Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Chromatography A, 1096 (2005) 177–186
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

Prediction of total green tea antioxidant capacity from chromatograms by multivariate modeling

A.M. van Nederkassel [a], M. Daszykowski [a,b], D.L. Massart [a], Y. Vander Heyden [a,*]

[a] *Department of Pharmaceutical and Biomedical Analysis, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, 1090 Brussels, Belgium*
[b] *Department of Chemometrics, Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland*

Available online 18 April 2005

Abstract

In this paper, a fast strategy for determining the total antioxidant capacity of Chinese green tea extracts is developed. This strategy includes the use of experimental techniques, such as fast high-performance liquid chromatography (HPLC) on monolithic columns and a spectrophotometric approach to determine the total antioxidant capacity of the extracts. To extract the chemically relevant information from the obtained data, chemometrical approaches are used. Among them there are correlation optimized warping (COW) to align the chromatograms, robust principal component analysis (robust PCA) to detect outliers, and partial least squares (PLS) and uninformative variable elimination partial least squares (UVE-PLS) to construct a reliable multivariate regression model to predict the total antioxidant capacity from the fast chromatograms.

Keywords: Green tea; Antioxidant capacity; Monolithic columns; Warping; Aligning; Multivariate calibration; PLS

1. Introduction

The dry leaves of the green tea plant *Camellia sinensis* (L.) are known to contain flavanols with antioxidant capacity such as catechin, (−)-epicatechin, (−)-epicatechin-3-gallate, (−)-epigallocatechin and (−)-epigallocatechin-3-gallate [1,2]. The consumption of green tea is therefore associated with protective effects against coronary heart diseases and cancers of the lung, forestomach, esophagus, duodenum, pancreas, liver, breast, colon, and skin, induced by chemical carcinogens [1]. The total antioxidant capacity of a green tea extract is an important quality criterion and can be determined with a decolorization assay, the Trolox equivalent antioxidant capacity (TEAC) method, in which the antioxidant capacity is expressed as the concentration of a Trolox solution with equal antioxidant capacity [3]. Besides the flavanols, the green tea may contain many other components, also contributing to the safety and quality of the extract. To avoid a time-consuming qualitative and quantitative analysis of each extract compound individually, fingerprint technology was introduced for quality control. Fingerprint chromatograms are developed and compared to that of a standardized extract to achieve authentication, identification and quality control of the herb [4,5]. Combining the information from the green tea fingerprint and the TEAC assay allows a good determination of the quality of the tea extract. An application can be found in the paper by Koleva et al. [6], where a post-column TEAC assay is presented to measure the antioxidant capacity of HPLC separated analytes of o.a. Rosemary extracts. However, for high-throughput applications simple and fast methods are preferred and in this paper, it is therefore investigated whether the total antioxidant capacity of green tea extracts can simply be predicted from their fast chromatograms, developed on monolithic silica columns. It was earlier shown that fast, robust and repeatable separations can be obtained at high flow rates (up to 9 ml/min) on these columns, allowing to speed up the analysis compared to conventional particle-packed HPLC columns [7].

A multivariate calibration model relating the chromatographic data with the TEAC data is built. A similar study but based on NIR spectra, instead of chromatograms,

* Corresponding author. Tel.: +32 2 477 47 34; fax: +32 2 477 47 35.
E-mail address: yvanvdh@vub.ac.be (Y. Vander Heyden).

0021-9673/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2005.03.102

is presented by Luypaert et al. [8]. However, chromatograms have the advantage that qualitative and quantitative information of given compounds can be retrieved, if needed.

In this paper, the total antioxidant capacity of 55 Chinese green tea extracts is determined using the TEAC assay. Simultaneously, these extracts were chromatographed with two fast HPLC methods (analysis times of 11 and 2 min), resulting in nearly baseline separated and only partially separated peaks, respectively. The aim of our work is to verify the potential use of fast chromatograms in the construction of multivariate models relating the chromatographic profiles with the antioxidant capacity of green tea. Multivariate calibration models such as PLS and UVE-PLS [9,10] were constructed for the short and long chromatograms. Due to instability of the HPLC instrument as well as small mobile phase variations, peak shifts were observed. Therefore, prior to model building, the chromatograms were aligned using correlation optimized warping (COW) [11–13]. In order to detect the presence of outliers in the data, robust principal component analysis (robust PCA) [14–16] is applied.

2. Theory

2.1. TEAC assay

In this study the Trolox equivalent antioxidant capacity (TEAC) assay, described by Re et al. [3], is used with slight modifications (see Section 3). The TEAC assay measures the capacity of a compound to scavenge the blue–green ABTS cationic radical resulting in a colorless product [3]. The acronym $ABTS^{\bullet+}$ denotes 2,2′-azinobis-(3-ethylbenzothiazoline-6-sulfonic acid). The amount of $ABTS^{\bullet+}$ scavenged by the antioxidants in the green tea extract is measured by the degree of decolorization, measured spectrophotometrically at 729 nm. In this study, the TEAC value reflects the scavenging capacity of 1% (m/v) green tea extract expressed as the equivalent concentration (in mM) of Trolox, a water-soluble Vitamin E analogue. More details about the TEAC assay can be found in ref. [3].

2.2. Correlation optimized warping

The correlation optimized warping, COW, aims to correct peak shifts along the time axis in chromatograms. There are a few reasons causing this phenomenon. Among them there are variations in mobile phase composition, column ageing and instrument instability. If peak shifts occur, the chromatograms must be aligned to ensure a proper chemometrical treatment of the data. This means that if the chromatograms are placed as rows of the data matrix, the apexes of the corresponding peaks ought to be in the same columns.

Several aligning (warping) techniques of analytical signals are described in the literature. Among them there are dynamic time warping [17], parametric time warping (PTW) [18,19], peak alignment by a genetic algorithm (PAGA) [20] and fuzzy warping [21]. In our application correlation optimized warping, COW, is used [11–13]. COW is a method which does not require peak detection. It aligns two signals (e.g., chromatograms) by means of piecewise linear stretching and compression of the chromatogram to align in order to match it as good as possible with a target chromatogram. At the beginning of the procedure both signals, the profile to be aligned, P, and the target profile, T, are divided into a user-specified number of sections, N. Each section in the profile P is warped, what means that its length is stretched or shortened by shifting the position of its section end point by a limited number of data points, defined by the slack parameter, t. The slack allows the section end points to shift from $-t$ to t points. For each section of P, the stretched or shortened sections are interpolated to the length of the corresponding section in T and the correlation coefficients between them are computed. The correlation coefficients allow to score a warping solution constructed as the cumulative sum of the correlation coefficients obtained for the previous sections. For every possible end point of a section (from $-t$ to t) always the highest value of the cumulative function is stored. After examining all possible end points of all sections, a global warping solution is constructed. It is found starting from the last section backwards, by determining the maximal value of the objective function and its corresponding end point for every section. A more detailed description of the COW algorithm can be found in refs. [11–13].

2.3. Leverage object and outlier detection

In order to detect objects with extreme characteristics in the space of chromatograms ($\mathbf{X}$) and outliers in TEAC values ($\mathbf{y}$), robust PCA and histograms are used, respectively. Extreme objects have to be identified and removed since they affect to a large extent the result of the data analysis. An important tool for visualizing data structure is principal component analysis [10], one of the most often applied dimensionality reduction methods. With PCA it is usually possible to project original data variables onto a set of a few new features, so-called principal components, that are mutually orthogonal and are linear combinations of the original variables [10]. The PCs are constructed in order to maximize the description of the data variance, and thus, the first few PCs represent the majority of the data variance, whereas the remaining ones are related to random noise. Because the extracted PCs maximize the data variance, PCA is sensitive to the presence of outlying objects, and often the PCs reveal the presence of atypical objects. In order to identify objectively atypical objects, one can apply robust PCA, obtained by substituting the variance criterion by a so-called robust scale [14–16]. Robust PCA is regarded as a model aiming to describe well the data majority. With this assumption outlier detection is possible based on the residuals from the robust PCA model. Taking into account the distance of an object from the data majority

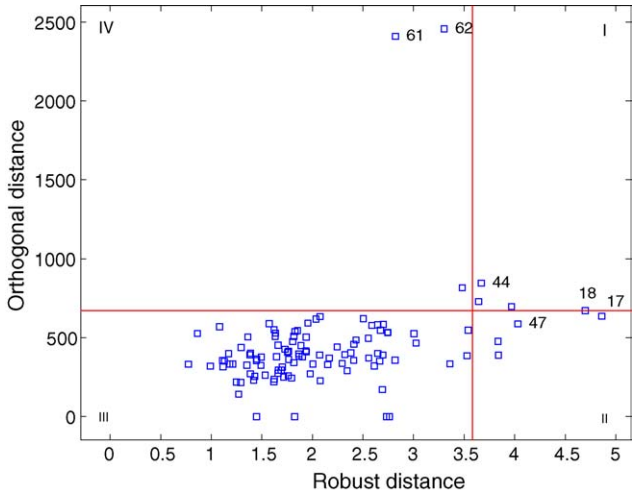

Fig. 1. The score diagnostic plot of the 110 long chromatograms. The orthogonal distance is plotted vs. the robust distance. The cut-off values are determined in the space of five rPC's.

(robust distance) and its distance from the robust PCA model space (orthogonal distance) two cut-off values are proposed [14,15], as illustrated in Fig. 1. Samples that exceed cut-off values for both distances are bad leverage objects (quadrant I in Fig. 1) because they are far from the data majority and also from the model's space. Objects that are only far from the data majority but close to model's space are good leverage objects (quadrant II), contrary to high-residual objects that do not fit the robust model and exceed the cut-off value for orthogonal distance (quadrant IV). Schematically, three types of leverage objects in **X**-space are presented in Fig. 1. A more detailed description of the robust PCA algorithm can be found in refs. [14–16].

2.4. Multivariate regression

2.4.1. Partial least squares (PLS)

In partial least squares regression, PLS, a dependent variable, **y**, is modeled using latent variables, maximizing the covariance between **X** and **y**. The PLS model can be presented as follows [22,23]:

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

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

where $\mathbf{X}(n, p)$ represents the data matrix, vector $\mathbf{y}(n, 1)$ is a dependent variable, $\mathbf{T}(n, n)$ is the score matrix, $\mathbf{P}^{\mathrm{T}}(n, p)$ denotes the transposed loading matrix, $\mathbf{q}(n, 1)$ is a loading vector, $\mathbf{E}(n, p)$ and $\mathbf{f}(n, 1)$ are the residuals.

In order to predict y_i for a new chromatogram $\mathbf{x}_i(1, p)$, the following equation can be used:

$$\hat{y}_i = \bar{y} + \mathbf{x}_i\mathbf{b} \tag{3}$$

where $\hat{y}_i$ is the predicted dependent value for the ith new sample, $\bar{y}$ denotes the mean of the dependent values for the calibration samples, and $\mathbf{b}(p, 1)$ is the vector of PLS regression coefficients computed as:

$$\mathbf{b} = \mathbf{P}\mathbf{q} \tag{4}$$

The optimal model complexity, i.e., the number of latent factors in the PLS model, can be determined by the leave-one-out cross-validation procedure (LOO-CV). During LOO-CV, each ith object of the data set is left out once, and for the remaining objects the PLS model is built. Then, the root mean squared error of cross-validation (RMSECV) is computed for PLS models with different numbers of latent factors [22]:

$$\mathrm{RMSECV}(f) = \sqrt{\sum_{i=1}^{N} \frac{(\hat{y}_{\mathrm{CV},i} - y_i)^2}{N}} \tag{5}$$

where y_i is the measured response of the ith sample, $\hat{y}_{\mathrm{CV},i}$ is a predicted response from a calibration equation obtained for the data without the ith sample, N is the number of calibration samples, and f denotes number of latent factors.

The optimal complexity of the PLS model corresponds to the number of latent factors resulting in the lowest RMSECV. Additionally, the PLS model can be validated with an independent test set for which the root mean squared error of prediction (RMSEP) is computed as [22]:

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

In Eq. (6), N_{t} is the number of test set samples, $\hat{y}_i^{\mathrm{t}}$ and y_i^{t} are the predicted and measured response values for the test set samples [23].

2.4.2. Uninformative variable elimination partial least squares (UVE-PLS)

The uninformative variable elimination-partial least squares approach, UVE-PLS, relies on the principle of the PLS method and aims to remove uninformative variables, i.e., the variables not more informative for modeling than noise. The main steps of UVE-PLS can be summarized as follows [9]:

(1) find the optimal complexity of the PLS model using the cross-validation procedure;
(2) simulate a matrix $\mathbf{R}(n, r)$ with r artificial variables as columns (where $r > 300$), with numbers drawn from a normal distribution and multiplied by a small constant 10^{-10};
(3) augment the original data $\mathbf{X}(n, p)$ with matrix **R** to form $\mathbf{XR}(n, p + r)$;
(4) construct n PLS models for the augmented data matrix with leave-one-out cross-validation, and store n vectors of regression coefficients, **b**, in a matrix **B** of a size $(n, p + r)$;

(5) for every regression coefficient, define its stability coefficient as the ratio of the column mean and the column standard deviation of **B**;
(6) define a cut-off value to distinguish between informative and uninformative variables as the absolute value of the maximal stability of the regression coefficients describing artificial variables;
(7) remove from the original data matrix **X** all the variables for which the absolute value of the stability coefficient is below the cut-off value;
(8) construct a final PLS model for the data containing informative variables only.

3. Experimental

3.1. Instruments, chemicals and mobile phases

3.1.1. Instruments

The high-performance liquid chromatography (HPLC) system consists of a L-7100 pump, L-7612 solvent degasser, L-7250 autosampler, L-7360 oven, L-7400 UV detector and a D-7000 interface from Merck-Hitachi (Tokyo, Japan). The volume and path length of the UV detector are 17 μl and 1 cm respectively. This system is operated with the LaChrom D-7000 HPLC Manager Software (Merck-Hitachi). The column oven temperature is 30 °C and the detection wavelength 280 nm. The injection volumes, flow rates and sampling rates are 15 μl, 2 ml/min and 200 ms and 10 μl, 5 ml/min and 100 ms for the methods with analysis times of 11 and 2 min, respectively.

A UV-2101 PC spectrophotometer (Shimadzu, Tokyo, Japan) at detection wavelength 729 nm is used for the determination of the antioxidant capacity with the TEAC assay. The cuvet path length is 1 cm.

3.1.2. Chemicals and reagents

Trolox (6-hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid 97%), $ABTS^{\bullet+}$ = 2,2′-azino-bis-(3-ethylbenzothiazoline-6-sulfonic acid), potassium persulfate and (−)-epigallocatechin gallate (EGCG) were purchased from Sigma-Aldrich (Steinheim, Germany). Caffeine is obtained from Fluka (Buchs, Switzerland) and ethanol (LiChrosolv quality) from Merck (Darmstadt, Germany). A 1.2 mg/ml caffeine and 0.2 mg/ml EGCG standard in Milli-Q water is prepared for peak identification purposes.

3.1.3. Mobile phases

The green tea extracts were separated using two gradient elution methods containing acetonitrile (ACN) (Hipersolv for HPLC quality, BDH Laboratory Supplies, Poole, England) and Milli-Q water (Milli-Q water purification system, Millipore, Molsheim, France) both with 0.05% trifluoroacetic acid (Sigma–Aldrich). In the fast method, the ACN changes from 5% to 26% within 0.7 min and then remains constant at 26%. In the long method the percentage ACN varies from 2% to 26% within 10 min and remains constant at 26%. The analysis times are 2 and 11 min, respectively.

3.1.4. Column

The stationary phase consists of a Chromolith SpeedROD (50 mm × 4.6 mm) and a Performance (100 mm × 4.6 mm) column both RP-18e from Merck which are coupled in line with a column coupler (Merck). A Chromolith guard column RP-18e (5 mm × 4.6 mm, Merck) was placed before the analytical ones.

3.1.5. Software

All data processing methods (COW, robust PCA, PLS and UVE-PLS) are done with subroutines developed under Matlab™ 5.3 software (The MathWorks Inc., Natick, MA). Computations were performed on a computer with a 1000 mHz Athlon processor and 256 MB RAM.

3.2. Preparation of the green tea extracts, $ABTS^{\bullet+}$ and Trolox solutions

The 55 Chinese green teas were purchased in several Chinese stores. The extracts were prepared as follows: 2.0 g of dry green tea leaves are milled during 3 × 10 s with a Moulinette 320 mixer (Moulinex, France) and sieved trough a 500 μm DIN 4188 sieve (Retsch, Haan, Germany). Then 0.5 g of the sieved tea is infused during 7 min in the dark with initially boiling Milli-Q water. The infusion occurred at room temperature in a 100.0 ml volumetric flask without mixing. The warm extracts were successively filtered through a 90 μm DIN 4188 sieve (Retsch) and a 0.2 μm membrane filter (Pall Gelman Laboratory, Karlstein/Main, Germany). The green tea extracts were stored in dark glass recipients at 9 °C before injection on the column, which was the day after preparation. Each extract was chromatographed in duplicate for both gradient elution methods. The same day, the extracts were also analyzed with the TEAC assay, after dilution (1/200) with Milli-Q water. The obtained TEAC value is multiplied by two to obtain a TEAC value equivalent to a 1% (m/v) green tea extract.

An aqueous solution containing 7.00 mM ABTS and 2.45 mM $K_2S_2O_8$ is prepared and stored in the dark during at least 12 h to produce the $ABTS^{\bullet+}$ radical. The solution is no longer used than three days after preparation. Before use, the solution is diluted about 1/40 with Milli-Q water to obtain a solution with absorbance between 1.5 and 1.7 at 729 nm.

A 5.00 mM Trolox solution in ethanol is prepared and stored at −7 °C. Daily a Trolox calibration line (35, 40, 45, 50 and 55 μM) in Milli-Q water is made.

3.3. TEAC assay

The antioxidant capacity of Trolox and of the green tea extracts is determined by measuring the decrease in absorbance at 729 nm (experimental λ_{max}), 60 s after addition of 0.3 ml

Trolox standard (or the 1/200 diluted green tea extract) to 1.0 ml diluted $ABTS^{\bullet+}$ solution with an absorbance between 1.5 and 1.7. The solvents in the cuvet are mixed by withdrawing and replacing 10 times a 300 μl volume of the solutions using a micropipet. The measurements are done in quadruplicate for each Trolox standard and green tea extract and the average decrease in absorbance is computed ($\overline{\Delta A}$). The calibration curve ($\overline{\Delta A}$ versus Trolox concentration) is then used to compute the equivalent Trolox concentration for each diluted green tea extract. This value is multiplied with 333 (300 μl from the 100 ml tea extract is analyzed) and with the dilution factor (200), and corrected for the amount tea weighted (in case different from 0.5000 g) and multiplied by 2 to obtain the TEAC value as for 1.0 g of green tea leaves.

Since caffeine is a major component of green tea, the $\overline{\Delta A}$ for a 4 mg/ml aqueous caffeine solution was measured and compared to the $\overline{\Delta A}$ for Milli-Q water. It was found that both $\overline{\Delta A}$ do not differ significantly from each other and thus it can be concluded that caffeine does not contribute to the TEAC value obtained for green tea extracts.

3.4. Precision of the TEAC assay

The precision of the reference method is determined for three teas with low (2999 mM), intermediate (3779 mM) and high (4058 mM) TEAC values, respectively. The standard deviation (SD) of the TEAC values, obtained for six replicate measurements of these three teas was found to be respectively 93 (5.62%), 104 (6.29%) and 204 (12.33%) mM. The pooled SD [10] is 143 (8.65%) and is strongly influenced by the tea with high TEAC value. The percentages between brackets are computed with 100% being the range of TEAC values (values between 2792 and 4446 mM after outlier removal, see Section 4.2).

4. Results and discussion

Prior to modeling, the 110 green tea extract chromatograms, obtained for each chromatographic method, are warped to correct the peak shifts. A proper alignment of chromatograms enables their use for multivariate regression purposes. The multivariate model aims relating chromatographic variables and the TEAC values. By use of robust principal component analysis and histograms it is possible to detect leverages in **X** and outliers in **y**, respectively. After the chromatograms alignment and outlier elimination, the mean of the replicates is computed and the data is divided into a calibration and test set. Finally, PLS and UVE-PLS methods are used to construct multivariate regression models.

4.1. Alignment of the chromatograms

Prior to the alignment of chromatograms, it is checked whether a baseline correction is necessary and uninformative baseline (200 sampling points) preceding the dead time is removed, resulting in signals of 3100 and 1000 points, for the long and short chromatograms, respectively. In both chromatographic profiles peak shifts occurred what is illustrated in Figs. 2a and 3a. In the long chromatograms the shifts vary between −8 and 47 data points (Fig. 2a), while in the short chromatograms the shifts vary from −11 up to 26 points (Fig. 3a). The negative sign refers to a forward shift, and a positive sign refers to a backward shift, compared to chromatogram one, selected as the most representative one for the alignment of the remaining chromatograms. The alignment with COW requires the optimization of two input parameters, N and t. The optimization is done as follows. First a reference signal is selected. It must be a representative chromatogram in which all peaks are clearly present. Chromatogram one fulfills these requirements for both data sets (long and short

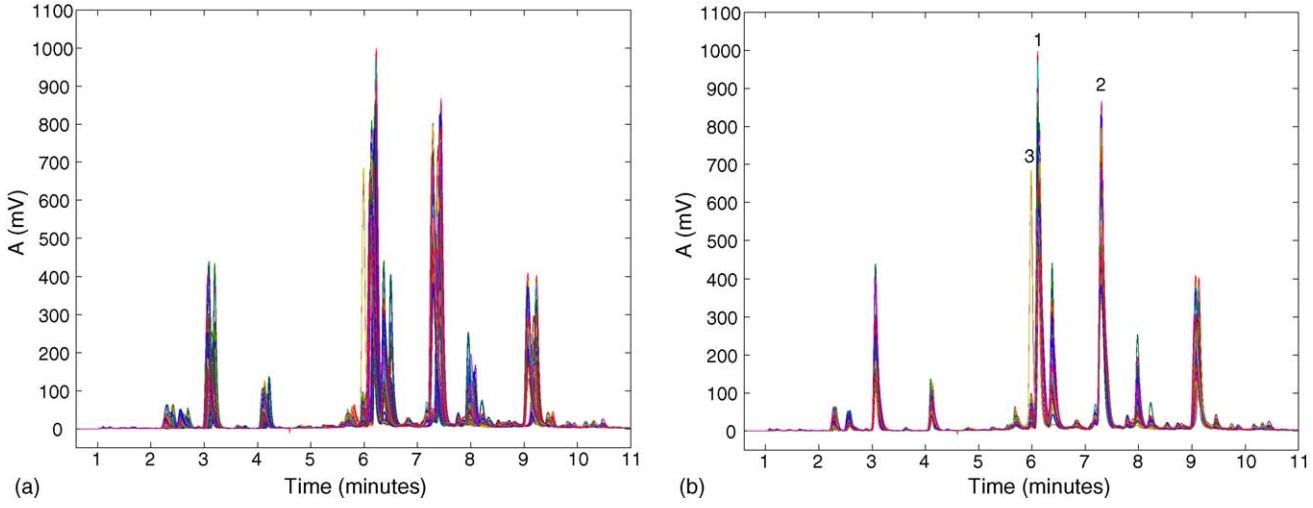

Fig. 2. One hundred and ten long chromatograms before (a) and after (b) warping ($N = 60$, $t = 3$) with peak 1 caffeine, peak 2 (−)-epigallocatechin gallate and peak 3 an unidentified substance, only in high concentrations present in one tea extract.

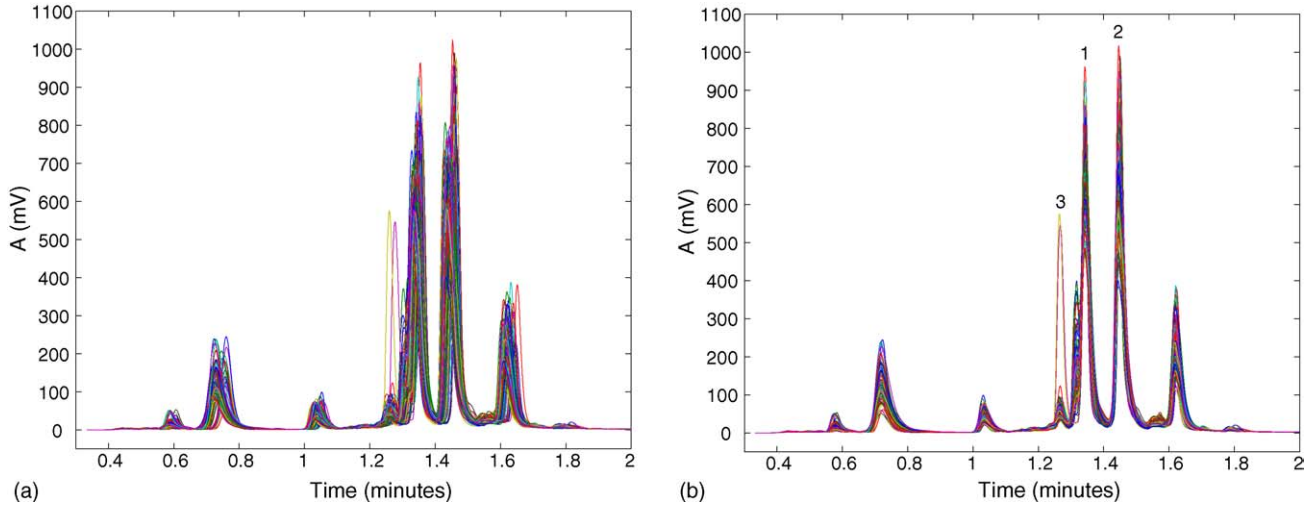

Fig. 3. One hundred and ten short chromatograms before (a) and after (b) warping ($N = 30$, $t = 3$) with peak 1 containing caffeine, peak 2 containing (−)-epigallocatechin gallate and peak 3 an unidentified substance, only in high concentrations present in one tea extract.

chromatograms) and is therefore chosen as reference. Then, chromatogram 17 is selected as the first chromatogram to align because it has most likely the largest peak shifts with respect to the reference. N and t are optimized by varying N from 10 to 70 and t from 1 to 6. With $N = 60$, $t = 3$ and $N = 30$, $t = 3$ the long and short chromatogram 17, respectively, were aligned. These parameters are then used to align the whole data set. If still some signals remain unaligned, N and t can be further optimized for the unaligned signals. However, it was not necessary in our case. The aligned signals are shown in Figs. 2b and 3b. The alignment of the 110 long and short chromatograms took 24 and 6 min, respectively.

4.2. Leverage objects and outliers

Before constructing a calibration model, the presence of leverages and outlying observations in the space of chromatograms, **X**, and TEAC values, **y**, is examined using robust PCA and histograms, respectively. In Figs. 1 and 4, plots of the orthogonal *versus* robust distances are shown for the long and short chromatograms for five and six robust principal components, respectively. The number of robust principal components was derived from so-called scree plots [14]. The cut-off values for the robust and orthogonal distances are indicated in the Figs. 1 and 4 by vertical and horizontal lines, respectively. These cut-off lines divide the space into four quadrants. The objects in the first quadrant are bad leverages, those in the second are good leverages, in the third quadrant are ordinary objects, and in the fourth quadrant are orthogonal outliers. In Figs. 1 and 4, both chromatograms 61 and 62 (located in the fourth quadrant) can be considered as orthogonal outliers. They are replicate chromatograms of one tea sample. A closer look at these chromatograms explains why they are identified as orthogonal outliers. Both have an exceptional high peak (labeled as peak 3) at 6 and 1.25 min, in the long and short chromatograms, respectively, which is at least five times higher than in all other chromatograms (see Figs. 2b and 3b). Besides this peak, they do not differ from the average chromatogram. Only one such tea extract was present in the data set and therefore this sample is considered 'atypical'. It would be interesting to construct a model, able to predict also atypical samples but building a model with only one such sample in the calibration set might damage the model with a consequence for future sample predictions. Moreover, having only one such sample prevents the evaluation of the prediction of new atypical samples and therefore, this tea sample is removed. In a situation where there are more atypical samples, they would not be called atypical anymore and we would keep them and divide them over the calibration and test set prior to modeling.

In Figs. 1 and 4, it is seen that chromatograms 17, 18, 29, 44 and 47 are situated in the first or second quadrant

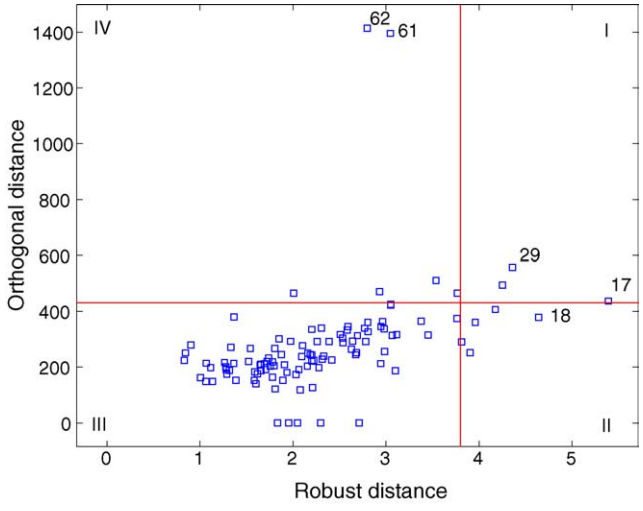

Fig. 4. The score diagnostic plot of the 110 short chromatograms. The orthogonal distance is plotted vs. the robust distance. The cut-off values are determined in the space of six rPC's.

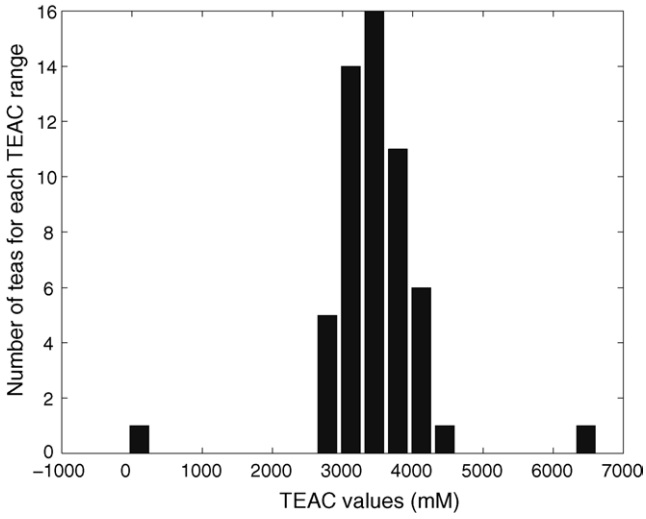

Fig. 5. Histogram of the 55 TEAC values.

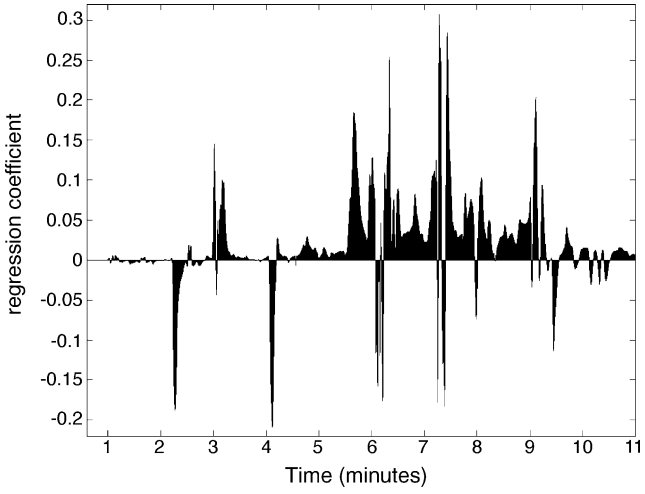

Fig. 6. Regression coefficients of the PLS model obtained for the long chromatograms.

and are either bad or good leverages. The distance of these objects from the majority of objects is not high compared to chromatograms 61 and 62. Therefore, the latter objects were kept in both data sets.

A histogram of the TEAC values of the 55 Chinese green tea extracts is shown in Fig. 5. It reveals that two tea extracts have extreme TEAC values, probably due to experimental errors. Since the experimental conditions did not allow to remake and remeasure the two samples, these two samples were removed. In total, three tea samples are removed from the data set. One tea sample is a leverage object in **X**, and two tea samples have atypical TEAC values. After removing outliers, the data sets contain 52 chromatograms obtained by averaging the replicates of the warped chromatograms.

4.3. Subset selection

The data sets with short and long chromatograms and their corresponding TEAC values are divided into a calibration set, to build to model, and a test set to validate the model. A calibration set was selected by uniform sampling of sorted, from low to high, TEAC values. The calibration set contains forty samples, and the remaining 12 objects are used as an external test set.

4.4. PLS and UVE-PLS models

The PLS model built for the long chromatograms contains eight PLS factors (see Table 1a). For this model a RMSECV value comparable to the precision of the reference method is obtained. The RMS, RMSECV and RMSEP of this model are respectively 81, 159 and 174 mM, what corresponds to about 5%, 10% and 11% of the total range of the TEAC values, respectively. By way of illustration, the regression coefficients of the model are shown in Fig. 6.

Using the UVE-PLS approach it is possible to remove the uninformative chromatographic variables, which have a high variance but a small covariance with the response values. Usually, the UVE-PLS approach yields less complex models in terms of number of latent variables compared to classical PLS, and also, offers an improvement of prediction abilities of the model. In our application, by means of the UVE-PLS, the complexity of the initial PLS model is reduced to seven factors, and the number of considered variables from 3100 to 142 only. In Fig. 7, the retained variables of the UVE-

Table 1
Models constructed for (a) the long, and (b) the short chromatograms and (c) the long chromatograms of which the number of sampling points is reduced by averaging

Model	Fn	RMS	RMSECV	RMSEP
(a) Models constructed for the long chromatograms				
PLS	8	80.53 (4.87%)	159.20 (9.63%)	173.76 (10.51%)
UVE-PLS	7	156.57 (9.47%)	114.28 (6.91%)	171.13 (10.35%)
(b) Models constructed for the short chromatograms				
PLS	3	177.03 (10.70%)	205.97 (12.45%)	176.74 (10.69%)
UVE-PLS	3	148.14 (8.96%)	165.34 (10.00%)	112.00 (6.77%)
(c) Models constructed for the reduced chromatograms				
PLS	7	106.31 (6.43%)	162.56 (9.83%)	123.59 (7.47%)
UVE-PLS	6	118.96 (7.19%)	147.83 (8.94%)	171.14 (10.35%)

Fn: model complexity, RMS: root mean squared error, RMSECV: root mean squared error of cross-validation and RMSEP: root mean squared error of prediction.

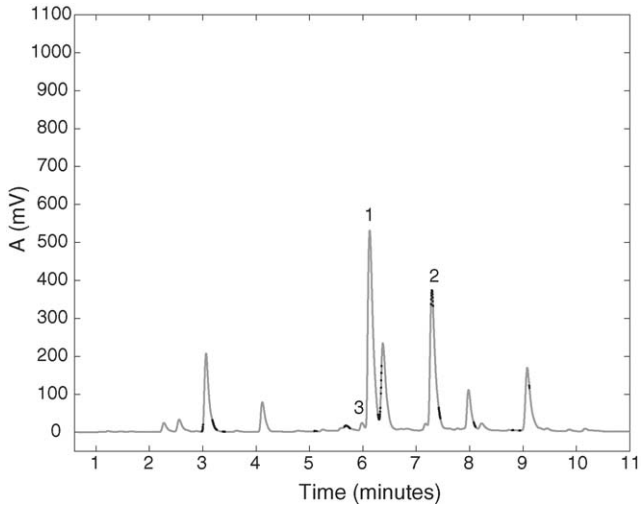

Fig. 7. Plot of the reference chromatogram for the alignment of the long chromatograms. Only the dark variables are retained in the UVE-PLS model. Peak 1 represents caffeine, peak 2 (−)-epigallocatechin gallate and peak 3 is an unidentified substance, which is only clearly present in chromatograms 61 and 62.

PLS model are indicated as dark spots on the reference chromatogram used for the alignment. From this Figure, it can be seen that the retained variables are selected from the peaks with retention times 3.0, 5.7, 6.4, 7.3, 8 and 9.1 min. Among the uninformative variables there are baseline variables but also variables from the caffeine peak and other peaks for which we did not have standards to identify them. It is not unexpected that variables from the large caffeine peak (peak 1 in Figs. 2b and 7) are not used in the UVE-PLS model since in Section 3.3 it was shown that caffeine has no antioxidant capacity. The removal of baseline variables is obvious. The UVE-PLS model seems to be more stable in terms of more constant RMS values (Table 1a) and thus is to be preferred.

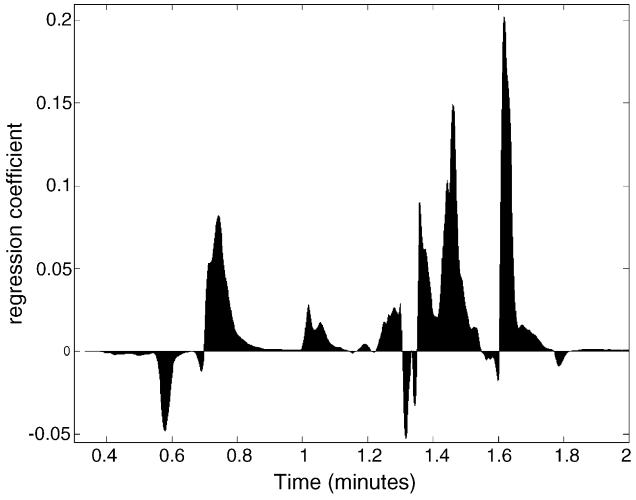

Fig. 9. Regression coefficients of the PLS model obtained for the short chromatograms.

The RMS, RMSECV and RMSEP of this model are respectively 157, 114 and 171 mM, what corresponds to about 9%, 7% and 10% of the total range of the TEAC values, respectively. Taking into account the range of TEAC values and the precision of the TEAC method, which is found to be 143 mM (pooled SD), these results can be considered satisfactory. The PLS and UVE-PLS models are shown in Fig. 8a and b, respectively. Calibration samples are indicated as (o) and test samples as (*). It is seen that the PLS model allows a slightly better prediction of calibration samples, but the prediction of test samples is comparable with the UVE-PLS model.

The results of modeling performed on the short chromatograms are shown in Table 1b, the PLS regression coefficients are shown in Fig. 9. Both models, PLS and UVE-PLS, have only three components. The UVE-PLS model, is based on 119 of the 1000 original variables and results in better

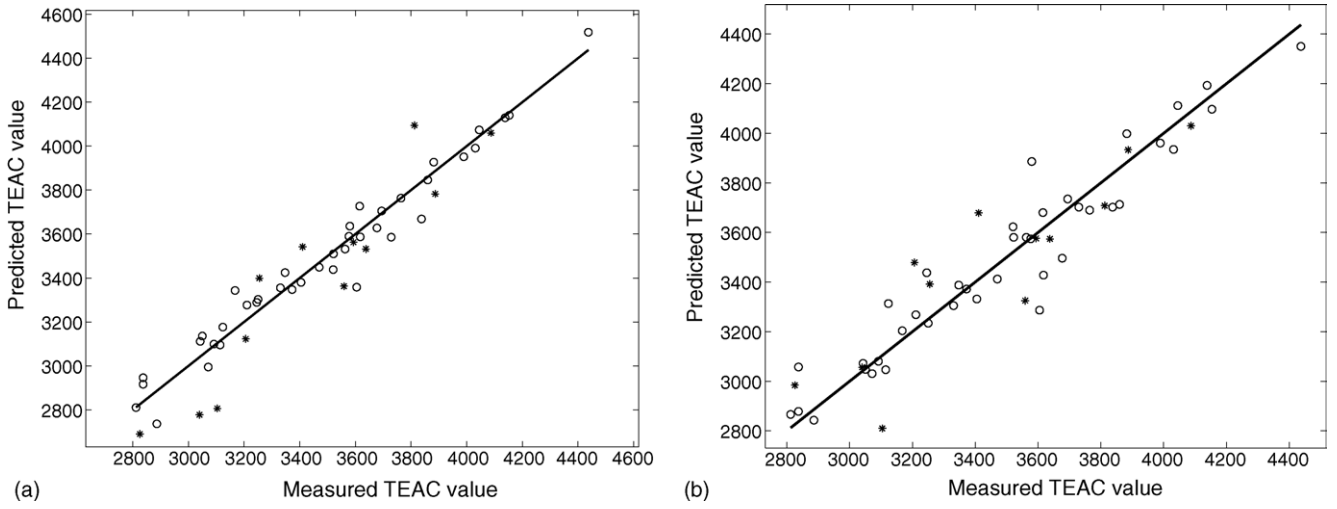

Fig. 8. Results of (a) the PLS model (eight factors) built with the 40 warped long calibration chromatograms, and (b) the UVE-PLS model (seven factors) built with the 40 warped long calibration chromatograms with 142 variables retained (◯: calibration samples, *: test samples).

predictions. The 119 retained variables are selected from the peak apexes at retention time 0.58 and 1.62 min and at retention times 0.95, 1.1, 1.3 and 1.42 min (Fig. 3b). Although the RMSECV and RMSEP of the PLS model are slightly higher, i.e., 206 mM (12%) and 177 mM (11%), it can be concluded that this model results in acceptable and stable predictions of the TEAC values (RMSEP = 11%). Generally, it is seen that with the short chromatograms a less complex model is obtained with a slightly higher cross-validation error than with the long chromatograms. Moreover, it can be concluded that a baseline separation of the green tea extract compounds is not required in order to obtain an acceptable calibration model between the chromatographic and TEAC method. A considerable gain of time can thus be achieved by using the fast chromatographic method on monolithic silica columns and the above calibration models for the prediction of the antioxidant capacity.

4.5. PLS models built with reduced chromatograms

The alignment of the long chromatograms took 13 s each signal, resulting in a total warping time of about 24 min. Considering the HPLC analysis time of one chromatogram (11 min), this computation time is acceptable but there might be cases where COW will require much higher computation times and then it can not be used for on-line applications anymore. Therefore, it was investigated whether a reduction of the chromatogram length, by averaging successive sampling points, results in a decreased warping time and in models with still acceptable prediction errors. The length of the long chromatograms was reduced six times (to 516 sampling points) and the chromatograms were warped using the earlier defined input parameters ($N = 60$, $t = 3$). The alignment took 19 min, a reduction of only 5 min. However, by reducing the signal length, smaller N values could be used. With $N = 30$ and $t = 3$, the computation time could be reduced to only 6 min. Then, a PLS and UVE-PLS model is built using the same calibration and test sets as above. The models contain seven and six factors, respectively, and in the UVE-PLS model, only 42 variables are retained from the 516. The plot with retained chromatographic variables is very similar to Fig. 7 and therefore not shown. The same peak parts as for the unreduced signals are found to be important for modeling. The RMS, RMSECV and RMSEP are shown in Table 1c. The models for the reduced chromatograms are somewhat less complex than for the original ones (Table 1a and c). Moreover, the prediction error of the PLS and UVE-PLS models is at least equally good than for the original chromatograms. It can thus be concluded that models built with reduced chromatograms can be used for the TEAC prediction as well.

Nevertheless, the computation time with COW remains rather long, even after data point reduction. Therefore it was investigated whether faster aligning methods, as for instance parametric time warping [18,19], could be used. With PTW, the chromatograms could be aligned within only 18 s. However, small differences in the warping quality were seen. A thorough comparison of the performance of these methods is being performed at the moment and will be reported later.

4.6. TEAC prediction of new tea samples

In a production application one is interested in a quick prediction of the antioxidant capacity of new tea samples. New samples will be treated as follows. After removal of the first 200 un-informative sampling points (dead time), the chromatograms of new samples will be aligned using the same reference signal and COW input parameters as for the alignment of the initial data set. As long as peak shifts in new chromatograms do not increase compared to chromatogram 17, one can conclude that column ageing is minimal and thus most likely the same N and t value can be used for the alignment of new samples. If not, one needs to re-optimize these input parameters for the new samples.

Then, the new chromatograms will be screened for leverage objects by robust PCA after adding them to the matrix of the 110 initial chromatograms. The objects falling in quadrants two and three are retained, while objects in quadrants one and four need further inspection before removing them. If their orthogonal distance is significantly higher than for the original objects, they must be removed and cannot be predicted with the above models. However, when among the new samples, there are many samples with high orthogonal distance, as seen for chromatograms 61 and 62 of the initial data (Figs. 1 and 4), one might consider to build a PLS model including these samples in the calibration and test sets to allow a precise TEAC prediction of these samples as well. In that case, such samples are not considered atypical anymore.

5. Conclusions

In this paper a stable and reliable model is built, able to predict the antioxidant capacity of green tea extracts (expressed as the TEAC value) from fast chromatograms with analysis times of 11 and 2 min obtained on monolithic silica columns. The chromatograms were successfully aligned with correlation optimized warping and used for multivariate calibration as if they were spectra. The models built with PLS and UVE-PLS resulted in acceptable predictions of the antioxidant capacity. However, with UVE-PLS, many uninformative chromatographic variables were eliminated and less complex models were obtained. It was found that the antioxidant capacity can also be predicted from fast, non-completely resolved chromatograms, or chromatograms with highly reduced sampling points, resulting in much shorter analysis times.

Acknowledgement

A.M van Nederkassel is grateful to the Fund for Scientific Research (FWO)-Flanders for financial support.

References

[1] H. Mukhtar, N. Ahmad, Am. J. Clin. Nutr. 71 (2000) 1698S.
[2] C. Rice-Evans, Exp. Biol. Med. 220 (1999) 262.
[3] R. Re, N. Pellegrini, A. Proteggente, A. Pannala, M. Yang, C. Rice-Evans, Free Radic. Biol. Med. 26 (1999) 1231.
[4] F. Gong, Y. Liang, P. Xie, F. Chau, J. Chromatogr. A 1002 (2003) 25.
[5] WHO, Guidelines for the Assessment of Herbal Medicine, Munich, World Health Organization, 1991.
[6] I.I. Koleva, H.A.G. Niederländer, T.A. van Beek, Anal. Chem. 73 (2001) 3373.
[7] A.M. van Nederkassel, A. Aerts, A. Dierick, D.L. Massart, Y. Vander Heyden, J. Pharm. Biomed. Anal. 32 (2003) 233.
[8] J. Luypaert, M.H. Zhang, D.L. Massart, Anal. Chim. Acta 478 (2003) 303.
[9] V. Centner, D.L. Massart, Anal. Chem. 68 (1996) 3851.
[10] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Data Handling in Science and Technology 20 A, Handbook of Chemometrics and Qualimetrics: Part A, Elsevier, Amsterdam, The Netherlands, 1997, pp. 28, 519.
[11] N.P. Vest Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17.
[12] V. Pravdova, B. Walczak, D.L. Massart, Anal. Chim. Acta 456 (2002) 77.
[13] G. Tomasi, F. van den Berg, C. Andersson, J. Chemom. 18 (2004) 231.
[14] M. Hubert, P.J. Rousseeuw, S. Verboven, Chemom. Intell. Lab. Syst. 60 (2002) 101.
[15] M. Hubert, S. Engelen, Bioinformatics 20 (2004) 1728.
[16] I. Stanimirova, B. Walczak, D.L. Massart, V. Simeonov, Chemom. Intell. Lab. Syst. 71 (2004) 83.
[17] A. Kassidas, J.F. MacGregor, P.A. Taylor, Am. Inst. Chem. Eng. 44 (1998) 864.
[18] P. Eilers, Anal. Chem. 76 (2004) 404.
[19] The Matlab Algorithms for Dynamic and Paramatric Time Warping, http://www.tipb.nl.
[20] J. Forshed, I. Schuppe-Koistinen, S.P. Jacobsson, Anal. Chim. Acta 487 (2003) 189.
[21] B. Walczak, W. Wu, Fuzzy warping of chromatograms, Chemom. Intell. Lab. Syst. (2005), in press.
[22] T. Næs, T. Isaksson, T. Fearn, T. Davies, A User-Friendly Guide to Multivariate Calibration and Classification, NIR Publications, Chichester, UK, 2002, pp. 27, 157.
[23] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Data Handling in Science and Technology 20 B, Handbook of Chemometrics and Qualimetrics: Part B, Elsevier, Amsterdam, The Netherlands, 1998, p. 331.