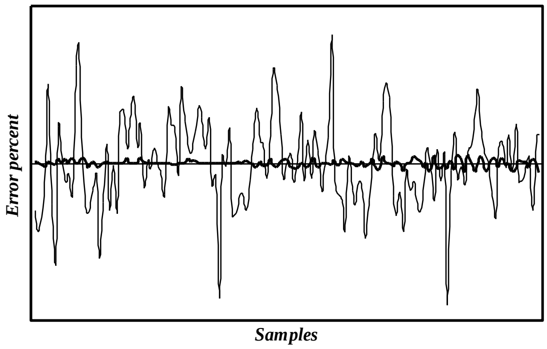

Fig. 2. The error percent obtained when one linear model represents data 3 (the light line) and when the data set was split into three subsets (the solid line).

References

[1] T. Næs and T. Isaksson, J. Chemometrics, 49, 5 (1991).

[2] A. Tishler and I. Zang, J. Am. Stat. Assoc., 980, 76 (1981).

[3] W.S. deSarbo, R.L. Oliver and A. Rongaswany, Psychometrika, 707, 54, 1989.

[4] J.C. Bezdek, C. Coray, R. Gunderson and J. Watson, SIAM. J. Appl. Math., 339, 40 (1981).

[5] J.C. Bezdek, C. Coray, R. Gunderson and J. Watson, SIAM. J. Appl. Math., 358, 40 (1981).

[6] S. Wold, J. Chromatogr. Sci.**,** 525, 13 (1975).

[7] J.H. Wang, Y.L. Xie and R.Q. Yu. J. Chemometrics, 373, 9 (1995).

[8] Tormod Næs, J. Chemometrics, 487, 5 (1991).

[9] J.H. Wang, Y.L. Xie and R.Q. Yu. J. Chemometrics, 373, 9 (1995).

[10] Nasser A.M. Barakat, Jian-Hui Jiang, Yi-Zeng Liang and Ru-Qin Yu, Chemometrics and Intell. Lab. Sys., 72, 1 (2004).

[11] M. Bos, Webber H.T., Anal. Chim. Acta, 97, 247 (1991).

[12] E. Fontain, Anal. Chim. Acta, 227, 265 (1992).

[13] D.B. Hibbert, Chemom. Intell. Lab. Syst., 277, 19 (1993).

[14] C.B. Lucasius, G. Kateman, Chemom. Intell. Lab. Syst. 1, 19 (1993).

[15] C. B. Lucasius, G. Kateman, Chemom. Intell. Lab. Syst., 99, 25 (1994).

[16] Korford, J.S. and G.F. Groner, IEEE Trans. Info. Theory, Jan., 42, IT-12 (1966).

[17] Richard O. Duda and Peter E. Hart, "Pattern Classification and Scene Analysis", John Wiley and Sons, New York, Chichester, Brisbane, Toronto, a. 178 b. 132, (1973).

[18] T. Næs and Matrens H., Communication in Statistics (Sim. And Comp.) 245, 14 (1985).

[19] Harald Martens and Tormod Næs, "Multivariate Calibration", John Wiley & Sons, New York, Singapore, Brisbane, Toronto, a.96, b.120 (1989).

[20] S. Wold, H.A. Martens and H. Wold, Proc. Conf. Matrix pencils (A. Ruhe, B. Kágström, eds), Lecture Notes in Mathematics, Springer Verlge, Heidelberg, 286 , March (1982).

[21] H. Martens and T. Næs, (ed. P.C. Williams. And K. Norris) Am. Assoc. Cereal Chem. St. Paul Minnesota, 57 (1987).

[22] Lu, Qing Zhang, Guo Li Shen, Ru Qin Yu, Journal of Computational Chemistry, 1365, 14, 135 (2003).

[23] Yi-ping Du, Yi-Zeng Liang, and Dong Yun, Computational Biology and Chemistry, 339, 27 (2003).

Received September 3, 2006
Accepted November 11, 2007