Contents lists available at ScienceDirect

Analytica Chimica Acta

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

# Improved variable reduction in partial least squares modelling by Global-Minimum Error Uninformative-Variable Elimination

Jan P.M. Andries [a], Yvan Vander Heyden [b], Lutgarde M.C. Buydens [c, *]

[a] Research Group Analysis Techniques in the Life Sciences, Avans Hogeschool, University of Professional Education, P.O. Box 90116, 4800 RA Breda, The Netherlands

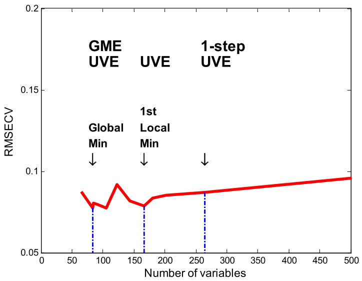
[b] Department of Analytical Chemistry and Pharmaceutical Technology, Center for Pharmaceutical Research, Vrije Universiteit Brussel-VUB, Laarbeeklaan 103, B-1090 Brussels, Belgium

[c] Radboud University Nijmegen, Institute for Molecules and Materials, Heyendaalseweg 135, 6525 ED Nijmegen, The Netherlands

## HIGHLIGHTS

- A modification of UVE-PLS is proposed, in which UVE-PLS is repeated until no further reduction in variables is obtained.
- The variable set with the global RMSECV minimum is selected and used for PLS modelling.
- The method is called Global-Minimum Error Uninformative-Variable-Elimination for PLS, denoted as GME-UVE-PLS.
- GM-UVE-PLS usually eliminates significantly more variables than the UVE-PLS method.

## GRAPHICAL ABSTRACT



## ARTICLE INFO

*Article history:*
Received 10 November 2016
Received in revised form
18 May 2017
Accepted 1 June 2017
Available online 16 June 2017

*Keywords:*
Variable elimination
Partial least squares regression
Uninformative-Variable Elimination PLS (UVE-PLS)
Global-Minimum Error Uninformative-Variable Elimination PLS (GME-UVE-PLS)
Paired *t*-test

## ABSTRACT

The calibration performance of Partial Least Squares regression (PLS) can be improved by eliminating uninformative variables. For PLS, many variable elimination methods have been developed. One is the Uninformative-Variable Elimination for PLS (UVE-PLS). However, the number of variables retained by UVE-PLS is usually still large.

In UVE-PLS, variable elimination is repeated as long as the root mean squared error of cross validation (RMSECV) is decreasing. The set of variables in this first local minimum is retained. In this paper, a modification of UVE-PLS is proposed and investigated, in which UVE is repeated until no further reduction in variables is possible, followed by a search for the global RMSECV minimum. The method is called Global-Minimum Error Uninformative-Variable Elimination for PLS, denoted as GME-UVE-PLS or simply GME-UVE. After each iteration, the predictive ability of the PLS model, built with the remaining variable set, is assessed by RMSECV. The variable set with the global RMSECV minimum is then finally selected. The goal is to obtain smaller sets of variables with similar or improved predictability than those from the classical UVE-PLS method.

The performance of the GME-UVE-PLS method is investigated using four data sets, i.e. a simulated set, NIR and NMR spectra, and a theoretical molecular descriptors set, resulting in twelve profile-response (X-y) calibrations. The selective and predictive performances of the models resulting from GME-UVE-PLS are statistically compared to those from UVE-PLS and 1-step UVE, one-sided paired t-tests.

The results demonstrate that variable reduction with the proposed GME-UVE-PLS method, usually eliminates significantly more variables than the classical UVE-PLS, while the predictive abilities of the

* Corresponding author.
*E-mail address:* chemometrics@science.ru.nl (L.M.C. Buydens).

http://dx.doi.org/10.1016/j.aca.2017.06.001

resulting models are better. With GME-UVE-PLS, a lower number of uninformative variables, without a chemical meaning for the response, may be retained than with UVE-PLS. The selectivity of the classical UVE method thus can be improved by the application of the proposed GME-UVE method resulting in more parsimonious models.

## 1. Introduction

Partial Least Squares is a commonly used multivariate regression technique, which is able to deal with a large number of noisy and correlated variables, as well as with rather small numbers of samples [1–3]. However, both theoretical [4–8] and experimental evidence [3,9–13] exist that elimination of uninformative variables improves the performance of the PLS calibration. Additionally, a better interpretation of the PLS models, lower measurement costs, or a reduced risk of overfitting may be obtained [10].

For PLS1, modelling one response **y**, many variable elimination methods have been published [3,9–11,14–17], among which the Uninformative-Variable Elimination for PLS [18]. UVE-PLS is a variable elimination method based on the significance of the PLS regression coefficients. The importance of each variable in the model is determined by its significance, being the ratio of the PLS regression coefficient and its standard deviation, estimated from jack-knifing. Variables below a cut-off value, calculated from artificial random variables added to the data set, are eliminated. The method has been widely applied in analytical chemistry [3,13,16–24].

Several modifications to UVE-PLS are also reported. Two robust modifications were already proposed in the original paper of Centner et al. [18]. In the first, called UVE-M, to estimate the significances of the variables, the mean PLS regression coefficient was replaced by the median and the standard deviation by the interquartile range, determined from jack-knifing. In the second, UVE-$\alpha$, the cut-off value corresponds to a user defined quantile of the ranked significances of the artificial added variables. Later modifications include (*i*) improvement of the calculation of the standard error of the regression coefficients [25], (*ii*) estimation of a cut-off threshold based on Monte Carlo (MC) selection of calibration samples rather than the addition of artificial random variables (MC-UVE) [13], or based on randomization of the response vector [26], (*iii*) the use of a new cut-off criterion [27], (*iv*) ensemble UVE-PLS (EUVEPLS), based on an ensemble of different calibration sets, which are randomly selected from the available calibration samples [28,29], and (*v*) wavelet transform (WT) techniques combined with UVE in Relevant Component Extraction for PLS (RCE-PLS) [30,31], and with MC-UVE in WT-MC-UVE [13].

However, the number of variables retained by UVE-PLS is rather large [16,32], probably because often variable elimination is stopped after only one elimination step instead of repeated steps, see Section 2.2. After UVE-PLS, the number of retained variables occasionally has been further reduced by a genetic algorithm (UVE-GA-PLS) [9,33–35], interval PLS (UVE-iPLS) [9], or a successive projections algorithm (UVE-SPA) [32,36,37].

Both in the original paper of Centner et al. [18] and in that of Westad and Martens [38], uninformative variables are removed iteratively as long as the root mean squared error of cross validation decreases and the set of variables in the thus found *first local RMSECV minimum* is retained. In this study, a modification of the UVE method is proposed in which uninformative variables are removed iteratively until no further reduction of variables is possible. Finally, the variable set corresponding to the *global RMSECV minimum* is selected, which may be smaller than that corresponding to the first local minimum. The goal is to obtain smaller sets of variables, resulting in more parsimonious models with similar or improved predictability, than those from the UVE-PLS method [18,38]. The method is called Global-Minimum Error Uninformative-Variable Elimination for PLS.

The utility and effectiveness of the GME-UVE-PLS and the original UVE-PLS method are tested and compared, using four data sets resulting in 12 **X**-**y** combinations (see Table 1). The X-profiles consist of simulated data, NIR and NMR spectra, and molecular descriptors used to build Quantitative Structure-Retention Relationship (QSRR) models for reversed-phase liquid chromatography (RPLC).

## 2. Theory

### 2.1. PLS model

The aim of PLS is to model the relationship between a data matrix **X** and a response vector **y** by using a set of latent variables or PLS factors that maximize the explained covariance between them. The PLS1 model is developed from a calibration set of *N* objects or observations with one response or dependent variable in the **y** vector and *K* predictor variables in the **X** matrix. The **y**($N$ x 1) vector consists of the *N* responses denoted by $y_i$ ($i = 1, \ldots, N$). The **X**($N$ x $K$) matrix consists of *K* column vectors of independent predictor variables denoted by $\mathbf{x}_k$ ($k = 1, \ldots, K$). The objective of PLS is to select the optimal number $A$ ($A \leq K$) of PLS factors, which are linear combinations of the original variables $\mathbf{x}_k$. The PLS model is given by Eqs. (1) and (2).

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E}_{\mathbf{A}} \tag{1}$$

$$\mathbf{y} = \mathbf{T}\mathbf{q}^{\mathbf{T}} + \mathbf{f}_{\mathbf{A}} \tag{2}$$

where **T**($N$ x $A$) is a score matrix, **P**($K$ x $A$) a matrix with the x-loading vectors $\mathbf{p}_a$ ($a = 1, 2, \ldots, A$) as columns, **q**($1 \times A$) the **y**-loading vector, $\mathbf{E}_{\mathbf{A}}$ and $\mathbf{f}_{\mathbf{A}}$ the residual matrix for **X** and the residual **y**-vector, respectively, after the extraction of *A* factors. The optimal number of PLS factors, *A*, can be determined using cross validation (CV). Further details on PLS can be consulted in Refs. [1,2,39].

### 2.2. Uninformative-Variable Elimination (UVE)

Uninformative-Variable Elimination for PLS is introduced in Ref. [18]. UVE-PLS determines the fitness of each predictor variable *k* in the **X**($N \times K$) matrix against those of *L* artificial random variables from a matrix **R**($N \times L$). **R** is added to **X**, resulting in an augmented matrix **[X R]** ($N \times (K + L)$). These added random variables have very small absolute values, of the order of magnitude of about $10^{-10}$, so that their influence on the PLS regression coefficients of the predictors is negligible.

The PLS regression coefficients $b_k$ are calculated in a vector **b** with $K + L$ coefficients, from

**Table 1**
Results of full spectrum modelling of all data sets. Abbreviations: see text.

| Data set | Response y | Model number | PLS complexity | Number of X variables | *RMSEP* |
|---|---|---|---|---|---|
| Simulated | y | 1 | 2 | 1000 | 0.090 |
| Corn | Moisture | 2 | 15 | 700 | 0.012 |
| | Oil | 3 | 11 | 700 | 0.060 |
| | Protein | 4 | 14 | 700 | 0.090 |
| | Starch | 5 | 15 | 700 | 0.170 |
| Alcohols | Propanol | 6 | 19 | 14000 | 0.823 |
| | Butanol | 7 | 19 | 14000 | 0.678 |
| | Pentanol | 8 | 19 | 14000 | 1.025 |
| Chromatographic | Log $k_w$ on Col1 | 9 | 7 | 1243 | 0.514 |
| | Log $k_w$ on Col2 | 10 | 7 | 1243 | 0.506 |
| | Log $k_w$ on Col3 | 11 | 7 | 1243 | 0.541 |
| | Log $k_w$ on Col4 | 12 | 7 | 1243 | 0.525 |

$$\mathbf{b} = \mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1}\mathbf{q} \tag{3}$$

where $\mathbf{W}((K+L)\times A)$ is the **[X R]** weight matrix, $\mathbf{P}((K+L)\times A)$ the **[X R]**-loading matrix, $\mathbf{q}(1 \times A)$ the y-loading vector, and $A$ the (optimal) number of PLS1 factors [1].

Influential predictor variables have large positive or negative coefficients $b_k$ with low uncertainties [38]. Therefore, their significance will be large. The significance of the PLS regression coefficient of variable $k$, denoted as $SIG_k$, is defined as the *t-value*, calculated from $n$ fold leave-more-out jack-knifing [40] as

$$SIG_k = t_k = \frac{\left|\bar{b}_k\right|}{s_{b_k}} \tag{4}$$

with $t_k$ the absolute *t-value* for variable $k$, $\bar{b}_k$ and $s_{b_k}$ the mean and standard deviation of the estimated coefficients $b_k$ for a given variable $k$.

A suitable cut-off value $SIG_{cut\text{-}off}$ is calculated from the significances of the $L$ artificial variables. Predictor variables $k$ with $SIG_k$ below the cut-off value $SIG_{cut\text{-}off}$ are classified as uninformative and eliminated. Using the remaining variables, a new PLS model is built and the predictive ability estimated by cross validation, resulting in a new RMSECV. If the new RMSECV is smaller than the RMSECV before the elimination step, a new **X** matrix is formed with the remaining variables and a new PLS model developed with complexity $A_{new} = A$-1. In UVE, variable elimination is repeated as long as the RMSECV decreases [18,38] and stops at an increase. Therefore, in UVE-PLS, the variable set corresponding to the first local RMSECV minimum is finally selected. In 1-step UVE, a simplified and pragmatic version of UVE, which is frequently encountered in the literature [13,31,41–44], variable elimination is conducted in only one step.

Several options exist to estimate a suitable cut-off value, such as ($i$) taking the maximum of the significances of the $L$ artificial variables, $\max(SIG_{art})$ [18], ($ii$) taking a fraction of $\max(SIG_{art})$ using a cut-off factor $f_{cut\text{-}off} \leq 1$, resulting in $SIG_{cut-off} = f_{cut-off} \cdot \max(SIG_{art})$ [13], ($iii$) taking the maximum of the $f_{cut\text{-}off} \cdot 100\%$ quantile of the $L$ artificial variables [27]. The latter is also implemented in the Matlab procedure uvepls.m of the ChemoAC Toolbox [45].

In UVE-PLS, the number of eliminated variables is variable because of the variability in the added artificial random variables. Consequently, upon replication of the procedure, different variable sets are retained.

### 2.3. Global-Minimum Error Uninformative-Variable Elimination (GME-UVE)

In GME-UVE, uninformative variables are groupwise eliminated iteratively until no more variables can be removed. In each iteration step, the number of added artificial random variables $L$ is equal to the number of remaining predictor variables. From all iterations, the global RMSECV minimum is determined and the corresponding variable set selected.

A first difference with the UVE method from Ref. [18], described in Section 2.2, is that the stop criterion is not an increase in RMSECV, but the fact that no more variables are eliminated. A second difference is that the global RMSECV minimum is used as criterion for the selection of variables and not the first local minimum. The reason is that possibly, besides the first local RMSECV minimum, a lower global one exists originating from a lower number of retained variables. A third difference is that after each iteration, a new PLS complexity is determined as described below, and not by taking $A_{new} = A$-1 as described in Section 2.2.

However, retaining the variable set from the global RMSECV minimum is the main characteristic of GME-UVE.

In Fig. 1, for the GME-UVE-PLS method, an example of a RMSECV plot as a function of the number of retained variables is shown. The curve stops when all uninformative variables are eliminated. In Fig. 1, the 1-step UVE, the first local RMSECV minimum, and the global RMSECV minimum are at clearly different numbers of variables.

The global RMSECV minimum corresponds to a lower number of retained variables than the first local minimum, which itself is at a lower number than 1-step UVE. The stop location of variable elimination, i.e. the end of the curve, is different from the location

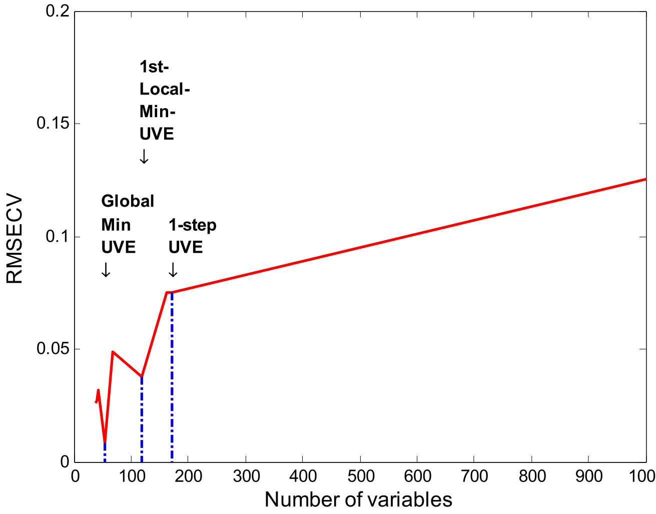


**Fig. 1.** RMSECV values as a function of the number of retained variables for the GME-UVE-PLS method on the Simulated set, with 1-step UVE, the first local minimum, and the global minimum approaches; cut-off = 1.00.

for the selection of variables, i.e. the global RMSECV minimum, although it is possible that these locations are identical.

The GME-UVE-PLS method consists of five steps. First, the data set is split into a training and a test set. After pre-processing, the predictive ability of the resulting full spectrum PLS model is assessed by internal validation with the training set, using segmented cross validation (SCV), see Section 2.5.1. The optimal number of PLS factors $A$ is determined by the application of the adjusted Wold's R criterion, $R_{adj} < 0.98$; see Section 2.5.3.

In the second step, a suitable cut-off factor, $f_{cut\text{-}off}$, for the groupwise elimination of uninformative variables is chosen, either after a few test runs, or in a systematic way, after the application of a range of cut-off factors, as discussed below.

In the third step, iteratively, variables with $SIG_k < SIG_{cut\text{-}off}$ are groupwise eliminated, until no more variables are removed. After each iteration step, using the retained variables in **X**, a new optimal PLS factor number $A'$ is determined by SCV, with a maximum equal to the number of remaining variables.

As an additional action to avoid over-fitting, during variable elimination, in the intermediate steps, a different criterion is used for the factor number determination. The minimal RMSECV is used instead of the R criterion $R_{adj} < 0.98$. This allows also a fast variable elimination process. The newly determined factor number $A'$ is used in the subsequent step.

In the fourth step, a graph of *RMSECV* against the number of remaining variables is made, and the global minimum determined, see the example in Fig. 1. The variable set corresponding to this global minimum is selected and considered the best remaining set.

In the fifth step, using the best remaining set, the PLS model is externally validated by the root mean squared error of prediction (*RMSEP*) using the test set. First a renewed determination of the optimal number of PLS factors $A''$, with a maximum equal to the number of remaining variables, is done by SCV and with the application of the R criterion $R_{adj} < 0.98$.

In this study $SIG_{cut\text{-}off}$ is taken as the maximum of the $f_{cut\text{-}off} \cdot 100\%$ quantile of the $L$ artificial variables. For the elimination of uninformative variables a suitable cut-off factor $f_{cut\text{-}off}$ can be chosen. Ideally, this is the cut-off factor corresponding to the minimal number of retained variables and minimal RMSECV. However, the cut-off factor is also suitable if it corresponds to the minimal number of retained variables for which the resulting RMSECV is considered suitable for the application at hand. The authors have the experience that, usually, the range $f_{cut\text{-}off} = 0.90–1.00$ is appropriate.

In this study, $f_{cut\text{-}off}$ is determined systematically, after the application of a range of cut-off factors (see Section 4.1). However, it is also possible to choose a suitable cut-off factor after a few test runs, applying some user defined cut-off values in the range $f_{cut\text{-}off} = 0.90–1.00$. Because UVE is in general a fast method, a suitable cut-off factor is found fast after a few test runs.

The number of eliminated variables by UVE depends on the applied cut-off, $SIG_{cut\text{-}off}$. Because of the random character of the added artificial variables, the number of remaining variables can vary between different replicates of the method. Therefore, to determine a suitable cut-off factor, the GME-UVE procedure was repeated eleven times for each cut-off and, to get robust results, the run with the median number of retained variables is used.

### 2.4. Relation between GME-UVE, 1-step UVE and UVE

In the graph of RMSECV vs. number of retained variables, three characteristic points can be observed: that after 1-step UVE, the first local RMSECV minimum, and the global RMSECV minimum, see Fig. 1.

The retained variable set in the first local RMSECV minimum corresponds to that obtained, often after several iterations, by the original UVE method [18]. The retained variable set in the point for 1-step UVE corresponds to that obtained after the application of only one iteration in the UVE method. The retained variable set in the global RMSECV minimum is selected by the GME-UVE method.

These three characteristic points can be used to compare the selective and predictive performances of the three corresponding methods: 1-step UVE, UVE, and GME-UVE. When two points coincide, the results of the corresponding methods are identical. For example, if the first local minimum coincides with the global, there is no difference in the retained variable sets of the UVE and GME-UVE methods.

## 2.5. Model validation

### 2.5.1. Internal validation

The predictive abilities of the PLS1 models are assessed by internal validation with the training set, using venetian blinds SCV, resulting in the root mean squared error of cross validation,

$$RMSECV = \sqrt{\frac{1}{N_{cal}} \sum_{i=1}^{N_{cal}} (y_i - \hat{y}_i)^2} \qquad (5)$$

where $y_i$ and $\hat{y}_i$ are the experimental and predicted responses, respectively, for the $i$th calibration sample when situated in a left-out segment of the training or calibration set, $N_{cal}$ is the number of calibration samples in the training set.

### 2.5.2. External validation

Before and after variable reduction, the predictive abilities of the PLS1 models, developed with the training set, are also assessed by external validation using a test set, resulting in the root mean squared error of prediction,

$$RMSEP = \sqrt{\frac{1}{N_{test}} \sum_{i=1}^{N_{test}} (y_i - \hat{y}_i)^2} \qquad (6)$$

where $y_i$ and $\hat{y}_i$ are the experimental and predicted responses, respectively, for the $i$th test-set sample, $N_{test}$ is the number of test-set samples.

### 2.5.3. Model complexity

Before and after variable reduction, the best complexity $A$ of a PLS1 model is determined by venetian blinds segmented ($n$-fold) cross validation. The complexity corresponding to the minimal *RMSECV* is used. However, to avoid overfitting, often a simpler model with a similar error is selected by the application of an adjusted Wold's R criterion, $R_{adj}$, [46,47]. The idea is that an additional PLS factor should only be included in the model if the *RMSCEV* improves with at least 2% (i.e. if $R_{adj} < 0.98$) [48]. First, the minimum in the *RMSECV* vs. model-complexity curve is determined. Thereafter, for models with less complexity $A$ than the one giving the minimal *RMSECV*, $A_{Min\_RMSECV}$, with $1 < A < A_{Min\_RMSECV}$, two successive values of *RMSECV*, i.e. $RMSECV_{A-1}$ and $RMSECV_A$, are pairwise compared using Eq. (7).

$$R_{adj} = \frac{RMSECV_A}{RMSECV_{A-1}} \qquad (7)$$

The maximal complexity $A < A_{Min\_RMSECV}$ for which $R_{adj} < 0.98$, is then considered as the best.

### 2.5.4. Comparison of methods

Because of the added random variables in the UVE method, the selective and predictive results can vary between different GME-UVE runs. Therefore, GME-UVE runs were repeated 100 times and, for each run, the characteristic points identified. Thereafter, the numbers of retained variables $K_{Best}$ and RMSEP's of the resulting PLS models are determined. They are used for the comparison of the selective and predictive performances of the three corresponding UVE methods, see Section 4.2. The comparisons for differences in methods are made by paired t-tests, using the Bonferroni correction for multiple testing [49].

## 3. Data and methodology

### 3.1. Data sets

#### 3.1.1. Simulated set

The first data set is simulated. It represents the profiles of mixtures of one main-compound and three interferents. The profile of the main-compound consists of five Gaussian peaks (A-E) $g(\mu,\sigma,h)$, with mean $\mu$, standard deviation $\sigma$ and height $h$ ($g_A$(100,8,0.1), $g_B$(235,8,0.3), $g_C$(250,8,0.25), $g_D$(265,8,0.3), $g_E$(400,10,0.5)). The profile of the main-compound is formed by $g_A + g_B + g_C + g_D + g_E$. The profiles of the three interferents 1, 2 and 3 consist of three individual Gaussian peaks (F-H), $g_F$(235,8,0.2), $g_G$(265,8,0.2), and $g_H$(400,10,0.5), respectively. The peaks A-E in the profile of the main-compound are shown in the top of Fig. 2 (Left) and the peaks of the interferents are shown in the bottom.

Three kinds of interactions between the main-compound and the interferents are built in the profile.

- Peak A in the main-compound profile is free of interactions with any interferent.
- Peaks B and D overlap with peak C in the peak complex BCD in the main-compound profile and interact with interferents 1 (peak F) and 2 (peak G), respectively. Peaks F and G of the interferents have the same mean and standard deviation as those of the main-component peaks B and D, respectively, but the heights are lower.
- Peak E in the main-compound profile interacts with interferent 3 (peak H). Mean, standard deviation and peak heights of peaks E and H are equal.

Uncorrelated random responses for the main component and the three interferents *Y(i,j)* between 0 and 1 were randomly generated, using the Matlab function for uniformly distributed pseudorandom numbers *rand*, where *i* and *j* are indices for the samples ($i = 1 \ldots 120$) and components ($j = 1 \ldots 4$), respectively. The analyte profiles of the mixture samples *i* were generated, combining the responses as weight factors with the above mentioned profiles, by $Y(i,1)\cdot(g_A + g_B + g_C + g_D + g_E) + Y(i,2)\cdot g_F + Y(i,3)\cdot g_G + Y(i,4)\cdot g_H$.

The analyte profiles in the mixtures cover a range of 500 *x* variables. These variables are informative. Additionally, 500 uninformative variables are added, consisting of random numbers between 0 and 0.8. The latter have high signal levels, comparable to that of the informative. This is to investigate whether the GME-UVE-PLS method is capable to find informative variables in profiles containing many uninformative variables at similar signal levels. Additionally, noise is added to the simulated profiles, consisting of random numbers in the range between 0 and 0.005, i.e. small compared to the signals of the main-compound. The data set is split into a training set of 100 and a test set of 20 samples using the duplex method [50], while 10-fold cross validation is conducted during model building.

In Fig. 2 (Right) the heat map is given for the correlation coefficients for the variable range [1, 500]. It shows the correlation structure of the informative variables. Only variables in the peak areas are correlated.

#### 3.1.2. Corn set

The second data set consists of NIR spectra of 80 corn samples with a wavelength range of 1100–2498 nm with 2 nm intervals, resulting in 700 predictor variables. This data set, labelled corn from the “m5” spectrometer, is provided by Eigenvector Research [51]. Moisture, oil, protein and starch contents of the samples are the responses. The data set is split into a training set of 60 and a test set of 20 samples using the duplex method. An 8-fold cross validation is conducted during model building.

#### 3.1.3. Alcohols set

The third data set is composed of 231 samples and contains $^1$H NMR spectra of mixtures of the alcohols propanol, butanol and pentanol, with chemical shifts from 0.65 to 3.85 ppm, resulting in 14000 predictor variables. The data set was downloaded from the website in Ref. [52]. Details are described in Ref. [53]. The propanol, butanol and pentanol percentages in the mixtures are used as responses. The data set is split into a training set of 171 and a test set of 60 samples using the duplex method. A 10-fold cross validation is conducted during model building.

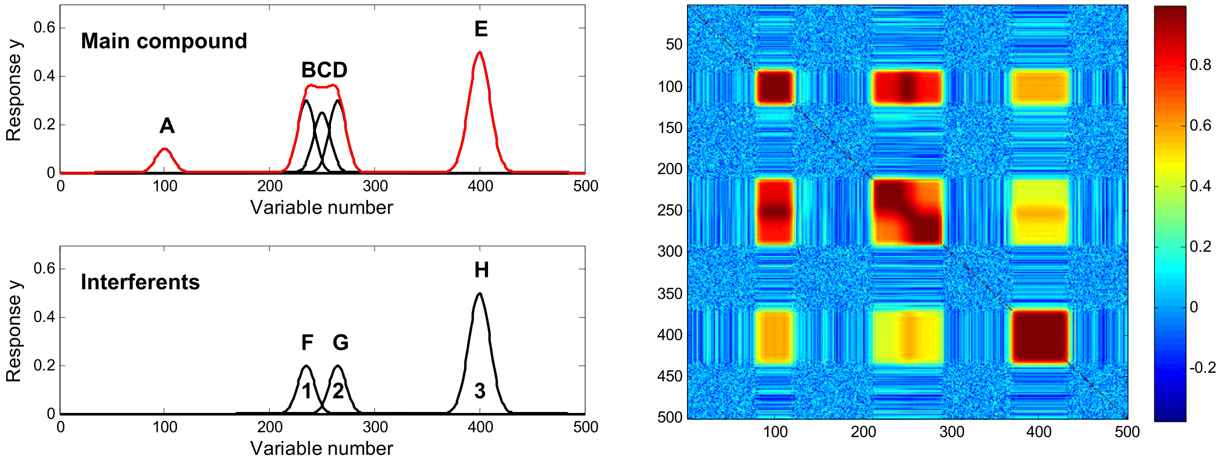


**Fig. 2.** Simulated set; (Left) Informative part; (Top) Profile of the main-compound (—) built up with five Gaussian peaks A-E; (Bottom) Profiles of the three interferents 1, 2 and 3, formed by three Gaussian peaks F-H; (Right) Heat map for the correlation coefficients between the informative variables.

### 3.1.4. Chromatographic set

The fourth data set consists of chromatographic data originating from Ref. [54]. Log $k_w$ values of 25 structurally diverse test analytes are used as response on four $C_{18}$ reversed-phase liquid chromatography (RPLC) columns, Zorbax RX-C18, Hypersil ODS, Polygosil-60-5-C18 and Alltima C18 5U. Experimental retentions were measured using as mobile phase methanol-buffer mixtures. The columns are numbered as Col1-Col4.

For the 25 analytes, conformational analysis was performed using the molecular dynamics module of the HyperChem V 7.5 Professional software [55]. Geometry optimization was performed by the molecular mechanics Force Field (MM+) method. The lowest-energy conformers from the MM + based simulations were found using the Polak–Ribière conjugate gradient algorithm (RMS gradient equal to 0.05 kcal/(Å mol) as stop criterion. Molecular descriptors were calculated for each analyte using Dragon V5 [56] and autoscaled. Descriptors with constant and near-constant values, and with at least one missing value were excluded, resulting in a data set with 1246 molecular descriptors. The data set for the four **X-y** combinations is split into a training set of 19 and a test set of 6 analytes using the Kennard and Stone method [57]. Because of the small calibration set, leave-one-out cross validation is conducted during model building. The molecular descriptors are used to build Quantitative Structure-Retention Relationships (QSRR) to model and predict the retention of analytes, log $k_w$, on the columns.

### 3.2. Software

All calculations are made with in-house programs developed in Matlab (V. 7.14) (The Math Works, Natick, MA, USA). The procedures for the duplex splitting and the Kennard Stone algorithm are from the ChemoAC Toolbox [45]. Molecular descriptors of the analytes in the chromatographic set are calculated with Dragon software V5 (Kode Chemoinformatics, Pisa, Italy) [56] and the geometrical representations of the analytes were obtained in Hyperchem 7.5 Professional software (Hypercube, Gainesville, Florida) [55]. Paired t-tests are conducted with the Statistics Toolbox of Matlab.

## 4. Results and discussion

The four data sets form 12 **X-y** combinations. The responses of all sets are pre-processed by mean centering. The independent variables of the chromatographic set are autoscaled while those of the other data sets are mean centered. First, for each of the 12 responses, a PLS1 model is developed. The optimal complexities are determined for the full spectrum models, by segmented cross validation and applying the criterion $R_{adj} < 0.98$. *RMSECV* and *RMSEP* are estimated as described in Section 2.5. The optimal PLS complexity, the number of variables and *RMSEP* values for the full spectrum models are shown in Table 1. Variable reduction by the GME-UVE-PLS method is then applied on the 12 **X-y** combinations.

### 4.1. Selection of a suitable cut-off factor

The applied cut-off factor has a large influence on the results of the GME-UVE method. For the reduction of a large number of uninformative variables, a high cut-off factor is preferred. However, if it is too high, informative variables may also be removed, resulting in reduced predictive abilities of the remaining variable sets, and hence higher RMSECV's. If a too low cut-off factor is applied, not all uninformative variables will be eliminated, resulting in larger remaining variable sets.

To investigate systematically the influence of the cut-off factor, repeated GME-UVE-PLS runs are conducted at different cut-off levels, ranging from 0.90 to 1.00, with steps of 0.01. Because the number of eliminated variables is variable, the GME-UVE runs are repeated eleven times at each cut-off level. For each run, the global RMSECV minimum is determined, together with the corresponding numbers of remaining variables. A suitable cut-off factor will result in a low number of remaining variables and a low corresponding RMSECV. To get robust results, at each cut-off level, the run with the median number of variables in the global RMSECV minimum is selected. The cut-off level with the minimal median number of remaining variables is then selected (Fig. 3).

### 4.2. Analysis of the data sets

For all **X-y** combinations, suitable cut-off factors are determined as described above, followed by GME-UVE variable reduction.

The number and identity of the remaining variables, and the RMSECV's of the resulting models in the characteristic points can vary between different GME-UVE runs because of different added random variables. Therefore, 100 repeated UVE runs are carried out for each **X-y** combination and the characteristic points were identified.

At each run, for the remaining variable sets in the characteristic points, the numbers of retained variables $K_{Best}$ are determined and the optimal number of PLS factors re-determined by SCV and the application of the criterion $R_{adj} < 0.98$. Additionally, in the characteristic points, *RMSEP* values are estimated using the remaining variable sets and these optimal number of PLS factors.

The characteristic points of two or even the three methods may coincide. For 100 runs and 3 methods, this results in a 100 × 3 matrix for both the numbers of retained variables and RMSEP's.

The numbers of retained variables $K_{Best}$ and RMSEP's of the resulting PLS models in the characteristic points are used for the comparisons of the selective and predictive performances of the three UVE methods. The comparisons are made by paired t-tests, using the Bonferroni correction for multiple testing, see Ref. [49].

The numbers of retained variables for GME-UVE are equal to or smaller than those of UVE, and the latter are equal to or smaller than those of 1-step UVE. Ideally, this will also apply for the corresponding RMSEP's. To test this statistically, one sided paired t-tests are carried out at the 95% confidence level for the pairs of $K_{Best}$ and of RMSEP's of (*i*) GME-UVE and UVE, (*ii*) GME-UVE and 1-step

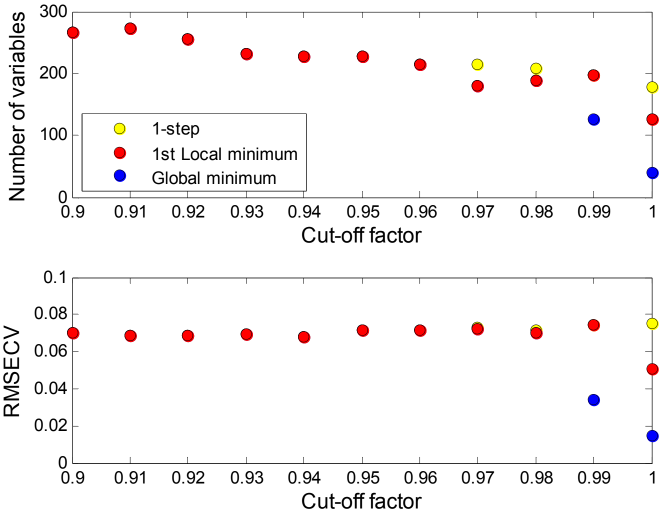


**Fig. 3.** Simulated set; Median numbers of variables (Top) in the remaining best sets for eleven times repeated GME-UVE-PLS runs and corresponding RMSECV (Bottom) both against cut-off factors.

UVE, and (*iii*) UVE and 1-step UVE, to find out whether the values of the first method are significantly lower than those of the second method.

### 4.3. Results for the data sets

In this section, an overview is given of the results after 100 repeated GME-UVE runs for each **X**-**y** combination. For each data set, the analysis of one **X**-**y** combination is fully described. For the other **X**-**y** combinations, only the main results are reported. The selective and predictive performances of the three UVE methods are compared by paired t-tests as described in Section 4.2.

Additionally, for the four discussed **X**-**y** combinations, the retained variable sets are investigated for the presence of variables with a chemical meaning relevant to the response, to evaluate the quality of the selective ability of the UVE methods.

#### 4.3.1. Simulated set

For the Simulated set, in Fig. 4 (Left) results are shown for the 100 repeated GME-UVE runs with a cut-off = 1.00. In the graph, the RMSEP as a function of the number of variables is given for the three methods. The points for 1-step UVE lie in the range [100, 200] for variables and those for the first local minima in a wider range [4, 150], covering both a large part of RMSEP range found. 67 runs have a global minimum with a lower number of retained variables than the corresponding first local minimum. The results for the global minima are concentrated in the bottom left corner with relatively small ranges for both the numbers of retained variables ([6, 64]) and the RMSEP's ([0.005, 0.037]).

In Table 2 (model 1) the t-values are shown for the paired t-tests comparing the numbers of variables for the pairwise comparison of the three methods. The t-values are in absolute values larger than the critical one-sided Bonferroni corrected value $|t| = 2.15$. This means that the remaining numbers of variables of GME-UVE are significantly smaller than those of UVE, and that the latter are significantly smaller than those of 1-step UVE. Therefore, for this data set, the selective ability of the GME-UVE method is significantly better than that of the UVE method, which in turn is significantly better than that of 1-step UVE.

The t-values are also shown for the paired t-tests on the RMSEP's for the three comparisons. They are also lower than the critical value. Thus the RMSEP's of GME-UVE are significantly smaller than those of UVE, which in turn are significantly smaller than those of 1-step UVE. Therefore, for this data set, the predictive ability of the GME-UVE method is better than that of the UVE method, and the latter is better than that of 1-step UVE.

Fig. 4 (Right) shows the profiles and retained variables after a GME-UVE example run with selections for 1-step UVE, in the first local and in the global RMSECV minimum. After 1-step UVE, all noisy variables are eliminated and correlated variables (see Fig. 2 (Right), belonging to peaks of the main-compound, are retained. All retained variables for the three UVE methods have a meaning because they are part of the peaks of the main-compound. No uninformative variables from the range [501, 1000] are retained. The GME-UVE method clearly results in the smallest number of retained variables.

#### 4.3.2. Corn set

For the Corn set with response Oil, in Fig. 5 (Left) results are given for 100 repeated GME-UVE runs with a cut-off = 0.95. Again, the RMSEP as a function of the number of retained variables are shown. The numbers of retained variables for 1-step UVE lie in the range [100, 300] and those of UVE in the range [10, 150]. 27 runs have a global minimum in the range [10, 100]. The entire RMSEP range observed seems to be rather independent of the numbers of retained variables.

In Table 2 (model 3) calculated t-values are shown for the paired t-tests for numbers of variables for the three pairs of UVE methods. The calculated t-values are significant. This means that the remaining numbers of variables of GME-UVE are significantly smaller than those of UVE, and that the latter are significantly smaller than those of 1-step UVE. Therefore, for this **X**-**y** combination, the selective ability of the GME-UVE method is significantly better than that of the UVE method, which in turn is significantly better than that of 1-step UVE.

For RMSEP only for the comparison of UVE and 1-step UVE the calculated t-value is borderline significant. Thus, the RMSEP's of UVE are borderline significantly smaller than those of 1-step UVE. The calculated absolute t-values for the comparion of RMSEP's of GME-UVE with those of UVE and 1-step UVE, 0.37 and −1.72, are smaller than the critical t-value. Therefore, for this **X**-**y** combination, the predictive ability (*i*) of the UVE method is significantly better than that of 1-step UVE, and (*ii*) of GME-UVE is similar to that of both UVE and 1-step UVE.

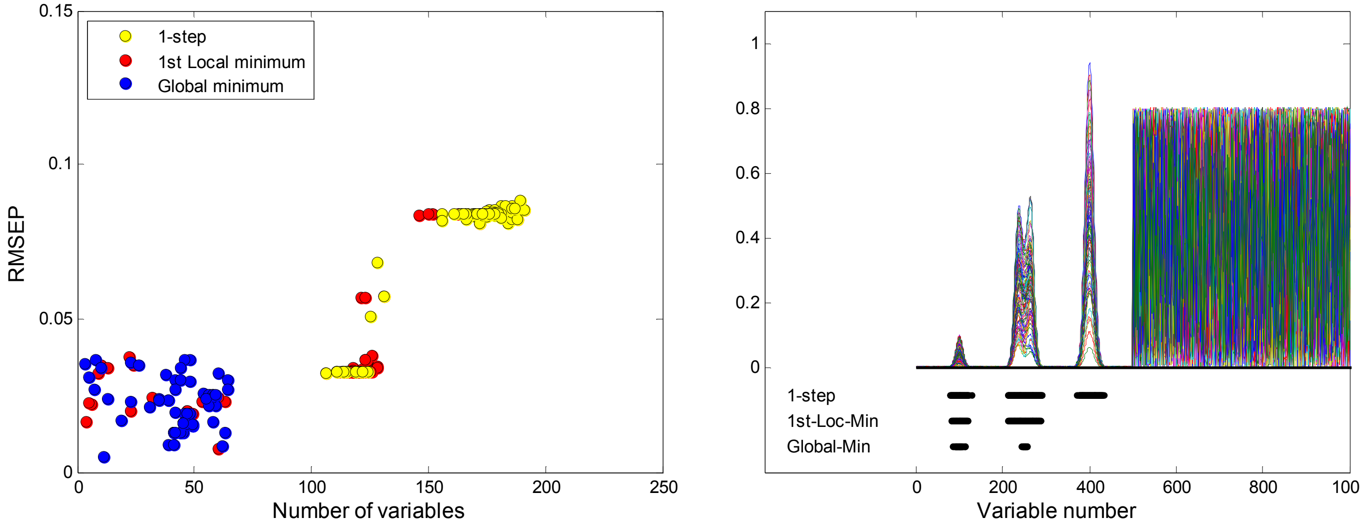


**Fig. 4.** Simulated set; (Left) RMSEP vs. number of retained variables for 100 repeated GME-UVE-PLS runs; (Right) Profiles and retained variables after a GME-UVE-PLS run with selections for 1-step UVE, in the first local RMSECV minimum, and in the global RMSECV minimum.

**Table 2**
Results for the 100 times repeated GME-UVE runs. Abbreviations: see text.

| Data set | Response | Model | Cut-off | Number of global minima | Method characteristics | global RMSECV minimum* (M1) | 1st local RMSECV minimum* (M2) | 1-step UVE* (M3) | t-values of paired t-tests for numbers of retained variables or RMSEP's | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | M1-M2 | M1-M3 | M2-M3 |
| Simulated | y | 1 | 1.00 | 67 | Range PLS complexity | [2,2] | [2,2] | [2,2] | | | |
| | | | | | number of variables | 55* | 108 | 167 | **−12.52** | **−16.14** | **−13.62** |
| | | | | | *RMSEP* | 0.027* | 0.037 | 0.078 | **−7.06** | **−23.20** | **−18.10** |
| Corn | Moisture | 2 | 1.00 | 1 | Range PLS complexity | [2,2] | [2,15] | [14,15] | | | |
| | | | | | number of variables | 8 | 8 | 82 | −1.00 | **−30.81** | **−30.77** |
| | | | | | *RMSEP* | 0.0005 | 0.0005 | 0.0038 | −1.00 | **−28.71** | **−28.71** |
| | Oil | 3 | 0.95 | 27 | Range PLS complexity | [5,11] | [5,11] | [5,11] | | | |
| | | | | | number of variables | 67 | 90 | 192 | **−3.16** | **−22.31** | **−11.56** |
| | | | | | *RMSEP* | 0.067 | 0.067 | 0.069 | 0.37 | −1.72 | **−2.20** |
| | Protein | 4 | 0.92 | 26 | Range PLS complexity | [6,14] | [7,14] | [9,14] | | | |
| | | | | | number of variables | 64 | 78 | 238 | **−4.38** | **−46.55** | **−33.23** |
| | | | | | *RMSEP* | 0.051 | 0.054 | 0.086 | **−2.69** | **−21.34** | **−19.92** |
| | Starch | 5 | 0.94 | 16 | Range PLS complexity | [8,15] | [8,15] | [11,15] | | | |
| | | | | | number of variables | 91 | 100 | 196 | **−2.68** | **−21.57** | **−18.94** |
| | | | | | *RMSEP* | 0.136 | 0.136 | 0.142 | 0.11 | **−2.25** | **−2.70** |
| Alcohols | Propanol | 6 | 0.90 | 22 | Range PLS complexity | [11,19] | [12,19] | [17,19] | | | |
| | | | | | number of variables | 446 | 642 | 1855 | **−4.08** | **−35.19** | **−23.38** |
| | | | | | *RMSEP* | 0.784 | 0.769 | 0.767 | **2.98** | 1.78 | 0.21 |
| | Butanol | 7 | 0.90 | 30 | Range PLS complexity | [9,19] | [12,19] | [15,19] | | | |
| | | | | | number of variables | 571 | 801 | 1657 | **−5.11** | **−22.18** | **−16.73** |
| | | | | | *RMSEP* | 0.778 | 0.763 | 0.735 | 2.11 | **4.75** | **4.58** |
| | Pentanol | 8 | 0.90 | 34 | Range PLS complexity | [11,19] | [12,19] | [16,19] | | | |
| | | | | | number of variables | 450 | 732 | 1886 | **−5.11** | **−26.97** | **−17.49** |
| | | | | | *RMSEP* | 0.951 | 0.940 | 0.984 | 2.10 | **−3.64** | **−5.30** |
| Chromatographic | Log $k_w$ Col1 | 9 | 0.99 | 59 | Range PLS complexity | [1,7] | [2,5] | [3,5] | | | |
| | | | | | number of variables | 45 | 112 | 145 | **−9.64** | **−16.05** | **−6.09** |
| | | | | | *RMSEP* | 0.295 | 0.368 | 0.382 | **−9.83** | **−11.76** | **−3.78** |
| | Log $k_w$ Col2 | 10 | 0.99 | 26 | Range PLS complexity | [1,7] | [2,5] | [3,5] | | | |
| | | | | | number of variables | 38 | 112 | 335 | **−4.12** | **−66.02** | **−33.00** |
| | | | | | *RMSEP* | 0.323 | 0.349 | 0.348 | **−4.85** | **−4.27** | 0.40 |
| | Log $k_w$ Col3 | **11** | 0.99 | 39 | Range PLS complexity | [2,7] | [3,5] | [5,5] | | | |
| | | | | | number of variables | 67 | 132 | 351 | **−5.64** | **−81.71** | **−20.56** |
| | | | | | *RMSEP* | 0.325 | 0.364 | 0.369 | **−6.33** | **−6.92** | **−2.21** |
| | Log $k_w$ Col4 | **12** | 0.99 | 32 | Range PLS complexity | [2,7] | [2,5] | [5,7] | | | |
| | | | | | number of variables | 71 | 114 | 338 | **−4.61** | **−56.37** | **−21.92** |
| | | | | | *RMSEP* | 0.312 | 0.341 | 0.372 | **−4.75** | **−8.86** | **−5.82** |

M1 corresponds to GM-UVE-PLS; M2 corresponds to UVE-PLS; M3 corresponds to 1-step UVE; * For methods M1, M2 and M3, mean values are given for numbers of variables and RMSEP's.
Critical one sided Bonferroni corrected |t|-value is 2.15; significant t-values in bold.

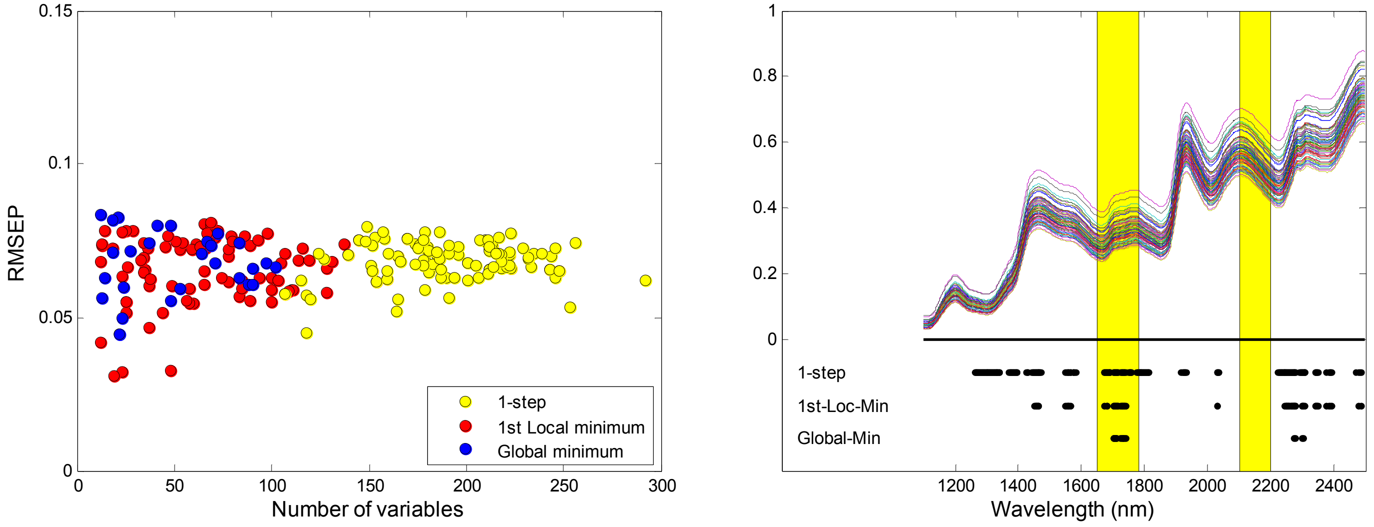


**Fig. 5.** Corn set, response oil; (Left) RMSEP vs. number of retained variables for 100 repeated GME-UVE-PLS runs; (Right) Spectra and selected variables after a GME-UVE-PLS run with selections for 1-step UVE, in the first local RMSECV minimum, and in the global RMSECV minimum; specific spectral absorption bands (see text).

Fig. 5 (Right) shows the corn spectra, the retained wavelengths by the three UVE methods after an example run, and the specific absorption bands for oil. These bands are at 1650–1780 nm and 2100–2200 nm [58,59]. The three UVE methods have retained variables with a chemical meaning, i.e. in the first absorption band. The graph shows that the selective ability is best for the GME-UVE method.

For the Corn set, responses moisture, protein and starch, for the three UVE methods, the ranges for the PLS complexities of the resulting models, the means of the numbers of retained variables and the RMSEP's, and the results of the paired t-tests are given in Table 2. It is also found that, for the 100 repeated runs, for each method, the PLS complexities of the resulting models can vary largely. For moisture, only one run with a global minimum, different from the local minimum, is observed. For the other responses, GME-UVE results in the smallest data sets, which however do not always lead to the best predictions (only the case for protein).

### 4.3.3. Alcohols set

For the Alcohols set with response propanol, in Fig. 6 (Left) results are given for the 100 repeated GME-UVE runs with a cut-off = 0.90. The graph shows that a cluster of retained variables exists for 1-step UVE in the range [1700, 2050]. The numbers of retained variables for the first local minima are in the range [100, 1000] well separated from 1-step UVE. 22 runs have a global minimum and are smaller than the first local minimum, located in the range [50, 300]. No relation seems to be observed between RMSEP and the number of retained variables.

In Table 2 (model 6) calculated t-values are shown for the paired t-tests for numbers of variables for the three pairs of UVE methods. The calculated t-values are significant, i.e. the remaining numbers of variables of GME-UVE are significantly smaller than those of UVE, which again are significantly smaller than those of 1-step UVE. Therefore, for this **X**-**y** combination, the selective ability of the GME-UVE method is significantly better than that of the UVE method, which in turn is significantly better than that of 1-step UVE.

The t-values for the paired t-tests for RMSEP's for three pairs of UVE methods, shown in Table 2 (model 6), are occasionally significant. For instance, the t-value (2.98) for the comparison of RMSEP's of GME-UVE and UVE is significant, but indicates that the RMSEP's of GME-UVE are significantly higher. The t-values for the comparion of RMSEP's of GME-UVE with those of 1-step UVE and of UVE with those of 1-step UVE, 1.78 and 0.21, are not significant. Therefore, for this **X**-**y** combination, the predictive ability of GME-UVE (*i*) is worse than that of UVE, and (*ii*) is similar to that of 1-step UVE. Additionally, the predictive ability of UVE is similar to that of 1-step UVE.

Fig. 6 (Right) shows the $^1$H NMR spectra of mixtures of alcohols and the retained variables by the three UVE methods after one example run. The retained variables for the three UVE methods are spread over the entire spectra, but the number of retained variables for GME-UVE is much lower than that of UVE and 1-step UVE, see also Table 2.

For the other responses, butanol and pentanol, it was also seen that GME-UVE results in the smallest variable sets, but not in better predictive models.

### 4.3.4. Chromatographic set

For the Chromatographic set, with response log $k_w$ on the Zorbax RX-C18 RPLC column (Col1), in Fig. 7 (Left) results are given for 100 repeated GME-UVE runs with a cut-off = 0.99. The graph shows that most points of 1-step UVE are above 100 variables. Below 100 variables, the points for the three methods are mixed. 59 runs have a global minimum, different from the first local minimum, located in the range [2, 61]. More global minima are found at low RMSEP values.

In Table 2 (model 9) the calculated t-values are lower than the critical value |t| = 2.15. The remaining numbers of variables of GME-UVE are significantly smaller than those of UVE, which are significantly smaller than those of 1-step UVE. Therefore, for this **X**-**y** combination, the selective ability of the GME-UVE method is best.

Simultaneously, the RMSEP's of the resulting models of GME-UVE are significantly lower than those of UVE, and those of UVE are significantly lower than those of 1-step UVE. Consequently, for this **X**-**y** combination, the predictive ability of the GME-UVE method is better than that of UVE, which in turn is significantly better than that of 1-step UVE. These observations can be made for all columns.

Fig. 7 (Right) shows the absolute values of autoscaled molecular descriptors, the retained variables by the three UVE methods after one example run, and a small band with molecular descriptors related to the logarithm of the *n*-octanol–water partition coefficient, log *P*. It is well known that the retention time in RPLC is affected by the log *P* of the compounds [60].

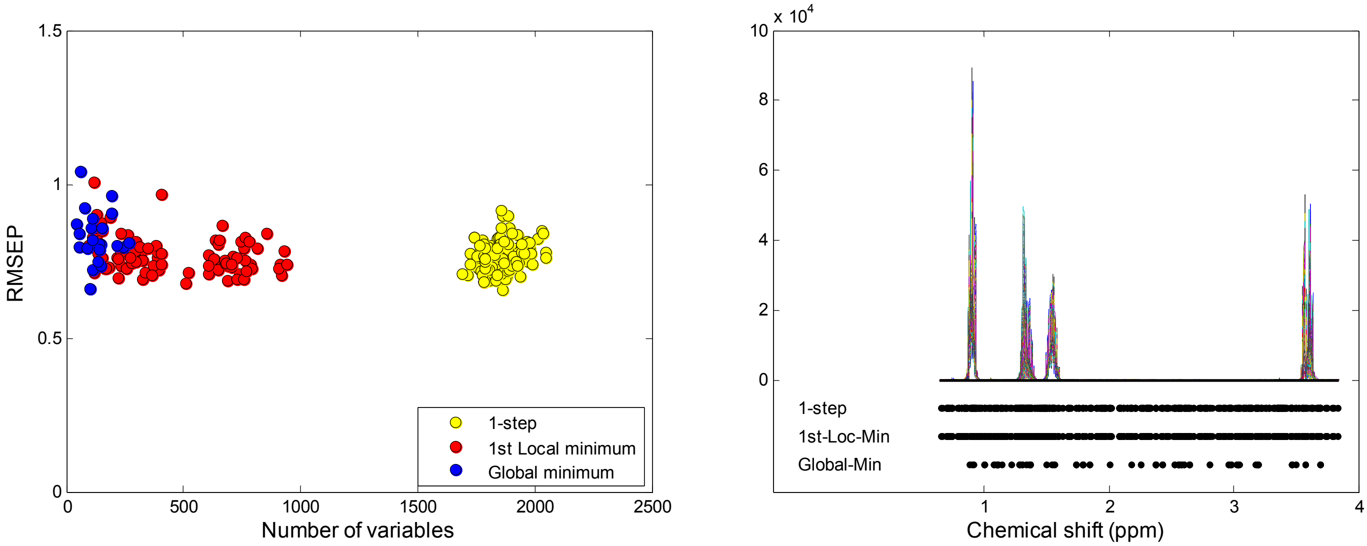


**Fig. 6.** Alcohols set, response propanol; (Left) RMSEP vs. number of retained variables for 100 repeated GME-UVE-PLS runs; (Right) Spectra and selected variables after a GME-UVE-PLS run with selections for 1-step UVE, in the first local RMSECV minimum, and in the global RMSECV minimum.

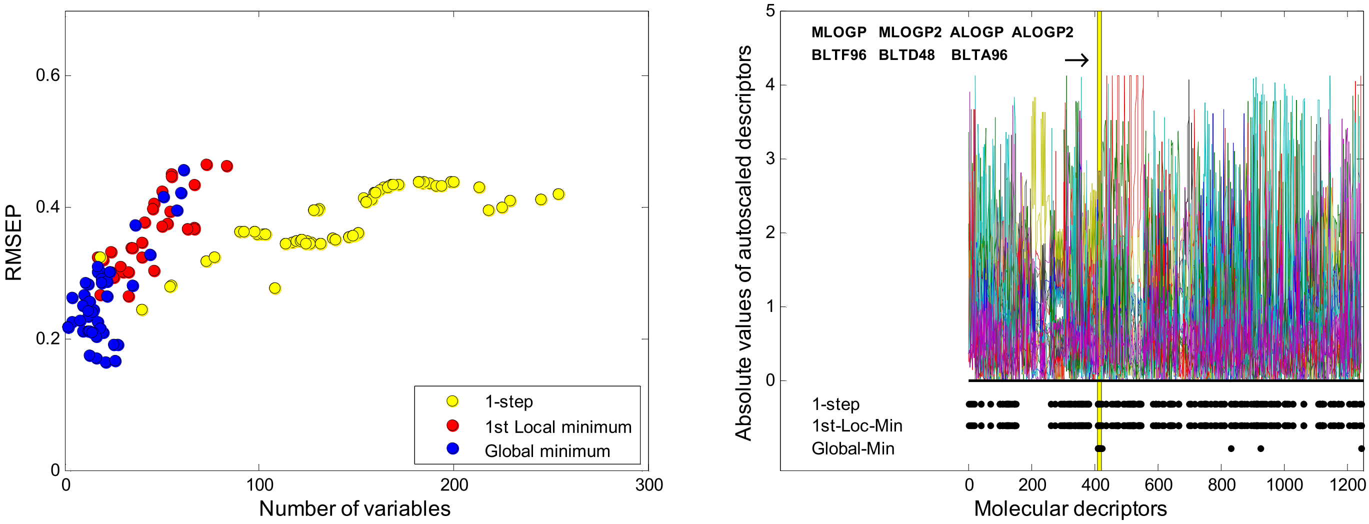


**Fig. 7.** Chromatographic set; (Left) RMSEP vs. number of retained variables for 100 repeated GME-UVE-PLS runs; (Right) Absolute values of autoscaled molecular descriptors and selected variables after a GME-UVE-PLS run with selections for 1-step UVE, in the first local RMSECV minimum, and in the global RMSECV minimum; Molecular descriptors related to log P (see text).

The number of retained variables for GME-UVE (11) in the example run is much lower than that of UVE and 1-step UVE. In the global minimum, among others, the following log *P* related molecular descriptors (see the Dragon molecular descriptor list [61]) are retained: MLOGP, MLOGP2, ALOGP, ALOGP2, BLTF96, BLTD48, and BLTA96. This indicates that the retained variables in the global minimum have a chemical meaning.

### 4.4. Comparison of 1-step UVE, UVE and GME-UVE

The retained variable sets for the three approaches are compared for their selective and predictive performances.

In Table 2, for the 12 **X-y** combinations, results are given for 100 repeated GME-UVE runs. Calculated t-values of paired t-tests are shown for differences in numbers of retained variables and RMSEP's for (*i*) GME-UVE and UVE (M1-M2), (*ii*) GME-UVE and 1-step UVE (M1-M3), and (*iii*) UVE and 1-step UVE (M2-M3).

The number of times that a global minimum different from the first local minimum is observed, varies strongly between the **X-y** combinations, from once for model 2 to 67 times for model 1. In 379 of 1200 runs (i.e. ±32%) a global minimum different from the first local minimum existed.

Eleven **X-y** combinations have a global minimum with significantly lower numbers of retained variables than in the corresponding first local minimum. From these 11 combinations, 6 have significantly lower RMSEP's at the global minima (|t|>2.15) and 4 have RMSEP's which are not significantly different (|t|<2.15). Thus, for 11 of 12 **X-y** combinations, GME-UVE has a significantly better selective ability than UVE, while the predictive ability of 6 combinations is significantly better and that of 4 it is similar.

For all **X-y** combinations, GME-UVE has a significantly better selective ability than 1-step UVE, while the predictive ability for 9 combinations is significantly better and of 2 it is not significantly different.

Classical UVE is beneficial compared to 1-step UVE, because for all **X-y** combinations, the selectivity of UVE is significantly better than for 1-step UVE, while the predictive ability for 9 combinations is significantly better and for 2 it is similar.

Therefore, it is advantageous to apply the global RMSECV minimum as selection criterion for UVE instead of the first local RMSECV minimum as in Ref. [18]. Because a global minimum, different from the local, does not occur in every run, it is recommended to apply a few repeated GME-UVE runs. This improves the chance to find smaller retained variable sets with better predictive abilities. GME-UVE is a fast method, therefore, replication of runs is not a serious drawback.

## 5. Conclusions

In this study, a modification of UVE-PLS, called GME-UVE-PLS or GME-UVE, in which UVE is repeated until no further elimination of variables is obtained, is proposed and investigated. The retained variables in the global RMSECV minimum are selected for GME-UVE-PLS. The predictive and selective abilities of GME-UVE, using the global RMSECV minimum, were compared statistically with those of the classical UVE method, using the first local RMSECV minimum, and 1-step UVE.

The number of retained variables for GME-UVE method is usually found significantly lower than that of the classical UVE method, and the latter is significantly lower than that of 1-step UVE, while the predictive abilities of the resulting models often are better or similar.

Usually, with 1-step UVE and UVE, a higher number of variables without a clear chemical meaning in relation to the response are retained than with GME-UVE.

A global minimum, different from the first local minimum, improves the chance to find smaller retained variable sets with better predictive abilities. Because a global minimum does not occur in every run, it is recommended to repeat a few GME-UVE runs. The selectivity of the UVE method thus can further be improved by the application of the GME-UVE method resulting in more parsimonious models.

## Acknowledgements

The authors thank the anonymous reviewers for their helpful comments and suggestions on an earlier version of this manuscript. T.H.M. Noij, PhD, of the Avans Hogeschool is gratefully thanked for his managerial support of this project.

## References

[1] H. Martens, T. Næs, Multivariate Calibration, second ed., Wiley, New York, 1993.

[2] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.

[3] M. Forina, S. Lanteri, M.C. Cerrato Oliveros, C. Pizarro Millan, Selection of useful predictors in multivariate calibration, Anal. Bioanal. Chem. 380 (2004) 397–418.

[4] C.H. Spiegelman, M.J. McShane, M.J. Goetz, M. Motamedi, Q.L. Yue, G.L. Coté, Theoretical justification of wavelength selection in PLS calibration: development of a new algorithm, Anal. Chem. 70 (1998) 35–44.

[5] S.P. Reinikainen, A. Höskuldsson, Covproc method: strategy in modelling dynamic systems, J. Chemom. 17 (2003) 130–139.

[6] A. Höskuldsson, H-methods in applied sciences, J. Chemom. 22 (2008) 150–177.

[7] L. Xu, I. Schechter, Wavelength selection for simultaneous spectroscopic analysis. Experimental and theoretical study, Anal. Chem. 68 (1996) 2392–2400.

[8] B. Nadler, R.R. Coifman, The prediction error in CLS and PLS: the importance of feature selection prior to multivariate calibration, J. Chemom. 19 (2005) 107–118.

[9] J.P.M. Andries, Y. Vander Heyden, L.M.C. Buydens, Elimination of uninformative variables for multivariate calibration, Anal. Chim. Acta 705 (2011) 292–305.

[10] C.M. Andersen, R. Bro, Variable selection in regression—a tutorial, J. Chemom. 24 (2010) 728–737.

[11] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression, J. Chemom. 23 (2009) 32–48.

[12] J.A. Hageman, M. Streppel, R. Wehrens, L.M.C. Buydens, Wavelength selection with Tabu search, J. Chemom. 17 (2003) 427–437.

[13] W. Cai, Y. Li, X. Shao, A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra, Chemom. Intell. Lab. Syst. 90 (2008) 188–194.

[14] J.P.M. Andries, Y. Vander Heyden, L.M.C. Buydens, Predictive-property-ranked variable reduction in partial least squares modelling with final complexity adapted models: comparison of properties for ranking, Anal. Chim. Acta 760 (2013) 34–45.

[15] J.P. Gauchi, P. Chagnon, Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data, Chemom. Intell. Lab. Syst. 58 (2001) 171–193.

[16] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta 667 (2010) 14–32.

[17] 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.

[18] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.G.M. Vandeginste, C. Sterna, Elimination of uninformative variables for multivariate calibration, Anal. Chem. 68 (1996) 3851–3858.

[19] C. Abrahamsson, J. Johansson, A. Sparén, F. Lindgren, Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets, Chemom. Intell. Lab. Syst. 69 (2003) 3–12.

[20] R. Put, M. Daszykowski, T. Baczek, Y. Vander Heyden, Retention prediction of peptides based on uninformative variable elimination by partial least squares, J. Proteome Res. 5 (2006) 1618–1625.

[21] R. Put, Y. Vander Heyden, The evaluation of two-step multivariate adaptive regression splines for chromatographic retention prediction of peptides, Proteomics 7 (2007) 1664–1677.

[22] A.M. van Nederkassel, M. Daszykowski, D.L. Massart, Y. Vander Heyden, Prediction of total green tea antioxidant capacity from chromatograms by multivariate modeling, J. Chromatogr. A 1096 (2005) 177–186.

[23] T. Hancock, R. Put, D. Coomans, Y. Vander Heyden, Y. Everingham, A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies, Chemom. Intell. Lab. Syst. 76 (2005) 185–196.

[24] H. Swierenga, F. Wülfert, O.E. de Noord, A.P. de Weijer, A.K. Smilde, L.M.C. Buydens, Development of robust calibration models in near infra-red spectrometric applications, Anal. Chim. Acta 411 (2000) 121–135.

[25] N.M. Faber, Improved computation of the standard error in the regression coefficient estimates of a multivariate calibration model, Anal. Chem. 72 (2000) 4675–4676.

[26] H. Xu, Z. Liu, W. Cai, X. Shao, A wavelength selection method based on randomization test for near-infrared spectral analysis, Chemom. Intell. Lab, Syst 97 (2009) 189–193.

[27] J. Moros, J. Kuligowski, G. Quintás, S. Garrigues, M. de la Guardia, New cut-off criterion for uninformative variable elimination in multivariate calibration of near-infrared spectra for the determination of heroin in illicit street drugs, Anal. Chim. Acta 630 (2008) 150–160.

[28] Q.J. Han, H.L. Wu, C.B. Cai, L. Xu, R.Q. Yu, An ensemble of Monte Carlo uninformative variable elimination for wavelength selection, Anal. Chim. Acta 612 (2008) 121–125.

[29] C. Tan, T. Wu, Z. Xu, W. Li, K. Zhang, A simple ensemble strategy of uninformative variable elimination and partial least-squares for near-infrared spectroscopic calibration of pharmaceutical products, Vib. Spectrosc. 58 (2012) 44–49.

[30] D. Jouan-Rimbaud, B. Walczak, R. Poppi, O.E. de Noord, D.L. Massart, Application of wavelet transform to extract the relevant component from spectral data for multivariate calibration, Anal. Chem. 69 (1997) 4317–4323.

[31] X. Shao, F. Wang, D. Chen, Q. Su, A method for near-infrared spectral calibration of complex plant samples with wavelet transform and elimination of uninformative variables, Anal. Bioanal. Chem. 378 (2004) 1382–1387.

[32] S. Ye, D. Wang, S. Min, Successive projections algorithm combined with uninformative variable elimination for spectral variable selection, Chemom. Intell. Lab. Syst. 91 (2008) 194–199.

[33] D. Jie, L. Xie, X. Fu, X. Rao, Y. Ying, Variable selection for partial least squares analysis of soluble solids content in watermelon using near-infrared diffuse transmission technique, J. Food Eng. 118 (2013) 387–392.

[34] L. Yuan, J. Cai, L. Sun, E. Han, T. Ernest, Nondestructive measurement of soluble solids content in apples by a portable fruit analyzer, Food Anal. Methods 9 (2016) 785–794.

[35] H. Yan, B. Han, Q. Wu, M. Jiang, Z. Gui, Rapid detection of Rosa laevigata polysaccharide content by near-infrared spectroscopy, Spectrochim. Acta Part A 79 (2011) 179–184.

[36] S. Ye, D. Wang, S. Min, Successive projections algorithm combined with uninformative variable elimination for spectral variable selection, Chemom. Intell. Lab. Syst. 91 (2008) 194–199.

[37] D. Wu, Y. He, P. Nie, F. Cao, Y. Bao, Hybrid variable selection in visible and near-infrared spectral analysis for non-invasive quality determination of grape juice, Anal. Chim. Acta 659 (2010) 229–237.

[38] F. Westad, H. Martens, Variable selection in near infrared spectroscopy based on significance testing in partial least squares regression, J. Near Infrared Spectrosc. 8 (2000) 117–124.

[39] P. Geladi, B.R. Kowalski, Partial least squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.

[40] B. Efron, G. Gong, A leisurely look at the bootstrap, the jackknife, and cross-validation, Am. Statistician 37 (1983) 36–48.

[41] H. Yan, B. Han, Q. Wu, M. Jiang, Z. Gui, Rapid detection of Rosa laevigata polysaccharide content by near-infrared spectroscopy, Spectrochim. Acta Part A 79 (2011) 179–184.

[42] L. Yuan, J. Cai1, L. Sun, E. Han, T. Ernest, Nondestructive measurement of soluble solids content in apples by a portable fruit analyzer, Food Anal. Methods 9 (2016) 785–794.

[43] D. Wu, X. Chen, X. Zhu, X. Guan, G. Wu, Uninformative variable elimination for improvement of successive projections algorithm on spectral multivariable selection with different calibration algorithms for the rapid and non-destructive determination of protein content in dried laver, Anal. Methods 3 (2011) 1790–1796.

[44] V. Pravdova, B. Walczak, D.L. Massart, S. Kawano, K. Toyoda, R. Tsenkova, Calibration of somatic cell count in milk based on near-infrared spectroscopy, Anal. Chim. Acta 450 (2001) 131–141.

[45] http://www.vub.ac.be/fabi/research/chemoac/toolbox.html (accessed February 1, 2017).

[46] B. Li, J. Morris, E.B. Martin, Model selection for partial least squares regression, Chemom. Intell. Lab. Syst. 64 (2002) 79–89.

[47] S. Wold, Cross-validatory estimation of the number of components in factor and principal components models, Technometrics 24 (1978) 397–405.

[48] B.M. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Windig, R. Scott Koch, PLS_ Toolbox Version 4.0, Eigenvector Research, Wenatchee.

[49] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Part a, Elsevier, Amsterdam, 1997.

[50] R.D. Snee, Validation of regression models: methods and examples, Technometrics 19 (1977) 415–428.

[51] http://www.eigenvector.com/data/Corn/index.htmlhttp://software.eigenvector.com/ (accessed on February 1, 2017).

[52] http://www.models.kvl.dk/datasets (accessed on February 1, 2017).

[53] H. Winning, F.H. Larsen, R. Bro, S.B. Engelsen, Quantitative analysis of NMR spectra with chemometrics, J. Magn. Reson. 190 (2008) 26–32.

[54] R. Kaliszan, M.A. van Straten, M. Markuszewski, C.A. Cramers, H.A. Claessens, Molecular mechanism of retention in reversed-phase high-performance liquid chromatography and classification of modern stationary phases by using quantitative structure–retention relationships, J. Chromatogr. A 855 (1999) 455–486.

[55] http://www.hyper.com/ (accessed February 1, 2017).

[56] https://chm.kode-solutions.net/products dragon.php (accessed February 1, 2017).

[57] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148.

[58] T. Sato, S. Kawano, M. Iwamoto, Near infrared spectral patterns of fatty acid analysis from fats and oils, JAOCS 68 (1991) 827–833.

[59] J.A. Panford, J.M. deMan, Determination of oil content of seeds by NIR: influence of fatty acid composition on wavelength selection, JAOCS 67 (1990) 473–482.

[60] R. Put, Y. Vander Heyden, Review on modelling aspects in reversed-phase liquid chromatographic quantitative structure–retention relationships, Anal. Chim. Acta 602 (2007) 164–172.

[61] http://www.talete.mi.it/products/dragon_molecular_descriptor_list.pdf (accessed February 1, 2017).