Talanta 121 (2014) 105–112

Contents lists available at ScienceDirect

Talanta

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

Performance evaluation of preprocessing techniques utilizing expert information in multivariate calibration

Sandeep Sharma, Mohammad Goodarzi, Herman Ramon, Wouter Saeys *

BIOSYST-MeBioS, KU Leuven, Kasteelpark Arenberg 30, 3001 Leuven, Belgium

ARTICLE INFO

Article history:
Received 7 October 2013
Received in revised form
18 December 2013
Accepted 24 December 2013
Available online 2 January 2014

Keywords:
Pure component spectrum
Glucose
Extended Multiplicative Signal Correction
Spectral Interference Subtraction
External Parameter Orthogonalization
Generalized Least Squares Weighting

ABSTRACT

Partial Least Squares (PLS) regression is one of the most used methods for extracting chemical information from Near Infrared (NIR) spectroscopic measurements. The success of a PLS calibration relies largely on the representativeness of the calibration data set. This is not trivial, because not only the expected variation in the analyte of interest, but also the variation of other contributing factors (interferents) should be included in the calibration data. This also implies that changes in interferent concentrations not covered in the calibration step can deteriorate the prediction ability of the calibration model. Several researchers have suggested that PLS models can be robustified against changes in the interferent structure by incorporating expert knowledge in the preprocessing step with the aim to efficiently filter out the spectral influence of the spectral interferents. However, these methods have not yet been compared against each other. Therefore, in the present study, various preprocessing techniques exploiting expert knowledge were compared on two experimental data sets. In both data sets, the calibration and test set were designed to have a different interferent concentration range. The performance of these techniques was compared to that of preprocessing techniques which do not use any expert knowledge. Using expert knowledge was found to improve the prediction performance for both data sets. For data set-1, the prediction error improved nearly 32% when pure component spectra of the analyte and the interferents were used in the Extended Multiplicative Signal Correction framework. Similarly, for data set-2, nearly 63% improvement in the prediction error was observed when the interferent information was utilized in Spectral Interferent Subtraction preprocessing.

1. Introduction

NIR spectroscopy is routinely used for lab analysis in chemical, pharmaceutical, petrochemical and food processing industries [1–3]. It offers many advantages over conventional analytical methods such as no need for sample preparation, simultaneous measurement of multiple components and its non-destructive nature [4]. However, the interpretation of measured signals using NIR spectroscopy is not straightforward since the NIR absorption bands are *n*th order harmonics and combinations of the fundamental absorption bands of C–H, N–H and O–H bonds which are broad, relatively weak and overlapping. NIR spectroscopy is a secondary reference method where the measured NIR spectra are related to the reference values obtained using a primary reference method through multivariate calibration approaches such as Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) [5]. In addition, NIR data might contain systematic variations which are coincidentally correlated to the variation in the concentration of the analyte of interest [6–8]. There can also be variation in chemical interferents which could deteriorate the prediction performance of linear regression models if not properly accounted for. The most obvious way to account for all possible variation and to obtain good estimates for the regression coefficients is to calibrate the linear regression model on a 'representative' calibration set. This is not trivial because all the expected variation in the component of interest and the variation of other contributing factors (interferents) should be included in the calibration step. In the past, many studies have reported that the prediction ability of an inverse model deteriorates with changes in the interferent structure caused by temperature effects, season to season variation, cultivar effects, different tablets for the same active component, batch effects in industrial production processes, etc. [6,8–10]. This underscores the importance of developing robust regression models with the ability to predict the concentration of the analyte of interest with desired level of accuracy even when the interferent structure changes.

Conventional approaches in such situations include augmenting the calibration matrix to include variability, correcting non-relevant variation from new spectra, orthogonalization, etc. However, a more explicit separation of the analyte and the interferent signal can be achieved by utilizing expert knowledge in the multivariate

* Corresponding author. Tel.: +32 16 328527; fax: +32 16 328590.
E-mail address: Wouter.Saeys@biw.kuleuven.be (W. Saeys).

0039-9140/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2013.12.053

modeling. In the past, we demonstrated the use of prior information to build robust calibration models by adopting various calibration approaches [11]. In this study, we explore the potential of using prior information in the preprocessing step.

The aim of the preprocessing step is to transform or filter the matrix of measured spectra in such a way that the perturbations caused by the interferents are either removed or substantially reduced. Subsequently, linear regression models can be built on the preprocessed 'linearized' data set. However, using conventional preprocessing techniques might not result in the prediction improvement, especially when the interferent concentrations in the test set vary differently from the ones in the calibration set. In such cases, the preprocessing techniques which use expert knowledge in the preprocessing step may outperform their conventional counterparts. The expert knowledge may include the pure component spectrum of the analyte of interest and/or the known interferents, which are either available or can be acquired.

The most popular preprocessing techniques which allow utilizing expert knowledge are External Parameter Orthogonalization (EPO), Generalized Least Squares Weighting (GLSW), Extended Multiplicative Signal Correction (EMSC) and Spectral Interference Subtraction (SIS). These techniques offer the flexibility in choosing the amount of pure component information to be supplied. This is the most important element in fine-tuning the preprocessing step as providing the accurate information of interferent(s) helps to efficiently filter their spectral influence. However, providing inaccurate or excess interferent information might lead to removal of part of the analyte specific information. This is also interesting from chemometrics point of view as the fine-tuning step exploits the expert knowledge available to the analyst which in conventional PLS calibration is seldom utilized.

Including expert knowledge to filter the measured spectra has potential to build robust PLSR models for analytical systems where the concentration of known interferents can vary in an unknown fashion. One such example can be blood serum solution where the expert knowledge regarding potential interferents is available but their concentration ranges can vary depending on the physical state of individuals. In such situations, the expert knowledge can be utilized to build PLSR models which are insensitive to concentration fluctuations of known interferents.

In this study, the performance of preprocessing techniques which make use of expert knowledge is benchmarked. To avoid drawing conclusions based on a single data set, the study is performed on two data sets where the data was deliberately split in a calibration and test set with a different concentration range of interferents to test the robustness of calibration models against such changes. The effect of choosing interferent information on the performance of preprocessing techniques is also investigated.

2. Preprocessing techniques exploiting expert knowledge

2.1. Extended Multiplicative Scatter Correction

In analyzing complex mixtures, uncontrolled variations such as light scattering might dramatically reduce the predictive ability of the regression models. As the scattering may vary depending on particle size and shape, sample packing and sample surface [12], the extent of scattering is hardly controllable in practice. Multiplicative Scatter Correction (MSC) is widely used in such cases to compensate for the nonspecific additive and multiplicative effects introduced by uncontrolled light variations. However, in spectral regions where the analyte of interest or interferents absorb strongly, MSC may confuse the chemical absorption with physical light scattering effects resulting into removal of analyte information. To avoid losing useful information from the measured spectra, the EMSC technique was proposed by Martens et al. [12]. The EMSC model was further extended to include prior knowledge about the chemical constituents, while estimating the parameters for the EMSC scatter correction [13]. When the pure component spectra are available, EMSC uses the spectral information of chemical constituents to downweigh those wavelength regions where chemical species exhibit strong chemical absorption.

According to the EMSC model, a measured spectrum $\mathbf{x_i}$ can be decomposed in different contributions:

$$\mathbf{x_i} = a_i + b_i\mathbf{x_{i,chem}} + d_i\boldsymbol{\lambda} + e_i\boldsymbol{\lambda}^2 + \boldsymbol{\varepsilon_i} \quad (1)$$

where a_i and b_i represent the additive (e.g. baseline) and multiplicative effects (e.g. optical path length or light scattering level), respectively; $\mathbf{x_{i,chem}}$ is the contribution of known pure components to the spectrum $\mathbf{x_i}$ of the ith sample; d and e allow for the correction of wavelength dependent spectral variations from sample to sample; $\boldsymbol{\lambda}$ is the vector of wavelengths at which the spectrum has been acquired and $\boldsymbol{\varepsilon}$ is the vector containing residual error.

When expert knowledge about the constituents is available, the parameters in EMSC are estimated based on a modified Beer–Lambert law, which can be expressed as in the following equation:

$$\mathbf{x_{i,chem}} = (c_{i1}\mathbf{k}_1 + c_{i2}\mathbf{k}_2 + \ldots + c_{ij}\mathbf{k_j}) \quad (2)$$

where $c_1, c_2, \ldots, c_j$ and $\mathbf{k_1}, \mathbf{k_2}, \ldots, \mathbf{k_j}$ are respectively the concentrations and the corresponding pure component spectra (expert information) for components 1–j. However, the concentration of every chemical constituent in the mixture is often not known and has to be estimated. As this would involve the estimation of products of unknown coefficients, the $\mathbf{x_{i,chem}}$ part is typically rewritten as a deviation from a reference spectrum $\mathbf{x_{ref}}$ (Eq. (3)).

$$\mathbf{x_{i,chem}} = \mathbf{x_{ref}} + \Delta c_{i1}\mathbf{k}_1 + \Delta c_{i2}\mathbf{k}_2 + \ldots + \Delta c_{ij}\mathbf{k_j} \quad (3)$$

where Δc_{ij} represents the difference in concentration of the jth component between the reference $\mathbf{x_{ref}}$ and ith sample. Typically, the mean spectrum of the calibration set is used as the reference spectrum $\mathbf{x_{ref}}$.

Now Eq. (1) gets a purely additive term $b_i\mathbf{x_{ref}}$.

$$\mathbf{x_i} = a_i\mathbf{1} + b_i\mathbf{x_{ref}} + h_i\mathbf{k_i} + d_i\boldsymbol{\lambda} + e_i\boldsymbol{\lambda}^2 + \boldsymbol{\varepsilon_i} \quad (4)$$

where $h_i = b_i\Delta c_i$, and $\mathbf{k}$ is the chemical variation spectrum estimated as the difference of pure component spectrum of the chemical constituents.

The parameters $\hat{a}_i$, $\hat{b}_i$, $\hat{h}_i$, $\hat{d}_i$ and $\hat{e}_i$ can then be estimated by least squares regression of each input spectrum. The EMSC corrected spectrum for each input spectrum is obtained as

$$\mathbf{x_i^*} = (\mathbf{x_i} - \hat{a}_l - \hat{h}_l\mathbf{k_i} - \hat{d}_l - \hat{e}_l^2)/\hat{b}_l \quad (5)$$

The matrix containing EMSC corrected spectra $\mathbf{X}^*$ can then be utilized for regression modeling using PCR or PLS.

2.2. Spectral Interference Subtraction

Spectral Interference Subtraction was proposed by Martens et al. [12] in conjunction with EMSC for the cases where light scattering effects are present in the data set. In such cases, EMSC is first applied to correct for the light scattering effects. Subsequently, SIS is applied to the EMSC corrected data [12]. However, SIS can also be applied as an independent preprocessing technique when light scattering is not a serious issue in the measured spectra. When the pure component spectra of interferents are available, SIS proposes to subtract their spectral contribution from the measured spectra. This allows filtering of the interferent contributions without removing the information about the analyte of interest.

If $\mathbf{k_P}$ and $\mathbf{K_I}$ represent the pure component contribution of the analyte of interest and the known interferents, respectively, then the measured absorbance $\mathbf{x_i}$ for a given sample i can be expressed

as follows:

$$\mathbf{x_i} = \mathbf{k_P} c_{P,i} + \mathbf{K_I c_{I,i}} + \mathbf{e_i} \quad (6)$$

where $\mathbf{c_P}$ is the concentration vector for the analyte and $\mathbf{C_I}$ is the concentration matrix of known interferents. The term $\mathbf{e_i}$ represents the residual error. In most of the cases, the interferent concentrations $\mathbf{C_I}$ are unknown and have to be estimated from the pure interferent spectra $\mathbf{K_I}$ using least squares estimation:

$$\hat{\mathbf{C}}_\mathbf{I} = [\mathbf{K}_\mathbf{I}'\mathbf{K_I}]^{-1}\mathbf{K}_\mathbf{I}'\mathbf{X} \quad (7)$$

In the next step, the spectral contribution of the interferents is subtracted from the measured spectra **X**.

$$\mathbf{X}^* = \mathbf{X} - \mathbf{K_I}\hat{\mathbf{C}}_\mathbf{I} \quad (8)$$

The SIS corrected spectra $\mathbf{X}^*$ can then be used to build a calibration model using PCR or PLS.

2.3. External Parameter Orthogonalization

The prediction accuracy of linear regression models can reduce dramatically due to variation of an external parameter such as sample temperature or other spectrally active components (interferents), which have an impact on the measured signals [8–10]. In principle, it would be possible to measure the external parameter and use this information to correct the sample spectra, but in practice, this is often not feasible. For such cases, Roger et al. [6] introduced the External Parameter Orthogonalization preprocessing technique with an aim to remove the part of measured spectra **X** which is mostly influenced by an external parameter. EPO models the measured spectra **X** as comprised of two orthogonal subspaces: one containing the useful information, and the other containing the parasitic variation which is largely unrelated to the variation in the concentration of the analyte of interest. Mathematically, the underlying model can be expressed as represented in the following equation:

$$\mathbf{X} = \mathbf{XP} + \mathbf{XQ} + \mathbf{E} \quad (9)$$

where **P** is the projection operator of the component of interest, **Q** is the projection operator for the external parameter, and **E** is the residual matrix. EPO proposes to remove a part of the perturbations caused by **Q** by projecting the spectra onto the subspace orthogonal to **Q**. However, the influence of **Q** on the spectra is not known in advance and hence needs to be estimated by decomposing the measured spectra as presented in Eq. (9). In practice, the parasitic subspace **Q** is estimated by performing Principal Component Analysis (PCA) on a small set of spectra measured on the same object when the external parameter is varying. When the pure component spectra of the interferents are available, they can be used to estimate the parasitic subspace $\hat{\mathbf{Q}}$.

The original spectra are then corrected by orthogonal projection onto the estimated $\hat{\mathbf{Q}}$ as presented in the following equation:

$$\mathbf{X}^* = \mathbf{X}(\mathbf{I} - \hat{\mathbf{Q}}) \quad (10)$$

This corrected matrix X* can then be used to build a calibration model using PCR or PLS.

2.4. Generalized Least Squares Weighting

Generalized Least Squares Weighting is a covariance weighted preprocessing technique [14], which performs a 'pre-whitening' of the spectra by deflating the variables that are unrelated to the variation in the concentration of the analyte of interest. In this way, GLSW aims at finding a balance between the relevant part and the noise component of each variable. In practice, the weighting can be expressed as follows:

$$\mathbf{X}^* = \mathbf{XG} \quad (11)$$

where **X** is the matrix of input spectra, and **G** is a scaling matrix which can be defined as the inverse square root of the uncertainty or noise covariance matrix $\mathbf{\Sigma}$:

$$\mathbf{G} = \mathbf{\Sigma}^{-1/2} \quad (12)$$

The uncertainty or noise covariance matrix can be defined either as the total standard deviation in the available variables, or the expected standard deviations in their errors.

When prior information such as pure component spectra of known interferents is available, $\mathbf{\Sigma}$ can be defined as follows:

$$\mathbf{\Sigma} = \mathbf{K_I}\,\mathbf{cov(D)}\mathbf{K}_\mathbf{I}' + \mathbf{cov(E)} \quad (13)$$

where $\mathbf{K_I}$ is the matrix of interferent spectra and '**cov(D)**' is an estimate for the variance–covariance matrix of the interferent concentrations, which is typically unknown to the analyst. In practice, this can be estimated as $d^2\mathbf{I}$, with d^2 being the expected average variance of the interferent concentrations. Similarly, '**cov(E)**' is the covariance matrix of the other unidentified error patterns and noise, which is assumed to be uncorrelated. This can be estimated as $s^2\mathbf{I}$, where s^2 is the average uncertainty variance of all **X** variables [14].

This leads to a simplified form of Eq. (13), which can be written as

$$\mathbf{\Sigma} = d^2\mathbf{K_I}\mathbf{K}_\mathbf{I}' + s^2\mathbf{I} \quad (14)$$

As d^2 is the general scaling factor, which balances the contribution of interferent spectra to the contribution of the uncertainty variance of the **X** variables, Eq. (14) can further be simplified as

$$\mathbf{\Sigma} = d^{*2}\mathbf{K_I}\mathbf{K}_\mathbf{I}' + \mathbf{I} \quad (15)$$

Thus, by defining the scaling factor d^{*2} (the relative variance in the interferent concentrations) sufficiently large, GLSW preprocessing can make the least squares modeling of **X** completely insensitive to the concentration variations of the unknown interferents. This should lead to a reliable estimation of the concentration of the analyte of interest [14].

3. Experimental

In this section, the experimental details of the data sets including the spectral acquisition and the estimation of pure component spectra of analyte and known interferents are presented.

Data set-1 consists of NIR spectra of three parallel sets of aqueous solutions containing similar concentrations for glucose (1, 3, 7, 12, 15, 22 and 30 mM), urea (5 and 6 mM) and sodium (Na) D-lactate (1 and 5 mM). A full factorial design of these concentrations was prepared resulting in 28 ($=7\times2\times2$) samples for every set of aqueous solutions. In total, 84 samples were produced for three sets. NIR spectra in the range 800–2500 nm were acquired for these samples with a Bruker MPA FT-NIR spectrometer (Bruker, Ettlingen, Germany) with a 1 mm transflectance probe. All the measurements were carried out in a temperature controlled facility at 37 °C ± 1.0 °C. The spectra for each sample were recorded in triplicate resulting in 252 spectra for 84 aqueous solution samples.

To estimate the pure component spectra, the method described by Amerov et al. [7] was followed. As the molar absorptivity of glucose, urea and lactate in the NIR region is very low (~the order of 10^{-4}–10^{-5} mM^{-1} mm^{-1}), relatively high concentration solutions of glucose (120 mM), urea (100 mM) and lactate (100 mM) were prepared to obtain good quality estimates of the pure component spectra [7]. The measured spectra were corrected for reflective losses, the dispersion effect and the water displacement effect by meticulously following the procedure described by Amerov et al. [7].

The wavelength-dependent values of the refractive index for the 'Quartz SUPRASIL cuvette' used in the spectral measurement were estimated using the Sellmeier equation [15]. The change in refractive index with glucose concentration was calculated as

$$\eta = 1.325 + 2.73 \times 10^{-5}[c_g] \quad (16)$$

where η is the refractive index of the solution and c_g is the concentration of glucose in milli-moles.

Data set-2 consists of NIR spectra for a triangular mixture design of glucose, casein and lactate powders analyzed by Naes et al. [16] and Saeys et al. [17]. The spectra were measured in a closed cup on a monochromator instrument (Technicon InfraAlyzer 500) in the wavelength range from 1100 to 2500 nm. The powder mixtures were prepared by mixing the appropriate amounts of casein, glucose and lactate powders. Since these powders contain some amount of moisture and ash, the weight percentage of casein, glucose and lactate was measured and the true content in the triangular design was calculated. The samples at the extremes in the triangular design represent the pure component spectra of three chemical constituents. It should be noted that the impurities, i.e., moisture and ash, also will have some contribution in the measured spectra and the pure component spectra used here are the NIR spectra of the powders containing some moisture, ash and the traces of impurities. In addition, the measured 'pure component spectra' still include the matrix effect, i.e., the variations arising from varying particle size, particle shape and sample packing.

4. Data analysis

For both data sets, the effect of using expert knowledge in preprocessing on the prediction ability of a PLS regression model was evaluated and compared against the performance of PLS regression combined with preprocessing techniques which do not use expert knowledge. Among the most popular preprocessing techniques, Detrend [18], EPO [6], GLSW [14,19], MSC [20,21], Standard Normal Variate (SNV) [18], Normalization using Euclidean Norm [22], Orthogonal Signal Correction (OSC) [23–25] and EMSC [12,13,26,27] were used to filter the spectra without providing any expert knowledge. The prior information was utilized in the preprocessing step using EMSC [13,28], EPO [6], GLSW [14,19] and SIS [12] framework. In Fig. 1, the mathematical basis of the preprocessing techniques used in this study is presented.

Using above preprocessing techniques, a total of seventeen calibration models were built for both data sets. The optimal number of Latent Variables (LVs) to be used in the calibration was chosen based on the reduction in root mean square error of cross validation (RMSECV). The optimal number was defined as the minimum of LVs in the calibration after which further addition of LVs yielded no significant improvement in the RMSECV. In most cases, this led to the number of LVs corresponding to (nearly) minimum RMSECV [29]. Finally, the performance of different preprocessing techniques was evaluated based on the improvement in root mean square error of prediction (RMSEP) in comparison to the performance of a PLS calibration built on the mean-centered NIR spectra.

4.1. Selection of calibration data

For **data set-1**, the calibration was performed for glucose concentration using the spectral region from 1525–1825 nm, known as the first overtone band of glucose absorption in the NIR region [7]. The data set was split based on the lactate concentration. The calibration set contained the samples with 1 mM lactate concentration whereas the samples with 5 mM lactate concentration were kept in the test set. The pure component spectra of glucose (the analyte signal), urea and Na-lactate (the interferent signals) were used as the expert knowledge. Cross-Validation was performed using Contiguous Blocks with 14 splits to ensure that the three spectral replicates belonging to the same physical sample were always kept together either in the calibration or the test set. The effect of prior information was demonstrated by building PLS models with varying amounts of prior information.

For **data set-2**, the measured intensities were converted to absorbance scale using the log(1/R) transform. The PLS calibration was performed for glucose weight percent. The experimental design and the validation scheme for data set-2 are illustrated in Fig. 2. The data set consists of 231 samples of which the three samples at the extremes of the triangular design correspond to the pure component spectra of glucose, casein and lactate. Since these spectra are used as expert knowledge, they were not included in the calibration or test set. The remaining 228 powder mixture

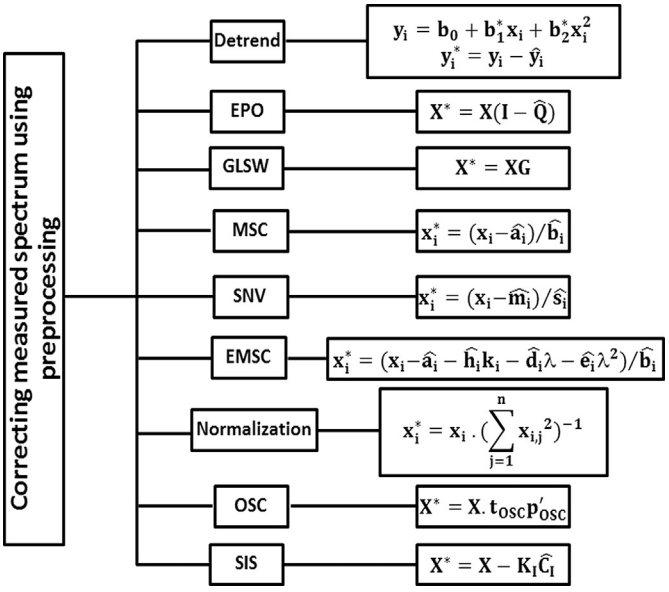

Fig. 1. Schematic presentation of mathematical basis of the preprocessing techniques used for correcting the measured spectrum.

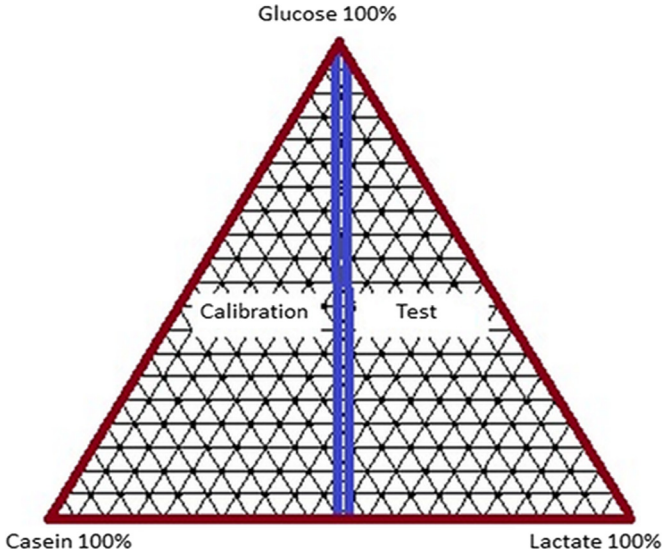

Fig. 2. Illustration of the splitting scheme for data set-2 (one sample for each intersection of lines) with marking of the calibration and test sets.

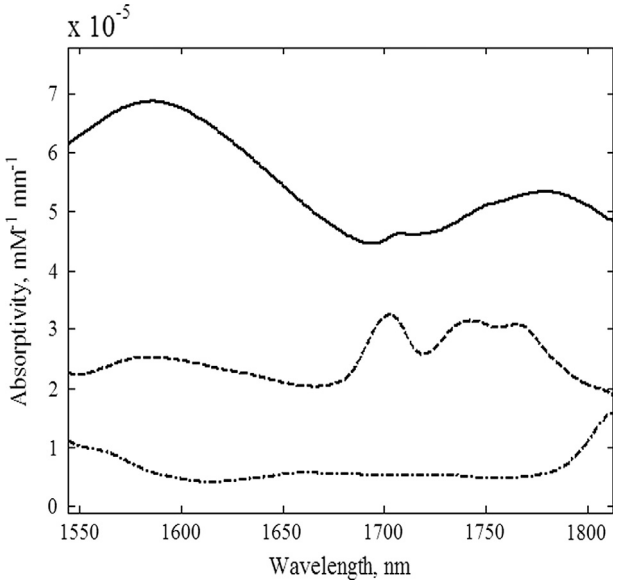

Fig. 3. Pure component spectra of glucose (solid), lactate (dash) and urea (dash-dot) in the first overtone band of glucose absorption in the NIR region.

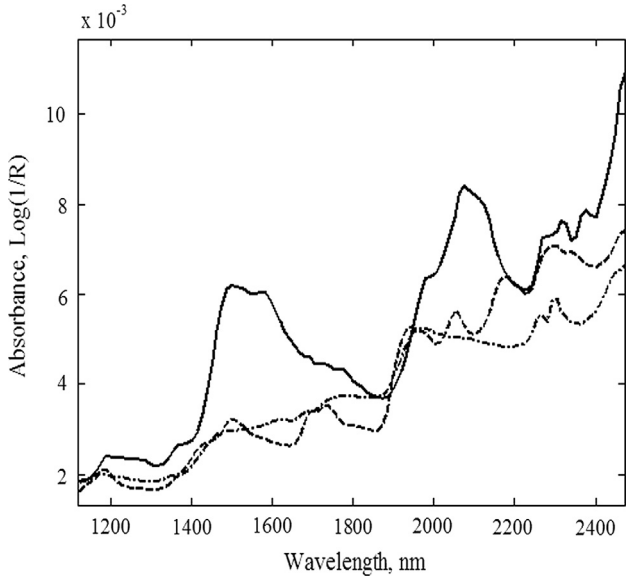

Fig. 4. Pure component spectra of glucose (solid), casein (dash) and lactate (dash-dot) in the NIR region.

spectra were split into calibration and test set in such a way that the glucose concentration covers a similar range in the calibration and test set (0–86.9%), while the ranges of the interferents, i.e., lactate and casein are different. The Cross-Validation strategy used for selecting the optimal number of latent variables was Random Splits with 10 splits.

4.2. Inclusion of expert knowledge

The effect of using selective interferent information was investigated for both data sets. For data set-1, it can be observed from Fig. 3 that urea has no absorption in the wavelength range used in this study, whereas the absorption bands of lactate are very similar to glucose absorption. In addition, the calibration and test sets are designed to contain different lactate concentrations. Thus, lactate is treated as the key interferent for data set-1. For data set-2, all three chemical constituents, glucose, casein and lactate, are spectrally active (Fig. 4). However, lactate is the key interferent as the test set contains higher concentration range of lactate than the calibration set.

Based on varying amounts of expert knowledge, two PLS models were built for EPO, GLSW and SIS preprocessing (Fig. 5). In the first case, the pure component information of key interferent (lactate for both datasets) was used as the interferent signal. In the second case, the pure component spectra of both interferents (lactate and urea for data set-1; lactate and casein for data set-2) were utilized in preprocessing. For EMSC, a total of four PLS calibrations were built (Fig. 5). In the first calibration, only the pure spectrum of the analyte (glucose for both data sets) was used in the EMSC framework. In the second calibration, the pure component spectrum of the key interferent alone was utilized. In the third calibration, the analyte and the key interferent spectra were supplied in the EMSC step. In the fourth calibration, the analyte spectra of glucose and both the interferents were utilized in the preprocessing step.

4.3. Performance comparison

A two-way ANOVA test was performed on the absolute prediction errors to detect whether the observed difference in prediction performance was statistically significant. The preprocessing technique was taken as the first ANOVA factor whereas the sample number

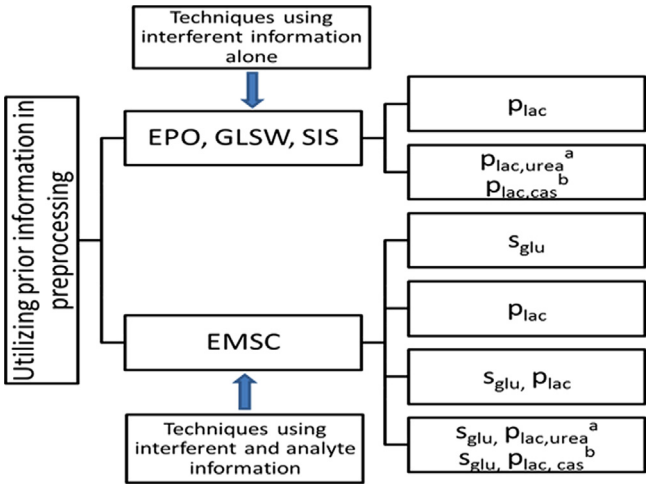

Fig. 5. Schematic presentation of using pure component spectra of the analyte (s) and the known interferents (p) in EPO, GLSW, SIS and EMSC preprocessing. The abbreviations lac, urea, cas, and glu refer to lactate, urea, casein and glucose, respectively. Superscripts 'a' and 'b' indicate data set-1 and 2, respectively.

was added as the second ANOVA factor to make the test paired [30]. The preprocessing technique was treated as the 'fixed factor' whereas the sample number was treated as a 'random factor'. The different preprocessing techniques were compared using Tukey Honestly Significant Difference (HSD) multiple comparison to ascertain whether a given preprocessing had a significant influence ($\alpha \leq 0.05$) on the prediction performance of the PLS regression model. It was observed that the inclusion of all preprocessing techniques in the Tukey HSD test resulted into a large number of sub-groups with very little analytical significance. Hence, the ANOVA and Tukey HSD tests were limited to the preprocessing techniques which resulted in RMSEP values less than or equal to the RMSEP obtained when the PLS calibration was performed without using any preprocessing except mean-centering.

All calibrations were performed in MATLAB®, 7.10.0 (R2010a) (The Mathworks, Natick, MA, USA). For EMSC, the EMSC toolbox (Eigenvector Research, Wenatchee, WA, USA) was used whereas other preprocessing techniques and the PLS calibration were

calculated using the PLS toolbox (Eigenvector Research, Wenatchee, WA, USA).

5. Results and discussions

5.1. Data set-1

For data set-1, the performance statistics for the different PLS calibrations are summarized in Table 1, where the preprocessing techniques are grouped together based on whether or not they exploit the expert knowledge. For the PLS model built without using any preprocessing, the RMSECV and RMSEP values were 1.53 mM and 2.72 mM, respectively for 7 LVs in the calibration. This result was used for benchmarking and performance evaluation of the preprocessing techniques used in this study. As it can be observed from Table 1, the preprocessing techniques which do not use expert information were not able to improve the prediction ability of the PLS calibration. EPO, GLSW and OSC were found to have no effect on the PLS model performance and their RMSECV and RMSEP values were similar. On the other hand, Detrend, MSC, SNV, Normalization using Euclidean Norm and EMSC (no prior information) were detrimental as they resulted in higher RMSECV and RMSEP values. This can be explained partly by the fact that glucose (analyte) and lactate (key interferent) have a very similar spectral signature, which could have resulted into partial removal of analyte specific information during preprocessing.

In the next step, the performance of various preprocessing techniques using expert information was evaluated. When the pure component spectrum of lactate alone was utilized in EPO and GLSW to define the noise clutter, it did not result in an improvement of the PLS performance. Similarly, when the spectra of both the interferents, lactate and urea, were supplied to define the clutter signal in GLSW no improvement in prediction ability was obtained. EPO resulted into an RMSEP of 2.94 mM indicating that the use of the urea spectrum to define the clutter signal has an adverse effect on the prediction performance of the PLS calibration model. However, the difference in RMSEP was not statistically different in comparison to when no expert information was used. Similar observations were made when the expert knowledge was used in EMSC framework. Using the glucose spectrum alone in EMSC resulted into an RMSEP value of 3.01 mM. However, when the lactate spectrum alone was utilized in EMSC, significant improvement in the prediction performance was obtained with a reduction of the RMSEP to 1.87 mM. Using the glucose spectrum in addition to lactate further improved the RMSEP to 1.84 mM. However, when the pure component spectrum of urea was used alongwith lactate and glucose, the RMSEP deteriorated dramatically to 4.41 mM. Increase in the RMSEP was also observed when the pure component spectrum of urea was used in SIS preprocessing (RMSEP=4.20 mM). However, when the lactate spectrum alone was utilized in SIS, the RMSEP improved to 2.17 mM.

Deterioration of the prediction performance upon including the urea spectrum as interferent highlights the importance of using accurate pure component information. Urea was not an active interferent as it does not have any absorption peak in the studied wavelength region. Its inclusion in the preprocessing step as an interferent might have resulted in excess signal removal including a part of analyte specific information. As these preprocessing techniques rely heavily on pure component information to define the noise signal, the quality and accuracy of the pure component information has a significant effect on the prediction performance.

The results of 2-way ANOVA and the Tukey Honestly Significant Difference multiple comparison tests are also presented in Table 1. The preprocessing techniques which did not show statistically significant difference in RMSEP values have been grouped together. This resulted into three groups. The RMSEP obtained for EMSC (using lactate spectrum alone, and while using glucose and lactate spectra together) and SIS (using lactate spectrum alone) preprocessing was found to be significantly different from that obtained with a PLS model built without using any expert knowledge. In both cases the reduction in the prediction error was only significant when only the lactate spectrum was used to define the clutter signal. Inclusion of the glucose spectrum in EMSC in addition to the lactate spectrum did not result in a significantly different RMSEP.

Table 1
Overview of the prediction ability of PLS models and preprocessing techniques using expert information for the prediction of glucose concentration in aqueous glucose solutions (data set-1).

Preprocessing[1]	LVs	R^2 (%)	RMSEC[2]	RMSECV[2]	RMSEP[2]
No expert information					
No Preprocessing[a]	7	96.8	0.88	1.53	2.72[b,c]
Detrend	5	96.2	1.11	1.80	3.12
EPO[a]	6	96.8	0.88	1.55	2.72[b,c]
GLSW[a]	7	96.8	0.88	1.53	2.72[b,c]
MSC	4	96.0	1.52	1.85	2.83
SNV	4	96.0	1.52	1.85	2.83
Normalization (Euclidean Norm)	5	96.0	1.44	1.74	2.84
OSC[a]	6	96.8	0.88	1.53	2.72[b,c]
EMSC	3	95.6	1.62	1.84	2.98
Exploiting expert information					
EPO[3,a]	7	96.8	0.88	1.53	2.72[b,c]
EPO[4]	6	96.5	1.03	1.57	2.94
GLSW[3,a]	7	96.8	0.88	1.53	2.72[b,c]
GLSW[4,a]	7	96.8	0.88	1.53	2.72[b,c]
EMSC[5]	3	95.6	1.61	1.79	3.01
EMSC[4,5]	5	85.0	2.92	4.24	4.41
EMSC[3,b]	4	96.3	1.89	2.31	1.87[a]
EMSC[3,5,b]	4	96.3	1.90	2.28	1.84[a]
SIS[3,c]	7	97.3	0.99	1.57	2.17[a]
SIS[4]	7	92.1	1.52	2.81	4.20

[a–c]Superscript letters present the results of Tukey HSD multiple comparison test; in the first column of the table, superscript letters indicate the group in which the preprocessing techniques belong while in the RMSEP column, different superscript letters represent significantly different groups(a£0.05).

[1] Data were mean centered prior to PLS modeling in all the cases.
[2] In mM units.
[3] Using lactate spectrum as interferent signal.
[4] Using lactate and urea spectra together as interferent signal.
[5] Using glucose spectrum as analyte signal.

5.2. Data set-2

The performance statistics of PLS calibration models trained on the original and preprocessed spectra are summarized in Table 2. Using different concentration ranges of lactate in calibration and test sets had a dramatic impact on the prediction performance of the PLS model resulting in RMSECV and RMSEP values of 0.70% and 2.77%, respectively for 10 LVs. Among the preprocessing techniques which do not use prior information, Detrend and EPO were found to improve the prediction performance of the PLS model. Building the PLS model using Detrend preprocessed data resulted in a reduction in RMSEP to 2.01% for 9 LVs used in the calibration. When EPO (without supplying prior information) was used to preprocess the data, the RMSEP improved to 2.62% for 9 LVs used in the PLS model. However, the improvement was not statistically significant compared to when no prior information was utilized. OSC did not have any effect on the prediction performance and the RMSEP was 2.78% for 9 LVs in the calibration. Using GLSW, MSC, SNV, Normalization (Euclidean Norm) and EMSC preprocessing for filtering the data apparently deteriorated the prediction performance of the corresponding calibration models; the RMSEP for these techniques was 3.03%, 4.25%, 2.85%,

Table 2
Overview of the prediction ability of PLS models and preprocessing techniques using expert information for the prediction of glucose concentration in powder mixture data set (data set-2).

Preprocessing[1]	LVs	R^2 (%)	RMSEC[2]	RMSECV[2]	RMSEP[2]
No expert information					
No Preprocessing[a]	10	99.5	0.62	0.70	2.77[b,d,f]
Detrend[b]	10	99.7	0.56	0.69	2.01[a,f]
EPO[c]	9	99.5	0.63	0.72	2.62[d,f]
GLSW	10	99.4	0.60	0.70	3.03
MSC	8	98.8	0.62	0.73	4.25
SNV	8	99.4	0.68	0.76	2.85
Normalization (Euclidean Norm)	9	99.4	0.56	0.66	3.08
OSC	9	99.5	0.62	0.71	2.78
EMSC	8	99.0	0.74	0.89	3.78
Exploiting expert information					
EPO[3,d]	9	99.8	0.80	0.90	1.63[a,c]
EPO[4,b]	8	99.8	0.83	0.92	1.71[a,f]
GLSW[3,d]	9	99.8	0.80	0.90	1.63[a,c]
GLSW[4,d]	9	99.8	0.80	0.90	1.63[a,c]
EMSC[5]	10	98.8	0.84	1.10	3.97
EMSC[4,5]	9	–	3.09	5.76	–
EMSC[3]	8	97.4	0.85	1.02	5.83
EMSC[3,5]	9	66.5	1.00	1.40	24.16
SIS[3,e]	8	99.7	0.82	0.92	1.82[f]
SIS[4,f]	11	99.9	0.52	0.63	1.02[a,b,c,e]

[a–f]Superscript letters present the results of Tukey HSD multiple comparison test; in the first column of the table, superscript letters indicate the group in which the preprocessing techniques belong while in the RMSEP column, different superscript letters represent significantly different groups(a£0.05).

[1] Data were mean centered prior to PLS modeling in all cases.
[2] In percentage (%) composition.
[3] Using lactate spectrum as interferent signal.
[4] Using lactate and casein spectra together as interferent signal.
[5] Using glucose spectrum as analyte signal.

3.08% and 3.78%, respectively for all calibration models using 8–10 LVs in the calibration.

Next, the expert knowledge was utilized in the preprocessing step. It was observed that all preprocessing techniques using expert knowledge except EMSC improved the prediction performance of PLS calibration models. When EMSC utilizing the analyte spectrum alone was applied, the RMSEP increased to 3.97%. Using the pure component spectrum of the key interferent (i.e. lactate) alone in EMSC deteriorated the RMSEP to 5.83%. The EMSC preprocessing was also performed using the glucose spectrum as the analyte information and (a.) the lactate spectrum alone, and (b.) the lactate and the casein spectra together, to define the interferent information. In the first case, EMSC predictions were unstable with the RMSECV and RMSEP being 1.40% and 24.16%, respectively. When the lactate and casein spectra were together utilized to define the interferent signals, the RMSECV rose to 5.76%, whereas the RMSEP values were strangely ranging around 200%. Further investigation revealed that the samples having casein concentration equal to zero in the test set destabilized the PLS calibration and resulted into abnormally high RMSEP values. Removing these samples from the test set improved the RMSEP values, but still the RMSEP was ranging around 4–5% indicating EMSC was not able to filter the analyte and interferent signal efficiently. The probable reason for this non-performance might be the spectral similarity between glucose, casein and lactate spectra in the investigated wavelength region which could lead to collinearity during the EMSC parameter estimation.

Using the lactate spectrum alone in EPO and GLSW significantly improved the prediction performance. For both preprocessing techniques, the RMSEP was equal to 1.63% for 9 LVs used in the calibration. Using lactate and casein spectra together to define the clutter signal in GLSW did not improve the predictions further and the RMSEP was still 1.63% for 9 LVs. For EPO using lactate and casein spectra together, the RMSEP increased to 1.71% although in this case only 8 LVs were used in the calibration.

Subtracting the spectral contribution of interferents using SIS prior to the PLS calibration improved the prediction performance. Using the lactate spectrum alone in SIS followed by building the PLS calibration resulted into RMSECV and RMSEP of 0.92% and 1.82% for 8 LVs in the calibration. When lactate and casein spectra were together utilized in SIS, the RMSEP further improved to 1.02% for 11 LVs in the calibration.

The results of the 2-way ANOVA and the Tukey Honestly Significant Difference multiple comparison test are also presented in Table 2. Grouping the preprocessing techniques with no statistically significant difference in the RMSEP values resulted into six groups. Among the preprocessing techniques which do not use any prior information, Detrend resulted in a significant improvement in the RMSEP (=2.01%). Among the preprocessing techniques using prior information, EPO, GLSW and SIS resulted in a significant improvement in the RMSEP in comparison to the PLS calibration model built without using any preprocessing apart from mean centering. No statistically significant difference was observed among the RMSEP values obtained with PLS calibration models built with EPO using lactate spectrum, GLSW using lactate spectrum and GLSW using lactate and casein spectra. SIS, which gave the lowest RMSEP value (=1.02%), also showed significant improvement in the RMSEP in comparison to the PLS model built without using any preprocessing and the PLS calibration built using Detrend preprocessing. An interesting comparison with our previous work on using expert information directly into calibration steps [11] revealed that none of the preprocessing techniques reported in this study outperformed the Augmented Classical Least Squares calibration model incorporating expert knowledge although the performance of SIS preprocessing was quite close.

In this data set, the primary reason for poor performance of PLS calibration without prior preprocessing was the difference in concentration ranges of the interferents in calibration and test sets. Successful PLS modeling in such situation needed the effective filtering or down-weighting of those wavelengths where lactate had absorption peaks. EMSC was not at all effective for this data set, which could be attributed to the spectral similarities between glucose and lactate. All other preprocessing techniques were effective in this data set as they were able to either down-weight (EPO, GLSW) or remove the chemical interference (SIS) of casein and lactate.

6. Conclusions

In this study, the performance of preprocessing techniques exploiting expert knowledge was evaluated and compared against the ones which do not utilize expert knowledge. Pure component spectra of the analyte of interest and/or known interferents were used in EMSC, EPO, GLSW and SIS to filter the measured spectra. Filtered spectra were fed to the PLS models with the aim to obtain calibration models which are (more) robust against changes in the interferent concentrations. To avoid drawing inference from a single data set, the study was performed on two data sets. In both cases, the data was split in a way to create calibration and test sets with different concentration range of interferents.

None of the preprocessing techniques was able to improve the PLS predictions unless the expert knowledge about the analyte and/or the know interferents was utilized in the preprocessing. The preprocessing techniques which use expert knowledge clearly outperformed the other methods and resulted in a significant improvement in the RMSEP values. Among the preprocessing techniques using expert information, SIS was found to be an effective technique for both the data sets. Other techniques, i.e.,

EPO, GLSW and EMSC were effective in one of the data sets, but not in the other one. This can be explained based on the fact that the success of these techniques depends on the physical characteristics of samples, measurement conditions and the spectral signature of analyte and interferent species. However, it should be noted that most of these methods are based on a multicomponent Beer′s law and thus assume that the interferents have an additive effect.

Altogether, the present study has demonstrated the potential of preprocessing techniques using expert knowledge to produce PLS calibration models that are (more) robust against changes in the interferent levels which have not been covered in the calibration set.

Acknowledgments

The authors gratefully acknowledge the Institute for the Promotion of Innovation through Science and Technology (IWT-Vlaanderen) for the financial support through the GlucoSens project (SB-090053). The authors also thank Dr. Tormod Naes, Bjorg Narum and Dr. Tomas Isaksson, Nofima, Ås, Norway for providing the powder mixture data set.

References

[1] J.M. Henshaw, L.W. Burgess, K.S. Booksh, B.R. Kowalski, Anal. Chem. 66 (1994) 3328–3336.
[2] M.H. Rhiel, M.B. Cohen, M.A. Arnold, D.W. Murhammer, Biotechnol. Bioeng. 86 (2004) 852–861.
[3] M. Ventura, A. de Jager, H. de Putter, F.P.M.M. Roelofs, Postharvest Biol. Technol. 14 (1998) 21–27.
[4] H.W. Siesler, in: H.W. Siesler, Y. Ozaki, S. Kawata, H.M. Heise (Eds.), Near-Infrared Spectroscopy: Principles, Instruments, Applications, Wiley-VCH Verlag GmbH, Weinheim, Germany, 2007, pp. 1–10.
[5] R.P. Codgill, C.A. Anderson, J. Near Infrared Spectrosc. 13 (2005) 119–131.
[6] J.-M. Roger, F. Chauchard, V. Bellon-Maurel, Chemom. Intell. Lab. Syst. 66 (2003) 191–204.
[7] A.K. Amerov, J. Chen, M.A. Arnold, Appl. Spectrosc. 58 (2004) 1195–1204.
[8] V.H. Segtnan, B.H. Mevik, T. Isaksson, T. Naes, Appl. Spectrosc. 59 (2005) 816–825.
[9] A. Peirs, J. Tirry, B. Verlinden, P. Darius, B.M. Nicolai, Postharvest Biol. Technol. 28 (2003) 269–280.
[10] B.J. Kemps, W. Saeys, K. Mertens, P. Darius, J.G. De Baerdemaeker, B. De Ketelaere, J. Near Infrared Spectrosc. 18 (2010) 231–237.
[11] S. Sharma, M. Goodarzi, L. Wynants, H. Ramon, W. Saeys, Anal. Chim. Acta 778 (2013) 15–23.
[12] H. Martens, E. Stark, J. Pharm. Biomed. Anal. 9 (1991) 625–635.
[13] H. Martens, J.P. Nielsen, S.B. Engelsen, Anal. Chem. 75 (2003) 394–404.
[14] H. Martens, M. Hoy, B.M. Wise, R. Bro, P.B. Brockhoff, J. Chemom. 17 (2003) 153–165.
[15] I.H. Malitson, J. Opt. Soc. Am. 52 (1962) 1377–1379.
[16] T. Naes, T. Isaksson, B. Kowalski, Anal. Chem. 62 (1990) 664–673.
[17] W. Saeys, K. Beullens, J. Lammertyn, H. Ramon, T. Naes, Anal. Chem. 80 (2008) 4951–4959.
[18] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Appl. Spectrosc. 43 (1989) 772–777.
[19] B.M. Zorzetti, J.M. Shaver, J.J. Harynuk, Anal. Chim. Acta 694 (2011) 31–37.
[20] H. Martens, S.A. Jensen, P. Geladi, in: O.H.J. Christie (Ed.), Proceedings of the Nordic Symposium on Applied Statistics. Stavanger, Stokkland Forlag, Norway, 1983, pp. 205–234.
[21] P. Geladi, D. MacDougal, H. Martens, Appl. Spectrosc. 39 (1985) 491–500.
[22] A. Rinnan, F. van den Berg, S.B. Engelsen, TrAC – Trends Anal. Chem. 28 (2009) 1201–1222.
[23] J.A. Westerhuis, S. de Jong, A.K. Smilde, Chemom. Intell. Lab. Syst. 56 (2001) 13–25.
[24] S. Wold, H. Antti, F. Lindgren, J. Ohman, Chemom. Intell. Lab. Syst. 44 (1998) 175–185.
[25] J.-C. Boulet, J.-M. Roger, Chemom. Intell. Lab. Syst. 117 (2012) 61–69.
[26] M. Decker, P.V. Nielsen, H. Martens, Appl. Spectrosc. 59 (2005) 56–68.
[27] N.K. Afsetha, A. Kohler, Chemom. Intell. Lab. Syst. 117 (2012) 92–99.
[28] S. Ottestad, T. Isaksson, W. Saeys, J.P. Wold, Appl. Spectrosc. 64 (2010) 795–804.
[29] M. Goodarzi, S. Funar-Timofei, Y.V. Heyden, TrAC – Trends Anal. Chem. 42 (2013) 49–63.
[30] H.R. Cederkvist, A.H. Aastveit, T. Naes, J. Chemom. 19 (2005) 500–509.