Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Chromatography A, 998 (2003) 155–167
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

Two-step multivariate adaptive regression splines for modeling a quantitative relationship between gas chromatography retention indices and molecular descriptors

Qing-Song Xu[a], D.L. Massart[a,*], Yi-Zeng Liang[b], Kai-Tai Fang[c]
[a]*ChemoAC, FABI, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium*
[b]*College of Chemistry and Chemical Engineering, Central South University, Changsha, China*
[c]*Department of Mathematics, Hong Kong Baptist University, Kowloon, Tong, Hong Kong*

Received 5 December 2002; received in revised form 31 March 2003; accepted 1 April 2003

Abstract

The relationship between retention indices and molecular descriptors of alkanes is established by two-step multivariate adaptive regression splines (TMARS). TMARS combines linear regression with multivariate adaptive regression splines (MARS). It is demonstrated for the present data set that using linear regression or MARS modeling alone causes lack of fit. TMARS avoids lack of fit and appreciably improves the prediction ability for the model. The use of this combined approach permits the development of additional understanding of the adaptive nature in MARS modeling.

Keywords: Regression analysis; Retention indices; Molecular descriptors; Multivariate adaptive regression splines; Alkanes

1. Introduction

Constructing quantitative relationships between molecular structure and gas chromatographic retention indices has been studied repeatedly [1–4]. The main goal is to develop a suitable model to predict the retention behavior and to explain the molecular mechanisms in gas chromatography.

The common approach for building a structure–retention relationship consists of the following steps: (1) to develop (or to select) the descriptors for the molecular structure; (2) to use proper mathematical methods to set up the model; (3) to evaluate the model built. This study is concerned with the latter two steps.

Linear methods, such as multiple linear regression (MLR), partial least squares and principal component regression are the more evident ones when searching for a relationship between molecular structure and gas chromatographic retention. The descriptors are included into the multiple linear regression model using variable selection procedures such as best subset, backward and stepwise selection [5,6] or more sophisticated ones that use genetic algorithms and simulated annealing [7,8]. Then the multiple correlation coefficient R and F-test value are computed to evaluate the model built with the selected descriptors. If R is very close to unity (for instance

*Corresponding author. Tel.: +32-2-477-4737; fax: +32-2-477-4735.
E-mail address: fabi@vub.vub.ac.be (D.L. Massart).

0021-9673/03/$ – see front matter
doi:10.1016/S0021-9673(03)00604-6

$R>0.99$) and the F-test value is larger than several hundreds or thousands, the model is regarded as very good. However, many examples show that there can still be unacceptably large residuals for some chromatographed substances compared to measurement errors. This means that there is a lack of fit for the model. One can increase the fit by including more descriptors into the model, but this does not give a better prediction. Indeed, when the root mean square error of cross validation (RMSECV) serves as a criterion to determine the model dimension, it will be found that more descriptors may lead to poorer prediction performance.

One possible reason for lack of fit is that the available descriptors give an incomplete description of molecular structure. Thus, seeking more informative descriptors for chemical structure has long been the aim of many researchers, and has led to the development of many molecular descriptors.

Another reason is that the linear model has limited flexibility to characterize the relationship between molecular structure and gas chromatographic retention index. It is the simplest and the most popular, but nonlinear methods are more general. There are two well-known methods, which have been used to a large extent in various disciplines during the last decade. One is neural networks [4,9]. The other is multivariate adaptive regression splines (MARS) [10–14]. While neural networks have been studied extensively in chemometrics, this is not the case for MARS.

The aim of this study is to develop a new strategy of MARS modeling which is called two-step MARS (TMARS). TMARS attempts to build a model between retention index and molecular descriptors based on linear modeling in a first step. In a second step, when it is found that the model shows lack of fit, splines are added to the model.

2. Methods

The general model to be constructed is:

$$y = f(\mathbf{x}) + e \tag{1}$$

where y is the retention index, e the measurement error and $\mathbf{x}$ a vector of molecular descriptors. As stated above, the exact form of $f(\mathbf{x})$ is not known. The goal of regression analysis is to build a function $\hat{f}(\mathbf{x})$ based on the available data set as an approximation to $f(\mathbf{x})$ that can perform well over the domain of interest.

2.1. Multiple linear modeling

The most prevalent way to approximate the relationship is to use a linear function:

$$y = f(\mathbf{x}) = a_0 + \mathbf{x}^t\mathbf{a} + e \tag{2}$$

where a_0 is the intercept, $\mathbf{a}$ the $p \times 1$ vector of regression coefficients, $\mathbf{x}$ the $p \times 1$ descriptor vector, p the number of descriptors and the superscript t stands for transpose. If there is more than one compound,

$$\mathbf{y} = f(\mathbf{X}) = \mathbf{1}a_0 + \mathbf{X}\mathbf{a} + \mathbf{e} \tag{3}$$

where $\mathbf{y}$ is the $n \times 1$ retention index vector for n compounds, $\mathbf{X}$ is the corresponding $n \times p$ descriptor matrix, $\mathbf{1}$ is the vector of ones, and $\mathbf{e}$ is the error vector.

In this study, the forward stepwise algorithm is used to select the descriptors included in the model.

2.2. MARS modeling

MARS uses left-sided (Eq. (4)) and right-sided (Eq. (5)) truncated power functions as spline basic functions

$$b_q^-(x-t) = [-(x-t)]_+^q = \begin{cases} (t-x)^q, \text{ if } x < t, \\ 0, \text{ otherwise} \end{cases} \tag{4}$$

$$b_q^+(x-t) = [+(x-t)]_+^q = \begin{cases} (x-t)^q, \text{ if } x > t, \\ 0, \text{ otherwise} \end{cases} \tag{5}$$

where q (≥ 0) is the power to which the splines are raised in order to manipulate the degree of smoothness of the resultant function estimate. When $q = 1$, which is the case in this study, a simple linear spine is applied, t is called the knot location. Fig. 1 shows a pair of spline functions when $q = 1$ at $t = 0.5$.

For model (3), a total of np pairs of spline basic functions, $\{[+(x_j - t)]_+, [-(x_j - t)]_+\}$ corresponding to the knot location $t = x_{ij}$ ($i = 1, 2, \ldots, n$,

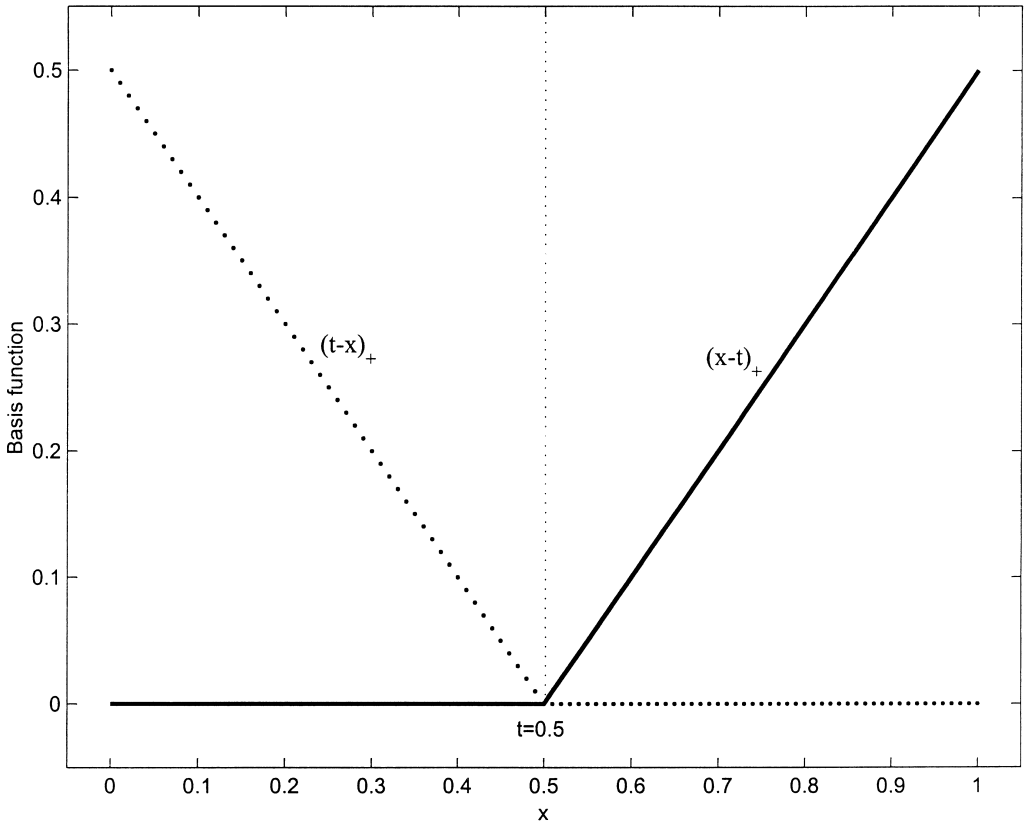

Fig. 1. A pair of one-sided spline basis functions $(0.5 - x)_+$ and $(x - 0.5)_+$.

$j = 1, 2, \ldots, p$), where x_j is the jth descriptor in Eq. (3).

The basis functions for MARS consist of either one single spline function or the product of two (or more) spline functions of different descriptors. The fundamental idea of MARS is to use the combination of basis functions to approximate model (1)

$$\hat{f}_M(\mathbf{x}) = a_0 + \sum_{m=1}^{M} a_m B_m(\mathbf{x}) \tag{6}$$

where a_0 is the coefficient of the constant basis function, $B_m(\mathbf{x})$ the mth basis function which may be a single spline function or product (interaction) of two (more) spline basic functions, a_m the coefficient of the basis function and M the number of basis functions included into the model.

MARS first uses a ''two at a time'' forward stepwise strategy to select a pair of basis functions into the model. The pair of basis functions is the one that fit the model best at the current stage. When the model has become excessively large and obviously overfits the data, MARS then uses a ''one at a time'' backward stepwise strategy to prune the basis functions. The generalized cross validation (GCV) is the mean squared residual of fit to the data divided by a penalty to account for the increased model complexity. This criterion is used to avoid an excessive number of spline basis functions

$$\mathrm{GCV}(M) = \frac{1}{n} \cdot \frac{\sum_{m=1}^{n} [y_i - \hat{f}_M(\mathbf{x}_i)]^2}{[1 - C(M)/n]^2} \tag{7}$$

where $C(M)$ is a complexity penalty function which increases as the number of terms. It is defined as

$$C(M) = M + dc \tag{8}$$

M is the number of terms in Eq. (6), c is the number of basis functions that consist of spline functions (or nonlinear terms). In this study, the parameter $d = 2$, the maximum interaction order of the spline functions is restricted to 3.

As more spline basis functions are included into the model, the bias of model estimates decreases, but the variance increases. The GCV could supply a suitable data-dependent estimate of future prediction error if the penalty function is well defined and this prediction error estimate is minimized with respect to the parameters of the strategy. In summary, MARS yields a model for the response that automatically selects the spline basis functions included into the final model. This model balances GCV against the bias of model estimates. Further details on MARS modeling are given in Ref. [10].

2.3. A two-step modeling procedure

As stated above, the linear model fits well the relationship between retention index and molecular descriptors, but some residuals are still too large. In other words, the relationship appears highly linearly correlated, but the linear model shows some lack of fit. The reason why some residuals are too large may be that some intrinsic relation hidden in the high dimensional data may not be characterized by linear function. In order to deal with such a problem, a two-step modeling procedure based on linear regression and MARS was developed.

The procedure is as follows.

2.3.1. First step

1. The multiple linear model is constructed as described in Section 2.1

$$f_L(\mathbf{x}) = a_0 + \sum_{i=1}^{L} a_{n_i} x_{n_i} \tag{9}$$

where x_{n_i} is the n_ith descriptor, L is the number of descriptors included into the model.

2. The residual sum of squares of model (9) is decomposed into pure error sum of squares and lack of fit sum of squares:

$$\boxed{\text{Residual sum of squares}} = \boxed{\text{Pure error sum of squares}} + \boxed{\text{Lack of fit sum of squares}}$$

or

$$\sum_{i=1}^{N}\sum_{j=1}^{m_i}(y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^{N}\sum_{j=1}^{m_i}(y_{ij} - \bar{y_i})^2 + \sum_{i=1}^{N} m_i(\hat{y_i} - \bar{y_i})^2 \tag{10}$$

where N stands for the total number of different objects, $y_{i1}, y_{i2}, \ldots, y_{im_i}$ are m_i replicate observations of the ith object $\mathbf{x}_i$ ($i = 1, 2, \ldots, N$), $\hat{y}_i$ and $\bar{y_i}$ are the estimate and the mean of these m_i observations, respectively.

3. A test for lack of fit is carried out with the F-ratio of mean squares for lack of fit (the sum of squares of lack of fit divided by its number of the degrees of freedom) and mean squares for pure error (the sum of squares of pure error divided by its number of the degrees of freedom). If the test is significant, the model is inadequate. Go to the second step. If not, the linear model is adequate and the second step is not required.

2.3.2. Second step

1. For Eq. (9), use forward stepwise procedure to determine whether some descriptors x_{n_i} ($i = 1, 2, \ldots, L$) should be replaced by a pair of one-sided linear splines $[\pm(x_{n_i} - x_{jn_i})]_+$ ($j = 1, 2, \ldots, n$). Thus the following equation is obtained

$$\hat{f}_1(\mathbf{x}) = c_0 + \sum_{k=1}^{K} c_k g_k(\mathbf{x}) \tag{11}$$

where the basis function $g_k(\mathbf{x})$ is either one of the descriptors x_{n_i} or a pair of the spline basic functions $[\pm(x_{n_i} - x_{jn_i})]_+$.

2. On the basis of Eq. (11), a combined model of form

$$\hat{f}_{cf}(\mathbf{x}) = \hat{f}_1(\mathbf{x}) + \hat{f}_f(\mathbf{x}) = c_0 + \sum_{k=1}^{K} c_k g_k(\mathbf{x}) + \sum_{m=1}^{M} a_m B_m(\mathbf{x}) \tag{12}$$

where

$$\hat{f}_f(\mathbf{x}) = \sum_{m=1}^{M} a_m B_m(\mathbf{x}) \tag{13}$$

can be fit to the data by applying the ‘‘two at a time’’ forward stepwise procedure. The coeffi-

cients c_k in Eq. (12) are jointly adapted alone with the parameters of the resulting model in forward stepwise procedure.

3. The equation can be rewritten as

$$\hat{f}_{cf}(\mathbf{x}) = b_0 + \sum_{j=1}^{M+K} b_j H_j(\mathbf{x}) \quad (14)$$

where $H_j(\mathbf{x})$ is the basis function $g_j(\mathbf{x})$ or $B_j(\mathbf{x})$. A ''one at a time'' backward stepwise deletion is applied to Eq. (13). Both basis functions $g_j(\mathbf{x})$ and $B_j(\mathbf{x})$ can be deleted within this stage.

For more details on forward and backward stepwise procedures see Ref. [10].

The first step tries to use a linear function to describe the relationship between the response and descriptors. Of course other regression subset selection strategies can be used to build the model. If the linear model shows lack of fit, the second step is started. In that second step a descriptor in the linear model is replaced by a pair of spline basic functions, if the resulting model is improved. Then the usual MARS procedure is completed based on the resulting model. The descriptors that remained in Eq. (8) after the deletion procedure are important factors and are needed to make the model accurate.

It should be pointed out that the final TMARS model has the form of what is called ''semi-parametric model'' in Ref. [10], but that the TMARS procedure is different from the semi-parametric modeling presented there.

3. Data

This data set contains retention indices of 149 alkanes including straight chain and branched alkanes, which were measured on squalane as stationary phase and at a column temperature of 333 K. They were collected and collated from 1587 retention index records in a GC retention index database [15]. The 1587 retention indices of alkanes were measured on the same column of squalane, but at different temperatures and in different laboratories. Thus the calibration of temperature for some compounds is necessary. We used regression between the retention index and temperature to compute retention indices at a fixed temperature of 333 K. The model used is as follows [15]:

$$I = a + \frac{b}{T} \quad (15)$$

where I is the retention index of the compound measured at temperature T (K), a and b are constants. Fig. 2a shows one result obtained. For most compounds, the linear regression lines fit the retention indices very well.

However, mistakes both in data transformation and in the reference sources may appear in some retention indices. In order to estimate the variance of the error and test lack of fit for consequent modeling, the retention indices of 14 alkanes are selected as a repeat set. For each compound in this set, the retention indices at different temperatures are partitioned into 2–4 groups. Each group yields a regression line. The retention indices predicted at 333 K by these lines are considered replicate observations of this compound. Fig. 2b shows an example. In this way retention indices of in total 173 alkanes were collected, out of which 149 are different alkanes and 24 are repeat measurements of these alkanes. The 173 retention indices are listed in Appendix A.

Two kinds of descriptors are calculated for the molecules, namely topological and quantum chemical descriptors. The first are the Kier and Hall [16] molecular connectivity indices ${}^1\chi$, ${}^2\chi$, ${}^3\chi_p$, ${}^3\chi_c$; kappa indices [17] ${}^0\kappa$, ${}^1\kappa$, ${}^2\kappa$, ${}^3\kappa$; path count indices [18] p_1, p_2, p_3, p_4; walk count indices [18] w_1, w_2, w_3, w_4; path/walk count indices [18] pw_1, pw_2, pw_3, pw_4, the indices proposed and used by Schultz et al. [19,20]: molecular topological index (MTI); the principal eigenvalue of the distance matrix (PED); the principal eigenvalue of the adjacency-plus-distance matrix (PEAD); the logarithm of determinant of the adjacency-plus-distance matrix (DET), the indices proposed by Xu et al.: *Yx* and EAID [21,22]. The quantum chemical descriptors are: heat of formation (HEAT), electronic energy (ELE), core–core repulsion energy (CORE), dipole moment (DIP), ionization potential (ION), LUMO energy (LUMO). The six quantum chemical descriptors were calculated using the MOPAC method in the Chem3D software. The software performed geometry optimization, followed by quantum chemical

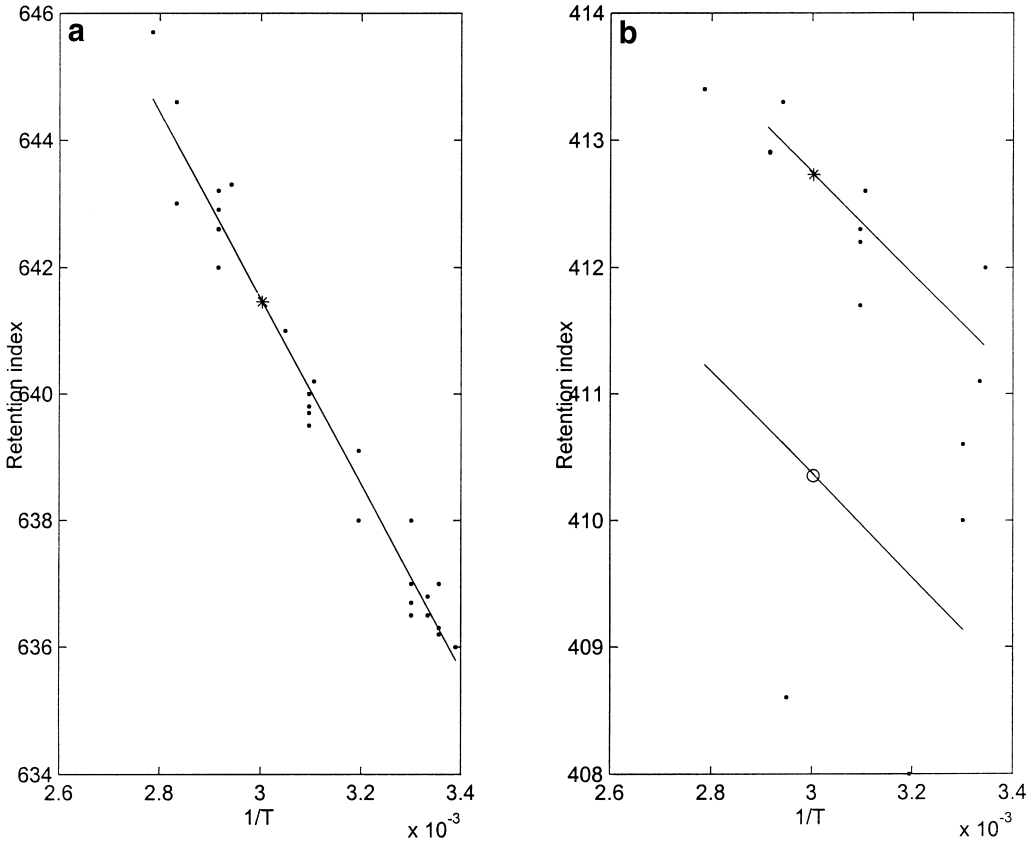

Fig. 2. (a) The regression line of the retention indices at different temperatures for 2,2,3-trimethyl-pentane. The vertical coordinate of the point "*" is the retention index used in the data set. (b) The regression lines of the retention indices at different temperatures for neopentane. The vertical coordinates of the points "*" and "○" are repeat observations (retention indices) of neopentane used in the data.

calculations according to the semiempirical AM1 method. The solvent probe radius used is 1.4 Å, which is the default value for water. The routines for calculating the other 26 topological indices were programmed using MATLAB language of version 5.3.

All the descriptors are labeled x_1–x_{32} according to the order in which they are described in the above paragraph.

4. Results and discussion

4.1. Multiple linear regression—step 1

Using forward stepwise procedures with GCV (Eq. (7)) as the criterion, the following linear model is obtained:

$$I = 74.87 + 171.19x_1 - 4.13x_3 + 6.03x_7 - 1.82x_8 + 20.64x_{11} + 9.07x_{12} + 2.07x_{16} - 53.21x_{20} - 0.17x_{21} + 16.03x_{23} - 187.42x_{24} + 164.45x_{25} - 0.64x_{27} - 0.05x_{29} - 25.72x_{30} + 153.87x_{31} + 42.57x_{32}$$

$$R^2 = 0.9994; \quad F = 14051; \quad s = 5.09 \qquad (16)$$

where R^2 is the multiple correlation coefficient, s is the standard error and F is the F-ratio for overall regression. In total, 17 descriptors, $\chi(x_1)$, ${}^3\chi_p(x_3)$, ${}^2\kappa(x_7)$, ${}^3\kappa(x_8)$, $p_3(x_{11})$, $p_4(x_{12})$, $w_4(x_{16})$, $pw_4(x_{20})$, MTI(x_{21}), PEAD(x_{23}), DET(x_{24}), $Yx(x_{25})$,

HEAT(x_{27}), CORE(x_{29}), DIP(x_{30}), ION(x_{31}), LUMO(x_{32}), are included in the linear model.

The values of R^2 and F indicate that the relationship between retention index and molecular descriptors is highly linearly correlated. Fig. 3 shows the results of predictions (cross validation) of retention. The root mean square error of cross validation (RMSECV) for the models is 5.54. This implies that the prediction ability for the models built is not bad. However, further investigation of these results indicates that the model is not good enough yet. Fig. 4a shows the residuals for the model. It is seen that there are many samples with residuals that are larger than 8 index units, much larger than the normal measurement errors. Table 1 describes the prediction behavior of a group of compounds. The absolute prediction errors are larger than 8 index units. To obtain clearer evidence, the lack of fit test is performed. Table 2 lists the analysis of variance (ANOVA) results. The F ratio is 3.34, clearly larger than 1.95 (the $\alpha = 0.01$ level of confidence that there is no lack of fit). The significant lack of fit indicates that the resulting model is inadequate.

4.2. Updating the model—step 2

In the second step, modeling the linear model is first updated by using forward stepwise procedure. Each descriptor x_{n_i} ($i = 1, 2, \ldots, L$) in Eq. (16) is replaced by a pair of one-sided spline functions of itself, if this improves the model. Then, the backward stepwise procedure is performed resulting in the equation:

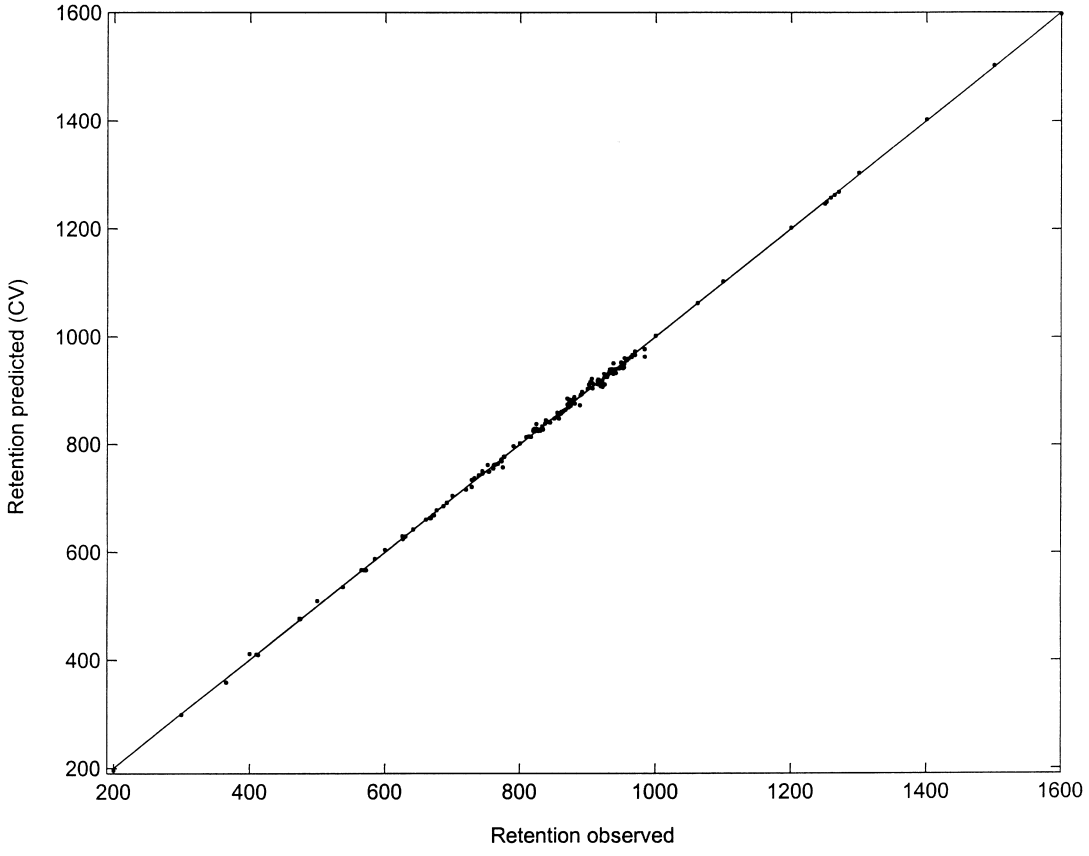

Fig. 3. The prediction behavior for the linear model.

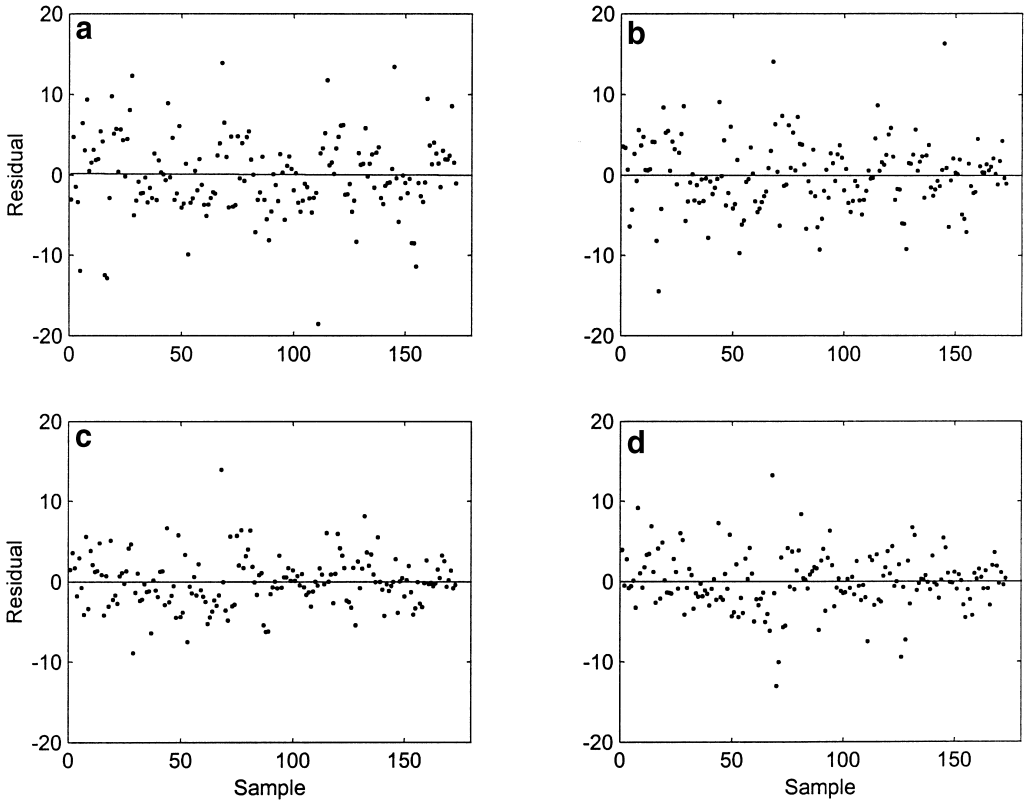

Fig. 4. The residual for the models. (a) The linear model. (b) The intermediate model. (c) The TMARS model. (d) The MARS model.

Table 1
The prediction behavior of a group of compounds by four models

Compound[a]	Prediction error			
	Linear model	Intermediate model	MARS model	TMARS model
22344m5C5	−14.760	−6.262	−1.329	6.803
2235m4C6	9.937	5.971	10.111	6.228
224m3-3eC5	12.100	9.450	7.965	5.456
22m2-3eC6	8.793	5.285	6.594	4.437
22m2-3eC5	13.528	8.887	5.495	4.906
233m3C5	9.198	9.602	7.898	7.246
2m-3eC7	−8.097	−9.645	−6.383	−6.522
3344m4C6	−20.747	6.294	−9.597	−4.851
33m2-4eC6	12.571	9.230	3.575	6.527
3m-3eC7	−9.830	−10.580	−8.817	−6.378
4eC7	−9.634	−5.413	0.120	−1.412
4eC8	−9.608	−5.914	−3.157	−4.415
4ipC7	−13.237	−7.844	−4.873	−3.532
C4	11.152	5.336	0.431	3.209
C5	9.392	4.637	1.355	1.591
Absolute mean	11.506	7.357	5.180	4.901

[a] 22344m5C5 is 2,2,3,4,4-pentamethylpentane and 224m3-3eC5 is 2,2,4-trimethyl-3 ethylpentane.

Table 2
ANOVA table for the linear and the intermediate model

Source	df	SS	Mean	F ratio
Linear model $R^2=0.9994$ RMSECV=5.54				
Regression	17	6.547×10^6	3.851×10^5	14 973
Residual	156	4012.6	25.72	
Lack of fit	132	3805.5	28.83	3.34
Pure error	24	207.05	8.63	Significant at $\alpha=0.01$
Total	173	6.551×10^6		
Intermediate model $R^2=0.9995$ RMSECV=4.86				
Regression	25	6.548×10^6	261 928	12 405
Residual	148	3124.8	21.114	
Lack of fit	124	2917.8	23.530	2.73
Pure error	24	207.05	8.63	Significant at $\alpha=0.01$
Total	173	6.551×10^6		

SS, sum of squares; df, the number of degrees of freedom.

$$\begin{aligned} I = {} & 559.91 + 169.92x_1 - 2.01x_8 + 17.29x_{11} \\ & + 16.65x_{12} - 38.33x_{20} - 0.19x_{21} + 16.76x_{23} \\ & + 143.69x_{25} - 160.34x_{24} - 0.048x_{29} \\ & - 33.90x_{30} + 165.91x_{31} + (x_7 - 3.41)_+ \\ & + (x_3 - 3.37)_+ + (107 - x_{16})_+ + (x_{16} - 107)_+ \end{aligned}$$

$$R^2 = 0.9995; \quad F = 12405; \quad s = 4.60 \qquad (17)$$

This model is called the intermediate model. It is a non-linear model. For the sake of comparison, the complexity penalty is used as a criterion to account for the degrees of freedom for the MARS regression model [10]. That is, it is used as the number of degrees of freedom based on which the standard errors, the F-ratio for overall regression F and the lack of fit test are calculated.

Comparing the intermediate model with the linear one, it is observed that the descriptors ${}^3\chi_p(x_3)$, ${}^2\kappa(x_7)$ and $w_4(x_{16})$ in the model are replaced by spline functions for the same descriptors. The descriptors HEAT(x_{27}) and LUMO(x_{32}), which were present in (16) are no longer in (17). The number of terms in (17) is one less than in (16). Although the value of R^2 is a little higher and the F value is smaller, the improvement of fit is clear since s is much smaller. RMSECV is 4.86. This indicates that the prediction ability of the updated model is better owing to better fit. Table 1 lists the improvements in prediction for the group of compounds for which the prediction errors are larger than 8 index units in the linear model. Fig. 4b shows the improvement of the residuals visually.

However, the lack of fit test again reveals the inadequacy of the updated model. Table 3 lists the ANOVA results. The F ratio for testing lack of fit is 2.73, and is significant at $\alpha=0.01$.

The two-step MARS goes on by expanding, on the basis of the intermediate model, with basis functions two at a time that fit the data best. In this study, the number of basis functions is predefined to be 50. The complexity of the model is of course too large and would lead to overfitting. Then, a backward stepwise procedure starts to delete the excessive basis functions one at a time.

First the GCV criterion was used. However, the results obtained were not as good as expected. The pruned model contains 33 basis functions, which still seems excessive. Furthermore, RMSECV for the model is 9.41, but s is very low (2.91). This indicates that some basis functions in the pruned model have a negative influence on the prediction ability. They are included into the model only because they improve the fit of the model. This typically leads to overfitting, that is, the model fits well, but predicts poorly. The GCV criterion optimizes fitting rather than prediction, and therefore the backward deletion procedure does not remove

Table 3
ANOVA table for the TMARS and the MARS model

Source	df	SS	Mean	F ratio
TMARS model $R^2=0.9997$ RMSECV=4.19				
Regression	45	6.550×10^6	1.456×10^5	10 849
Residual	128	1717.3	13.42	
Lack of fit	104	1510.2	14.52	1.76
Pure error	24	207.05	8.63	Not significant at $\alpha=0.01$
Total	173	6.551×10^6		
MARS model $R^2=0.9997$ RMSECV=4.36				
Regression	67	6.549×10^6	97 750	5036.1
Residual	106	2057.5	19.41	
Lack of fit	82	1850.5	22.57	2.73
Pure error	24	207.05	8.63	Significant at $\alpha=0.01$
Total	173	6.551×10^6		

enough basis functions. For the sake of the flexibility of the MARS modeling, the criterion GCV in (7) can be replaced by one that minimizes another loss function. Cross validation, for instance, would provide better prediction performance for the model. However, it is difficult to use during the forward selection because too many descriptors must be considered, but it can be used in the backward selection where less descriptors are involved. Thus in backward stepwise stage, 10-fold cross validation [23] is used after GCV. The model obtained is as follows:

$$\begin{aligned} I = {} & 372.05 - 1.79x_8 + 29.55x_{11} - 37.21x_{20} \\ & - 0.24x_{21} + 48.48(x_2 - 4.90)_+(107 - x_{16})_+ \\ & + 20.22x_{23} - 0.31(4.90 - x_2)_+(107 - x_{16})_+ \\ & + 118.84x_{25} + 0.09(0.82 - x_4)_+(10.74 - x_{30})_+ \\ & - 0.02x_{29} - 114.67(37 - x_{15})_+(x_{20} - 1.42)_+ \\ & - 27.79x_{30} + 2.46(x_{16} - 107)_+(3.48 - x_{19})_+ \\ & + 48.33(0.82 - x_4)_+ - 29.65(x_4 - 0.82)_+ \\ & - 6.69(107 - x_{16})_+ + 6.89(x_{16} - 107)_+ \\ & - 21.86(x_{15} - 37)_+ + 22.55(37 - x_{15})_+ \\ & + 87.012(x_3 - 3.37)_+ \end{aligned}$$

$$R^2 = 0.9997; \quad F = 10849; \quad s = 3.66 \qquad (18)$$

The ANOVA results listed in Table 4 show that the lack of fit F-ratio = 1.76 is not significant. RMSECV is 4.19, which indicates that the prediction ability has been improved. Fig. 4c shows the residuals of the model are closer to zero. Furthermore, we can see from Table 1 the manifest improvement by the TMARS model in prediction for the group of compounds that are not well predicted by the linear model.

Exploring the model of Eq. (18) provides some insight into the nature of these improved results. It includes basis functions involving four new descriptors: ${}^2\chi(x_2)$, ${}^3\chi_c(x_4)$, $w_3(x_{15})$ and DIP(x_{29}), and it no longer uses the descriptors ${}^1\chi(x_1)$, ${}^2\kappa(x_7)$, $p_4(x_{12})$, DET(x_{24}) and ION(x_{31}). The model also contains five interacting spline functions. These non-linear

Table 4
The prediction behavior of a group of compounds by MARS and TMARS

Compound	Prediction error	
	MARS model	TMARS model
2235m4C6	10.111	6.228
24m2-3ipC5	−14.218	−4.333
24m2-4eC6	−10.477	−5.309
25m2-3eC6	9.611	6.735
3344m4C6	−9.597	−4.851
34e2C6	−13.129	−3.893
3m-3eC7	−8.817	−6.378
C16	−9.344	1.226
Absolute mean	10.663	4.869

terms of higher order can tackle the variation in the data set that could not be handled by model (17).

4.3. Comparison of MARS and TMARS

In order to obtain a fair comparison, the MARS method is accomplished on the data under the same conditions. The ANOVA results are shown in Table 2. The lack of fit test for the MARS model is significant. The standard error s and RMSECV are 4.41 and 4.36, respectively. Fig. 4d shows the residuals for the model. The performance of this model is distinctly better than the linear model and the intermediate model. Table 1 also shows the clear improvement by the MARS model for the group of compounds that are not well predicted by the linear model. However, the MARS model is worse than the TMARS model. Table 4 gives more evidence. The prediction behavior of this group of compounds is not acceptable using the MARS model, but it is acceptable when using the TMARS model. The MARS model is shown in Eq. (19). It is more complex than the TMARS model. It uses more descriptors, more basis functions and higher orders of interaction of spline functions, but it does not give better results than the TMARS model. Since the TMARS model is less complex than the MARS model, the former should be preferred

$$
\begin{aligned}
I = {} & 912.27 + 1.48(x_{16} - 108)_+ - 17.78(14 - x_{11})_+ \\
& - 1.36(108 - x_{16})_+ - 0.03(-5901.9 - x_{28})_+ \\
& + 31.06(x_{11} - 14)_+ + 8.38(x_4 - 1.21)_+(x_{12} - 7)_+ \\
& + 18.37(x_{23} - 16.35)_+ + 25.17(1.21 - x_4)_+ \\
& - 15.87(16.35 - x_{23})_+ + 0.03(x_{28} + 5901.9)_+ \\
& - 194.28(6.99 - x_5)_+(x_{24} - 4.90)_+ \\
& - 1.98(1.21 - x_4)_+(x_{22} + 15.41)_+ \\
& - 0.02(108 - x_{16})_+(x_{27} + 67.18)_+ \\
& + 2.02(x_{23} - 16.35)_+(3.86 - x_2)_+ \\
& - 4.75(1.21 - x_4)_+(15.41 - x_{22})_+ \\
& + 1.75(4.48 - x_8)_+(1.61 - x_{20})_+ \\
& - 25.76(x_4 - 1.27)_+(0.86 - x_{20})_+ \\
& - 35.58(x_4 - 1.27)_+(x_{20} - 0.86)_+ \\
& - 44.35(2.85 - x_3)_+(3.71 - x_{32})_+ \\
& - 1.67(x_{23} - 16.35)_+(x_2 - 3.86)_+ \\
& - 0.32(2.85 - x_3)_+(4.76 - x_7)_+(x_{27} + 50.73)_+ \\
& + 6.53(1.27 - x_4)_+(x_{22} - 15.41)_+(x_4 - 0.70)_+ \\
R^2 = 0.9997; \quad & F = 5036.1; \quad s = 4.41
\end{aligned} \tag{19}
$$

5. Conclusion

The proposed modeling method, TMARS, combines linear regression and MARS. It consists of a linear and a nonlinear part. The results show that the TMARS model performs better than either the linear or the MARS model.

In situations where the linear model produces fits with good quality, but still is inadequate, the MARS model may fail because it is a completely nonlinear modeling method. TMARS is intermediate between the two methods, and can be expected to achieve better results.

Acknowledgements

The first author is on leave from Hunan University, China.

Appendix A

Retention table

No.	Compound	Retention	No.	Compound	Retention	No.	Compound	Retention	No.	Compound	Retention
1	22334m5C5	953.4	45	233m3C5	761.51	89	2m-3eC7	941.00	133	3mC7	772.67
2	2233m4C4	728.69	46	2344m4C6	935.00	90	2m-3eC6	844.75	134	3mC6	676.6
3	2233m4C6	928.80	47	234m3-3eC5	969.40	91	2m-3eC5	762.57	135	3mC9	969.62
4	2233m4C5	855.13	48	234m3C6	850.88	92	2m-3ipC6	915.50	136	3mC12	1270.1
5	22344m5C5	921.70	49	234m3C5	744.31	93	2m-4eC7	907.40	137	3mC5	584.7
6	2234m4C5	821.90	50	234m3C5	754.45	94	2m-4eC6	824.88	138	3mC8	870.35
7	2234m4C5	825.26	51	234m3C5	753.91	95	2m-5eC7	924.00	139	3eC7	867.45
8	2235m4C6	873.30	52	235m3C6	813.05	96	2mC3	365.61	140	3eC6	773.1
9	223m3-3eC5	965.70	53	236m3C7	919.00	97	2mC4	475.00	141	3eC5	686.8
10	223m3C4	641.46	54	23m2-3eC6	949.40	98	2mC4	473.85	142	3eC8	964
11	223m3C7	914.40	55	23m2-3eC5	875.00	99	2mC4	475.41	143	44m2C7	828.71
12	223m3C6	823.30	56	23m2-4eC6	930.60	100	2mC7	764.95	144	44m2C8	918
13	223m3C6	823.16	57	23m2C4	568.80	101	2mC10	1062.3	145	4pC7	906
14	223m3C6	819.74	58	23m2C4	564.92	102	2mC6	667.00	146	4m-3eC7	940.5
15	223m3C5	738.98	59	23m2C4	568.14	103	2mC6	668.12	147	4m-4eC7	937.6
16	2244m4C6	888.60	60	23m2C7	855.34	104	2mC6	666.74	148	4mC7	767.48
17	2244m4C5	774.77	61	23m2C6	760.79	105	2mC9	963.9	149	4mC9	960
18	2245m4C6	872.10	62	23m2C5	672.55	106	2mC12	1264.1	150	4mC12	1258.3
19	224m3-3eC5	903.90	63	23m2C5	671.74	107	2mC5	570.00	151	4mC8	863.30
20	224m3C7	875.70	64	23m2C5	671.00	108	2mC5	571.79	152	4mC8	861.52
21	224m3C6	790.60	65	23m2C8	952.10	109	2mC5	569.93	153	4eC7	857.82
22	224m3C5	691.55	66	244m3C7	889.40	110	2mC8	864.86	154	4eC8	951.5
23	2255m4C6	820.20	67	244m3C6	809.56	111	3344m4C6	983.7	155	4ipC7	925
24	225m3C7	878.10	68	246m3C7	870.10	112	334m3C7	936.6	156	5mC9	957.4
25	225m3C6	777.07	69	24m2-3eC5	838.17	113	334m3C6	855.25	157	5mC12	1252.4
26	226m3C7	873.00	70	24m2-3ipC5	915.10	114	335m3C7	907.7	158	6mC12	1249.9
27	22m2-3eC6	902.10	71	24m2-4eC6	920.70	115	33m2-4eC6	937.8	159	C3	300
28	22m2-3eC5	824.28	72	24m2C7	821.31	116	33m2C7	837.09	160	C4	400
29	22m2-4eC6	881.30	73	24m2C7	829.98	117	33m2C6	744.81	161	C7	700
30	22m2C3	412.73	74	24m2C7	829.80	118	33m2C5	660.39	162	C10	1000
31	22m2C3	410.35	75	24m2C6	732.69	119	33m2C8	932	163	C6	600
32	22m2C4	537.77	76	24m2C5	630.00	120	33e2C6	954.1	164	C9	900
33	22m2C7	816.50	77	24m2C5	625.65	121	33e2C5	880.34	165	C12	1200
34	22m2C7	814.61	78	24m2C5	630.32	122	344m3C7	932.2	166	C16	1600
35	22m2C6	720.17	79	24m2C8	915.80	123	34m2-3eC6	964.6	167	C13	1300
36	22m2C5	626.55	80	255m3C7	891.70	124	34m2C7	859.56	168	C14	1400
37	22m2C8	914.90	81	25m2-3eC6	891.40	125	34m2C6	771.84	169	C15	1500
38	2334m4C6	949.10	82	25m2C7	833.21	126	34e2C6	945.8	170	C11	1100
39	2334m4C5	861.15	83	25m2C6	728.82	127	35m2C7	834.26	171	C5	500
40	2335m4C6	903.30	84	25m2C8	921.80	128	3m-3eC7	953	172	C8	800
41	233m3C7	931.70	85	26m2C7	827.46	129	3m-3eC6	855.42	173	C2	200
42	233m3C6	841.89	86	26m2C8	931.50	130	3m-3eC5	776.13			
43	233m3C5	761.86	87	27m2C8	928.50	131	3m-4eC6	856.16			
44	233m3C5	752.32	88	2m-33e2C5	984.00	132	3m-5eC7	924			

References

[1] R. Kaliszan, Quantitative Structure–Chromatographic Retention Relationships, Wiley, New York, 1987.
[2] D. Amic, D. Davidovic-Amic, N. Trinajstic, J. Chem. Inf. Comput. Sci. 35 (1995) 136.
[3] R. Dias de Mello Castanho Amboni, B. da Silva Junkes, R.A. Yunes, V.E. Fonseca Heinzen, J. Mol. Struct. (Theochem) 586 (2002) 71.
[4] A. Yan, Z. Hu, Anal. Chim. Acta 433 (2001) 145.
[5] N.R. Draper, H. Smith, Applied Regression Analysis, Wiley, New York, 1981.

[6] A. Miller, in: 2nd ed., Subset Selection in Regression, Chapman and Hall/Crc, 1995.
[7] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, New York, 1989.
[8] J.M. Sutter, T.A. Peterson, P.C. Jurs, Anal. Chim. Acta 342 (1997) 113.
[9] F. Despagne, D.L. Massart, Analyst 123 (1998) 157.
[10] J.H. Friedman, Ann. Statistics 19 (1991) 1.
[11] R.D. De Veaux, D.C. Psichcgios, H. Ungar, Comput. Chem. Eng. 17 (1993) 819.
[12] D. Rogers, A.J. Hopfinger, J. Chem. Inf. Comput. Sci. 34 (1994) 854.
[13] Y. Fan, L. Shi, K.W. Kohn, Y. Pommier, J.N. Weinstein, J. Med. Chem. 44 (2001) 3254.
[14] http://www.salford-systems.com, citations of MARS in the literature.
[15] Y.P. Du, Y.Z. Liang, C.J. Wu, in: Proceedings of the 8th Chinese Computers and Applied Chemistry Conference, Huangshan, 2001, p. 147.
[16] L.B. Kier, L.H. Hall, Molecular Connectivity in Chemistry and Drug Research, Academic Press, New York, 1976.
[17] L.B. Kier, L.H. Hall, in: J. Devillers, A.T. Balaban (Eds.), Topological Indices and Related Descriptors in QSAR and QSPR, Gordon and Breach Science, The Netherlands, 1999.
[18] M. Randic, J. Zupan, J. Chem. Inf. Comput. Sci. 41 (2001) 550.
[19] H.P. Schultz, J. Chem. Inf. Comput. Sci. 29 (1989) 227.
[20] H.P. Schultz, E.B. Schultz, T.P. Schultz, J. Chem. Inf. Comput. Sci. 30 (1990) 27.
[21] L. Xu, W.J. Zhang, Anal. Chim. Acta 446 (2001) 477.
[22] C.Y. Hu, L. Xu, J. Chem. Inf. Comput. Sci. 36 (1996) 82.
[23] S. Geisser, J. Am. Statist. Assoc. 70 (1975) 320.