Optimizing Savitzky-Golay parameters for improving spectral resolution and quantification in infrared spectroscopy

Boris Zimmermann[1], Achim Kohler[2,3]

[1]Department of Organic Chemistry and Biochemistry, Ruđer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia
[2]Department of Mathematical Sciences and Technology (IMT), Norwegian University of Life Sciences, 1430 Ås, Norway
[3]Nofima AS, Centre for Biospectroscopy and Data Modelling, Osloveien 1, 1430 Ås, Norway

Corresponding author:
Boris Zimmermann, Department of Organic Chemistry and Biochemistry, Ruđer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia
Fax:+385 1 4680 195; Tel:+385 1 4571 220; E-mail: bzimmer@irb.hr

Abstract

Calculating derivatives of spectral data by the Savitzky-Golay (SG) numerical algorithm is often employed as a preliminary preprocessing step in order to resolve overlapping signals, enhance signal properties, and to suppress unwanted spectral features that arise due to non-ideal instrument and sample properties. Addressing these issues, the study on the simulated and the measured infrared data by partial least squares regression has been conducted. The simulated data sets were modelled by considering a range of undesired chemical and physical spectral anomalies and variations that can occur in a measured spectrum, such as baseline variations, noise and scattering effects. The study has demonstrated the importance of optimization of the SG parameters during the conversion of spectra into derivative form, specifically window size and polynomial order of the fitting curve. A specific optimal window size is associated with an exact component in the system being estimated, and this window size does not necessarily apply for some other component present in the system. Since the optimization procedure can be time consuming, as a rough guideline spectral noise level can be used for assessment of window size. Moreover, it has been demonstrated that, when the extended multiplicative signal correction (EMSC) is used alongside the SG procedure, the derivative treatment of data by the SG algorithm must precede the EMSC normalization.

Index Headings: **Pre-processing; FTIR spectroscopy; Savitzky-Golay, Derivative; Extended multiplicative signal correction; EMSC; Partial least squares; PLS; Regression.**

INTRODUCTION

Infrared (IR) spectroscopy has been extensively applied in characterizing and identifying inorganic, organic and biological constituents in complex mixtures and matrices. The basis for quantifying various components within a mixture is a separation of the corresponding chemical analyte signals from unwanted interferants. Typical sources of interferant signals are: 1) accompanying chemical constituents in the measured system (including contaminants), 2) collateral chemicals present in the experimental setting (such as atmospheric CO_2 and H_2O), 3) nonchemical effects (such as refractive index variation and scattering), and 4) instrumental anomalies (including white noise). These interferant signals have deleterious effects on the interpretation of the spectra and on the success of the spectral data analysis. It has been shown that the multivariate regression methods, such as principal components regression (PCR) and partial least-squares regression (PLSR), result in simpler and often better models when applying them on pre-processed data.[1,2] Therefore, it is advantageous to employ spectral preprocessing in order to suppress interferant signals and/or to enhance analyte signals.

Since the bandshape of a chemical analyte signal is quite specific in IR spectroscopy, it can often be successfully discerned from a majority of interferant signals, which differ considerably from the signatures of chemical analytes. The line profile of absorption bands in the IR signals of condensed phase is usually close to the Voigt profile.[3] A Voigt profile can be approximated as a linear combination of the Gaussian profile (modelling the Doppler broadening), and the Lorentzian profile (modelling the collision broadening). The difference between spectra of liquid and solid phases is quite small: Spectra of liquids exhibit predominantly Lorentzian bandshapes, while the bands of solids are usually narrower and are often represented with the increased Gaussian fraction in the simplified Voigt model.[4]

Calculating derivatives of spectral data is often a first preprocessing step since it in one step may acquire all of the mentioned benefits. The derivatives emphasize band widths, positions and separations, while simultaneously reducing or eliminating baseline and background effects.[5,6] The derivative values are usually assessed by the Savitzky-Golay (SG) numerical algorithm that approximates spectrum by polynomial (typically quadratic or quartic) least square fitting inside a moving window. [7-9] The least-square value of a given point is calculated as a weighted combination of that point and m points on each side of it, sequentially acquiring a mowing least-squares fit (within $2m+1$ points) across the data. The polynomial order, and especially the window size, can strongly influence the properties of the derivated curve, and consequently the result of the multivariate analysis. Although it is often presumed that the differentiation of spectra increases the noise, the least-square fit of the SG procedure acts as a low-pass filter and hence can actually reduce the level of homoscedastic spectral noise.[10] In fact, the SG procedure is the most widely used smoothing algorithm in spectroscopy since it attenuates high-frequency signals, such as noise, while at the same time tends to preserve important features of the chemical analyte signals, such as relative maxima, minima, height and width. Furthermore, the intrinsic property of any derivative procedure is higher suppression of a broad signal than a narrow one, providing that the signals are of equal amplitudes.[6] Therefore, when provided for calculating the derivative spectra, the SG procedure functions as a high-pass filter, basically reducing the amplitudes of broad spectral features,

such as curving baseline, that predominantly have detrimental effect in data analysis. Concerning baseline, the practical effect of using derivative data is removal of a baseline vertical shift by the first derivative and slope by the second derivative. The mentioned low-pass and high-pass properties effectively achieve that the SG derivative procedure acts as a band-pass filter that can be tuned to suppress all of the unwanted signals except those with the similar bandshape properties as the measured analyte. In essence, by optimizing the SG procedure according to the bandshape properties of the analyte of interest, the interferant signals, either high-frequency ones (usually referred to as "noise") or broad spectral features (including curving baseline) can be greatly reduced.

However, signal bandshapes belonging to the same analyte can differ markedly in different spectral regions, and hence it should be taken into consideration that an optimal SG window size is often limited to an exact spectral region. Why specific parameters are chosen is seldom reported in scientific studies, and it is uncommon to find that several parameters were taken into consideration within the same study.

Along with differentiation, data analysis of vibrational spectroscopic data often includes scatter-corrective preprocessing methods such as Multiplicative Signal Correction (MSC), Extended MSC (EMSC), and Standard Normal Variate (SNV).[11] EMSC, the most universal of the mentioned methods, is a model-based preprocessing method that allows estimation and removing of additive and multiplicative effects.[2,12] The physical "interferant" information is explicitly parameterized, and thus the separation of physical and chemical information from the spectral data is enabled. When the EMSC method is performed after the conversion of spectral data into a derivative form it principally has the feature of removing the multiplicative effect, since broad baseline structures are effectively suppressed by derivatives. When EMSC is applied directly to raw spectra, it also eliminates baseline effects. Both approaches, applying EMSC on raw data and applying EMSC on derivated data, are regularly employed.[13-17] It is important to demonstrate the effects and differences of these two approaches, since it is not readily apparent in scientific literature which is the right one to use.

The main objective of this study was to investigate and assess the preprocessing methods for analysis of spectral data. The effect of the SG data manipulation regarding differentiation order (zero, first and second) was reported recently, indicating some risks of incorrect preprocessing of near-IR data.[18] In the presented study, the particlar attention was given to the optimization of size of the moving window employed in the SG algorithm during the conversion of recorded spectra into derivative form. Moreover, since the SG and the EMSC preprocessing procedures are often used cooperatively,[14-17] the optimal sequence of the preprocessing procedures was assessed. The preprocessing methodologies were studied using a real data set and a simulated one where all constituent effects are known. For both data sets the effect of the different preprocessing strategies, such as an optimum window size and a polynomial order for the SG algorithm, was evaluated by estimating the predictive ability by PLSR.[19]

SIMULATED DATA

The simulated spectra were built up to comprise analyte signals (i.e. the interesting components to be quantified by the regression model), and the interferant signals in the form

of additional constituents, baseline shifts and noise. The analyte signals were modelled by Lorentzian functions, which offer fairly good approximation for IR bands of liquids and biological samples in general. Simulated data consisted of several data sets, with 100 spectra each, differing in the quality and the quantity of the interferant signals. Each spectrum was composed of two or more of the following contributions: 1) *A components* - analyte signals, 2) *B components* - broad interferant signals (two groups with five signals each), 3) *Av components* - the "average" spectrum, 4) *Bv components* - vertical baseline shift, 5) *Bl components* - linear baseline shift, and 6) *N components* - high-frequency random noise (Overview of simulated signals is presented in Table 1).

Analyte signals (***A components***)**:** The properties of the modelled chemical analyte signals were selected according to the characteristics of the standard vibrational bands in the fingerprint mid-IR spectral region (500-2000 cm^{-1}). The analyte signals were modelled by combining Lorentzian curves, with their signal maxima within 1725-1650 cm^{-1} spectral region. A set of 30 Lorentzians (***A^1 components***) with constant height (0.01) was randomly created, with independent variations in their positions and bandwiths between each data set: the distance between adjacent signals varied from 1 to 3 cm^{-1}, and widths from 5 to 10 cm^{-1}. An example of the 30 simple analyte signals, simulated with randomly selected parameters of signal positions and widths, is shown in Fig. 1a. Complex analyte signals (***A^5 components***) were created by randomly combining five *A^1 components*, each scaled independently with the random factor ranging from 0.7 to 1.4. Each data set comprised four of these complex analyte signals, and an example is shown in Fig. 1b. Each *A^5 component* was scaled independently for each spectrum with the scaling factor y_s (ranging from 0 to 0.2) within each data set to simulate concentration variation of the analyte.

Broad interferant signals (***B components***)**:** The analyte signals were overlapped by impeding broad signals, simulating spectral distortions and anomalies due to accompanying chemical constituents in a sample such as water, or nonchemical effects such as Mie scattering. Each set of these broad interferant signals was modelled by the two groups, each comprising five Lorentzians, in two separate spectral regions (roughly centred at 1000 and 2400 cm^{-1}). Analogue to *A components*, all signals had constant height (0.1), while their positions and bandwiths varied independently between each data set: the distance between adjacent signals varied from 50 to 100 cm^{-1}, and widths from 350 to 450 cm^{-1}. An example of the ten broad interferant signals, simulated with randomly selected parameters of signal positions and widths, is shown in Fig. 1c. The signals were scaled independently for each spectrum (from 0 to 0.2) within each data set. Fig. 1d shows the two resulting curves when the ten signals from Fig. 1c are randomly scaled and overlapped with one another, and with the analyte signals from Fig. 1b.

The "average" spectrum (***Av components***) (with 0.3 maximum absorbance value), obtained from spectra of pollen samples,[20] was added to the modelled signals (Fig. 2a). The spectrum is representative for biological samples, having typical bands associated with lipids (2950-2800 and 1730-1760 cm^{-1}), proteins (1650-1500 cm^{-1}), carbohydrates (1200-900 cm^{-1}) and water (3000-3600 cm^{-1}). This "average" spectrum was added in order to provide a realistic data set with slight inter-spectral variations that vary around an average spectrum, as we use to have in IR spectroscopy of biological systems (Fig. 2b). It should be stressed out that simple pre-processing methods such as MSC and SNV are efficient only if this precondition is met;

That is, the methods postulate that variations in analyte and other chemical signals have minor effect on the overall spectral set compared to the physical variations.

Adding multiplicative effects and baseline effects: In addition, modelled data (the sum of *components A, B* and *Av*) was scaled by the multiplicative parameter b_s that varied independently (from 0.8 to 2.0) for each spectrum. The multiplicative parameter was added to simulate variations in sample thickness and/or concentration (Fig. 2c).[2,12] In addition to the broad interferant signals, several other hindrances were added to the spectra: vertical (***Bv components***) and linear baseline shifts (***Bl components***) (Fig. 2d). Baseline variations were generated by changing the shift (from -0.2 to 0.2) and the baseline slope (from 0 up to 0.0005 cm) independently for each spectrum.

High-frequency random noise: As the last component, white Gaussian noise was added to simulate high-frequency homoscedastic noise (***N components***). Maximum amplitude value of random noise was constant within each data set, amounting for approximately 10 % of maximum height values of the analyte signals. It should be noted that, although the noise level was kept constant within each data set, the signals on the other hand were scaled independently, and therefore the signal-to-noise ratio varied slightly from spectrum to spectrum.

A random number generator was used to vary the parameters of the simulated signals. The simulated spectral range covered region from 800 to 4000 cm^{-1}, with a digital resolution of 1.9 cm^{-1}, and the data was analysed by PLSR in the whole data range (Fig. 2d).

SPECTRAL DATA

The real data set consisted of FTIR spectra of 219 milk samples. The milk samples (see details in [21]) were recorded as thin dried films using a high-throughput screening eXTension (HTS-XT) unit coupled to a Tensor 27 spectrometer (both Bruker Optik GmbH, Germany). Spectra were recorded in transmission mode in the spectral region from 4000 to 500 cm^{-1} with a resolution of 6 cm^{-1} (digital resolution: 1.9 cm^{-1}) and an aperture of 5.0 mm. Background spectra of the silicon substrate were collected before each sample measurement to account for variation in water vapour and CO_2. For each spectrum, 64 scans were collected and averaged (apodization: Blackmann–Harris 3 term). Modified data sets were created by adding the simulated random noise of varying amplitude to the measured data. The spectral region of 720 to 3200 cm^{-1} was used as predictor variables for the PLSR regression.

Alongside the FTIR measurement, the milk samples were subjected to reference analysis by gas chromatograph with flame ionization detection (Agilent 6890N GC, Agilent Technologies, USA) (see details in [21]). The concentration of individual fatty acids (such as palmitic acid; C:16) was expressed in percent of total fatty acids present (on a fatty acid methyl ester basis). The complete list of measured fatty acids can be found in [15]. Summed fatty acid parameters were calculated directly from the GC-FID results (PUFA - summed polyunsaturated fatty acids; SAT - summed saturated fatty acids; MUFA - summed monounsaturated fatty acids). Thus obtained chemical reference values of fatty acid composition were used as regressors in the PLSR modelling of measured milk data (in the following referred to as y_m-variables).

DATA PREPROCESSING AND ANALYSIS

Simulated and the measured data sets were pre-processed identically prior to calibration: for the non-derivated (original) spectra the raw spectra were pre-processed by the EMSC only, while for the derivative spectra the raw data were subjected to second-derivative treatment by the Savitzky–Golay algorithm followed by the EMSC normalization. For the SG preprocessing a polynomial of degree two (quadratic) or four (quartic) was used. The lower-degree polynomials were used since for any filter length chosen, low degree is preferred for noise reduction over higher polynomials.[22] The cubic polynomial was omitted since for second-derivative treatment the SG algorithm produces equivalent results for central point using even polynomial degree and the next higher odd degree (i.e. the same second-derivative values are obtained for quadratic and cubic polynomials) .[22] In order to assess the influence of the SG window size on the data analysis the extremely broad range of window sizes was examined, from 3 up to 201 points. This number of window points is way over the standard window range, which in the typical IR data analysis is 5 to 23 points.

Following the second-derivative treatment by the SG algorithm the data were pre-processed by the EMSC normalization with linear and quadratic component. The EMSC model used in the pre-processing is defined by Eq. 1, and the EMSC corrected spectra by Eq. 2.

$$\mathbf{z}_i = b_i\mathbf{m} + a_i\mathbf{1} + d_i\tilde{\boldsymbol{\nu}} + e_i\tilde{\boldsymbol{\nu}}^2 + \boldsymbol{\varepsilon}_i \quad (1)$$

$$\mathbf{z}_{i,\mathrm{Corr}} = (\mathbf{z}_i - a_i\mathbf{1} - d_i\tilde{\boldsymbol{\nu}} - e_i\tilde{\boldsymbol{\nu}}^2) / b_i \quad (2)$$

where $\mathbf{z}_i$ is a 'measured' spectrum (where $i = 1,\ldots, N$; number of spectra in a data set N is 100), $\mathbf{z}_{i,\mathrm{Corr}}$ is a corrected spectrum, b_i is a multiplicative parameter, $\mathbf{m}$ is a reference spectrum often presented by the mean spectrum for a given data set, $\mathbf{1}$=[1, 1, 1, . . ., 1], a_i, d_i and e_i are constant, linear and quadratic parameters respectively, $\boldsymbol{\varepsilon}_i$ is a residual term, $\tilde{\boldsymbol{\nu}}$ is spectral range (wavenumbers).

Pre-processed spectra were used to develop multivariate regression models based on PLSR. Chemical reference values of fatty acid composition were used in the PLSR modelling of measured milk data (y_m-variables), while the scaling factors y_s, used for simulating different amounts of analytes as described above, served the same purpose for the simulated data sets. The optimal number of PLSR factors of the calibration models was determined using segmented cross-validation using 10 segments. The reference values, y_{mi} or y_{si}, and the predicted value, $\hat{y}_i$, of every sample were used to calculate the prediction error of the cross-validated calibration model, expressed as the root mean square error of cross-validation (RMSECV). Both the RMSECV and the coefficient of determination (R^2) between the reference and predicted values were used to evaluate the calibration models. Although RMSECV is a better indicator for the evaluation of a calibration model, the general progressions of both of these parameters had no noticeable differences, and the R^2 were used in the graphical representation of the results. It should be noted that optimal number of PLSR factors can vary depending on window sizes; however, for simpler viewing the number of factors represented in Figs. 4, 6 and 8 were kept constant, at values obtained for an optimal window sizes (except for the values obtained for the analysis on the original data). Residuals in Fig. 11 were calculated as normalized (by the number of data points) sum of squares of residuals $\boldsymbol{\varepsilon}_i$. The data analyses were performed by algorithms written for the setting of Matlab, V. 7.10 (The Mathworks Inc., Natick, MA).

RESULTS AND DISCUSSION

Vibrational spectra of biological samples (Fig. 3) are often rich in information, which present not only an opportunity in data analysis but also a great challenge. In most cases different spectral bands can only be discriminated after application of resolution-enhancement techniques, such as conversion of data into second derivatives. Despite the complexity of the spectral information, the data can be readily used in a qualitative and quantitative analyses. However, spectral properties have to be well understood so that chemical analyte signals can be separated from all possible anomalies and interferants. One of the most important properties, of both chemical and physical interest, is signal bandwidth. Signal bandwidths (usually defined as *full width at half maximum*, FWHM) always vary within a certain range of values, depending on the spectroscopic method, sample and sampling configuration. Mid-IR spectroscopy is probably a technique with one of the widest range of possible signal bandwidths, spanning approximately two orders of magnitude. For instance, the typical FWHM for IR bands of a solid sample can be as small as 5 cm^{-1}. On the other hand, for the stretching vibration of hydrogen bonded –OH and –NH group the corresponding absorption band is broadened, and the observed feature may encompass overlap from several hydrogen-bonded species. The FWHM of these absorption envelopes are often extending over several hundred cm^{-1}. Such broadened bands, caused by vibrations of water molecules, can be seen in a spectrum of dried films of milk spectrum (Fig. 3) as broad underlying features (centred at 3490 and 685 cm^{-1}).

If spectral bands with a vast scope of different bandwidths are present in the measured data this can cause extreme complication during the SG pre-processing. As already mentioned, prior to obtaining derivative values the SG algorithm performs smoothing (i.e. polynomial fit) which can result with significant distortion of the signal of interest. If all of the bands being smoothed are of equal FWHM, then the bands will be distorted proportionally. However, since this is rarely the case in a typical IR spectrum we can expect that each band will be distorted to a different extent. Therefore, if parameters for calculation of derivative spectra are ill-chosen, there is always a great risk that a substantial part of potentially valuable information is lost.

Simulated data. The simulated data sets were modelled by considering a range of undesired chemical and physical spectral anomalies and variations that can occur in measured spectrum as described above. The chemical analyte signals (*A components*) were modelled by Lorentzian functions parameterized according to the characteristics of the standard Mid-IR vibrational bands. Lorentzian functions were used also for the modelling of broad interferant signals (*B components*). It should be pointed out that these broad signals were simulating not only unusually broad vibrational bands caused by hydrogen bonding, but also spectral anomalies that can arise due to non-ideal instrument and sample properties. For example, a strong Mie scattering of the source radiation by the measured sample can also produce broad spectral features that can significantly interfere with and distort the signals of chemical analytes.[23,24] In transmission measurements excessive scatter can also cause sloping (linear) baseline that decreases with decreasing wavenumber (*Bl components*). Vertical baseline shift (*Bv components*), another baseline effect that was simulated as well, usually occurs when the light source between the background and the sample spectrum varies, thus hindering

quantitative analysis of the analyte. Finally, a high-frequency homoscedastic noise was modelled in the data (*N components*).

The calibration results for the estimation of the analyte signals (A^5 *components*) obtained from second derivative (quadratic polynomial) of simulated data are shown in Fig. 4. For the calibration results R^2 is presented as a function of the window size used for SG. For the simulated data containing analyte signals only (*components* A^5, *Av, Bv* and *Bl*; i.e. containing neither broad interferent signals nor noise) the optimal window size is three, which is the theoretically minimal value of window size for a quadratic polynomial (Fig. 4a). This is a direct consequence of the fitting procedure applied in SG algorithm, since a small window size enables near-perfect fit of the polynomial on a reduced number of data points. Another important aspect is that the predictive value of the calibration model that is based on the original data set is almost as good as for the model based on the second derivative data (see Fig. 4a; the window size zero marks the analysis on the non-derivated data set). However, when broad interferent signals (*components B*) are incorporated into simulated data, containing analyte signals and all interferants except noise (*components N*), the non-derivated data set has significantly inferior predictability than the derivative data. In both cases an optimal window size is not even relevant since calibration models are artificially accurate ($R^2 > 0.99999$) within broad range of window sizes. This situation, however, is not corresponding to a real data set, which will always contain a certain amount of noise.

The two curves in Fig. 4b refer to simulated data that contains all components (the results in Fig. 4b were obtained for the two A^5 *components* represented with the same colours in Fig. 1b.). The most striking result is that the addition of the random noise (*components N*) to the simulated data results in a higher optimal window size. This is due to an already mentioned property of the SG procedure to act as a low-pass filter. As the size of the SG window increases the suppression of the high-frequency signals becomes more effective (i.e. random noise in a spectrum is smoothed). However, while with the increase of the window size the level of noise will be more effectively reduced, at the same time the distortion of the analyte signals will be increased. The size of the window size should not be increased indiscriminately since the distortion of the analyte signal will hinder the analysis of a studied chemical system. Therefore, the exact influence of these two conflicting factors will determine the value of the optimal window size. For the simulated data the optimal window size mostly varied within 9-25 range.

A band-pass filtering property of the SG procedure, i.e. when it is used for conversion of spectra into derivative data, is illustrated in Fig. 5. Fig. 5a shows the simulated data comprising one A^5 *component* (centred at 1600 cm^{-1}) and one set of *B components* (two groups, each with five broad Lorentzians, centred at 1000 and 2400 cm^{-1}), while the effect of the SG window size on the derivative data is shown in Figs. 5b and 5c. It is obvious that small window sizes effectively reduce strong broad signals (Fig. 5b). If these broad signals, however, contain important information (for example associated with a chemical constituent of interest, such as water) then extremely large window sizes must be employed (Fig. 5c). On the other hand, this setting will reduce the information-rich lineshapes of narrow analyte bands into a handful of indistinctive signals. Therefore, if we want to establish a model with good predictive value for both type of information, the one contained in the narrow signals (*A components*) and the other

that is contained in the broad ones (*B components*), it is more appropriate to use two distinct data sets that were pre-processed optimally for these tasks.

As shown in Fig. 4b, the optimal window size for narrow analyte signals (that are standard in Mid-IR) is quite small, usually between 9 and 25 depending on the noise level. It was mentioned previously that using the original data is not an option for estimation of these narrow analyte signals (A^5 *components*). However, it is a viable option for the broad ones (*B components*, Fig. 6). For the small window sizes the derivatives of these broad signals are completely obscured by high-frequency noise (*N components*), and the predictive value of the calibration model is restored to the value obtained by the original data only when extremely large window sizes are used.

Alongside window size, the additional concern in the SG preprocessing is the polynomial order of the fitting curve. If higher order polynomials are used in the SG differentiation, overfitting can be expected for small window sizes due to the excessive number of parameters relative to the number of data points. Therefore the ratio of the order of polynomials to the window-size should match the noise level in the data. For a high order of polynomial and a relatively low window size overfitting will artificially create occurrence of the additional high-frequency interferants in the second-derivative data (Fig. 7b). However, these additional interferant signals are seldom noticed in a real measured data due to a presence of random noise. As can be seen in Fig. 7c, when the high-frequency random noise is present in the spectrum, the noise level of the differentiated data is more or less independent of wavelength. Nevertheless, if higher order polynomials are utilized for differentiation of data one should avoid using windows of comparable size as the polynomial for the sake of evading overfitting.

Measured data. Analyses of the measured FTIR spectra, belonging to 219 milk samples, show the same general result as the one obtained for simulated data. The calibration results for prediction of fatty acids in milk samples (represented as R^2 values), as a function of the SG window size, are shown in Fig. 8. The results indicate that a choice of window size in the SG preprocessing step can greatly affect the estimation errors. Therefore, how good the calibration model is will depend on the window size, while the polynomial order of a SG fitting curve will affect the results only slightly. As previously reported,[10] quartic polynomials give better results than quadratic. Although for accurate reproduction of signals in a data higher polynomial degree is superior, it is inferior to quadratic polynomial when considering smoothing of a high-frequency random noise in a data.[22] Because of this inferior noise filtering property, the optimal window size for quartic polynomial is always larger than for quadratic one (Fig. 8).

It is important to notice that a specific optimal window size is associated with an exact component (analyte) in the system being estimated, and does not necessarily applies for some other component. For example, in the case of milk spectra the optimal window sizes for the estimation of palmitic acid (C16:0) is significantly smaller than the one for estimation of summed polyunsaturated fatty acids (PUFA) (Fig. 8). Partial reason for this considerable difference in the optimal window sizes can be deduced from the regression coefficients of the calibration models for mentioned components (Fig. 9). It is apparent that the dominant spectral features of the two models are in the completely different spectral regions (Fig. 9a). For the

C16:0, strong signals are founded in the region between 900 and 1300 cm^{-1}, which is a region that is mainly related to C–H deformations and C–O stretches (Fig. 9b). On the other hand, for the PUFA model, the main spectral feature is the *cis* =C–H stretch situated at 3012 cm^{-1}. Considering the complex spectral features with numerous overlapped bands in the former region, and practically simple band overlapping in the latter region, it is quite straightforward why the SG window size influences the data analysis. The discrimination of the overlapped bands, important for the measurement of C16:0, is attained only at relatively small window sizes; for the quadratic polynomial the optimal window size is five (Fig. 8). In contrast, the discrimination of significant spectral features for the analysis of PUFA is achieved at relatively large window sizes; for the quadratic polynomial the optimal window size is 17 (Fig. 8). The analysis of PUFA shows the importance of the SG preprocessing on the prediction model. For the window sizes smaller than 13 the suppression of the random noise is ineffective, while for the window sizes larger than 19 the distortion of analyte signals is too large for a reliable analysis. Only if the data are pre-processed within a very limited interval of window sizes, from 13 to 19, will the regression model enable reliable estimation of PUFA. Basically, flexible band-pass properties of the SG procedure enable targeted optimization of the preprocessing for each component in the system of interest. It should be stressed out again that, since the SG window size is affecting the estimation error of the calibration model, it is most advantageous to use data that is pre-processed specifically for the optimal estimation of one particular component in the measured system. For example, the use of pre-processed data which is optimized for the estimation of C16:0, will result in an extremely unreliable prediction model for PUFA.

Noise. High-frequency noise sources in FT-IR measurements are mostly associated with non-ideal instrument properties, such as light source and electrical current fluctuations, and other imperfections shared by all electronic devices. Level of noise is an important quantity for characterizing the quality of recorded spectrum, and since noise level is a function of wavelength, this value is calculated from a defined spectral region. The region 2200–2100 cm^{-1} is often chosen as it is near the maximum transmitted wavelength for most FTIR spectrometers. For example the noise level, as defined by diagnostic algorithm of Opus software (Bruker Optics), is the difference between maximum and minimum of the first derivative spectra within the 2200–2100 cm^{-1} spectral region, obtained by SG procedure using quadratic polynomial and window of size nine.

In order to determine influence of noise level on an optimum window size the modified data sets were created by adding the simulated random noise of varying amplitude to the measured data of milk samples. The noise level was calculated by the definition of Opus software, mentioned in the previous paragraph. The impact of noise level on the preprocessing of spectral data is shown in Fig. 10. As expected the optimum window size increases with the increase in amplitudes of noise, while simultaneously the predictive value of the calibration model rapidly decreases. In other words, as the level of random noise increases in spectra, larger windows are needed to suppress the noise, resulting with unwelcomed deformation of analyte signals, and hence unsatisfactory predictive value of the model.

The optimization of the pre-processing, with regard to window size as depicted in Fig 8, can be extremely time-consuming. However, by measuring noise level for a measured

spectral data, an optimum window size for the preprocessing procedure can be estimated promptly.

Correct sequence of the preprocessing procedures. In the data analyses so far the data were pre-processed with the exact order of the preprocessing procedures: the derivative treatment of data by the SG algorithm followed by EMSC normalization. Although this is inherently the correct order in which these two methods should be applied on any given data set, it is worth explaining why this order should be respected. Moreover, the difference between MSC and EMSC preprocessing methods is illustrated here since the analogous arguments can be applied to MSC preprocessing as well.

In case of wavelength-dependent baseline variations, MSC preprocessing results with insufficient spectral correction (Fig 11a). The reason is the lack of higher terms in the MSC model, as opposed to linear and quadratic terms in EMSC model (see Eq.1), The difference between MSC and EMSC preprocessing can be estimated by substracting the MSC residuals from the EMSC ones. The residual term ($\boldsymbol{\varepsilon}_i$ in Eq. 1) comprise chemical information as well as interferant signals such as noise and baseline variations. Therefore, the more of an interferent signal is removed from a spectrum by preprocessing method (and the smaller is a residual term) the better is a data analysis. Figs 11c and 11d show the differences between MSC and EMSC preprocessed original data and second derivative data respectively. It should be noted that this difference is always positive, i.e. residuals are always smaller with EMSC preprocessing, thus resulting with better estimate of analyte signals. Although the difference between the two methods is smaller when spectral correction is preceded by the SG differentiation, it is still present due to a better correction of complex interferent signals (such as *B components*) by EMSC. It is worth noting that obtaining the SG differentiation prior to MSC preprocessing is quite valuable since the differentiation will result with at least partial removal of baseline variations. Thus SG differentiation suppresses broad underlying baselines, but it does not totally remove them. Therefore, in general EMSC performs better on SG differentiated FTIR data than MSC.

One of the parameters estimated by MSC and EMSC preprocessing is the multiplicative parameter *b* (see Eq. 1.). In IR spectroscopy this parameter accounts for the differences in the spectral absorbance due to variations in the effective optical path length. Erroneous estimation of this scaling factor results in over- or underestimated total absorbance. It is therefore of outmost importance that the multiplicative parameter is estimated as accurately as possible.

In order to show the influence of the exact order of the SG and the EMSC preprocessing procedures on the estimation of the multiplicative parameter the data set was pre-processed in the two opposite sequences: 1) the SG differentiation followed by the EMSC, and 2) the EMSC on the raw data followed by SG differentiation. The procedure success was based on the difference between the estimated multiplicative parameter *b*, as obtained by the EMSC, and the actual simulated value b_s. As can be seen on Fig. 11e, the sequence of differentiation followed by EMSC results in estimates that are very close to the true values of *b*, and much closer than is the case for the *b* values obtained from the EMSC on the original data. Since the FTIR bands become very sharp by taking the second derivative, the average spectrum can be nearly considered as independent from the other model spectra and, therefore, the calculation of the parameters is very precise when the right order is chosen, i.e. derivative first and then EMSC.

It is therefore apparent that this pair of operations, SG differentiation and EMSC (or MSC) pre-processing, are not-commutative, and that the correct sequence of these procedures must be obeyed.

CONCLUSION

Considering the vast range of possible signal bandwiths encountered within a typical Mid-IR spectrum it is not possible to provide general method with specific parameters for spectral preprocessing by the SG procedure. Bandshape properties in NIR spectra are rather uniform, and in this case some general guidelines can be made, however such investigation was not an objective of this study. It is important to notice that a specific optimal window size is associated with an exact component in the system being estimated. This window size does not necessarily apply for some other component present in the measured system. Therefore, it is preferred that each case is studied independently, i.e. by optimizing preprocessing of spectral data for each component. This can significantly impede a multivariate analysis of spectral data when a regression model for predicting a number of variables (components) is demanded. In such case the combination of several pre-processed data can be utilised in data analysis; for example the more reliable data set can be created by concatenating data sets that were pre-processed using different SG parameters. The optimization of the SG preprocessing parameters can be time-consuming, and when time is in short supply spectral noise level can be used for their rough assessment. Contrary to previous notion,[11] it has been demonstrated that, when the EMSC (or the MSC) is used alongside the SG procedure, the derivative treatment of data by the SG algorithm must precede the EMSC (MSC) normalization.

ACKNOWLEDGMENTS

Support of this research by the Norwegian Food Research Foundation , by the Research Council of Norway (grants 199581/I10 and 173321/I10), and by the Unity Through Knowledge Fund (grant 92/11) is gratefully acknowledged. The authors wish to thank Bjørg Narum and Nils Afseth for their help in the FT-IR measurements.

REFERENCES

1. A. Kohler, U. Böcker, J. Warringer, A. Blomberg, S.W. Omholt, E. Stark, H. Martens. ”Reducing Inter-replicate Variation in Fourier Transform Infrared Spectroscopy by Extended Multiplicative Signal Correction”. Appl. Spectrosc. 2009. 63(3): 296-305.
2. A. Kohler, C. Kirschner, A. Oust, H. Martens H. ”Extended Multiplicative Signal Correction as a Tool for Separation and Characterization of Physical and Chemical Information in Fourier Transform Infrared Microscopy Images of Cryo-Sections of Beef Loin”. Appl. Spectrosc. 2005. 59(6): 707-716.
3. A. Belafhal. ”The Shape of Spectral Lines: Widths and Equivalent Widths of the Voigt Profile”. Opt. Commun. 2000. 177(1-6): 111-118.
4. P.R. Griffiths. ”Introduction to Vibrational Spectroscopy”. In J.M. Chalmers, P.R. Griffiths, editors. Handbook of Vibrational Spectroscopy. Chichester, UK: John Wiley and Sons, 2002. Pp. 33-70.

5. L.K. DeNoyer, J.G. Dodd. ”Smoothing and Derivatives in Spectroscopy”. In J.M. Chalmers, P.R. Griffiths, editors. Handbook of Vibrational Spectroscopy. Chichester, UK: John Wiley and Sons, 2002. Pp. 2173-2183.
6. C.D. Brown, L. Vega-Montoto, P.D. Wentzell. ”Derivative Preprocessing and Optimal Corrections for Baseline Drift in Multivariate Calibration”. Appl. Spectrosc. 2000. 54(7): 1055-1068.
7. A. Savitzky, M.J.E. Golay. ”Smoothing and Differentiation of Data by Simplified Least Squares Procedures”. Anal. Chem. 1964. 36(8): 1627-1639.
8. T.H. Edwards, P.D. Willson. ”Digital Least Squares Smoothing of Spectra”. Appl. Spectrosc. 1974. 28(6): 541-545.
9. J. Steinier, Y. Termonia, J. Deltour. ” Smoothing and Differentiation of Data by Simplified Least Square Procedure”. Anal. Chem. 1972. 44(11): 1906-1909.
10. H. Ziegler. ”Properties of Digital Smoothing Polynomial (Dispo) Filters”. Appl. Spectrosc. 1981. 35(1): 88-92.
11. Å. Rinnan, F. Van der Berg, S. Balling Engelsen, ”Review of the Most Common Pre-Processing Techniques for Near-Infrared Spectra”. TrAC, Trends Anal. Chem. 2009. 28(10): 1201-1222.
12. H. Martens, E. Stark. ”Extended Multiplicative Signal Correction and Spectral Interference Subtraction - New Preprocessing Methods for Near-Infrared Spectroscopy”. *J. Pharm. Biomed. Anal.* 1991. 9(8): 625-635.
13. S. Ottestad, T. Isaksson, W. Saeys, J.P. Wold. ”Scattering Correction by Use of a Priori Information”. Appl. Spectrosc. 2010. 64(7): 795-804.
14. K.H. Norris, G.E. Ritchie. ”Assuring Specificity for a Multivariate Near-Infrared (NIR) Calibration: The Example of the Chambersburg Shoot-Out 2002 Data Set”. *J. Pharm. Biomed. Anal.* 2008. 48(3,4): 1037-1041.
15. S.W. Bruun, I. Søndergaard, S. Jacobsen “Analysis of Protein Structures and Interactions in Complex Food by Near-Infrared Spectroscopy. 1. Gluten Powder”. *J. Agric. Food Chem.* 2007. 55(18): 7234-7243.
16. J. Vongsvivut, P. Heraud, W. Zhang, A. Jaroslav, J.A. Kralovec, D. McNaughton, C.J. Barrow. ”Quantitative Determination of Fatty Acid Compositions in Micro-Encapsulated Fish-Oil Supplements Using Fourier Transform Infrared (FTIR) Spectroscopy”. Food Chem. 2012. 135(2): 603-609.
17. P. Heraud, E. Ng, S. Caine, Q. Yu, C. Hirst, R. Mayberry, A. Bruce, B. Wood, D. McNaughton, E. Stanley, A. Elefanty. “Fourier Transform Infrared Microspectroscopy Identifies Early Lineage Commitment in Differentiating Human Embryonic Stem Cells”. Stem Cell Res. 2010. 4(2): 140–147.
18. S.R. Delwiche, J.B. Reeves. ”A Graphical Method to Evaluate Spectral Preprocessing in Multivariate Regression Calibrations: Example with Savitzky-Golay Filters and Partial Least Squares Regression”. Appl. Spectrosc. 2010. 64(1): 73-82.
19. S. Wold, H. Martens, H. Wold. ”The Multivariate Calibration-Problem in Chemistry Solved by the PLS Method”. Lect. Notes Math. 1983. 973: 286-293.
20. B. Zimmermann. ”Characterization of Pollen by Vibrational Spectroscopy”. Appl. Spectrosc. 2010. 64(12): 1364-1373.

21. N.K. Afseth, H. Martens, L. Giskehaug, B. Narum, K. Jørgensen, S. Lien, A. Haug, A. Kohler. ”Predicting the Fatty Acid Composition of Milk: A Comparison of Two Fourier Transform Infrared Sampling Techniques”. Appl. Spectrosc**.** 2010. 64(7): 700-707.
22. P. Barak. ”Smoothing and Differentiation by an Adaptive-Degree Polynomial Filter”. Anal. Chem. 1995. 67(17): 2758-2762.
23. A. Kohler, J. Sulé-Suso, G.D. Sockalingum, M. Tobin, F. Bahrami, Y. Yang, J. Pijanka, P. Dumas, M. Cotte, H. Martens. ” Estimating and Correcting Mie Scattering in Synchrotron-Based Microscopic Fourier Transform Infrared Spectra by Extended Multiplicative Signal Correction”. Appl. Spectrosc. 2008. 62(3): 259-266.
24. M. Romeo, B. Mohlenhoff, M. Diem. ”Infrared Micro-Spectroscopy of Human Cells: Causes for the Spectral Variance of Oral Mucosa (Buccal) Cells”. Vib. Spectrosc. 2006. 42(1): 9-14.

Figures

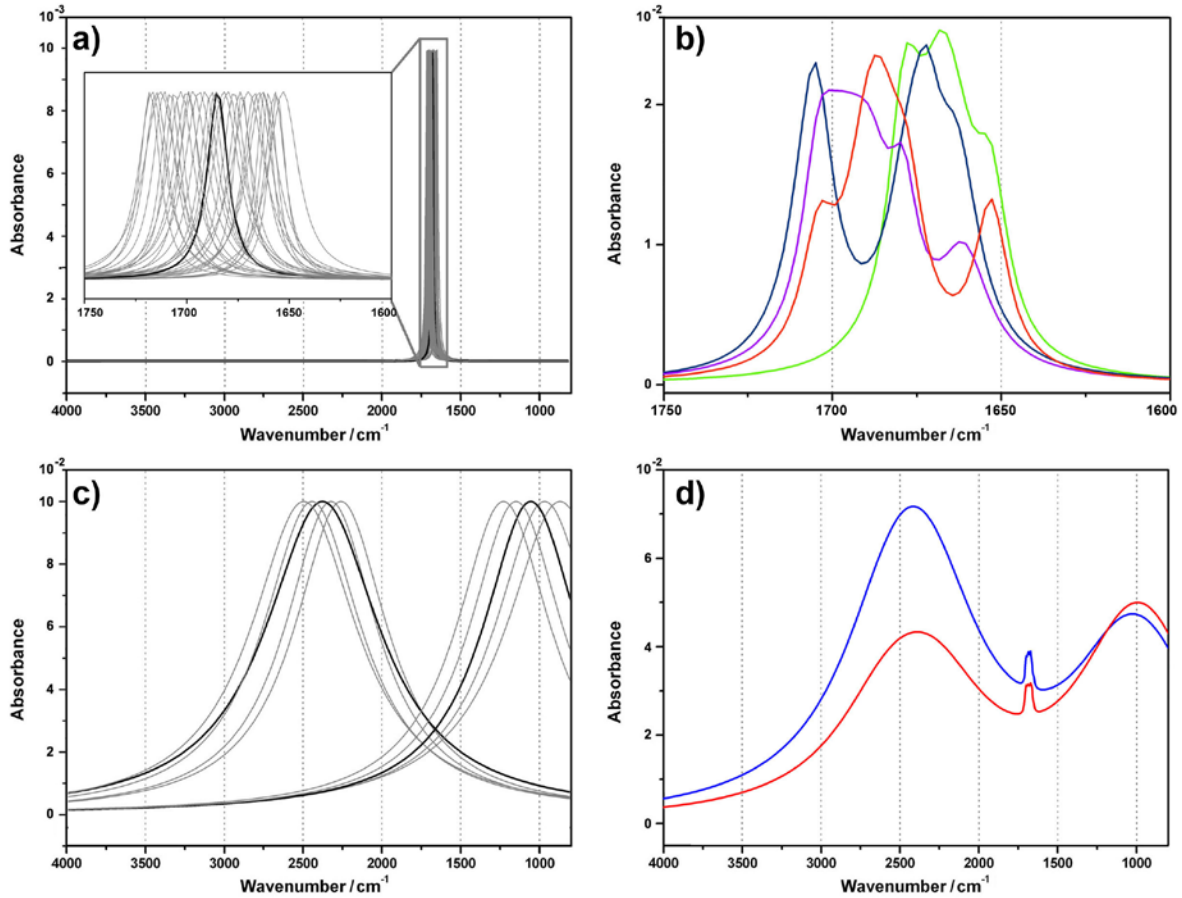

Figure 1. Simulated data: a) 30 modelled simple analyte signals (*A^1 components*); b) four modelled complex analyte signals (*A^5 components*), composed of five *A^1 components* each; c) two sets of broad interferant signals, each comprising five curves (*B components*); d) simulated spectra containing scaled and overlapped *A^5* and *B components*; for better viewing only 2 out of 100 spectra in the data set are shown.

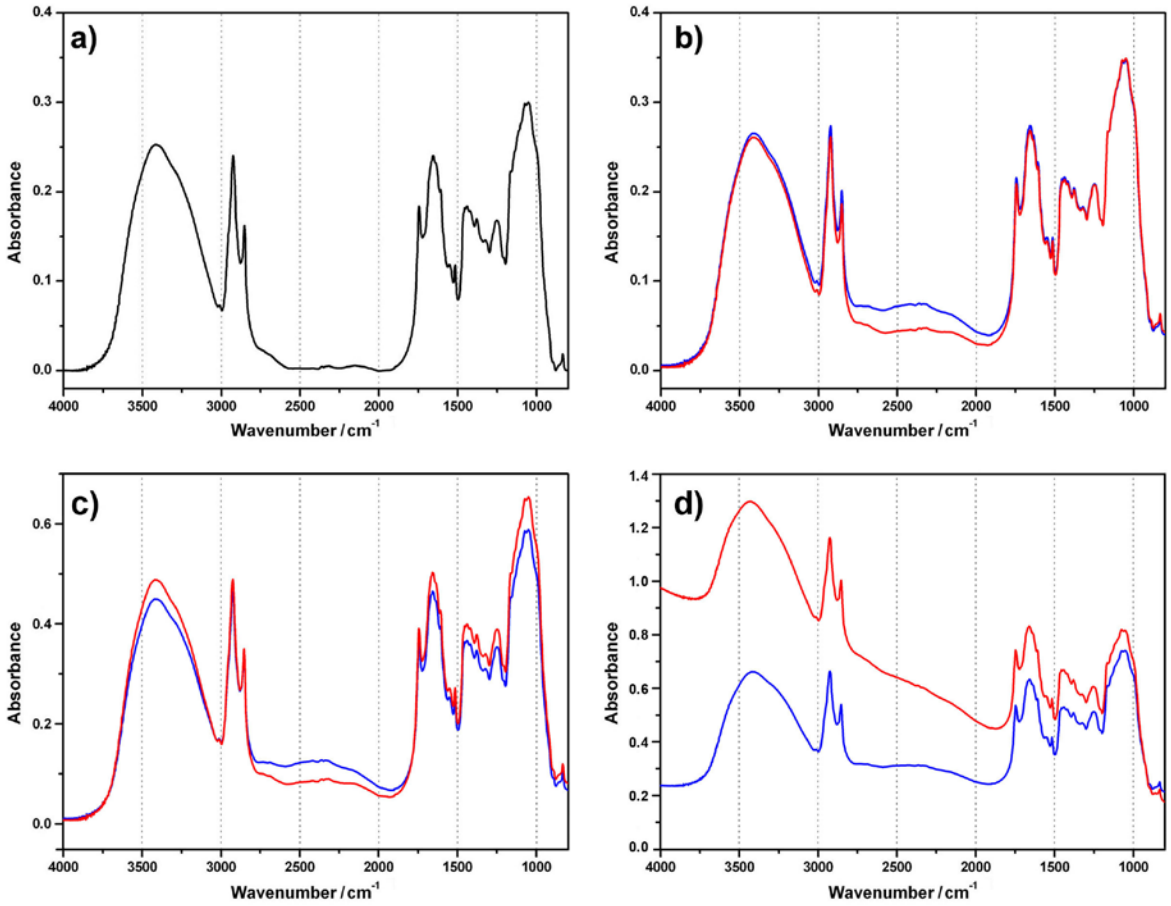

Figure 2. Simulated data: a) the "average" spectrum (*Av components*); b) simulated spectra containing overlapped *components* A^5, *B* and *Av*; c) simulated spectra containing overlapped components A^5, *B* and *Av* after scaling with the multiplicative parameters *b*; d) simulated spectra containing overlapped *components* A^5, *B*, *Av*, *Bv*, *Bl* and *N* (scaling with the multiplicative parameters *b* included). For better viewing only 2 out of 100 spectra in the data set are shown.

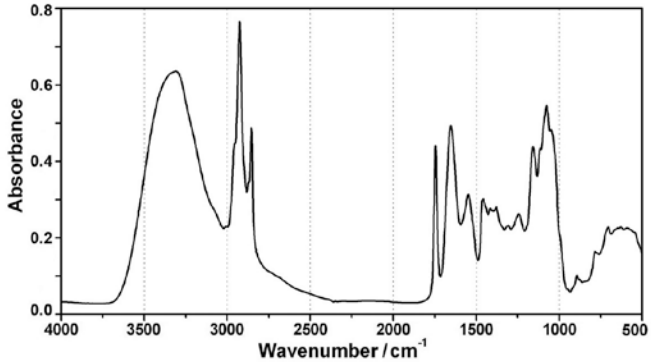

Figure 3. FTIR spectrum belonging to one of the measured milk samples; extensive range of bandwidths as well as widespread band overlapping is clearly seen.

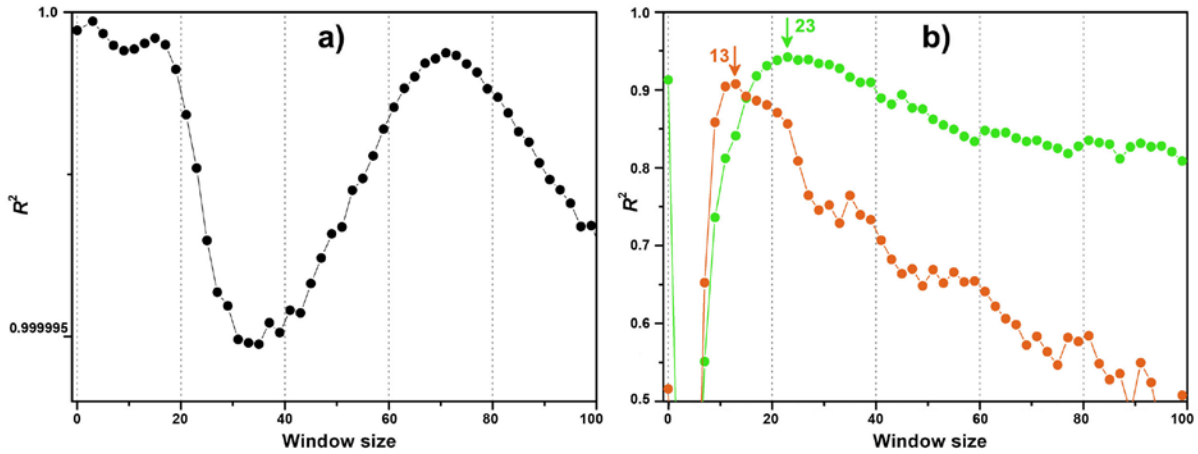

Figure 4. Calibration results for the estimation of the analyte signals (the two A^5 *components* represented with the same colours in Fig. 1b) obtained from second derivative (quadratic polynomial) simulated data, containing all *components*. Window size zero marks the analysis on the original (non-derivative) data.

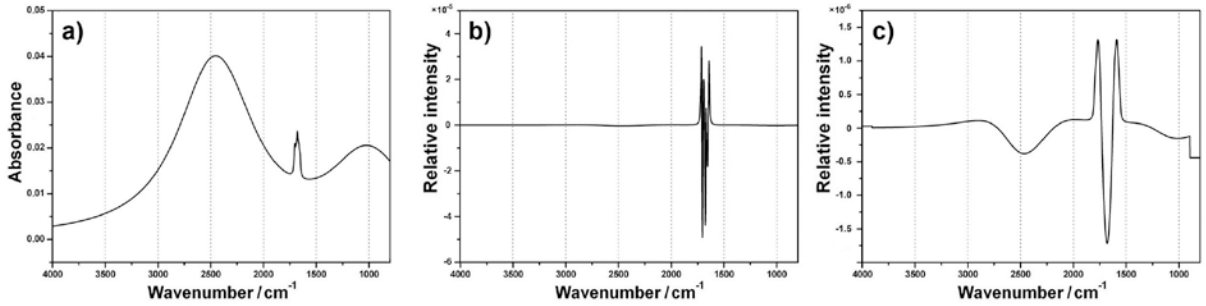

Figure 5. Simulated data containing *components* A^5 and *B*: Effect of window size utilized in the SG algorithm on the profiles of second-derivative data (quadratic polynomial): a) original simulated data, b) second-derivative data obtained by window size 11, c) second-derivative data obtained by window size 101.

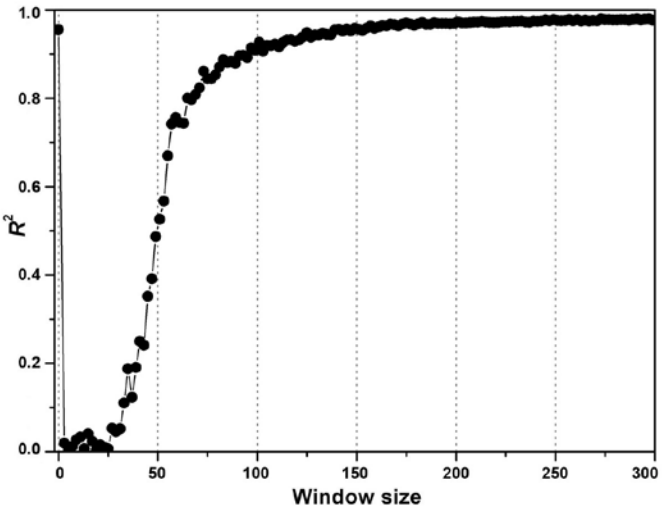

Figure 6. Calibration results obtained from the second derivatives (quadratic polynomial) of the simulated data set with all *components* for estimation of broad interferant signals (*B components*). Window of zero size marks the original (non-derivative) data.

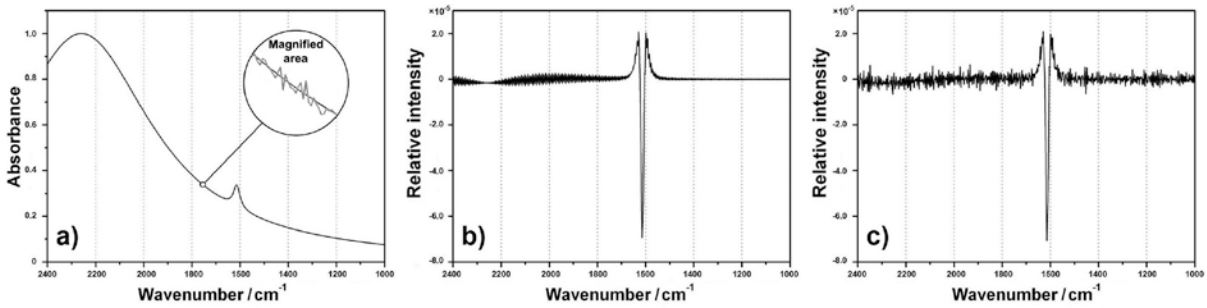

Figure 7. Effects of small window size utilized in SG algorithm on the profiles of second derivative data (obtained by quartic polynomial and window of size 7): a) original simulated data consisting of only one *component A*[1] and one *component B*, with (gray) and without (black) *component N* (see magnified area), b) second-derivative of the simulated data which contained no component *N*, c) second-derivative of the simulated data that contained *component N*.

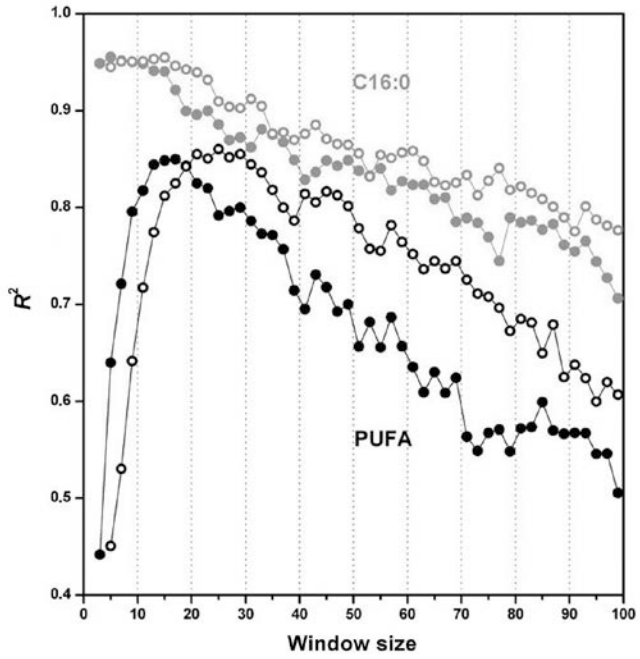

Figure 8. Calibration results for estimation of palmitic acid (C16:0, grey) and summed polyunsaturated fatty acids (PUFA, black) obtained from the second derivative of measured data set; order of the SG fitting polynomial: quadratic (full circle) and quartic (hollow circle).

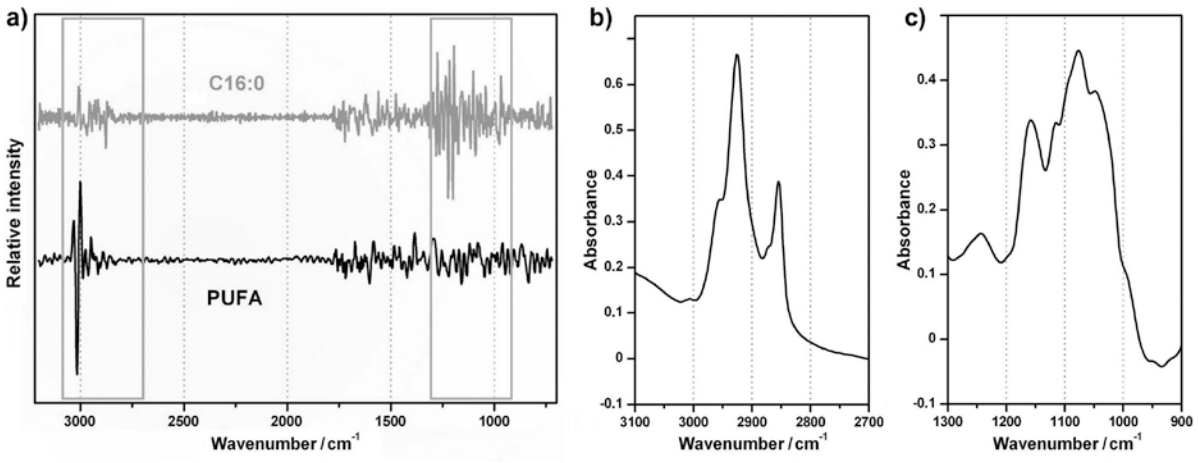

Figure 9. a) Regression coefficients obtained for the optimum window size: 5 for palmitic acid (C16:0) and 17 for polyunsaturated fatty acids (PUFA). b) and c) Details of FTIR spectrum belonging to one of the measured milk samples (see Fig. 2 for the whole spectrum); corresponding wavenumber regions for the regression coefficients are represented within the gray rectangles.

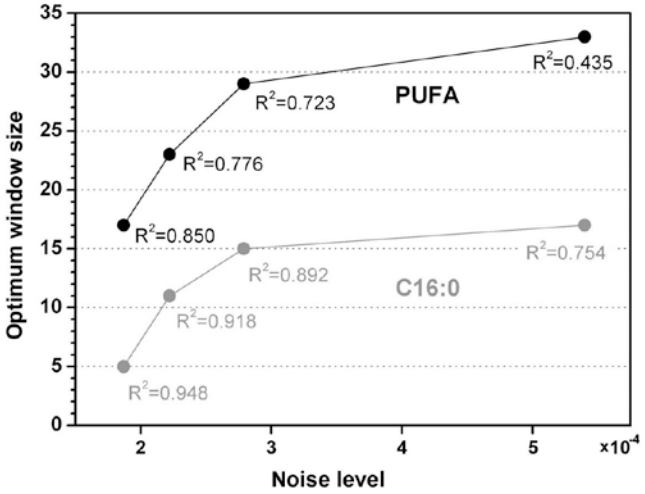

Figure 10. Influence of noise level on optimum window size (second derivatives obtained by quadratic polynomial), calculated for the measured milk data with simulated increase in noise level.

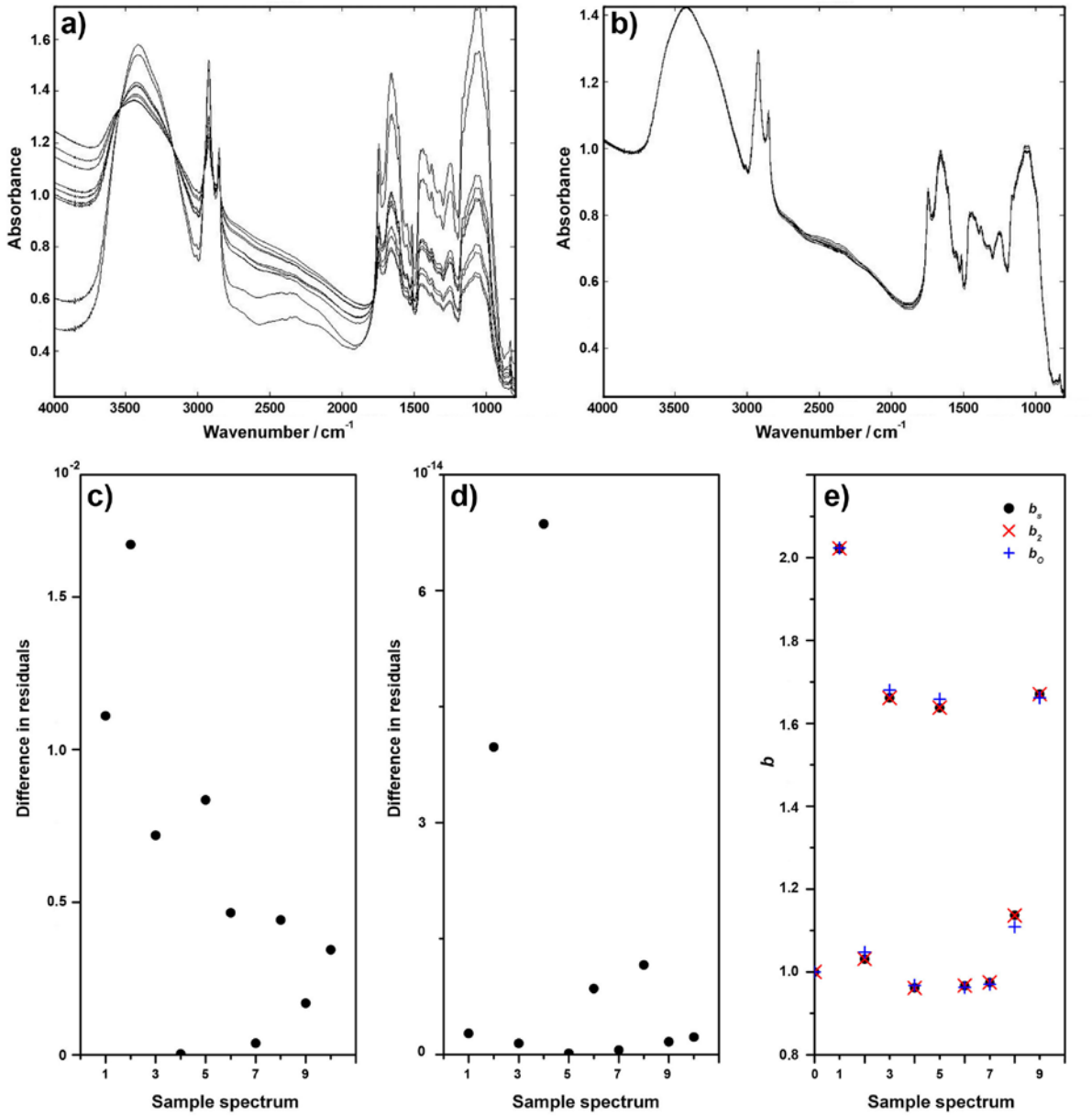

Figure 11. a) MSC and b) EMSC preprocessing of ten simulated spectra. Difference in residuals: c) between MSC and EMSC preprocessed original data d) between MSC and EMSC preprocessed second derivative data. e) Values of the multiplicative quantities *b* for spectra containing all *components* (b_s – simulated values; b_2 – estimated values obtained from the EMSC on the second derivative data; b_O – estimated values obtained from the EMSC on the original data); all *b* values are normalized to the value of the first sample (designated as sample spectrum 0).

Tables

Table 1. The components of the simulated data sets; Typical data set had 100 spectra, each composed of 4 *A^5 components*, 10 *B components*, 1 *Av component*, 1 *Bv components*, 1 *Bl component*, and 1 *N component.*

Simulated signal	Signal form	Parameters
A^1 components: Simple analyte signals	Lorentzian: 30 curves (centred at 1675 cm^{-1})	Height: 0.01 Width: 5 - 10 cm^{-1} [a]Distance: 1 - 3 cm^{-1}
A^5 components: Complex analyte signals	Random combination of five *A^1 components*	[b]Scaling of *A^1*: 0.7-1.4 [c]Scaling: 0 - 0.2
B components: Broad interferants	Lorentzian: 2x 5 curves (centred at 1000 and 2400 cm^{-1})	Height: 0.1 Width: 350 - 450 cm^{-1} [a]Distance: 50 - 100 cm^{-1} [c]Scaling: 0 - 0.2
Av components: "Average" spectrum	Recorded spectra of pollen	[d]Height: 0.3
Bv components: Baseline vertical shift	Constant	Shift value: -0.2 - 0.2
Bl components: Baseline slope	Line	Slope value: 0 - 0.0005 cm
N components: Random noise	White Gaussian noise	[e]Height: 0.001

[a]Distance between adjacent signals.
[b]Each *A^1 component* was scaled independently with the random factor.
[c]The signals were scaled independently for each spectrum (from 0 to 0.2) within each data set.
[d]Maximum absorbance value.
[e]Approximately 10 % of maximum height values of the analyte signals.