Pre-processing of spectral data in the extended multiplicative signal correction framework using multiple reference spectra

September 28, 2018

Abstract

Extended Multiplicative Signal Correction (EMSC) is a widely used framework for pre-processing spectral data. In the EMSC framework spectra are scaled according to a given reference spectrum. Spectra that are far from collinear with the selected reference spectrum may not be scaled appropriately. An extension of the EMSC framework that allows for the incorporation of multiple reference spectra in the EMSC model is proposed to remedy this issue. Useful candidate reference spectra can be obtained from the dominant right singular vectors associated with the matrix of spectra, but any desired reference spectra can be used. As a part of this extension we propose to change the basis used in the EMSC pre-processing to an orthonormal basis. Using an orthonormal basis will remove confounding issues between the basis vectors and make the obtained EMSC model simpler to interpret. We discuss the proposed modification theoretically and demonstrate its use with two data sets of Raman spectra and modelling with Partial Least Squares regression and Tikhonov Regularization. The data sets used are Raman spectra of oil samples from salmon with iodine value as the response, and Raman spectra of an emulsion of water, whey protein, and different oils with polyunsaturated fatty acids as response (both as percentage of total fat content and total weight).

Keywords: Extended Multiplicative Signal Correction (EMSC); Raman spectroscopy; Pre-processing; Modeling

1 Introduction

Since raw spectral data often contain unwanted artefacts and noise that make modelling and interpretation difficult, some kind of pre-processing is often required.[1–4] The goal of pre-processing spectral data is to transform the raw data into a form that is more suitable for modelling or interpretation. A vast amount of pre-processing methods for spectral data are available. The most widely used pre-processing methods include the standard normal variate (SNV),[5] the Savitzky-Golay filter,[6] various baseline correction algorithms,[7] and other methods.[8]

In the present work we consider the Extended Multiplicative Signal Correction (EMSC),[3] which is a model-based pre-processing framework that corrects for both unwanted additive and multiplicative effects in data.[1] The EMSC is flexible in the sense that it is possible to include a priori knowledge about chemical and non-chemical patterns in the pre-processing model to improve the data quality.[9]

The additive corrections are obtained by orthogonalizing the spectra with respect to the directions representing irrelevant additive trends in the data. The multiplicative corrections are based on a chosen reference spectrum, and each original spectrum is appropriately scaled so that it can be expressed as a sum of the reference spectrum and a residual part representing the spectral information of actual interest.[1] Such scaling usually works quite well for most of the spectra in a data set, but particular spectra that are far from collinear with the reference spectrum may not be scaled appropriately.[10]

In the present work we propose an extension of the EMSC framework that allows for the inclusion of multiple reference spectra to estimate scaling coefficients for the spectra to be corrected. The proposed extension is particularly useful when dealing with datasets containing one or several outlier spectra. By including additional reference spectra that better accounts for the chemical profiles of the outlier spectra, the pre-processing step may obtain more useful estimates of the EMSC scaling coefficients.

The structure of the present work is as follows: First we review the traditional EMSC framework for pre-processing of spectral data. Then we motivate and discuss how multiple reference spectra can be incorporated in a useful extension of the EMSC framework. Finally we demonstrate the suggested extension for two applications with data sets of Raman spectra.

2 Review of EMSC pre-processing

When modelling by the traditional EMSC pre-processing framework the spectra are scaled according to a pre-specified reference spectrum, and irrelevant polynomial trends are subtracted from the data.[1] In the following we assume that $\boldsymbol{X}$ is an $n \times p$ data matrix with n samples and p predictor variables, $\boldsymbol{r}$ is the chosen reference spectrum (typically the mean spectrum[1,2]) and d is the degree of the polynomial trends to be corrected for. The vectors spanning the subspace of the adverse polynomial trends are denoted by $\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d$. In the traditional EMSC framework a spectrum $\boldsymbol{x}$ is projected onto the subspace spanned by the vectors in the basis

$$B_{EMSC} = \{\boldsymbol{r}, \boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d\}. \tag{1}$$

Note that the exact choice of basis vectors in (1) is unfortunately not specified when the EMSC framework is described, and in practice it has been most common to use a basis that is not orthogonal (the choice of basis will be discussed in more detail later). The associated representation of a spectrum $\boldsymbol{x}$ in B_{EMSC} is:

$$\boldsymbol{x} = b\boldsymbol{r} + \sum_{i=0}^{d} (c_i \boldsymbol{v}_i) + \boldsymbol{e}, \tag{2}$$

where the scalars are obtained by least squares regression and $\boldsymbol{e}$ is the residual spectrum orthogonal to the subspace spanned by B_{EMSC}. The notation $\boldsymbol{e}$ will be used regardless of which EMSC model is applied later in this article. The EMSC corrected spectrum is defined as:

$$\boldsymbol{x}_{cor} = \frac{\boldsymbol{x} - \sum_{i=0}^{d} (c_i \boldsymbol{v}_i)}{b} = \boldsymbol{r} + \frac{1}{b}\boldsymbol{e}. \tag{3}$$

The purpose of the polynomial trends in B_{EMSC} is to model and subtract the expected effects of additive noise, whereas the b-coefficient is used to obtain an appropriate scaling of the residual $\boldsymbol{e}$ to obtain the corrected spectrum $\boldsymbol{x}_{cor}$. The EMSC model can be justified from the Beer-Lambert law, exploiting that chemical spectra are basically non-negative linear combinations of pure component spectra (including interferents) for vibrational spectroscopy techniques.[1] The special case when the polynomial degree is 0, so that only constant trends are corrected, is referred to as the Multiplicative Scatter Correction (MSC).[11] The EMSC is thus a direct extension of theMSC.

Several extensions of the traditional EMSC model have been proposed in the literature. If any known interferents also are present, these can be included to extend the basis B_{EMSC} and handled in the same way as the polynomial trends.[1,2] In applications including replicated measurements of the spectra it is sometimes useful to include additional terms representing inter-replicate variance.[12] The EMSC model has also been extended to correct for the so-called Mie-scattering effects.[9]

Suppose we have n_{intf} interferents, and let $\boldsymbol{w}_i$ denote the i-th interferent. To incorporate the interferents in the model we extend the basis given in (1) to include the vectors representing the interferents. This results in the following extended set of basis vectors:

$$B_{EMSC} \cup \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{n_{intf}}\}. \tag{4}$$

The correction of a spectrum $\boldsymbol{x}$ is obtained by subtracting its projection onto the subspace spanned by the interferents in (4) and the following scaling:

$$\boldsymbol{x}_{cor} = \frac{\boldsymbol{x} - \sum_{i=0}^{d}(c_i\boldsymbol{v}_i) - \sum_{i=1}^{n_{intf}}(d_i\boldsymbol{w}_i)}{b} = \boldsymbol{r} + \frac{1}{b}\boldsymbol{e}. \tag{5}$$

In the following we will use (4) and (5) as our starting point. Since the spectra corrected with EMSC are written as deviations from the reference spectrum, the corrected spectra will typically be quite similar to the reference spectrum. This means that any unwanted artefact in the reference spectrum might also be present in the corrected spectra. Some examples of such effects could be fluorescence in Raman spectroscopy,[1] and Mie Scattering in FTIR[9]. These types of artefacts are usually not a problem for the predictive modelling because the corrected spectra will not vary in the direction spanned by the reference spectrum.

3 EMSC pre-processing with multiple reference spectra

The purpose of the reference spectrum in the EMSC pre-processing is to facilitate the estimation of multiplicative effects for transforming the measured spectra to a common scale. It is known that the MSC can accentuate outliers when the outliers and the selected reference spectrum are poorly correlated.[10] Because the EMSC employs the same scaling strategy as the MSC, it can be expected that the EMSC can also accentuate outliers. The most extreme case would be a spectrum that is orthogonal to the reference spectrum, in which case the reference spectrum would give no indication of how to scale the spectrum. This scaling problem can be alleviated by introducing multiple reference spectra for estimating the scaling coefficients.

The practical use of this idea requires (i) a strategy for deriving more than one reference spectrum, and (ii) a generalization of the EMSC-correction given in (5) to allow for multiple reference spectra. To obtain multiple reference spectra we propose considering the most dominant right singular vectors from the (reduced) singular value decomposition (SVD) of the matrix of the measured spectra. The right singular vectors can be viewed as an ordered list of orthogonal directions in the sample space sorted by the magnitude of joint signal strength in each direction. The ordering emphasizes the first few dominant right singular vectors as natural candidate reference spectra because they represent the part of the information that is most common across the entire collection of measured spectra. If these vectors describe signals in the data having a chemical origin, it can be expected that the measured spectra will appear similar in the subspace spanned by these vectors. As the right singular vectors are only uniquely defined up to sign, it may be required to change the signs for visualization purposes. A practical method for checking this is to calculate the correlation between the mean spectrum and the first right singular vector and change signs if the correlation is negative. Note that the first right singular vector is often highly correlated to the mean spectrum for spectral data. Therefore using the first right singular vector as a reference spectrum will often give a pre-processing result that is quite similar to the result obtained by using the mean spectrum as the reference.

In the traditional EMSC pre-processing a non-orthogonal basis is typically used, and the correction of additive trends in the scaling is done implicitly when projecting a spectrum onto the subspace spanned by the basis in (4). This basis is not appropriate when employing multiple reference spectra because of the interactions between the reference spectra and the polynomial trends (and possibly the other interferents). However, the problem is easily dealt with by employing an orthonormal basis eliminating any ambiguities in the regression coefficients (and the associated EMSC model interpretations) resulting from some particular choice of non-orthogonal basis.

A good and practical procedure for obtaining an orthonormal basis is to collect the EMSC basis vectors as columns in a matrix and calculate its QR-factorisation. We recommend the columns in this matrix to be ordered as follows: Start with the polynomial trends followed by the interferents (if any), and finally include the reference spectra. The reason for suggesting this ordering is that it makes more sense to first eliminate the irrelevant effects of the polynomial trends and the interferents from the reference spectra, rather than the other way around which would result in using reference spectra being contaminated by both additive (polynomial) effects and the other interferents that one wants to avoid. To obtain the i-th polynomial vector representing a polynomial trend of degree $i-1$ we sample the function x^{i-1} uniformly over p points (the number of features) in the interval $[-1, 1]$. The QR-factorisation used to obtain an orthonormal basis will then produce the associated Legendre polynomials.[13] To distinguish between the traditional non-orthogonal EMSC basis and the orthonormal basis introduced here, the superscript ^{o} is used to denote spectra that are part of an orthonormal basis that has been obtained using a QR-factorisation as described above. Let n_{ref} be the total number of reference spectra (identified by the SVD or some other insights), and denote the i-th reference spectrum by $\boldsymbol{r}_i$. For the orthonormal basis of the suggested modified EMSC-framework we use the notation:

$$B^o_{EMSC} = \{\boldsymbol{v}^o_0, \boldsymbol{v}^o_1, \boldsymbol{v}^o_2, \ldots, \boldsymbol{v}^o_d, \boldsymbol{w}^o_1, \boldsymbol{w}^o_2, \ldots, \boldsymbol{w}^o_{n_{intf}}, \boldsymbol{r}^o_1, \boldsymbol{r}^o_2, \ldots, \boldsymbol{r}^o_{n_{ref}}\}. \tag{6}$$

Because the basis is constructed to be orthonormal, the coefficients (the α_i's, the δ_i's and the γ_i's) for the projection of a particular spectrum onto the subspace spanned by B^o_{EMSC} can be calculated directly by taking the inner products between each of the basis vectors and the spectrum, i.e. $\alpha_i = (\boldsymbol{v}^o_i)^t\boldsymbol{x}$, $\delta_j = (\boldsymbol{w}^o_j)^t\boldsymbol{x}$ and $\gamma_k = (\boldsymbol{r}^o_k)^t\boldsymbol{x}$. Expressing a spectrum $\boldsymbol{x}$ with respect to this basis therefore yields:

$$\boldsymbol{x} = \sum_{i=0}^{d} \alpha_i \boldsymbol{v}^o_i + \sum_{j=1}^{n_{intF}} \delta_j \boldsymbol{w}^o_j + \sum_{k=1}^{n_{ref}} \gamma_k \boldsymbol{r}^o_k + \boldsymbol{e}, \tag{7}$$

where $\boldsymbol{e}$ is the resulting residual not accounted for by B^o_{EMSC}.

The corrected version of $\boldsymbol{x}$ is obtained by subtracting its projection onto the subspace spanned by the polynomial trends (the $\boldsymbol{v}^o_i$'s) and the interferents (the $\boldsymbol{w}^o_j$'s), and scaling by the inverse of the norm of its projection onto the subspace spanned by the reference spectra (the $\boldsymbol{r}^o_k$'s), i.e:

$$\begin{aligned} \boldsymbol{x}_{cor} &= \frac{1}{\sqrt{\sum_{k=1}^{n_{ref}} \gamma_i^2}} \cdot \left(\boldsymbol{x} - \sum_{i=0}^{d} \alpha_i \boldsymbol{v}^o_i - \sum_{j=1}^{n_{intf}} \delta_i \boldsymbol{w}^o_i \right) \\ &= \frac{1}{\sqrt{\sum_{k=1}^{n_{ref}} \gamma_i^2}} \cdot \left(\sum_{i=1}^{n_{ref}} \gamma_i \boldsymbol{r}^o_i + \boldsymbol{e} \right) \\ &= \boldsymbol{r}_{\boldsymbol{x}} + \frac{1}{\sqrt{\sum_{k=1}^{n_{ref}} \gamma_i^2}} \cdot \boldsymbol{e}, \end{aligned} \tag{8}$$

where the reference combination $\boldsymbol{r}_{\boldsymbol{x}} = \frac{1}{\sqrt{\sum_{k=1}^{n_{ref}} \gamma_i^2}} \cdot \sum_{i=1}^{n_{ref}} \gamma_i \boldsymbol{r}^o_i$ depends on the original spectrum $\boldsymbol{x}$. Note that in the special case with $n_{ref} = 1$ (one reference spectrum $\boldsymbol{r}^o$) the above correction simplifies to

$$\boldsymbol{x}_{cor} = \boldsymbol{r}^o + \frac{1}{|\gamma_1|} \cdot \boldsymbol{e}, \tag{9}$$

where the reference $\boldsymbol{r}^o$ is common for all the spectra subject to correction. The residual term in (9) will be similar but not identical to the residual obtained from standard EMSC pre-processing, as the reference spectrum in (9) is initially corrected for the polynomial trends and interferents.

Note that for the traditional EMSC pre-processing with a single reference spectrum there is no variation across the samples in the subspace spanned by the reference spectrum. The regression coefficients derived in the the subsequent regression modelling can therefore be chosen orthogonal to $\boldsymbol{r}^o$. When including multiple reference spectra, equation (8) implies that this is no longer the case, and one should expect the regression coefficients to be non-orthogonal to the $\boldsymbol{r}_{\boldsymbol{x}}$'s. More specifically, suppose we have some regression coefficients $\boldsymbol{\beta}$ (obtained by Partial least squares regression[14] regression or otherwise). The prediction based on the corrected spectrum $\boldsymbol{x}_{cor}$ is then given by

$$\boldsymbol{x}_{cor}\boldsymbol{\beta} = \left(\boldsymbol{r}_{\boldsymbol{x}} + \frac{1}{\sqrt{\sum_{k=1}^{n_{ref}} \gamma_i^2}} \cdot \boldsymbol{e} \right) \boldsymbol{\beta} = \boldsymbol{r}_{\boldsymbol{x}}\boldsymbol{\beta} + \frac{1}{\sqrt{\sum_{k=1}^{n_{ref}} \gamma_i^2}} \cdot \boldsymbol{e}\boldsymbol{\beta}. \tag{10}$$

The vectors $\boldsymbol{r}_{\boldsymbol{x}}$ can therefore be viewed as correctives term for the spectra.

It should be noted that if a spectrum has a very high correlation with one of the reference spectra provided in B^o_{EMSC}, then it must necessarily be nearly orthogonal to the others. Thus, the projection of the spectrum onto $(n_{ref} - 1)$ of the reference spectra will be close to zero, and just one reference spectrum will have a noticeable impact on the pre-processing. This property makes the use of multiple reference spectra particularly attractive for data sets containing a low number of spectra that are very different from the primary desired reference spectrum, as only these spectra will be noticeably affected by the inclusion of additional reference spectra.

Any choice of multiple reference spectra requires certain knowledge about their representation of particular chemical information in the data. If some unwanted artefact, not picked up by the polynomial trends, is present in a candidate reference spectrum, it should either be included as an interferent in B^o_{EMSC} or ignored completely. The practical estimation of interferents can be handled in several ways. One possibility is to use a strategy based on difference spectra.[1,2] Alternatively, if there are no difference spectra that appropriately model the unwanted trends, then the interferent can be modelled from the data. This can for example be done using the approach proposed by Beattie[15], which is mentioned below.

Pre-processing approaches based on the SVD are well known from the literature. Beattie has used a particular SVD loading for collagen and heme was used for scaling spectra.[15] This approach is similar to our scaling using a single reference spectrum obtained from the SVD. Beattie also suggested using selected SVD loadings to estimate non-Raman background effects.[15,16] This was done by utilizing the fact that Raman peaks typically are quite narrow so that high bandwidth features in the right singular vectors indicate non-Raman phenomena. The non-Raman phenomena can then be estimated from the right singular vectors.[15] To correct the spectra, these estimates can be scaled and subtracted from the spectra, or the approximations can be added as interferents to an EMSC model. This approach is general and can be very useful for obtaining estimates of unwanted additive trends in candidate reference spectra not accounted for by the polynomial trends.

Because of the choice of an orthonormal basis in (6) the spectra pre-processed according to (8) are not directly suitable for visualization, peak quantification or peak ratio calculations without some modifications. This is because an ideal reference spectrum will not be orthogonal to all polynomial trends. But for the pre-processing it is computationally advantageous to use an orthogonal basis. For plotting one should therefore consider adding back the projection of the first reference spectrum onto the polynomial trends, which will have no effect on modelling.

From a mathematical point of view, the polynomial terms in the EMSC basis will eliminate any baseline effect for modelling purposes, but if a baseline is present in the first reference spectrum then it will in general also be present in the corrected spectra. Such a baseline can be removed by for example finding a baseline correction for the first reference spectrum and subtracting this baseline from all the spectra.

Prototype MATLAB code implementing the suggested modification of the EMSC pre-processing is included in the Appendix.

4 Examples

In this section we will compare using the traditional EMSC pre-processing method using the mean spectrum as reference to the proposed modification of the EMSC framework using the first $1 - 3$ right singular vectors as reference spectra. Correction of polynomial trends up to the 6th degree is included for all the pre-processing alternatives. No interferents will be added to the pre-processing models. The traditional EMSC framework using the mean spectrum as the reference spectrum will be referred to as simply (standard) EMSC pre-processing. For the modified EMSC framework we will use parentheses to denote the number of right singular vectors used as reference spectra, so that, for example, EMSC(3) refers to the modified EMSC framework using the first three right singular vectors as reference spectra. We consider modelling with Partial Least Squares (PLS) regression[14] and Tikhonov Regularization (TR)[17]. The following two data sets will be considered:

1. *Fish oil data.*[18] This is a data set consisting of Raman spectra measured on oil samples from salmon. There are $n = 45$ measured samples, and the spectra are truncated to the range $790cm^{-1} - 3052cm^{-1}$. This truncation has been used before when the data set has

been analysed.[7] After truncation there are $p = 2263$ wavenumbers. The response is the associated measured iodine values. The raw spectra are shown in Fig. 1.

2. *Emulsion data.*[19] This data set consists of Raman spectra measured on an emulsion of water, whey protein, and different oils. The oil types used were refined olive oil, refined coconut oil, soy oil, cod oil with omega 3 fatty acids, and salmon oil. A mixture design was used to create the samples.[19] The responses are Polyunsaturated Fatty Acids (PUFA) quantified as percentage of total weight, and PUFA as percentage of total fat content. The spectra are truncated to the wavenumbers $675cm^{-1} - 1770cm^{-1}$. This truncation has been used before when the data set has been analysed.[19,20] There are a total of $n = 69$ measured samples in the data set, and after truncation there are $p = 1096$ wavenumbers. The raw truncated spectra are shown in Fig. 2.

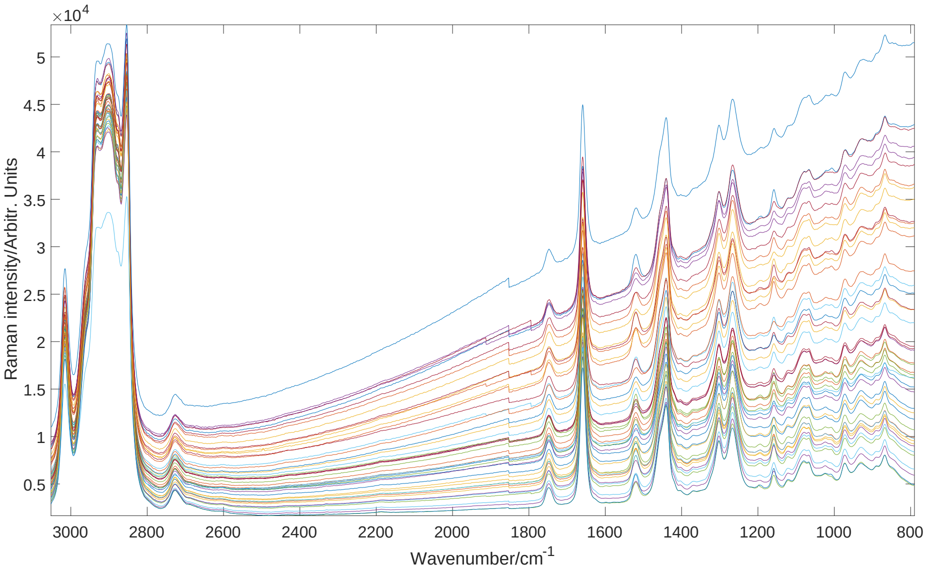

Figure 1: *Fish oil data: Raw Raman spectra.*

For modelling, the following procedure was used: A nested cross-validation strategy was employed to separate pre-processing and parameter optimization from model validation. The outer validation loop was a repeated two-fold (50:50) shuffle-split, while the inner optimisation loop was a Leave-one-out Cross-Validation (LooCV). For each outer split, the first half of the samples were used to create pre-processing models and subsequently estimate model parameters (using LooCV) for TR and PLS on the pre-processed data. The second half of the outer split was pre-processed correspondingly and its response values predicted using optimal parameter values from the first half. For PLS up to 15 components were considered, and the number of components minimising the root mean squared error of cross-validation (RMSECV) was selected. For TR L_2 regularization as well as discrete first and second derivative regularization were used.[21] 1000 values of the regularization parameter were selected uniformly on a log scale, and the parameter value minimising the RMSECV was selected. Note that there is some data leakage for the LooCV in the inner loop as the data was pre-processed based on all the training samples. This may have caused a small bias in the model selection, but not in the prediction as an independent test set was used for model evaluation. An outer shuffle-split was repeated 500 times, and in every iteration a new random split of the data was created. The average root mean squared errors of prediction (RMSEP) over these 500 iterations are reported in Table 1 for the fish oil data, and Table 2 and Table 3 for the emulsion data.

From Fig. 1 we see that most samples of the the fish oil data appear to be very similar. Although the intensity of the fluorescence background varies between samples, the relative sizes of the different peaks appear similar for all samples. The fluorescence background will be removed when the spectra are corrected for polynomial trends, so for this data set we can expect one

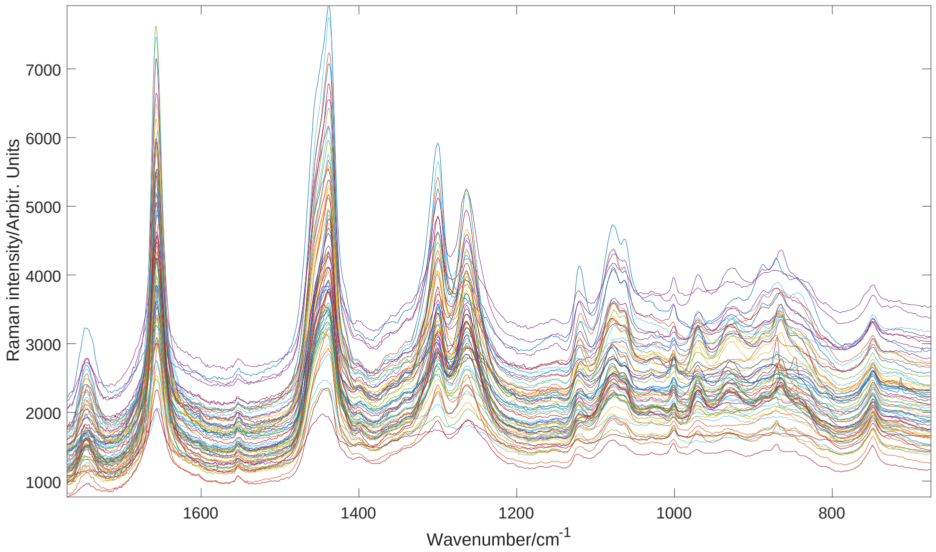

Figure 2: *Emulsion data: Raw Raman spectra. The spectra have been truncated to the range* $675cm^{-1} - 1770cm^{-1}$.

reference spectrum to be sufficient to obtain an appropriate scaling. Inspecting the first three right singular vectors of the fish oil data plotted in Fig. 3 we see that the differences between the right singular vectors can be attributed mostly to the baseline in the data. After removing the projection onto the polynomial trends from the data and the first right singular vector, it can be verified that the maximum angle between a sample and the first right singular vector is 1.6° (alternatively, the lowest correlation between a sample and the first right singular vector is 0.9996). If the first right singular vector is used as a reference spectrum then the spectra will necessarily be nearly orthogonal to any other reference spectrum. Thus for the fish oil data it is sufficient to use a single reference spectrum. This is also supported by Table 1, from which it is clear that all the different pre-processing alternatives give roughly the same prediction errors for the subsequent regression modelling.

For the emulsion data the situation is different. In this dataset there is much more variation between the spectra, and not all the spectra are that highly correlated with the first right singular vector if we compare with the fish oil data. After correcting for polynomial trends, the angle between the first right singular vector and more than 50% of the samples is larger than 10° (corresponding to a correlation lower than 0.9848). For 6 of the samples the angle between the sample and the first right singular vector is between $20° - 35°$ (corresponding to correlations in the range $0.8192 - 0.9393$). In Fig. 4 the first three right singular vectors of the emulsion data are plotted. Unlike the fish oil data, the differences between the right singular vectors cannot be attributed to any baseline or unwanted additive effect. The reference spectra do not appear to contain any unwanted effect that is not accounted for by the polynomial trends, making them appropriate reference spectra candidates.

The pre-processed emulsion spectra are plotted in Fig. 5 and Fig. S2 (Supporting Information). In Fig. 5 there is no apparent visual difference between the two pre-processing alternatives, except for the scale difference between the standard EMSC pre-processed spectra and the modified EMSC pre-processed spectra. The similarities between the standard EMSC and EMSC(1) is supported by Fig. S1 (Supporting Information), from which it is clear that the mean spectrum and the first right singular vector are very similar. Since the two spectra are that similar, we expect the standard EMSC and EMSC(1) pre-processed spectra to be highly similar as well. The scale difference is irrelevant for the subsequent regression modelling as it will be accounted for by the regression coefficients. When including 2 and 3 reference spectra, we can see from Fig. S2 (Supporting Information) that this does not result in a huge visual impact on the spectra, with the notable

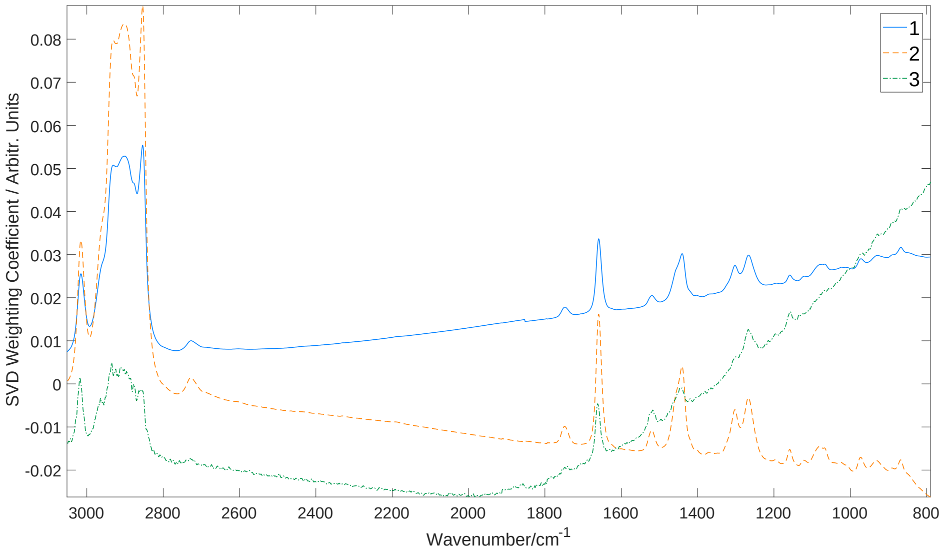

Figure 3: *Fish oil data: The first three right singular vectors.*

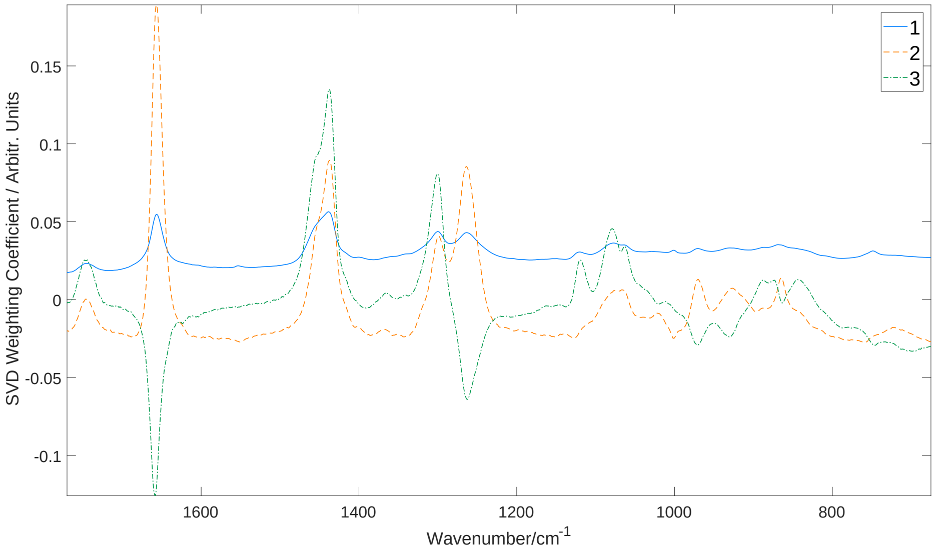

Figure 4: *Emulsion data: The first three right singular vectors.*

exception of one spectrum (see in particular the peak at about $1445 cm^{-1}$).

From Table 2 it follows that for the response of fatty acids measured as the % of total weight, modelling based on the raw data gives the best prediction results, and the differences between the other pre-processing alternatives are relatively small. In Table 3 the situation is changed, and regression models based on the raw data are the poorest by a huge margin. From both Tables the RMSEP obtained using standard EMSC pre-processing is approximately the same as the RMSEP obtained from the EMSC(1) pre-processed data. In Table 3 the RMSEP decreases when the number of reference spectra is increased. The best prediction results are obtained when using the first three right singular vectors as reference spectra. In Fig. 6 and Fig. 7 we plot RMSECV and RMSEP as a function of the model selection parameter for TR and PLS for the response considered in Table 3 and one particular split of the data into a training set and a test set. The RMSECV and

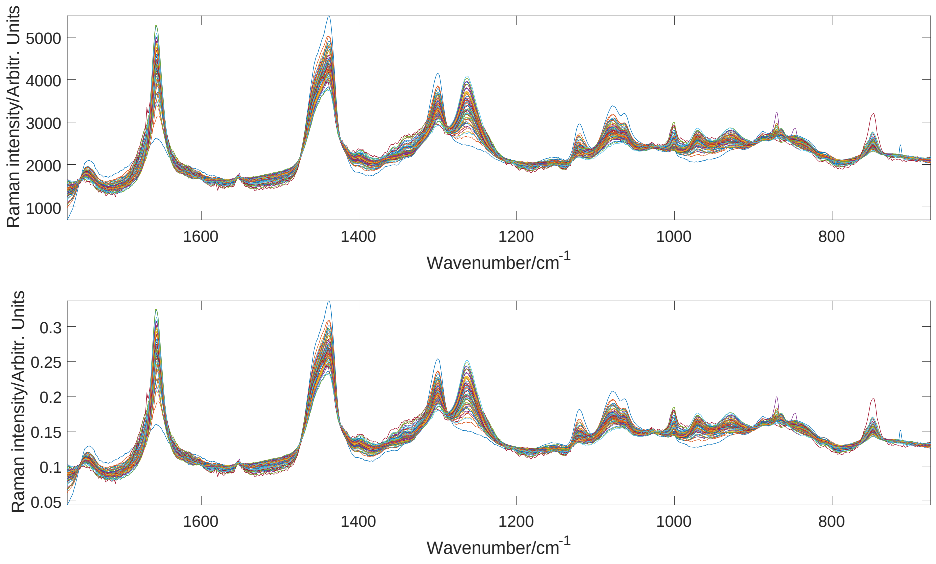

Figure 5: *Emulsion data: Pre-processed Raman spectra using different pre-processing methods. Top: standard EMSC. Bottom: EMSC(1).*

RMSEP curves are very similar, and we see that increasing the number of reference spectra seems to increase the prediction performance independent of the choice of the TR model parameter or number of PLS components.

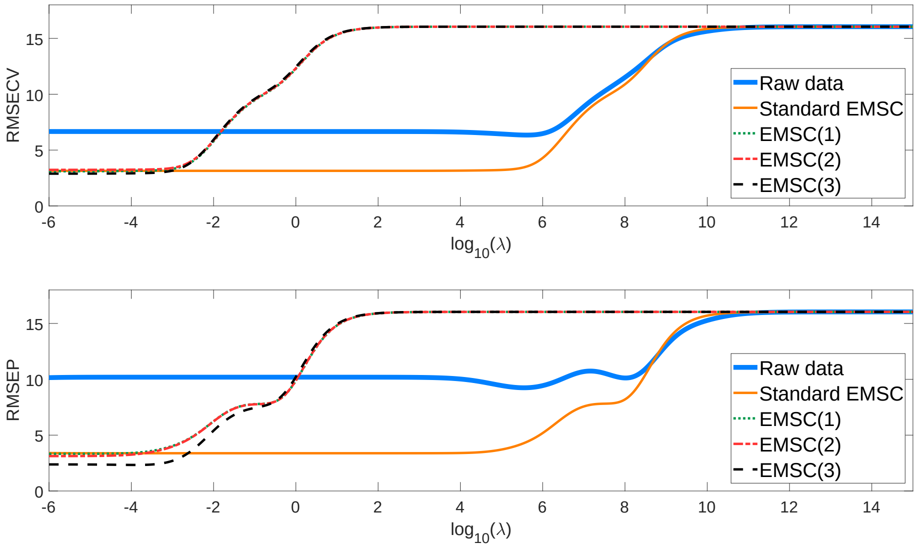

Figure 6: *Emulsion data: TR modelling (L_2 regularization) for the response PUFA as % of total fat content for a particular split of the data. Top: RMSECV. Bottom: RMSEP. In the top plot we see that the RMSECV curves for the modified EMSC pre-processing are overlapping. In the bottom plot we see that the RMSEP curves for the modified EMSC using* 1 *and* 2 *reference spectra are overlapping.*

The prediction errors for the response PUFA as % of total fatty acids were inspected for every sample to study the differences in prediction between the different pre-processing methods in more detail. Most samples obtain a lower prediction error when using 3 reference spectra compared to

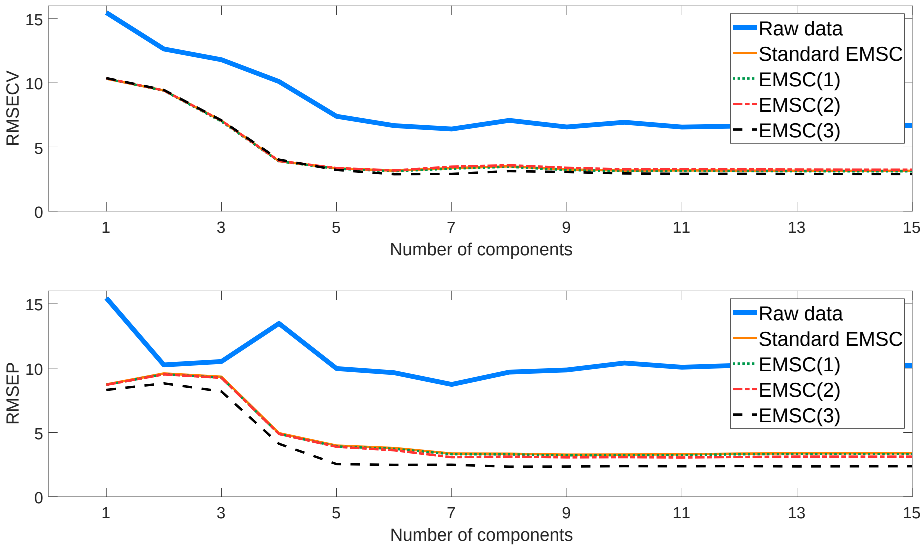

Figure 7: *Emulsion data: PLS modelling for the response PUFA as % of total fat content for a particular split of the data. Top: RMSECV. Bottom: RMSEP. In the top plot the RMSECV curves for all pre-processing alternatives are overlapping. In the bottom plot the RMSEP curves are overlapping for all pre-processing alternatives except for EMSC(3) pre-processing.*

using 1 reference spectrum, but just a few of the samples are responsible for the larger part of the difference in prediction. The three samples most poorly predicted when using only one reference spectrum are plotted in Fig. 8 together with the mean spectrum. We observe that there are obvious differences between at least two of these spectra and the mean spectrum, confirming that the mean spectrum does not work as a useful reference spectrum for all the samples. By including additional reference spectra in the pre-processing, much better scaling estimates are obtained for these spectra.

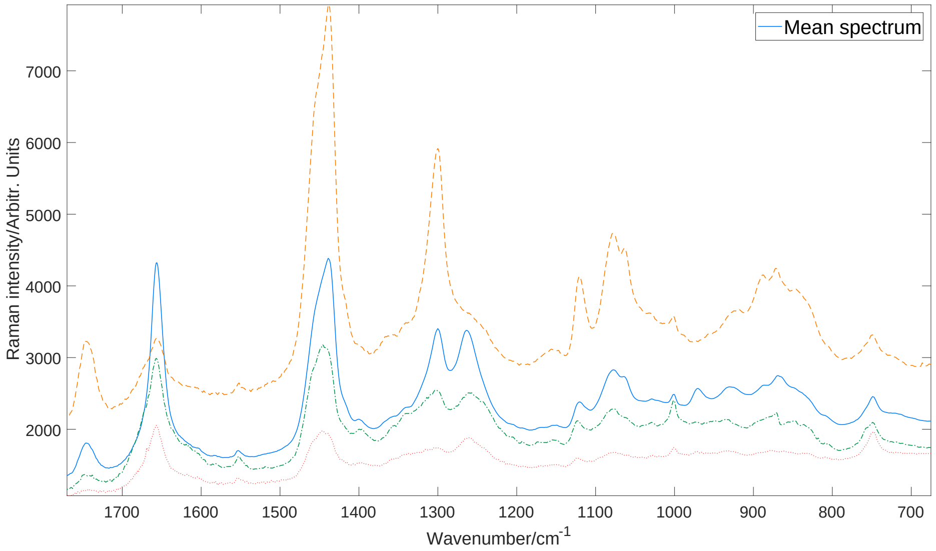

Figure 8: *Emulsion data: Mean spectrum together with the three spectra with worst cross-validated prediction errors when using standard EMSC pre-processing.*

5 Conclusions

The traditional EMSC framework is very flexible, and it is simple to extend the basic correction model to account for additional unwanted additive effects in the data. In the present work we have proposed how the framework can be extended further when it is appropriate to utilise multiple reference spectra to obtain proper scaling coefficients. When using multiple reference spectra it is necessary to use an orthogonal basis (consisting of polynomials, interferent spectra, and reference spectra) in the pre-processing because of the interactions between the different basis vectors. The use of an orthogonal basis is also advantageous because it eliminates any possible confounding between the different basis vectors. For the fish oil data only one reference spectrum was required to obtain a satisfactory pre-processing, but we observed that the inclusion of additional reference spectra did not cause the subsequent regression models to be poorer. For the emulsion data there were some spectra that were very different from the first (traditional) reference spectrum, and pre-processing the data with multiple reference spectra caused the subsequent regression model to predict considerably better for one of the responses. Considering the first right singular vectors of the uncorrected spectra as candidate reference spectra is often a sensible alternative as these vectors describe the most dominant directions in the data. The candidate reference spectra should be inspected visually to make sure they describe relevant chemical variation, rather than interferents or physical phenomena. Candidates with contaminations should be discarded, while more or less pure interferent spectra should be exploited as such in the EMSC.

A Prototype MATLAB code

```
function [XCor,basis,coefs] = EMSCmod(X,polDeg,nRef,intF)
%% Modified EMSC using multiple reference spectra
%% Input
% X      — Data set
% polDeg — Degree of polynomial trends to correct for
% nRef   — Number of reference spectra to use
% intF   — Interferent spectra
%% Output
% XCor  — Corrected spectra with polynomial trends "added back"
% basis — Basis used for correction (polynomial trends, interferents,
%         reference spectra)
% coefs — Projections onto basis
%% Code

if nargin < 2; polDeg = 2; end
if nargin < 3; nRef = 1; end
if nargin < 4; intF = []; end

% Finding reference spectrum/a from SVD:
[∼,∼,V] = svd(X,'econ');
refSpec = V(:,1:nRef)';

[n,p] = size(X);
nintF = size(intF,1);
tot   = polDeg + 1 + nintF;

P = zeros(polDeg+1,p);
for i=0:polDeg; P(i+1,:) = linspace(—1,1,p).ˆi; end

[basis, R] = qr([P' intF' refSpec'],0); % Finding orthonormal basis

coefs = X ⋆ basis; % Projections onto basis
mult  = sqrt(sum(coefs(:,tot+1:end).ˆ2,2));
XCor  = X — coefs(:,1:tot) ⋆ basis(:,1:tot)';
XCor  = bsxfun(@rdivide,XCor,mult);

% Adding back polynomial trends for better visualisation when plotting:
refPol = R(tot+1,1:tot) ⋆ basis(:,1:tot)' / R(tot+1,tot+1);
XCor   = bsxfun(@plus,XCor,refPol);
```

B Acknowledgements

The Research Council of Norway (project number 239070) provided financial support for this work.

References

[1] N. K. Afseth, A. Kohler, *Chemom. Intell. Lab. Syst.* **2012**, 117.

[2] K. H. Liland, A. Kohler, N. K. Afseth, *J. Raman Spectrosc.* **2016**, *47*, 6.

[3] H. Martens, E. Stark, *J. Pharm. Biomed. Anal.* **1991**, *9*, 8.

[4] A. Kohler, C. Kirschner, A. Oust, H. Martens, *Appl. Spectrosc.* **2005**, *59*, 6.

[5] R. J. Barnes, M. S. Dhanoa, S. J. Lister, *Appl. Spectrosc.* **1989**, *43*, 5.

[6] A. Savitzky, M. J. Golay, *Anal. Chem.* **1964**, *36*, 8.

[7] K. H. Liland, T. Almøy, B.-H. Mevik, *Appl. Spectrosc.* **2010**, *64*, 9.

[8] Å. Rinnan, F. van den Berg, S. B. Engelsen, *TrAC Trends Anal. Chem.* **2009**, *28*, 10.

[9] A. Kohler, J. Sule-Suso, G. Sockalingum, M. Tobin, F. Bahrami, Y. Yang, J. Pijanka, P. Dumas, M. Cotte, D. Van Pittius, *Appl. Spectrosc.* **2008**, *62*, 3.

[10] T. Fearn, C. Riccioli, A. Garrido-Varo, J. E. Guerrero-Ginel, *Chemom. Intell. Lab. Syst.* **2009**, *96*, 1.

[11] P. Geladi, D. MacDougall, H. Martens, *Appl. Spectrosc.* **1985**, *39*, 3.

[12] A. Kohler, U. Böcker, J. Warringer, A. Blomberg, S. Omholt, E. Stark, H. Martens, *Appl. Spectrosc.* **2009**, *63*, 3.

[13] E. Kreyszig, *Introductory functional analysis with applications*, *Vol. 81*, wiley New York, **1989**.

[14] S. Wold, M. Sjöström, L. Eriksson, *Chemom. Intell. Lab. Syst.* **2001**, *58*, 2.

[15] J. R. Beattie, *J. Raman Spectrosc.* **2011**, *42*, 6.

[16] J. R. Beattie, J. J. McGarvey, *J. Raman Spectrosc.* **2013**, *44*, 2.

[17] J. H. Kalivas, *J. Chemom.* **2012**, *26*, 6.

[18] N. K. Afseth, J. P. Wold, V. H. Segtnan, *Anal. Chim. Acta* **2006**, *572*, 1.

[19] N. Afseth, V. Segtnan, B. Marquardt, J. Wold, *Appl. Spectrosc.* **2005**, *59*, 11.

[20] T. Næs, O. Tomic, N. K. Afseth, V. Segtnan, I. Måge, *Chemom. Intell. Lab. Syst.* **2013**, 124.

[21] P. Hansen, *Discrete Inverse Problems*, Society for Industrial and Applied Mathematics, **2010**.

Pre-processing Model	Raw spectra	EMSC	EMSC(1)	EMSC(2)	EMSC(3)
TR (L_2)	3.63	2.88	2.87	2.87	2.87
TR (D_1)	4.34	3.20	3.20	3.21	3.21
TR (D_2)	4.55	3.41	3.41	3.40	3.40
PLS	3.91	2.95	2.95	2.95	2.95

Table 1: *Fish oil data: Average RMSEP over* 500 *random data splits.*

Pre-processing Model	Raw spectra	EMSC	EMSC(1)	EMSC(2)	EMSC(3)
TR (L_2)	0.84	1.07	1.07	1.06	1.09
TR (D_1)	1.03	1.09	1.09	1.10	1.13
TR (D_2)	1.30	1.26	1.15	1.20	1.22
PLS	0.86	1.12	1.12	1.10	1.13

Table 2: *Emulasion data: Average RMSEP over* 500 *random data splits for the response fatty acids as % of total weight.*

Pre-processing Model	Raw spectra	EMSC	EMSC(1)	EMSC(2)	EMSC(3)
TR (L_2)	8.33	3.42	3.38	3.10	2.56
TR (D_1)	8.83	3.08	3.04	2.95	2.59
TR (D_2)	11.3	3.39	3.20	3.14	2.82
PLS	8.59	3.45	3.42	3.14	2.59

Table 3: *Emulsion data: Average RMSEP over* 500 *random data splits for the response PUFA as % of total fat content.*