Research article

Journal of RAMAN SPECTROSCOPY

Received: 17 August 2015 Revised: 10 December 2015 Accepted: 4 January 2016 Published online in Wiley Online Library: 10 February 2016

(wileyonlinelibrary.com) DOI 10.1002/jrs.4886

Model-based pre-processing in Raman spectroscopy of biological samples

Kristian Hovde Liland,[a,b]* Achim Kohler[c] and Nils Kristian Afseth[a]

Model-based pre-processing has become wide spread in spectroscopy and is the standard procedure in Fourier-transform infrared spectroscopy. It has also been shown to give valuable contributions in Raman spectroscopy. Extended multiplicative signal correction is flexible enough to handle varying fluorescence background and take into account individual variations in baselines while still keeping enough rigidity through reference spectra and model fitting to avoid degenerate solutions and overfitting, when used correctly. We demonstrate the basic extended multiplicative signal correction method and some extensions, including a novel shift correction, on real Raman data to demonstrate effects on visual appearance, replicate variation and prediction. Comparisons with other standard correction methods are also shown and discussed. © 2016 The Authors. *Journal of Raman Spectroscopy* Published by John Wiley & Sons, Ltd.

Additional supporting information may be found in the online version of this article at the publisher's web site.

Keywords: extended multiplicative signal correction (EMSC); Raman spectroscopy; baseline correction; normalisation; shift correction

Introduction

Various physical effects and even interferents hamper the interpretation of Raman spectra of biological samples and constituents. Fluorescence, which is a process that usually 'competes' with Raman scattering, will in some cases even render the collection of Raman scattering impossible. While there are both chemical and instrumental ways to reduce the effect of interferents in biological Raman spectra, mathematical pre-processing is in many cases the only practically feasible way to generate reproducible qualitative and quantitative data. It is generally agreed that two basic pre-processing steps are needed for feasible quantitative Raman spectroscopic analysis[1,2]: (1) baseline corrections to remove the effect of fluorescence and other additive features in the spectra and (2) a normalisation procedure to remove multiplicative effects related to for instance uncertainties in reproducible focusing and to laser intensity fluctuations. Baseline and noise removal techniques[3,4] and normalisation procedures[5] have thus been extensively discussed in the literature. In addition, shifts in the wavenumber axis often occur in Raman spectra because of temperature drift or hardware replacements, and there is a general lack of standardisation procedures.[6] Recently, Beattie *et al*. introduced a novel approach for background estimation and removal based on multivariate loadings from singular value decomposition.[7] This approach provided both qualitatively and quantitatively interesting results when applied to data from pathology.[8]

Whereas standard pre-processing techniques like derivatives and vector normalisation are used to remove undesired interferents in the spectra, so-called model-based pre-processing techniques allow for quantifying and separating different types of physical and chemical variations in the spectra. Multiplicative scatter correction (MSC),[9] and later extended MSC (EMSC),[10] was developed in the 80s for applications in near-infrared (NIR) spectroscopy in food science, and today, EMSC is one of the major frameworks for model-based pre-processing in vibrational spectroscopy. In the last years, EMSC has particularly attracted attention within IR spectroscopy for selective correction of features like sample thickness and temperature, water vapour, carbon dioxide and salt concentration.[11–14] An EMSC-based algorithm for estimating and correcting the contribution of Mie scattering effects in Fourier-transform (FT) IR microspectroscopy of cells has also recently gained attention.[15]

An intriguing aspect concerning model-based pre-processing techniques is to extend the basic EMSC algorithm to correct for specific interferents. Within Raman spectroscopy, however, these possibilities have so far not been extensively explored. The most natural extension of the EMSC algorithm for Raman correction is to add polynomial extensions to the basic EMSC algorithm to correct for fluctuating baseline features. This approach was suggested already in 2006,[1] and it has been implemented by several authors.[2,16,17] Another possibility is to extend the EMSC model by adding constituent spectra. This approach has especially been evaluated for bacterial Raman spectra,[18] where spectral features of fatty acids and other metabolites have been extracted from the Raman spectra. For feasible applications of this extension,

* *Correspondence to: Kristian Hovde Liland, Nofima AS – Norwegian Institute of Food Fisheries and Aquaculture Research Osloveien 1, N-1430 Ås, Norway. E-mail: kristian.liland@nofima.no*

a *Nofima AS – Norwegian Institute of Food, Fisheries and Aquaculture Research, Osloveien 1, N-1430, Ås, Norway*

b *Norwegian University of Life Sciences, Department of Chemistry, Biotechnology and Food Science, Postboks 5003, N-1432, Ås, Norway*

c *Norwegian University of Life Sciences, Department of Mathematical Sciences and Technology, Postboks 5003, N-1432, Ås, Norway*

however, care must be taken to avoid using constituent spectra of high similarity, thus introducing rank problems during the least squares estimation of EMSC parameters.[2,19] This risk of rank problems is reduced when constituent spectra are introduced as difference spectra.[20] As alternative, rank problems could also be reduced by introducing constituent or interferent spectra in so-called orthogonal subspace models [i.e. estimations of constituent or interferent effects based on for instance principal component analysis (PCA)] as extensions to the EMSC model. This approach was utilised in the so-called EMSC replicate correction, originally developed to reduce the between-replicate variation in FT-IR spectra of microorganisms.[21] The approach has also proved positive in background correction of biological Raman spectra.[2]

Even though pre-processing techniques of Raman spectra have been extensively studied in recent years, in applied Raman spectroscopy, there is still a need for user-friendly 'all-in-one' approaches that facilitate both spectral interpretability and good quantitative correction of the Raman spectra.[22] EMSC is a pre-processing framework satisfying these requirements, but a thorough quantitative evaluation of EMSC and feasible extensions comparing with standard pre-processing approaches has yet to be published. Thus, in the present study, extensions of EMSC are for the first time evaluated qualitatively and quantitatively for different real-world data sets (i.e. Raman spectra of dried milk samples, adipose tissue and muscle tissue respectively). In the study, EMSC modelling will be compared with traditional pre-processing using polynomial baseline correction[23] in combination with the standard normal variate (SNV). In addition, a novel approach for Raman shift correction using the framework of EMSC is introduced. The success criteria for choosing pre-processing will here be threefold: visual inspection of spectra, predictive ability and model simplicity in regression and mean replicate variation. There are other ways of judging the success of the procedures, but these span a relevant set of applications that most readers will be interested in.

Materials and methods

Milk data

Two-hundred and sixty-four milk samples were obtained from a feeding experiment designed for evaluating two major aspects related to dairy cow feed and their effect on milk production and milk quality of dairy cows.[24] These samples have previously been analysed by FT-IR spectroscopy[25] in an experiment where different IR sampling techniques were compared for prediction of fatty acid composition. The milk samples were frozen and kept at −80 °C until spectroscopic analysis.

Milk samples were thawed over night at approximately 2 °C, and the samples (100 μl) were applied on custom-made aluminium discs (38 mm diameter, 3 mm thickness). The samples were dried for approximately 1 h under low pressure (approximately 0.5 bar) using anhydrous Silica Gel (Prolabo, France) in a desiccator. In this way, semi-aqueous films of concentrated milk-samples were obtained. The aluminium discs were then placed beneath the laser beam onto a custom-built stepless rotation device, and the samples were rotated during spectral acquisition. The rotation velocity was set at about 60–90 rpm. The sample position with respect to the non-contact objective was focused and aligned before the first analysis. Then, the distance was fixed, and no focusing was performed during the rest of the analyses. Raman spectra were collected using a Kaiser Optical Systems Raman RXN1 Analyzer (Ann Arbor, MI, USA) consisting of a holoprobe transmission holographic spectrograph and a charge-coupled device detector with a working temperature of −40 °C. The spectrograph was connected with fibre optics to a Kaiser multireaction filtered probehead, and the system was equipped with a 785-nm stabilised external cavity diode laser. All Raman spectra were obtained using a non-contact objective (f/2, 2.5 inch working distance) connected to the probehead. Six Raman spectra of 30 s exposure times each were obtained for each sample, and between every third spectral acquisitions, the horizontal position of the sample was slightly changed to increase the sampling area. The average laser power was approximately 150 mW at the sample. Because of limited amounts of each sample, 232 samples of the 264 samples included in the experimental design were analysed using Raman spectroscopy. Two chemical replicates were analysed for each sample.

Reference analyses of the fatty acid composition of the milk samples have been described earlier.[25] In the present study, only two fatty acid features, namely the iodine value (i.e. a measure for the total fatty acid unsaturation in a lipid sample, expressed in g I_2/100 g fat) and the concentration of conjugated linoleic acid (CLA, expressed in percent of total fatty acids present), were used for prediction purposes. These two parameters were selected to comprise both a major and a minor fatty acid feature, respectively.

Adipose data

Seventy-seven samples of fat from pork-back fat adipose tissue[26] were cut in pieces (approximately 20 × 20 × 60 mm), homogenised with a mixer, heated for 30 minutes at 75 °C and centrifuged at 22 700 g for 10 min at 40 °C (Beckman J2-MS centrifuge, Palo Alto, CA, USA). Raman measurements were performed using a Raman instrument (RamanRXN1, Kaiser Optical Systems, Inc., MI, USA) equipped with at near-infrared external-cavity-stabilised diode laser (Invictus, Kaiser Optical Systems, Inc.) with a wavelength of 785 nm, an air-cooled charge-coupled device detector, and a ball probe (Matrix Solutions, Bothell, WA, USA) (Ø = 13 mm) utilising a sapphire spherical lens (Ø = 6 mm). The fat was measured at 47–50 °C with the probe in direct contact with the sample for 20 s. Reference analyses were performed by dissolving the samples in toluene and methylating them by adding potassium methylate/methanol before they were analysed in a gas chromatograph (Perkin–Elmer Auto system XL; Perkin–Elmer Analytical Instruments, Shelton, USA).

Muscle data

One-thousand eighty Raman spectra were obtained from a previous study concerning beef muscles subjected to brining.[27] Samples of beef muscle (longissimus dorsi) were taken from four Norwegian Red Cattle 48 h post rigour, and from each animal, two muscle blocks were excised and placed in each of 18 different salt brines. The salt brines comprise 6 pure and 12 mixed NaCl, KCl and $MgSO_4$ solutions, made in 1.5%, 6% and 9% total salt weight percentage concentration, respectively. The samples were kept in brines at 4 °C for 48 h. Subsequently, two muscle blocks were excised from each of the muscle samples, consecutively embedded in O.C.T. compound (Tissue-Trek, Electron Microscopy Sciences, Hatfield, USA) and snap-frozen in liquid N_2. Cryo-sectioning was performed transversely to the fibre direction on a Leica CM 3050 S cryostat (Leica Microsystems Wetzlar GmbH, Wetzlar, Germany). From each of the snap-frozen meat pieces, two cryo-sections were excised and cut in 20 μm thickness, thaw-mounted on CaF_2 slides and subsequently stored in a desiccator before acquisition of the Raman spectra. Raman spectra were recorded by a LabRam HR

800 Raman microscope (Horiba Scientific, France). The excitation wavelength of 632.8 nm was generated by a He–Ne laser. A 100X objective (Olympus, France) was used for focusing and collecting scattered Raman light. The laser power was approximately 15 mW on the sample surface. The confocal hole was set at 200 μm, and an exposure time of 4 × 15 s was used. The Raman scattering was dispersed with a 300-lines/mm grating, which resulted in spectra in the range 408.9–2611.1 cm^{-1}. Data acquisition and instrument control was carried out using the LABSPEC software (version 5.45.09, Horiba Scientific, France). The resulting Raman data set consisted of nine single-myofibre spectra per experimental treatment. The final data set consisted of 1080 spectra (4 animals × 2 muscles blocks × 18 brines × 9 replicate spectra – infeasible and unsuccessful mixtures).

Basic pre-processing approaches

Polynomial background correction

Among the many available types of background corrections, the polynomial ones are quick to fit, extensively applied and often give satisfactory results. Here, we have chosen to use the standard iterative approach of Lieber and Mahadevan-Jansen,[23] which was originally developed for Raman spectra. For each spectrum, the procedure fits a polynomial to the spectrum, calculates a new spectrum as the minimum of the original spectrum and the polynomial, fits a polynomial to the new spectrum and repeats until convergence. Usually, a few iterations are enough to estimate the baseline of the spectrum. Finally, the corrected spectrum is found by subtracting the baseline from the spectrum.

Standard normal variate

The SNV is a simple, but effective procedure for making spectra comparable. It works independently on each spectrum by subtracting the spectrum mean and dividing by the spectrum standard deviation. As long as the original scale of the spectra is not interesting, this is an efficient way of removing constant baseline effects and scaling differences from spectra.

Extended multiplicative signal correction

The EMSC[10] model is an extension of MSC.[9] EMSC is highly efficient and adaptive because the main workhorse is a least square fit of single spectra against a few profile spectra. In addition to correcting the spectra with the desired model, model parameters are returned. These can give valuable information regarding the analysed samples.

Multiplicative signal correction

Multiplicative signal correction extends a Lambert–Beer-type model through an additive effect resulting in the model:

$$A(\tilde{v}) = a + \bar{x}(\tilde{v}) \cdot b + e(\tilde{v}) \quad (1)$$

where $A(\tilde{v})$ is the absorbance at wavenumbers $\tilde{v}$, a is an additive baseline constant, $x^{-}(\tilde{v})$ is the mean spectrum (or another chosen reference) and b is a multiplicative constant. Finally, $e(\tilde{v})$ is the residual vector containing the interesting chemical differences between the samples, i.e.:

$$e(\tilde{v}) = b \cdot \sum_{j=1}^{J} c_j \Delta k_j(\tilde{v}) \quad (2)$$

where c_j are concentrations of the species k_j and $\Delta k_j(\tilde{v})$ are the species' profile deviations from a mean profile or other reference.

Basic and polynomial extended multiplicative signal correction

For EMSC, Eqn (1) is extended with polynomial baseline profiles to handle more complex baseline changes from sample to sample:

$$A(\tilde{v}) = a + \bar{x}(\tilde{v}) \cdot b + d_1 \tilde{v} + d_2 \tilde{v}^2 + \cdots + d_n \tilde{v}^n + e(\tilde{v}) \quad (3)$$

where $\tilde{v}^j$ are polynomials of the wavenumbers with corresponding constants d_j. Different EMSC models can be derived from Eqn (3). The EMSC model where the polynomial in Eqn (1) is extended up to the quadratic term is often referred to as the basic EMSC model. A further extension of the basic EMSC model above quadratic terms is often denoted polynomial EMSC. The unknown parameters are estimated using an ordinary or weighted least squares estimation, and the spectra are corrected according to Eqn (4):

$$A_{corr}(\tilde{v}) = \frac{A(\tilde{v}) - a - d_1 \tilde{v} - d_2 \tilde{v}^2 - \cdots - d_n \tilde{v}^n}{b} \quad (4)$$

Replicate correction

Constituent or interferent spectra can be introduced as extensions to the EMSC model by so-called orthogonal subspace models (i.e. estimations of constituent or interferent effects based on for instance PCA). A special case of this approach is the so-called EMSC replicate correction,[21] originally developed to limit the effect of systematic inter replicate variation. This is done by first calculating individual EMSC models for each set of replicates, centring the results, collecting them and calculating the first A principal components from these. The loading weights $p_k(\tilde{v})$ are included in a global EMSC model for the uncorrected data:

$$A(\tilde{v}) = a + \bar{x}(\tilde{v}) \cdot b + d_1 \tilde{v} + d_2 \tilde{v}^2 + \cdots + d_n \tilde{v}^n + \sum_{k=1}^{A} g_k \cdot p_k(\tilde{v}) + e(\tilde{v}), \quad (5)$$

where the g_k are fitting parameters associated with the loadings. To avoid over fitting and subsequent reduction of relevant information from the corrected spectra, one should limit the number of subspace component included in the model.

Further extensions of extended multiplicative signal correction

There are also other ways of extending the EMSC model, which can be useful in more or less specialised situations, e.g. models of Mie scattering.[15] All the effects that are subtracted in Eqn (4) can be considered as interferents. Other interferents that one can include in the model are spectra o f compounds that are known to interfere with the measurements, spectra recorded in experiments where conditions that are usually out of control are changed systematically, and so on. To make the least squares estimation stable, it is important that the interferent spectra are as different as possible, preferably orthogonal. This can often be achieved simply by using the difference between the reference and the interferent instead of the interferent directly, but may also warrant compression or a more sophisticated orthogonalisation. Constituent spectra can also be included in the EMSC model for stabilising parameter estimation, but these will not be corrected for. The constituents may need the same orthogonalisations as the interferents.

Shift correction

An extension of the EMSC modelling, which does not extend the EMSC model itself, is a simple shift correction. One can either use whole spectra or one or more regions known to have peaks in fixed positions as basis for the correction. A maximum window of shift is defined before the procedure is started. First, an EMSC model is fitted to the whole data set. A single spectrum is set as the reference, e.g. the spectrum whose corrected version is most similar to the mean spectrum. Then, one goes through every spectrum systematically, applying EMSC correction to one and one uncorrected spectrum using all possible shifts within the window of shifts that was chosen. The shift giving the highest correlation of the current spectrum to the reference spectrum is selected.

Where traditional shift correction is done either before or after pre-processing, this procedure combines the two corrections in one. The effect is that pre-processing is performed on the optimally shifted spectra and vice versa, that optimal shift is found on pre-processed spectra that are comparable. The only type of shift that is handled is the global, linear shift typically resulting from drift in the instrument over time. This combination of optimally shifted reference-based pre-processing and rigidity from the global, linear shift makes the procedure robust.

Literature and software

The EMSC methodology and extensions have been described in the EMSC Tutorial of Afseth and Kohler.[2] Full MATLAB code for the tutorial and a MATLAB graphical user interface for EMSC with several extensions are freely available at http://nofimaspectroscopy.org. The user interface is described in (submit JSS, Liland 2015). Other implementations of EMSC are also available through several commercial software packages.

Partial least squares regression

As an objective criterion for the success of the pre-processing,[4] we will use partial least squares regression (PLSR)[28] combined with cross-validation.[29] PLSR decomposes the calibration data to create a component-based representation of the data similar to PCA.[30] The difference between PCA and PLS is that the former maximises the variation of the predictors per component, while PLS maximises the predictors' covariance to the response per component. PLSR is a powerful and easy-to-use method for compressing highly multicollinear data for regression purposes. Details can be found in the references.

The success of PLSR predictions are measured by the root mean squared error of cross-validation, $\mathrm{RMSECV} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2}$, and prediction, $\mathrm{RMSEP} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}\right)^2}$, where the predicted responses originate from the cross-validation ($\hat{y}_{(i)}$) or prediction ($\hat{y}$), respectively. A low value of either of these measures means that the difference between reference and the predicted values are similar on average.

Results

Milk data

The milk data are trimmed to the region between 3100 and 120 cm^{-1} and shown before pre-processing in Fig. 1 (top, left). It is easy to see the high degree of variation in the baseline offset, but there also seems to be different curvature in the spectra towards the low shift end. Baseline correction using a fifth order polynomial[23] removes the major baseline offset, although some

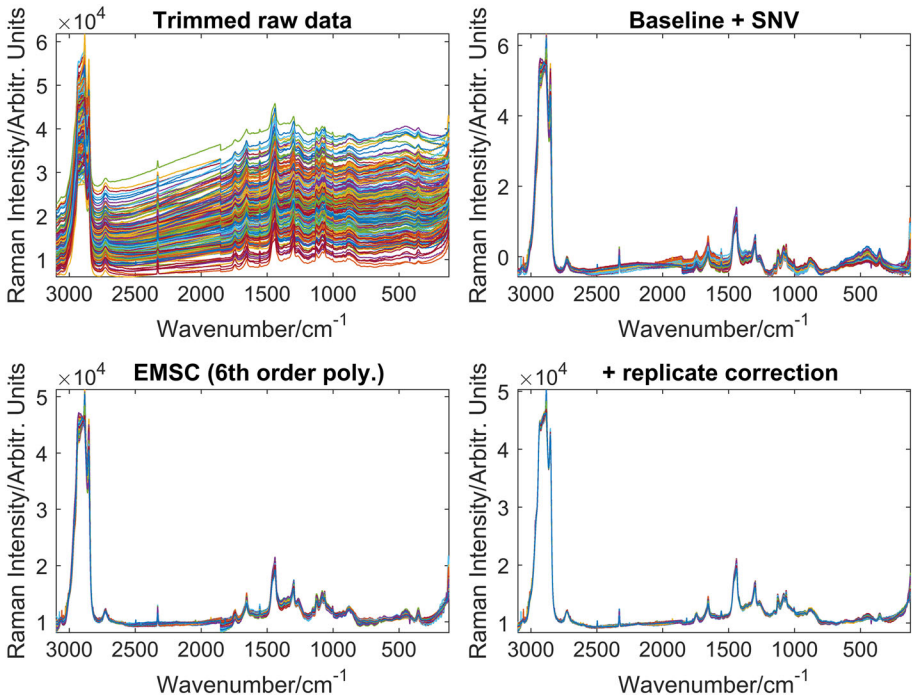

Figure 1. Top: Raw Raman spectra of milk samples after trimming to the region between 3100 and 120 cm^{-1} and spectra corrected by polynomial baselines and standard normal variate (SNV). Bottom: Raman spectra of milk samples corrected by extended multiplicative signal correction (EMSC). The left plot shows EMSC with polynomials up to the sixth order while the right plot shows results when replicate correction has been included in addition.

global intensity variation is still left (results not shown). Adding SNV correction to this improves the intensity variation in the broad peak at above 2750 cm^{-1}, while other parts of the spectra show higher variability in the baselines (Fig. 1; top, right).

Applying MSC to the milk spectra turns some of the spectra upside down, while many others are scaled very differently from the reference, as shown in Fig. S1 (Supporting Information). This effect is caused by the steep slant of the spectra resulting in difficulties for the least square estimation of the parameters for the constant baseline and reference scaling.

The basic EMSC model containing polynomials up to the second degree already shows some promise, as the resulting spectra are on par with baseline correction + SNV, visually. However, the curvature towards the low wavenumber end is not well handled (results not shown). Increasing the polynomial degree to six gives a major improvement to the spectra (Fig. 1; bottom, left). Now the instrument detector shift at around 1850 cm^{-1} is easy to spot. When adding replicate correction to the model (three subspace components), the detector shift is removed, and the spectra cluster more together (Fig. 1; bottom, right). The use of replicate correction is natural here because 12 Raman spectra per reference measurement have been recorded.

While visual inspection of spectra is, to some degree, a subjective approach, measuring the success of the pre-processing through the predictive properties of the spectra can be seen as more objective. In Fig. 2 and Fig. S2 (Supporting Information), prediction of the iodine number and CLA content of the samples have been summarised by RMSECV of (tenfold) cross-validation. A low RMSECV indicates that predictions are close to the reference values of the samples on average. Means across replicates were applied before the PLSR modelling was performed.

The regression results mostly confirm what we observed visually regarding the effect of the different correction methods. All EMSCs lead to low minimum RMSECV values, while raw spectra and spectra that are only baseline corrected lead to higher minimum RMSECV values. The combination of baseline correction and SNV had a minimum RMSECV at the same level as the EMSC did. For this data set, we observe that the more complex the EMSC method is, the more parsimonious the prediction model becomes (fewer components to reach minimum RMSECV).

The milk data were also split into a calibration set and a validation set by holding aside every third sample for validation starting from sample number two (resulting in an even spread in reference values between calibration and validation for both iodine values and CLA values). Models were built on the corrected calibration data, and the reference values of the corrected validation data were predicted using the number of PLSR components resulting in minimum RMSECV in the previous cross-validations. The results showed marginally lower RMSEP values than RMSECV values for the iodine values and very similar values for the CLA values. The same trend in error reduction with higher EMSC complexity was also achieved for the iodine values. The only exception was the data corrected with replicate correction, where the RMSEP was higher than when using the sixth order EMSC model for correction, although still lower than the corresponding RMSECV value (Fig. 2 and Fig. S2). This small decrease in performance improvement may be due to differences in the replicate variation between calibration and validation data or that there was an overfitting on the replicate variation in the calibration data, although these differences are so small that they should not be over interpreted.

The regression coefficients give indications of the underlying chemical features of the respective PLSR models and might serve as an interpretational validation of the approaches. Regression coefficients representing both PLSR models are presented in Fig. S3 (Supporting Information). For the iodine value PLSR model, three major Raman peaks are seen, corresponding well with results reported for other lipid-rich systems.[31] The C=C stretching mode is located at 1660 cm^{-1}, which corresponds to the *cis* C=C configuration. For the CLA PLSR model, however, the C=C stretching mode is located around 1653 cm^{-1}. According to published literature, a shift of the C=C stretching mode to lower wavenumbers can correspond to the presence of conjugated C=C configurations.[32]

A final measure of the success of the pre-processing is the repeatability of the sample replicates. In Fig. 3, this is visualised through the mean replicate variation. For each set of replicates, the standard deviation ($\sigma_{r,p}$) is calculated across replicates (r) for each wavenumber (s). The mean across these standard deviations is a measure of the variability among each set of replicates. Denoting the replicate variation for one replicate set by η_r, we have

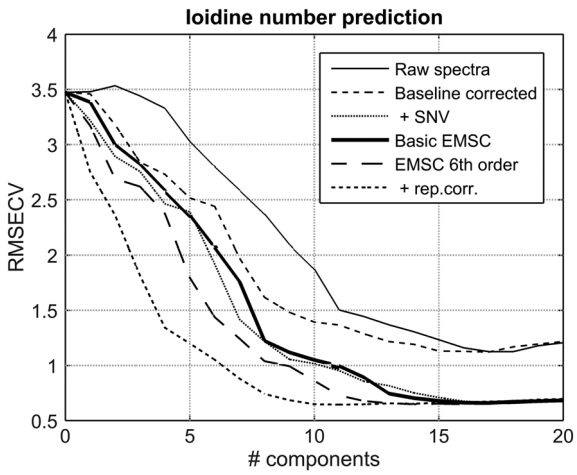

Figure 2. Iodine prediction using Raman spectra of milk corrected by various methods. Minimum RMSECV for the six strategies are 1.126 (17), 1.122 (17), 0.663 (17), 0.660 (17), 0.649 (16) and 0.645 (11), respectively (number of components in parentheses). Corresponding RMSEP for the validation data are 1.026, 0.999, 0.657, 0.640, 0.623 and 0.640. SNV, standard normal variate; EMSC, extended multiplicative signal correction; RMSECV, root mean squared error of cross-validation; RMSEP, root mean squared error of prediction.

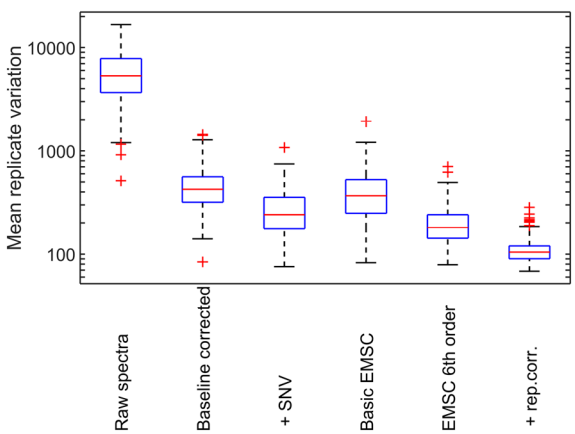

Figure 3. Replicate variation for each method (log scale). Variation is measured as the mean of the standard deviations of each Raman shift for a replicate set.

$\eta_r = 1/p\sum_{s=1}^{p}\sigma_{r,p}$. In the figure, these values have been summarised as boxplots for each pre-processing method.

As the main source of replicate variation is the vertical baseline offset in the spectra, the raw spectra are highly penalised by this repeatability measure, while each baseline correction and EMSC are able to bring replicates closer together. An increased order of the polynomials in EMSC has a positive effect on the repeatability. In addition, the replicate correction improves the repeatability significantly. Because the prediction results showed only minor signs of overfitting with the use of replicate correction, the improvements in replicate variation using replicate correction do not seem to be associated with a loss of fatty acid information in the corrected spectra.

Adipose data

In the spectra from the adipose data, there is no visible fluorescence background, but rather an elevation in the region between 2000 and 1000 cm^{-1} (Fig. 4; top, left). This interfering signal, which most likely is related to an optical effect due to the immersion probe used, seems to be quite consistent in shape, but has varying intensity from sample to sample. In a previous study, this phenomenon was handled by a customised baseline correction, which could yield varying flexibility in different spectral regions.[33] In (Fig. 5; top, right), we have plotted a difference spectrum calculated by subtracting a spectrum with no interfering signal from a spectrum with much interfering signal. These spectra have been chosen to have very similar reference values to avoid confounding with the response in the subsequent predictions.

In Fig. 5, the left, bottom part shows the adipose spectra corrected by SNV. We observe that the interfering signal introduces a baseline shift in the SNV corrected spectra. The right part of the figure is generated by including the difference spectrum as an interferent in an EMSC model with polynomials up to sixth degree. This simple addition to the EMSC model almost completely removes the interfering signal.

In Fig. S4 (Supporting Information), we show the leave-one-out cross-validated prediction errors of four different fat references from the adipose data. It is evident from the curves that including the interferent (difference spectrum) in the EMSC model simplifies the PLSR models. For the iodine number, it also produces the only stable model with a large improvement in prediction error compared with both SNV correction and EMSC without the interferent. Regression coefficients corresponding to the EMSC model with interferent spectrum can be found in Fig. S5 (Supporting Information).

This data set has also been split into calibration data and validation data, again holding aside every third sample from sample two to ensure even spread of the reference values. As the RMSECV values flatten out for most references around the minimum, the optimal number of components is highly affected by random variations. Thus, we have employed the technique of Indahl[34] using a chi-squared distribution test to limit the number of components to the number giving solutions that are not significantly worse than the best with regard to RMSECV. The RMSECV and RMSEP values are mostly similar to each other with a slight over-optimism for the Saturated Fatty Acids (SFA) and a slight under-optimism for the Monounsaturated Fatty Acids (MUFA) cross-validations. In general, the raw data are among the worst performers, baseline correction + SNV and basic EMSC are a bit unstable, while both the EMSC with sixth-degree polynomials and the EMSC with interferent correction are consistently the best or among the best. The exceptional performance of the latter model on the iodine number predictions is confirmed by the validation.

Muscle data

Measurements on beef muscles subjected to brining[27] demonstrate global linear shifts in the Raman spectra. In Fig. 5, the spectra

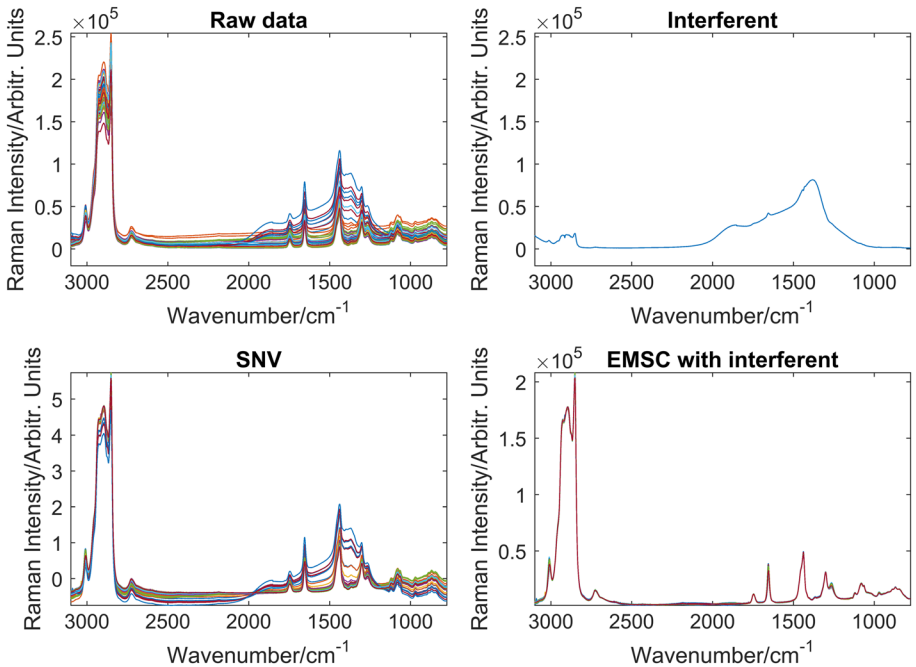

Figure 4. Model-based pre-processing in Raman spectroscopy of biological samples, Kristian Hovde Liland*, Achim Kohler and Nils Kristian Afseth. Model-based pre-processing through various forms of extended multiplicative signal correction is demonstrated on three different Raman data sets. Emphasis is given on visual improvements, improvements in prediction, removal of interferents and reduction of replicate variation. A novel robust global shift correction with optimised reference correction is applied to data with instrumental shift problems.

Journal of RAMAN SPECTROSCOPY

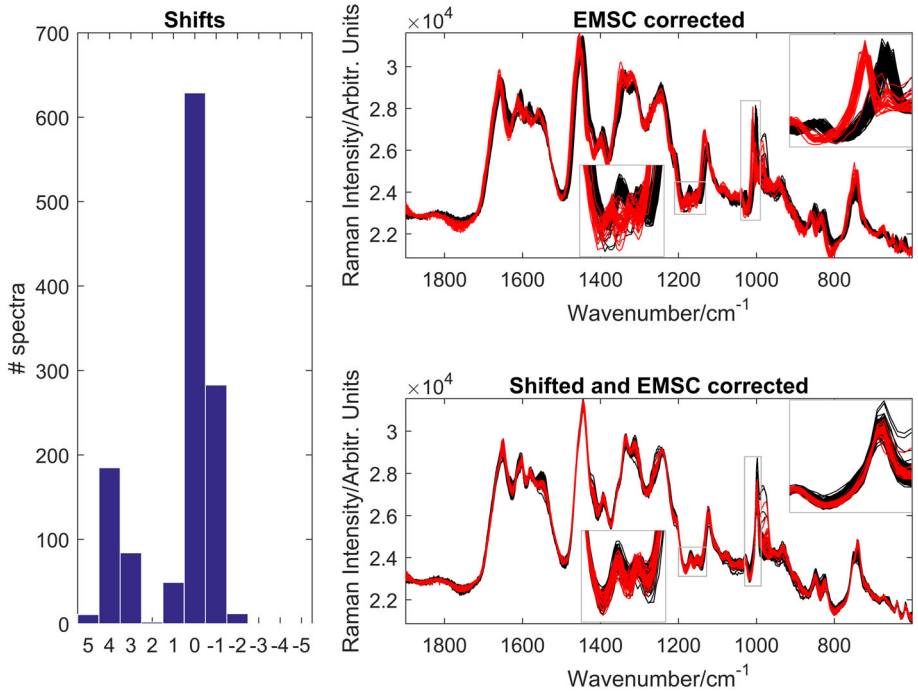

Figure 5. Top: Raw Raman spectra of adipose tissue after trimming to the region between 3100 and 775 cm^{-1} and difference spectrum between spectra affected by interferent and not. Bottom: Raman spectra of adipose tissue corrected by SNV and spectra corrected by extended multiplicative signal correction (EMSC) with polynomials up to the sixth order and an interferent spectrum (difference spectrum from Fig. 4).

have been split into two groups and coloured according to the amount of estimated mean shift. The shift effect is easily spotted by looking at the peak around 1000 cm^{-1}, i.e. the ring breathing mode of phenylalanine, which has been magnified in the upper right corner.

The EMSC included polynomials up to the sixth degree and replicate correction over the nine spectral replicates. After performing shift correction with EMSC, we observe that two rather distinct groups have formed in the histogram: spectra with a shift estimated at around +4 and spectra with a shift estimated around 0. The former group is dominated by brines having salt concentrations of around 1.5%, while the latter group is dominated by brines having salt concentrations of 6–9%. However, because of lack of total randomisation of the measurement order, the measurement day and salt concentrations are partly confounded in the experiment. As shown in Fig. 5, using the shift correcting EMSC approach thus enables Raman data evaluation of linearly shifted Raman spectra to be evaluated based on 'true' chemical variation and not based on physical or instrumental features. This means that the peaks will align nicely with the reference spectrum so that the spectrum intensities are not locally distorted by trying to make the spectra fit to false or absent peaks. This is especially important for the replicate correction that may end up including spectral shift-related information in its correction models if not properly aligned.

Discussion

From the literature and the examples in the Results section, we can summarise that extended multivariate signal correction is well suited for sorting various effects from Raman spectra and cleaning these before visual or analytical use. Based only on the minimum prediction error, there is no difference between using EMSC or applying baseline correction + SNV on the analysed milk data. However, if we take into account either the appearance of the spectra after correction, the parsimony of the models or the mean replicate variation, EMSC seems to be the better choice. SNV is still a valuable method when only mean and scaling show change from observation to observation. In the presence of baseline effects, however, SNV is highly dependent on an effective and objective baseline correction. The major benefit of EMSC in this respect is that it uses a reference spectrum and an overall fit to guide the baseline correction. MSC is by some scientists preferred over SNV because of its model-based approach. However, a steep baseline can lead to sign changes in the corrected spectra if no limitation or compensation is applied.

Because EMSC is model-based, one can also store the parameters of the corrections for further analysis. They can reveal systematic variations in the samples and capture various effects that are interesting in themselves. The amount of shift for each spectrum in the brining data is an example of parameters that can hold valuable information about the samples, but which needs to be corrected for to make the spectra more compatible in the data analysis.

As mentioned in the Introduction section, extensions of the basic EMSC algorithm to correct for varying chemical or physical interferents are advantage of the model-based methods. Such three extensions have been demonstrated here. Firstly, the polynomial was extended to the sixth degree to increase the flexibility of the baseline inherent in Raman spectra. In addition, the reference was chosen to be a spectrum with minimal baseline elevation. The effect was a flatter baseline in the corrected spectra and more parsimonious prediction models. Secondly, a replicate correction was performed to reduce the effect of systematic intra replicate variations in the data. Here, the effect was reduction of noise and removal of a sensor shift effect in addition to improved modelling. Finally, simple shift compensation was built into the EMSC modelling to estimate and correct for horizontal shifts of the spectra from beef muscle subjected to brining. Direct quantification of

constituents is sometimes performed on the Raman spectra by integrating peak volumes or by measuring peak heights or ratios. In such cases, it becomes very important that zero signals are truly zeros. For this, a customised reference can be used, where the reference itself has been baseline corrected.

Interferent spectra can be any type of effect that one wishes to remove from spectra, e.g. something that is interfering with sampling, but which is possible to measure accurately separately under controlled conditions. It can also be something much simpler, like including the shape of a detector shift or baseline distorting probe effect as interferent spectra. The latter was explored for the adipose tissue with great success and very little effort. The possibility of adding constituent spectra was not exploited in this analysis.

The model-based pre-processing methods share some characteristics with regression models. Firstly, the model used on calibration data needs to be stored so that it can be applied to validation data or newly acquired at a later time. Especially, the reference spectrum is important in this respect. Secondly, the model-based approach is prone to overfitting. The main sources of overfitting are the use of too high degrees of the included baseline polynomials or replicate models with too many subspace dimensions. The former can lead to removal of chemical information by the pre-processing, while the latter can lead to modelling of phenomena only found in the calibration data or modelling of chemical information. This means that if the scientist is not positive that new data will follow the pattern of previously corrected data, extra care must be taken when building a complex EMSC model. The EMSC model must be inspected, and possibly a calibration/validation splitting of the data should be performed to assess possible overfitting.

Free implementations of EMSC are available with highly efficient calculations and simple to use interfaces, e.g. the Open EMSC toolbox for MATLAB at http://nofimaspectroscopy.org/sofware-downloads.

Acknowledgements

We would like to thank the Research Council of Norway and the Foundation for Research Levy on Agricultural Products for financial support. We would also like to thank Elisabeth Fjærvoll Olsen for sharing her spectral data with us.

References

[1] N. K. Afseth, V. H. Segtnan, J. P. Wold, *Appl. Spectrosc.* **2006**, *60*, 12.
[2] N. K. Afseth, A. Kohler, *Chemom. Intell. Lab. Syst.* **2012**, 117.
[3] A. E. Kandjani, M. J. Griffin, R. Ramanathan, S. J. Ippolito, S. K. Bhargava, V. Bansal, *J. Raman Spectrosc.* **2013**, *44*, 4.
[4] K. H. Liland, T. Almoy, B.-H. Mevik, *Appl. Spectrosc.* **2010**, *64*, 9.
[5] P. Lasch, *Chemom. Intell. Lab. Syst.* **2012**, *117*, 100.
[6] B. T. Bowie, D. B. Chase, P. R. Griffiths, *Appl. Spectrosc.* **2000**, *54*, 5.
[7] J. R. Beattie, *J. Raman Spectrosc.* **2011**, *42*, 6.
[8] J. R. Beattie, J. J. McGarvey, *J. Raman Spectrosc.* **2013**, *44*, 2.
[9] H. Martens, S. A. Jensen, P.. Geladi, *Multivariate Linearity Transformation for Near-Infrared Reflectance Spectrometry*, (Eds: O. H. J. Christie), Stokkand Forlag, Stavanger, **1983**, pp. 208-234.
[10] H. Martens, E. Stark, *J. Pharm. Biomed. Anal.* **1991**, *9*, 8.
[11] H. C. Bertram, A. Kohler, U. Bocker, R. Ofstad, H. J. Andersen, *J. Agric. Food Chem.* **2006**, *54*, 5.
[12] S. W. Bruun, A. Kohler, I. Adt, G. D. Sockalingum, M. Manfait, H. Martens, *Appl. Spectrosc.* **2006**, *60*, 9.
[13] N. Perisic, N. K. Afseth, R. Ofstad, J. Scheel, A. Kohler, *J. Agric. Food Chem.* **2013**, *61*, 13.
[14] H. Martens, S. W. Bruun, I. Adt, G. D. Sockalingum, A. Kohler, *J. Chemom.* **2006**, *20*, 8–10.
[15] A. Kohler, J. Sule-Suso, G. D. Sockalingum, M. Tobin, F. Bahrami, Y. Yang, J. Pijanka, P. Dumas, M. Cotte, D. G. van Pittius, G. Parkes, H. Martens, *Appl. Spectrosc.* **2008**, *62*, 3.
[16] M. Scholtes-Timmerman, H. Willemse-Erix, T. B. Schut, A. van Belkum, G. Puppels, K. Maquelin, *Analyst* **2009**, *134*, 2.
[17] A. Kohler, N. K. Afseth, H. Martens, *Chemometrics in biospectroscopy*, (Eds: J. Chalmers, E. Li Chan, P. R. Griffiths), John Wiley & Sons Ltd, Chichester, UK, **2010**, pp. 89-106.
[18] J. De Gelder, K. De Gussem, P. Vandenabeele, P. De Vos, L. Moens, *Anal. Chim. Acta* **2007**, *585*, 2.
[19] P. Candeloro, E. Grande, R. Raimondo, D. Di Mascolo, F. Gentile, M. L. Coluccio, G. Perozziello, N. Malara, M. Francardi, E. Di Fabrizio, *Analyst* **2013**, *138*, 24.
[20] H. Martens, J. P. Nielsen, S. B. Engelsen, *Anal. Chem.* **2003**, *75*, 3.
[21] A. Kohler, U. Bocker, J. Warringer, A. Blomberg, S. W. Omholt, E. Stark, H. Martens, *Appl. Spectrosc.* **2009**, *63*, 3.
[22] H. G. Schulze, R. F. B. Turner, *Appl. Spectrosc.* **2015**, *69*, 6.
[23] C. A. Lieber, A. Mahadevan-Jansen, *Appl. Spectrosc.* **2003**, *57*, 11.
[24] A. T. Randby, M. R. Weisbjerg, P. Norgaard, B. Heringstad, *J. Dairy Sci.* **2012**, *95*, 1.
[25] N. K. Afseth, H. Martens, A. Randby, L. Gidskehaug, B. Narum, K. Jorgensen, S. Lien, A. Kohler, *Appl. Spectrosc.* **2010**, *64*, 7.
[26] E. F. Olsen, E. O. Rukke, A. Flatten, T. Isaksson, *Meat Sci.* **2007**, *76*, 4.
[27] N. Perisic, N. K. Afseth, R. Ofstad, S. Hassani, A. Kohler, *Food Chem.* **2013**, *138*, 1.
[28] S. Wold, H. Martens, H. Wold, *Lecture Notes in Math.* **1983**, *973*, 286.
[29] M. Stone, J. R. Stat, *Soc. Ser. B Stat. Methodol.* **1974**, *36*, 2.
[30] K. Pearson, *Philos. Mag.* **1901**, *2*, 11.
[31] N. K. Afseth, J. P. Wold, V. H. Segtnan, *Anal. Chim. Acta* **2006**, *572*, 1.
[32] I. Stefanov, V. Baeten, O. Abbas, E. Colman, B. Vlaeminck, B. De Baets, V. Fievez, *J. Agric. Food Chem.* **2011**, *59*, 24.
[33] K. H. Liland, E.-O. Rukke, E. F. Olsen, T. Isaksson, *Chemom. Intell. Lab. Syst.* **2011**, *109*, 1.
[34] U. Indahl, *J. Chemom.* **2005**, *19*, 1.

Supporting information

Additional supporting information may be found in the online version of this article at publisher's web site.