Baseline and interferent correction by the Tikhonov Regularization framework for linear least squares modeling

Joakim Skogholt[1], Kristian Hovde Liland[1] and Ulf Geir Indahl[1]

[1]Faculty of Science and Technology, Norwegian University of Life Sciences, °As, Norway

Abstract

Spectroscopic data is usually perturbed by noise from various sources that should be removed prior to model calibration. After conducting a pre-processing step to eliminate unwanted multiplicative effects (effects that scale the pure signal in a multiplicative manner), we discuss how to correct a model for unwanted additive effects in the spectra. Our approach is described within the Tikhonov Regularization (TR) framework for linear regression model building, and our focus is on ignoring the influence of non-informative polynomial trends. This is obtained by including an additional criterion in the TR problem penalizing the resulting regression coefficients away from a selected set of possibly disturbing directions in the sample space. The presented method builds on the Extended Multiplicative Signal Correction (EMSC), and we compare the two approaches on several real data sets showing that the suggested TR-based method may improve the predictive power of the resulting model. We discuss the possibilities of imposing smoothness in the calculation of regression coefficients as well as imposing selection of wavelength regions within the TR framework. To implement TR efficiently in the model building, we use an algorithm that is heavily based on the singular value decomposition (SVD). Due to some favourable properties of the SVD it is possible to explore the models (including their generalized cross-validation (GCV) error estimates) associated with a large number of regularization parameter values at low computational cost.

1 Introduction

Spectroscopic data are often contaminated by various sources of noise and disturbances making analysis and/or interpretations challenging. Pre-processing of spectroscopic data before building models may therefore be essential for obtaining both accurate predictions and useful interpretations.[1,2] The noise in spectroscopic data is typically caused by various physical effects, depending on the type of technology being used. Baseline shifts and various types of scatter effects are quite

common in spectroscopic data. Mathematically we often model the noise as multiplicative and additive effects, where we assume that the noisy part of each spectrum is unique.

The purpose of the present paper is to discuss how to eliminate the influence of additive effects in linear regression model building by utilizing the Tikhonov Regularization (TR) framework. The elimination part is attained by adding an extra criterion to the linear regression problem, forcing the regression coefficients to be orthogonal to the directions in the sample space spanned by the additive effects. By varying a tuning parameter, the directions corresponding to additive effects can be completely removed or allowed to contribute to the model in a restricted fashion if this contributes to improving predictive performance. The suggested method can be applied directly to the raw data, or subsequent to any data pre-processing step. See also Andries and Kalivas[3] for a theoretical discussion of this idea.

The focus of our work is on how to remove the influence of polynomial trends efficiently as an integrated part of the model building. We will also compare this approach with some existing pre-processing methods that correct for polynomial trends. This idea has been mentioned in papers by Kalivas[4] and Stout and Kalivas[5] in the context of Tikhonov Regularization, and discussed in Vogt et al.[6] in the context of principal component regression. The proposed method solves a penalized linear least squares problem by including additional penalty terms within the Tikhonov Regularization framework. The solution to this least squares problem will be orthogonal to unwanted polynomial trends in the data.

Using raw spectra as input to this TR problem will often produce subpar results. The reason for this is that spectral data often contains scattering effects that affect the spectra multiplicatively. These effects should be corrected in a pre-processing step prior to model building. Here we discuss using Extended Multiplicative Signal Correction (EMSC) and Standard Normal Variate (SNV) to pre-process data prior to model building. In the examples we will use EMSC to pre-process the spectra.

For regularization in the TR problem, we will discuss three different types of regularizations: (1) L_2 regularization, (2) discrete first derivative regularization, (3) discrete second derivative regularization. For L_2 regularization without any wavelength selection we will show that polynomial trends can be corrected for when pre-processing the data. We will also show that when using a type of derivative regularization or L_2 regularization with wavelength selection, an extra polynomial criterion in the TR problem is necessary for obtaining orthogonality between the unwanted polynomial trends and the regression coefficients. By using one of the above types of regularizations together with EMSC pre-processed spectra, we obtain regression models comparable to PLS models with EMSC pre-processed spectra.

In the following sections we give a short review of some common pre-processing methods for spectroscopic data, and of the TR-framework. Thereafter we introduce the baseline correcting approach as the main topic of this paper. The baseline correcting method is then compared to EMSC and some similarities and differences between the two approaches are discussed. Finally we show the results of applying the TR-method on two different data sets of Raman spectra.

2 Pre-processing of spectral data

2.1 Pre-processing

Pre-processing of spectral data is widely considered as necessary prior to regression model building.[1,7,8] There are different ways to describe noise and artifacts in spectroscopic data. One can, for example, distinguish between baseline, scatter, noise, and misalignments.[7] In Raman spectroscopy, fluorescence may cause large baseline effects[8,9] which can result in a vertical shift of the spectra. Many baseline correcting procedures rely on a baseline estimation and correction by fitting and subtracting low degree polynomials from the spectra. See for example Liland et al.[8] for a review of several baseline estimation algorithms, or Liland et al.[10] for a discussion of how to choose an appropriate baseline correction.

In NIR spectroscopy there may be variations in the spectra due to variable path lengths that light travels inside the samples, and/or scatter effects due to the particle size distribution.[11,12] Ambient light and light intensity of the radiation source can also affect the spectra.[13] Scatter effects can be caused by the particle size in a sample being similar in size to the wavelength of the light used, and it is often modeled by individual scaling factors adjusting each spectrum.[7] The most common scatter correction methods are *multiplicative scatter correction* (MSC) and *standard normal variate* (SNV),[7,14] as well as baseline correcting procedures.

The method suggested in this paper does not include the correction of multiplicative scatter effects, so such effects must be handled prior to solving the regression problem. A review of the two methods most commonly used to correct for multiplicative scatter effects is given in the next section.

2.2 Scatter correction by SNV and (E)MSC

The SNV was introduced in Barnes et al.,[11] where it is claimed that the main variation in near-infrared diffuse reflectance spectra are due to (1) scatter, (2) path length, and (3) chemical composition. The variations due to scatter and path length may corrupt the spectra by both an unwanted

vertical shift and an unwanted multiplicative effect (due to scatter rather than chemical information). The SNV is simply an auto scaling procedure correcting each spectrum individually as follows[14] : Suppose we have n spectra represented by the vectors $\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)}$. Then for $i = 1, \ldots, n$, we define the SNV-corrected spectra as

$$\boldsymbol{x}_{cor(i)} = \frac{\boldsymbol{x}_{(i)} - \bar{\boldsymbol{x}}_{(i)}}{sd(\boldsymbol{x}_{(i)})}, \tag{1}$$

where $\bar{\boldsymbol{x}}_{(i)}$ and $sd(\boldsymbol{x}_{(i)})$ denote the mean and standard deviation of the spectrum $\boldsymbol{x}_{(i)}$, respectively.

In Barnes et al.[11] the authors also suggest a baseline correcting procedure referred to as *detrending*. The detrending is obtained by regressing the spectra onto a polynomial evaluated at the measured wavelengths and returning the residual vectors from these regressions.

The MSC was introduced in Geladi et al.[12] to separate absorption in samples due to chemical content from the various sources of scatter. The idea behind the MSC is that scatter and light absorption due to chemical effects have different dependencies on electromagnetic wavelengths, and that this fact should enable the possibility of separating the scatter phenomena from the signal of interest. By using the MSC we model each spectrum as

$$\boldsymbol{x}_{(i)} = a \cdot \mathbf{1} + b \cdot \boldsymbol{x}_{ref} + \boldsymbol{e}_{mi} \tag{2}$$

where $\boldsymbol{x}_{ref}$ is a fixed reference spectrum and $\mathbf{1}$ is a vector of corresponding length. The scalars a, b are obtained by least-squares regression, and $\boldsymbol{e}_{mi}$ is the associated residual vector (where the subscript m is used to indicate MSC pre-processing). In the original description of the MSC, it is argued that one should be using an "ideal" sample as the reference spectrum $\boldsymbol{x}_{ref}$, and correct the other spectra 'so that all samples appear to have the same scatter level as the "ideal"'.[15] Choosing the reference spectrum to be the sample mean of the considered spectra is often considered a useful choice.[9,12] In the end, the MSC-corrected spectrum is given by the formula

$$\boldsymbol{x}_{m(i)} = \frac{\boldsymbol{x}_{(i)} - a \cdot \mathbf{1}}{b} = \boldsymbol{x}_{ref} + \frac{1}{b}\boldsymbol{e}_{mi}. \tag{3}$$

It is a simple task to extend the MSC by including additional terms in the representation of the spectrum $\boldsymbol{x}$, and the resulting correction method is usually referred to as the Extended MSC (EMSC).[16] The most basic version of the EMSC has the representation

$$\boldsymbol{x}_{(i)} = a \cdot \mathbf{1} + b \cdot \boldsymbol{x}_{ref} + c_1 \cdot \boldsymbol{v}_1 + c_2 \cdot \boldsymbol{v}_2 + \boldsymbol{e}_{ei}, \tag{4}$$

where the vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ represent the measured wavelength numbers and the square of these numbers, respectively. The subscript e in the residual $\boldsymbol{e}_{ei}$ is conventionally used to denote that EMSC pre-processing is taking place. The scalars a, b, c_1, c_2 are obtained by linear least squares fitting of $\boldsymbol{x}$ to the vectors $\mathbf{1}$, $\boldsymbol{x}_{ref}, \boldsymbol{v}_1$ and $\boldsymbol{v}_2$. The corrected spectra are given by (where the subscript e is used to indicate EMSC pre-processing):

$$\boldsymbol{x}_{e(i)} = \frac{\boldsymbol{x}_{(i)} - a \cdot \mathbf{1} - c_1 \cdot \boldsymbol{v}_1 - c_2 \cdot \boldsymbol{v}_2}{b} = \boldsymbol{x}_{ref} + \frac{1}{b}\boldsymbol{e}_{ei}. \tag{5}$$

The basic EMSC modeling described above can also be extended to include polynomials of an arbitrary degree.[9] Note that the scalars a, b to be estimated in both the MSC and EMSC formulas will in general not be identical because the vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are not required to be orthogonal to the vectors $\mathbf{1}$ and $\boldsymbol{x}_{ref}$.

In practice, this means that the estimated multiplicative effect (b) of a spectrum depends on whether the MSC or the EMSC is chosen for the pre-processing. This is also pointed out in Rinnan et al.,[14] and more details will be given below.

By using the EMSC pre-processing we are eliminating the components of the spectra associated with the subspace spanned by the vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. Note that the projection of a corrected spectrum $\boldsymbol{x}_{e(i)}$ onto this subspace is identical to the projection of the reference spectrum $\boldsymbol{x}_{ref}$ for all samples ($1 \leq i \leq n$), and that this projection in general will be non-zero. Therefore, the $\boldsymbol{v}_1, \boldsymbol{v}_2$-directions will not influence the later models obtained by methods such as PLS as the (corrected) data matrices are always centred prior to model building. As we will discuss later, these directions may or may not affect the regression coefficients in TR depending on the type of regularization used.

The MSC and SNV are often considered as similar for most applications when a representative spectrum is used as the reference spectrum,[14] as they both include a centering as well as a scaling part. However, the two methods may in some cases produce very different results as their centerings and scalings are calculated according to different strategies.[17]

It is worthwhile to note that the SNV operates on each spectrum completely individually, whereas the EMSC uses a reference spectrum based on all the available spectra to be included in the individual correction-models. This issue is relevant, for example, when using cross-validation strategies for model selection. If the (E)MSC pre-processing is used and the reference spectrum is taken as the mean spectrum of the training set, then strictly speaking a new (E)MSC model should be re-calculated for each choice of training set, whereas this challenge does not occur when the SNV-method is used.

There are also other pre-processing methods that can be used to estimate and correct for scatter effects. One example is the Optical Path-Length Estimation and Correction (OPLEC),[2] which allows for estimating scatter when the concentration of the components in a sample is known. When using OPLEC there is also a polynomial correction by projection.

3 Tikhonov Regularization

3.1 A brief overview of Tikhonov Regularization for linear least squares modeling

In this section we briefly review the Tikhonov Regularization framework for linear least squares modeling. We assume we have a data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ associated with n samples and p predictor variables, and a corresponding response vector $\boldsymbol{y} \in \mathbb{R}^n$. We also assume we have a matrix $\boldsymbol{L} \in \mathbb{R}^{p \times p}$, and a tuning parameter $\lambda > 0$. The TR problem is specified by the linear system

$$\begin{bmatrix} \boldsymbol{X} \\ \sqrt{\lambda} \cdot \boldsymbol{L} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix}. \tag{6}$$

The corresponding least squares problem to be minimized with respect to $\boldsymbol{\beta}$ is

$$\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 + \lambda \|\boldsymbol{L}\boldsymbol{\beta}\|^2, \tag{7}$$

where the regularization parameter λ is considered as fixed. The purpose of the regularization matrix $\boldsymbol{L}$ in (6, 7) is to impose additional constraints on the regression coefficients and to overcome problems with multicollinearity present in the ordinary least squares (OLS) formulation. The most common choice for $\boldsymbol{L}$ is the identity matrix ($\boldsymbol{I}$). Various discrete differential operators and diagonal matrices representing wavelength selections are other popular choices.[4,5] Note that the choice $\boldsymbol{L} = \boldsymbol{I}$ corresponds to the ordinary Ridge Regression problem[18] without variable standardization. As shown later in the examples, the choice of regularization may have a considerable impact on the resulting regression coefficients.

3.2 Regression coefficients

In the following we will assume that the regularization matrix $\boldsymbol{L}$ in (6) is invertible. If $\boldsymbol{L} \neq \boldsymbol{I}$ (the identity matrix), one can then transform the problem into standard form by considering $\boldsymbol{X}\boldsymbol{L}^{-1}$ in the place of the original $\boldsymbol{X}$ (see e.g. Stout and Kalivas[5] for a more thorough explanation).

Without loss of generality, we will therefore assume $\boldsymbol{L} = \boldsymbol{I}$ in the following. If $\boldsymbol{L}$ is not invertible the standardization process is a bit more involved. See e.g.[19] for details about this case.

The least squares solution of (6) can be obtained by solving the corresponding normal equations

$$(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}. \tag{8}$$

By considering the reduced singular value decomposition (SVD) of $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}'$ (here $\boldsymbol{S}$ is the diagonal matrix of non-zero singular values, $\boldsymbol{U}$ and $\boldsymbol{V}$ represent the corresponding left and right singular vectors), the solution $\boldsymbol{\beta}$ to (8) simplifies to

$$\boldsymbol{\beta} = \boldsymbol{V}(\boldsymbol{S}'\boldsymbol{S} + \lambda\boldsymbol{I})^{-1}\boldsymbol{S}\boldsymbol{U}'\boldsymbol{y}. \tag{9}$$

A derivation of this expression can be found in Hastie et al.[20]

The following properties of equation (9) should be noted:

1. The formula for the regression coefficients in (9) are only depending on λ in the inversion of a diagonal matrix. This implies that from the reduced SVD of a data matrix, the computation of the regression coefficients corresponding to any choice of the regularization parameter λ only requires multiplication of matrices and the inversion of a diagonal matrix. Thus, having calculated the reduced SVD of the data matrix, we can generate regression coefficients for any value of λ at a very low computational cost.

2. From equation (9) it is clear that the matrix $\boldsymbol{V}(\boldsymbol{S}'\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\boldsymbol{S}\boldsymbol{U}'$ linearly transforms (by left multiplication) *any* response vector $\boldsymbol{y} \in \mathbb{R}^n$ to be associated with the data matrix $\boldsymbol{X}$ into a corresponding vector $\boldsymbol{\beta} \in \mathbb{R}^p$ of regression coefficients.

The above remarks imply that once we have calculated the reduced SVD of the data matrix $\boldsymbol{X}$, the desired model for any value of λ and any choice of response vector $\boldsymbol{y}$, can be obtained directly by ordinary matrix multiplications. The only restriction with this approach is its reliance upon the SVD of $\boldsymbol{X}$. If $\boldsymbol{X}$ is large and calculating its reduced SVD is not computationally feasible one can solve the least squares problem (6) using alternative techniques, such as QR factorization.

3.3 Model selection

When using a regularized approach to linear modeling such as TR, choosing an appropriate value of the regularization parameter(s) can make or break the modeling process.[4] Thus having a good procedure for choosing the value(s) of the parameter(s) is essential.

Choosing an appropriate value of the regularization parameter is a trade-off between model fit and model complexity.[20] There is no known approach to this problem that always provides an objectively optimal solution.[19] Some alternatives include consideration of L-curves[19,21] or more statistically motivated techniques like cross-validation. In this paper we advocate for using the generalized cross-validation (GCV) proposed by Golub et al.[22] for the selection of an appropriate regularization parameter value. The reason for this is that, as explained below, this can be implemented very efficiently using the singular value decomposition of the data matrix $\boldsymbol{X}$. In the examples we will compare the TR models to PLS models. To make the comparison fair, we can of course use LOOCV for both TR and PLS. In our experience, minimisation of the LOOCV and GCV statistics results in comparable values of the regularization parameter, and it matters little which one is used. We indicate this in the examples by giving prediction results for TR solutions obtained from both LOOCV and GCV.

The primary motivation for preferring the GCV is that this method avoids some problems with leave-one-out cross-validation (LOOCV) as the GCV is a rotation-invariant version of the LOOCV.

The GCV statistic is defined as (our projection matrix differs from the one in Golub et al.[22] by a factor of n in the term with λ):

$$GCV(\lambda) = \frac{\|(\boldsymbol{I} - \boldsymbol{A}(\lambda))\boldsymbol{y}\|^2}{\left[\frac{1}{n}Tr(\boldsymbol{I} - \boldsymbol{A}(\lambda))\right]^2} \tag{10}$$

where $\boldsymbol{A}(\lambda) = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'$.

We now show how the GCV statistic can be calculated using matrix addition and multiplication only when the SVD of X is known. Note that the numerator in (10) is simply the squared norm of the residual. As discussed in the previous section, the regression coefficients (and hence the corresponding residuals) can be calculated using only matrix multiplications. Using the reduced SVD of $\boldsymbol{X}$ the matrix $\boldsymbol{A}$ can be expressed as

$$\boldsymbol{A}(\lambda) = \boldsymbol{U}\boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S} + \lambda\boldsymbol{I})^{-1}\boldsymbol{S}'\boldsymbol{U}' = \boldsymbol{U}[\boldsymbol{S}^2(\boldsymbol{S}^2 + \lambda\boldsymbol{I})^{-1}]\boldsymbol{U}'. \tag{11}$$

The matrix inside the brackets in (11) is diagonal, and can be calculated directly by simple scalar operations for any choice of λ.

It is therefore computationally "inexpensive" to compute the GCV statistic once the SVD of the data matrix $\boldsymbol{X}$ is available. Thus one way of finding a good value of the regularization parameter λ using GCV is to consider it as a function of λ, and plot the $GCV(\lambda)$-function for a "large" but finite set of well-spread λ-values. Finally, we choose the particular λ-value associated with the

smallest GCV-value. For a more genuine minimization of $GCV(\lambda)$, the minimizer obtained from the discrete procedure proposed above can be taken as a starting point for running a numerical optimization routine.

A MATLAB implementation of this approach using the `fminbnd`-function from MATLABs Optimization Toolbox is given in the appendix. The code in the appendix also include code for calculating the GCV statistic for a selected sample of values of the regularization parameter. In our experience this approach works equally well to using `fminbnd` to find the optimal value of the regularization parameter, assuming a sufficiently sized sample of values of the regularization parameter is selected in an appropriate range.

We note that the use of GCV here is primarily aimed at selecting an appropriate value of the regularization parameter λ rather than providing an accurate error estimate of the model. Once a good value for the regularization parameter has been found, the associated model may be validated with respect to its predictive performance using some appropriate cross-validation strategy or a separate test set.

3.4 Adding additional criteria to the model calibration

The basic formulation of the TR problem given in (6) is easily extended by including additional rows in the equation. Such inclusions correspond to imposing additional constraints on the desired regression coefficients.

The focus of this this paper is to eliminate the influence of additive effects in spectra by integrating additional constraints in the TR problem formulation. This can be done by inserting extra rows into the matrix on the left hand side in equation (6) and corresponding zeros on the right hand side. The extra rows should be chosen as set of basis vectors spanning the subspace of additive effects that are not supposed to influence our final model. In what follows we discuss primarily polynomial trends. For a more general theoretical discussion, see e.g. Andries and Kalivas.[3]

Additive effects are often modeled as lower-order polynomials. An orthogonal basis for such polynomial spaces can be obtained by considering the Legendre polynomials up to some desired degree.[2][3] More precisely we create a matrix with the polynomial trends evaluated evenly in the interval $[-1, 1]$ as columns. We then find a QR-decomposition of this matrix and use the resulting orthogonal vectors as rows in the matrix $\boldsymbol{P}$ (see the MATLAB-function Plegendre in the appendix implementing the details). By multiplying $\boldsymbol{P}$ with a huge constant $\sqrt{\mu}$, and inserting zeros in the corresponding rows of the response vector on the right hand side of (6), the updated equation becomes

$$\begin{bmatrix} \boldsymbol{X} \\ \sqrt{\mu} \cdot \boldsymbol{P} \\ \sqrt{\lambda} \cdot \boldsymbol{L} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}. \tag{12}$$

The least squares solution of (12) corresponds to finding the minimizer with respect to $\boldsymbol{\beta}$ of the expression

$$\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 + \mu\|\boldsymbol{P}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{L}\boldsymbol{\beta}\|^2, \tag{13}$$

where λ and μ are considered as fixed quantities. By selecting μ sufficiently large we can force the regression coefficients solving the least squares problem (12) to be numerically as close to orthogonal to the chosen $\boldsymbol{P}$-directions in the measured samples as we like. The resulting model will therefore ignore such polynomial trends directly, instead of deflating them off the spectra in a pre-processing step.

We note that this method is also applicable in correcting for arbitrary known interferents (not only polynomial trends) by specifying an appropriate set of basis vectors for the actual interferent-subspace.

In the limiting case when μ grows large, the suggested method corresponds to projecting the spectra onto subspaces orthogonal to the polynomial trends, but as we will show later, in the context of TR with $\boldsymbol{L} \neq \boldsymbol{I}$ the two approaches are not equivalent.

In the discussion above and what follows we suggest using a "hard-coded" large value for μ. In the code for the examples the value $\mu = 10^{24}$ is used. This value was chosen to be large enough to make the regression coefficients obtained orthogonal to the polynomial trends to machine precision. If the scale of the measurements is significantly different than for the examples used in the present work then a different value of μ may be chosen. The result of this choice is to completely remove the influence of the directions spanned by the rows in $\boldsymbol{P}$ on the regression coefficients. We note that it is also possible to treat μ as an ordinary regularization parameter that may be chosen by some model selection criterion. If this is done and the regularization parameter is not chosen too large, then the rows in $\boldsymbol{P}$ are allowed to contribute partially in the resulting regression coefficients.

In the practical calculations we first centre $\boldsymbol{X}$ and $\boldsymbol{y}$ with respect to their column means before appending $\sqrt{\mu}\boldsymbol{P}$ to $\boldsymbol{X}$ and calculating the singular value decomposition for the augmented matrix. When computing the GCV statistic as described in the previous section, it is therefore important to truncate the GCV calculations to only account for the upper n rows of the augmented $\boldsymbol{X}$ as it does not make sense to consider the rows in $\boldsymbol{P}$ for model selection. See the code in the appendix for the required details.

4 Comparison with EMSC

4.1 MSC and EMSC explained by linear algebra

EMSC pre-processing is used for both eliminating polynomial trends and correcting for scatter effects in spectroscopic data. By using the EMSC pre-processing with second order polynomial correction, the spectra are projected onto a 4-dimensional subspace (where 3 of the basis vectors are associated with the second degree polynomial subspace). In the present work we suggest including the correction of polynomial trends as an integrated part of the TR approach by considering the required equations enforcing the desired orthogonality properties. Because the EMSC as well as the proposed TR-approach are aiming at the same purpose, it is of interest to compare and contrast the two methods. Before comparing the two methods, we will briefly review the linear algebra required for describing the MSC and the EMSC pre-processing.

Recall that the rows of the matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and the vector $\boldsymbol{y} \in \mathbb{R}^n$ represent our spectra and associated response measurements. We also assume the reference spectrum $\boldsymbol{x}_{ref} \in \mathbb{R}^p$ to be known. For MSC and EMSC the two subspaces required for filtering the samples are given by the subspace bases $W_{MSC} = \{\mathbf{1}, \boldsymbol{x}_{ref}\} \subset \mathbb{R}^p$ and $W_{EMSC} = \{\mathbf{1}, \boldsymbol{x}_{ref}, \boldsymbol{v}_1, \boldsymbol{v}_2\} \subset \mathbb{R}^p$, respectively. According to Section 2.2, the formulae for MSC and EMSC pre-processing are given by the equations (3) and (5).

For both types of pre-processing, the scaled residuals $\frac{1}{b_{\cdot i}}\boldsymbol{e}_{\cdot i}$ are considered to be representative for the interesting chemical information of the associated samples $\boldsymbol{x}_{\cdot(i)}$. To make a direct comparison of $\boldsymbol{x}_{m(i)}$ and $\boldsymbol{x}_{e(i)}$, one needs to express these vectors with respect to a common basis. An appropriate basis can be obtained by extending W_{EMSC} into a complete basis for $\mathbb{R}^p$. Such a basis can be found by introducing a set of basis vectors $W_r = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{p-4}\} \subset \mathbb{R}^p$ that spans the orthogonal complement of $span(W_{EMSC})$, i.e. $span(W_r) = span(W_{EMSC})^{\perp}$ and $\mathbb{R}^p = span(W_{EMSC}) \oplus span(W_r)$.

With respect to the basis $W_{EMSC} \cup W_r$, the pre-processed spectra given in (3) and (5) can be represented as follows:

$$\boldsymbol{x}_{m(i)} = \frac{a_{ei} - a_{mi}}{b_{mi}} \cdot \mathbf{1} + \frac{b_{ei}}{b_{mi}} \cdot \boldsymbol{x}_{ref} + \frac{c_{i1}}{b_{mi}} \cdot \boldsymbol{v}_1 + \frac{c_{i2}}{b_{mi}} \cdot \boldsymbol{v}_2 + \frac{1}{b_{mi}} \cdot \sum_{j=1}^{p-4} \alpha_j \boldsymbol{r}_j \tag{14}$$

and

$$\boldsymbol{x}_{e(i)} = 0 \cdot \mathbf{1} + 1 \cdot \boldsymbol{x}_{ref} + 0 \cdot \boldsymbol{v}_1 + 0 \cdot \boldsymbol{v}_2 + \frac{1}{b_{ei}} \cdot \sum_{j=1}^{p-4} \alpha_j \boldsymbol{r}_j \tag{15}$$

for MSC and EMSC, respectively. The first of these equations is obtained by applying the MSC pre-processing to the sample $x_{(i)}$ with the basis $W_{EMSC} \cup W_r$. The differences between the scatter

correction scalars (the b_{mi} and b_{ei} in the above equations) will typically be small for MSC and EMSC. However, in some cases they may be noticeably different and the differences may affect the predictive power of the model (as is shown for the fish oil data in Section 5). Aside from the different estimates of the scatter correction scalars b_{mi} and b_{ei}, the differences between the MSC and EMSC pre-processed spectra are clearly located in the subspace spanned by the vectors $\{\mathbf{1}, \boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{x}_{ref}\}$.

4.2 MSC with trend correction versus EMSC

We will now compare the removal of polynomial trends by EMSC to the removal of such trends by including the required polynomial orthogonality as an additional constraint in the TR problem. Although we will limit investigation to considering polynomials of degree 2 or less, the given argument readily generalizes to the correction of polynomial trends of any degree. Consider the following two regression problems:

$$\begin{bmatrix} \boldsymbol{X}_{EMSC} \\ \sqrt{\lambda} \cdot \boldsymbol{L} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{y} \\ \mathbf{0} \end{bmatrix} \tag{16}$$

and

$$\begin{bmatrix} \boldsymbol{X}_{MSC} \\ \sqrt{\mu} \cdot \boldsymbol{P} \\ \sqrt{\lambda} \cdot \boldsymbol{L} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{y} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \tag{17}$$

where $\boldsymbol{P}$ is a matrix with 3 rows representing the space of polynomials of degree 2. First we consider the case when $\boldsymbol{L} = \boldsymbol{I}$ (this corresponds to putting restrictions on the L_2-norm of the solution vector $\boldsymbol{\beta}$) and the corresponding solution of (16).

Denote the reduced SVD of $\boldsymbol{X}_{EMSC}$ by $\boldsymbol{X}_{EMSC} = \boldsymbol{USV}'$. From (9) we see that the solution $\boldsymbol{\beta}$ to (16) is a linear combination of the columns in $\boldsymbol{V}$. From equation (15) we see that after centering $\boldsymbol{X}_{EMSC}$, the rows in $\boldsymbol{X}_{EMSC}$ will be orthogonal to the vectors in W_{EMSC}. By considering the the full SVD of $\boldsymbol{X}_{EMSC}$, all the vectors in W_{EMSC} can be expressed as linear combinations of the right-singular vectors associated with the singular value zero.

As the right singular vectors are orthogonal, it follows that the columns of $\boldsymbol{V}$ are orthogonal to W_{EMSC}. Therefore the solution of (16) will be orthogonal to the vectors in W_{EMSC}. Because we assume $\boldsymbol{L} = \boldsymbol{I}$ together with EMSC pre-processed spectra the solution vector will be orthogonal to the trends being corrected for in the EMSC pre-processing. Thus, in this case, adding an extra polynomial correction criterion to (16) will not affect the regression coefficients.

Now consider the solution of (17). From (3) and (14) we see that after centering, the rows in $\boldsymbol{X}_{MSC}$ will be orthogonal to the vectors in W_{MSC}. Without the inclusion of the additional polynomial criterion (represented by the matrix $\boldsymbol{P}$) the solution vector of (17) would in general only be orthogonal to the vectors $\boldsymbol{x}_{ref}$ and $\mathbf{1}$. However, the additional polynomial criterion forces the solution $\hat{\boldsymbol{\beta}}$ of (17) to also be as close to orthogonal to the vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ as we like by choosing $\sqrt{\mu}$ to be sufficiently large. The difference in the solutions of (16) and (17) is therefore explained by the difference in the estimated scatter coefficients. Such estimates will often be fairly similar, but as demonstrated in the fish oil example below, their differences may affect the predictive power of the model.

In the more general case with $\boldsymbol{L} \neq \boldsymbol{I}$, one can solve (16) and (17) by first transforming the data as indicated in Section 3.2. Such transformations will in general affect the right singular vectors of the data matrix. Therefore, the above argument based on $\boldsymbol{L} = \boldsymbol{I}$ to show that the solution to (16) is orthogonal to the vectors in W_{EMSC} is no longer valid. So when using a regularization matrix $\boldsymbol{L} \neq \boldsymbol{I}$, the resulting regression coefficients will <u>not</u> in general be orthogonal to the trends corrected for in the pre-processing. In this case adding the extra polynomial block $\sqrt{\mu}\boldsymbol{P}$ to (16) corresponding to the polynomial trends removed in the pre-processing may affect the resulting regression coefficients (this point is illustrated in the examples presented below). In the examples we will also in some cases add an extra criterion to the TR problem consisting of a diagonal matrix with large entries for wavelengths that are irrelevant for prediction. In this case, for the same reason as discussed above, it will be necessary to add an extra orthogonality condition to the TR problem to ensure orthogonality between the regression coefficients and the unwanted polynomial trends. We note that if SNV is used for pre-processing the data, the detrending described in Barnes et al.[11] will correspond to the polynomial trend correction proposed here if L_2 regularization is used together with a large "hard-coded" value of the μ parameter. We also note that if the μ parameter is chosen by validation instead of using a hard-coded value, then the method of removing polynomial trends discussed here will not be equivalent to other methods that removes the projection onto subspaces spanned by polynomials, such as e.g. EMSC and SNV with trend correction.

The regression coefficients (i.e. the model parameters) obtained when using EMSC pre-processing may sometimes represent information considered to be useful for interpretations.[24] When using both MSC pre-processing and correction of polynomial trends by the method suggested in this paper, we do not derive these coefficients explicitly, as we obtain regression coefficients that are orthogonal to the subspaces of interest without explicitly calculating the sample projections onto these subspaces (for prediction purposes these parameters are clearly irrelevant). The EMSC model parameters are the regression coefficients obtained by solving multiple OLS problems, so these pa-

rameters can always be calculated at the computational cost of solving the regression problem $\boldsymbol{A}'\mathbf{B} = \boldsymbol{X}'$, where $\boldsymbol{A}$ is a matrix with columns being the vectors in the basis W_{EMSC}.

5 Examples

Here we will study the practical side of the theoretical considerations discussed in this paper by applications to two data sets of Raman spectra. We will primarily use EMSC to pre-process the spectra. When using EMSC to correct Raman spectra it is common to use polynomials up to degree six or seven. This choice of polynomial degree can be justified as the chemical information in Raman spectra is generally contained in very steep peaks.[9] In both examples, unless otherwise stated, we use EMSC to pre-process the spectra and correct for polynomial trends up to and including degree 6, and we refer to this as EMSC(6) pre-processing.

In addition to TR models we also provide PLS models for comparisons. Selection of the PLS models are based on LOOCV. The regularization parameter values for the TR models shown in the tables are primarily obtained by LOOCV. The regularization parameter values for the associated models obtained by GCV are in most cases very similar to the LOOCV-results. The tables in the examples below also include prediction results from TR models obtained using GCV. This is included to illustrate that LOOCV and GCV typically performs very similarly for selecting the value of the regularization parameter in TR. As we have shown earlier, the GCV statistic can be calculated very efficiently. We can therefore safely recommend using GCV for estimating an appropriate value of the regularization parameter.

The `fminbnd` function from the MATLAB Optimisation Toolbox was used to determine the value of the regularization parameter giving the minimal GCV or RMSECV statistic. The `fminbnd`-function requires a lower and upper bound on the value of the regularization parameter. A minimum value of 0 was chosen as the regularization parameter is non-negative, and a maximum value of 10^{20} was chosen as for the examples considered here this is outside of the range of a reasonable value of the regularization parameter. For some models the minimizer failed and generated models with the maximum allowed value of the regularization parameter (10^{20}). In this case a lower maximum value of the regularization parameter was set, and the model calculation redone. This was repeated, lowering the maximum value each time, until a reasonable model was found. An alternative to using the `fminbnd` function which from our experience works equally well is to simply sample a range of values for the λ-parameter and calculate the GCV statistic or the RMSECV associated with these values. One can then simply choose the λ corresponding to the minimum GCV or RMSECV statistic. The code for this approach using GCV is integrated into

the MATLAB function given in the appendix.

Note that by following the above steps we are, strictly speaking, not calculating the LOOCV estimates and GCV statistic correctly, as we are not generating new EMSC models for each spectrum we remove from the model (which we should clearly do for LOOCV, and for GCV as GCV is LOOCV in a particular coordinate system). This should not have any significant impact as the only information we use from all the spectra in the training set is the mean of the spectra, but our estimates will have a small bias.

The optimal model in a model family is defined as the model with the value of the regularization parameter with the minimum RMSECV (or GCV) value.

5.1 Raman spectra of fish oil

First we look at a data set of Raman spectra of oil samples from salmon.[10,25,26] The response variable is the iodine value, which is used as a measure of unsaturation in the fat. This data set was also analyzed in Liland et al.,[10] using various baseline correction algorithms with PLSR. For comparison purposes, we use the same training/test set split and the same wavelength truncations as in Liland et al.[10] The data set consists of 45 spectra (30 samples used for training, 15 for testing) with 2263 wavelengths between $790cm^{-1}$ and $3050cm^{-1}$ (after truncation).

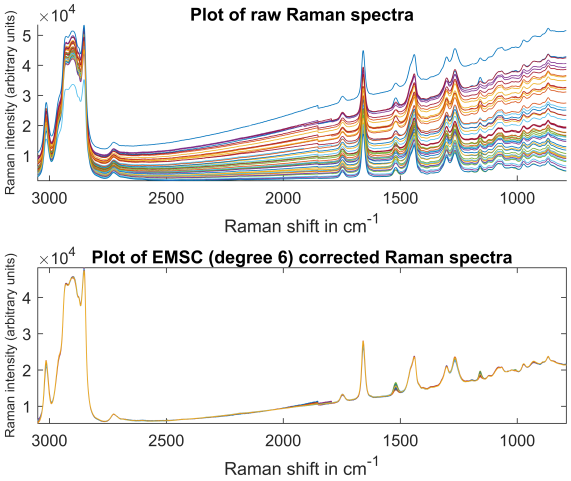

Figure 1: *Top: Raw Raman spectra of salmon oil. Notice in particular the non-linearities in the baseline. Bottom: EMSC(6) pre-processed Raman spectra of salmon oil.*

There are unwanted additive and multiplicative noise effects affecting the spectra, as well as an instrument detector shift at about $1800cm^{-1}$. Following the analysis in Afseth and Kohler,[9] we use EMSC including corrections for polynomials up to degree 6 to pre-process the spectra. The raw spectra and the EMSC(6) pre-processed spectra are shown in Figure 1. There is still a

clear baseline in the spectra, as can be seen in the corrected spectra in Figure 1, but most of the unwanted variation between the spectra has been removed. As we are centering the data prior to modeling this baseline will not affect the predictions. The test spectra were corrected using the reference spectrum obtained from the training spectra, i.e. the mean of the training spectra.

Following the steps given at the beginning of Section 5 we generated models for EMSC pre-processed spectra with L_2 regularization, discrete first derivative and second derivative regularization (hereafter referred to as D_1 and D_2 regularization, respectively).

For comparison PLS models were created with up to 20 components, using EMSC(6) pre-processed data for the results in Table 1, and using MSC pre-processing for the results in Table 2. For each PLS model the RMSECV was calculated using leave-one-out cross-validation. The optimal PLS model was selected as the model with the minimum RMSECV. This resulted in a PLS model with 2 components for the EMSC(6) pre-processed spectra, and a model with 3 components for the MSC pre-processed spectra.

The results are summarized in Tables 1 and 2.

The GCV statistic reported in the tables is the square root of the GCV statistic as defined earlier in the paper. This is done for easier comparison with the RMSEP values.

The rows of the matrix $\boldsymbol{P}$ with the polynomial trends used in this example are plotted in Figure 2.

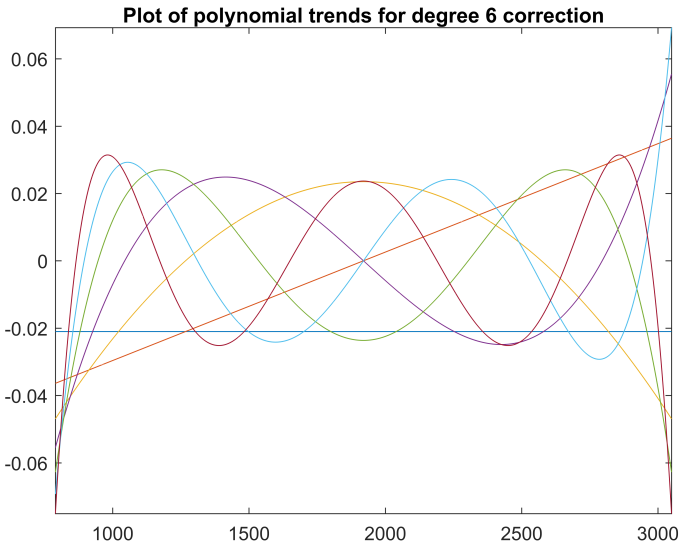

Figure 2: *Plot of the rows in the matrix* $\boldsymbol{P}$ *appended to the TR problem. There are 7 curves as we are correcting for polynomial trends up to and including degree 6.*

Notice from Table 1 that the performance increase on the test set by adding degree 6 orthogonalization to the TR problem using LOOCV for model selection is roughly 26%. This should be

Orthogonalization	Reg.	Optimal λ (LOOCV)	Min. RMSECV (LOOCV)	RMSEP (LOOCV)	RMSEP (GCV)
TR (No orth.)	L_2	$1.45 \cdot 10^7$	3.02	2.03	2.00
TR (Degree 6)	L_2	$1.45 \cdot 10^7$	3.02	2.03	2.00
TR (No orth.)	D_1	$9.56 \cdot 10^9$	3.12	2.15	1.99
TR (Degree 6)	D_1	$1.35 \cdot 10^9$	3.17	1.97	1.83
TR (No orth.)	D_2	$1.55 \cdot 10^{12}$	3.13	2.35	2.15
TR (Degree 6)	D_2	$2.74 \cdot 10^{13}$	3.36	1.74	1.83
PLS (2 components)	NA	NA	3.07	1.83	NA

Table 1: *Fish oil data with EMSC(6) pre-processing. Comparison of properties of the regression coefficients. The orthogonality column refers to which polynomials (if any) are added as an additional criterion to the TR problem.*

considered an extreme case, but it illustrates how adding an additional orthogonalization criterion to the TR problem can impact prediction even if 'the same correction' has been made in the pre-processing of the spectra. From Figure 3 we see that the model family generated by adding a degree 6 correction to the TR problem has better prediction in the region containing the λ-values that are likely to be chosen based on the RMSECV statistic. From the same Figure we also see that the curves for the training set does not give an indication that the model created with a degree 6 orthogonalization will be significantly better than the model with only D_2-regularization. We note that the corresponding curves for GCV look very similar to the LOOCV curves. This shows that using LOOCV and GCV for model validation can be problematic.

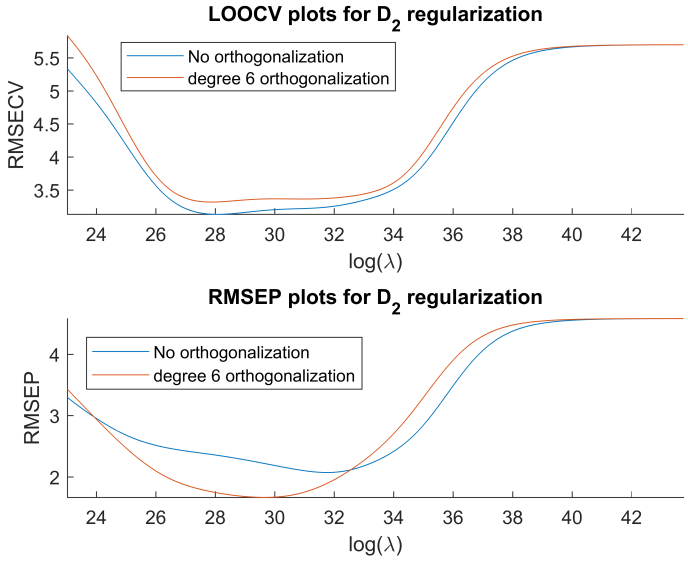

Figure 3: *Fish oil data with ESMC(6) pre-processing. LOOCV and RMSEP plots for models with D_2 regularization.*

The optimal regression coefficients for the models with an additional orthogonalization criterion are plotted in Figure 4. As can be seen from Table 1 the regression coefficients obtained

using derivative regularization and extra orthogonalization perform better on the test set than the regression coefficients obtained from L_2 regularization. This will clearly not be the case in general, but often the loss in prediction will be relatively small. For smaller data sets such as the one discussed here, the computation of the regression coefficients for the PLS models and the three regularization types considered does not take more than a minute on a personal computer. A possible strategy for modeling is thus to generate models from all families and select the final model based on the performance on e.g. a validation set. If this is done, then clearly a split into training, validation and test set is preferable if an estimate of predictive power is also wanted.

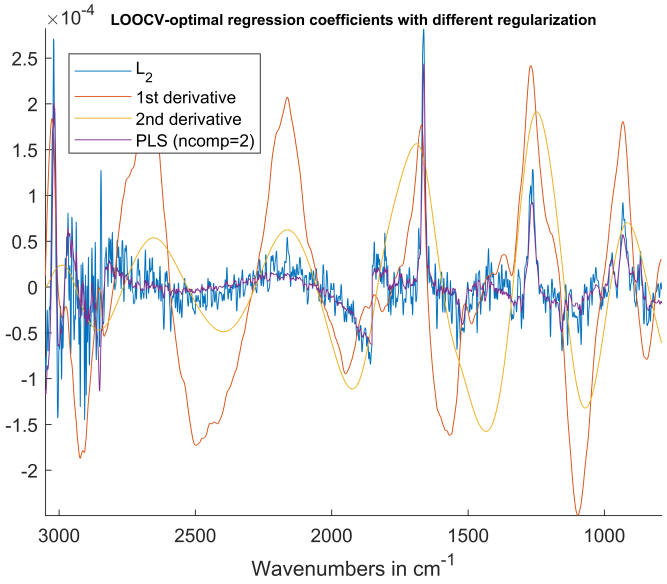

Figure 4: *Fish oil data with ESMC(6) pre-processing. LOOCV-optimal regression coefficients for different regularizations and an additional orthogonalization criterion in the TR problem (constant term omitted). See Table 1.*

One problem with the regression coefficients obtained using derivative regularization is that the extra criterion can force structure on the regression coefficients that is not supported by the data. From Figure 1 we can, for example, see that we do not expect non-zero regression coefficients in the area corresponding to roughly $1800cm^{-1} - 2600cm^{-1}$. Comparing this to the coefficients in Figure 4 we see that the coefficients with derivative regularization have non-zero coefficients in this area. There are several ways to remedy this problem if one wants smooth regression coefficients, and the easiest way is perhaps to use some form of wavelength selection.[4] One possibility is to add an additional criterion to the TR problem in the form of a diagonal matrix with large entries in the columns corresponding to the wavelengths we want to exclude. This results in regression coefficients with local norm smoothing in this area. The regression coefficients are shown in Figure 5. We see that that this results in regression coefficients that are

zero for wavenumbers $1800cm^{-1} - 2600cm^{-1}$ and continuous on the border of this region. On the test set the RMSEP of the model with a diagonal matrix added to the TR problem with 2nd derivative regularization is 1.73. For comparison, PLS coefficients with the same wavelength selection where also calculated. The calculation for PLS was done by excluding the columns of the data matrix corresponding to the wavenumbers we want to exclude from the regression problem, and afterwards inserting an appropriately sized zero vector into the obtained regression coefficients. For this data set the RMSECV-curve is very flat so that choosing the PLS model from the model with minimum RMSECV value results in a sub-optimal model with 4 components (with an RMSEP of 2.55). Manual inspection of the RMSECV-curve shows that a model with 2 components is much more reasonable (the resulting model has an RMSEP of 1.32). In Figures 4 and 5 we see that we can generate regression coefficients that have very different profiles but also have similar predictive power, showing that one should be very careful when interpreting regression coefficients. The problem of interpreting regression coefficients and how very different regression coefficients can have similar predictive power is a well-known problem. [27]

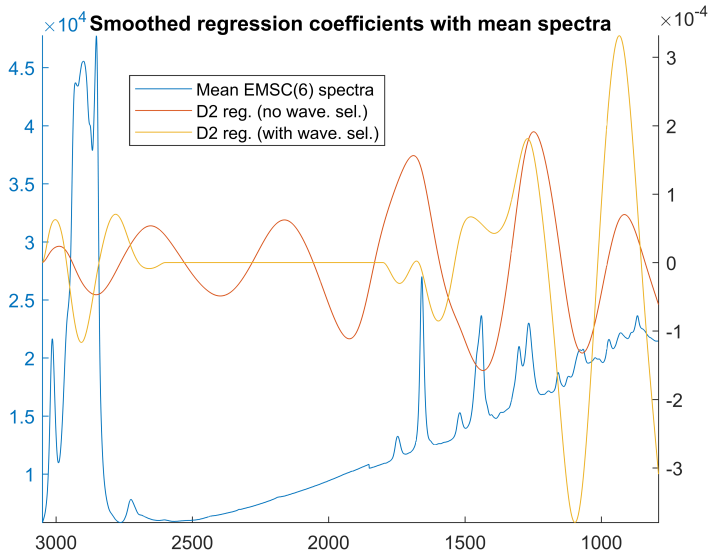

Figure 5: *Fish oil data with ESMC(6) pre-processing. Plot of mean EMSC(6) pre-processed spectra and regression coefficients with 2nd derivative smoothing (with an extra orthogonalization criterion in the TR problem) with and without wavelength selection (constant term omitted). We can make the regression coefficients zero in a region where we do not expect any chemical information by appending an extra criterion to the TR problem.*

Finally we consider using MSC to pre-process the spectra and create models as before with L_2 regularization. The results are summarized in Table 2. We can see that including a degree 6 orthogonalization improves the prediction, but the prediction is still different from the prediction from using EMSC pre-processing.

Orthogonalization	Reg.	Optimal λ (LOOCV)	Min. RMSECV (LOOCV)	RMSEP (LOOCV)	RMSEP (GCV)
TR (No orth.)	L_2	$4.53 \cdot 10^7$	3.72	2.39	2.70
TR (Degree 6)	L_2	$1.56 \cdot 10^7$	3.38	2.31	2.30
TR (No orth.)	D_1	$5.92 \cdot 10^9$	3.85	2.98	2.91
TR (Degree 6)	D_1	$3.52 \cdot 10^{10}$	3.70	2.21	2.13
TR (No orth.)	D_2	$6.81 \cdot 10^{13}$	3.90	3.04	2.97
TR (Degree 6)	D_2	$1.67 \cdot 10^{12}$	3.97	2.03	1.95
PLS (3 components)	NA	NA	3.71	2.21	NA

Table 2: *Fish oil data with MSC pre-processing. Comparison of properties of the regression coefficients. The orthogonality column refers to which polynomials (if any) are added as an additional criterion to the TR problem.*

The difference can mostly be explained by the different estimates of the multiplicative scalars. If we use MSC to pre-process the spectra and do TR with L_2 regularization, but replace the estimates of the multiplicative scalars with the ones obtained from the EMSC pre-processed spectra the RMSEP of the models obtained using GCV falls from 2.70 to 2.01, which is reasonably close to the estimate obtained using EMSC pre-processing. This example shows that the polynomials chosen in the EMSC pre-processing not only affect the regression coefficients by what is subtracted from the spectra, but can also impact the prediction by affecting the estimates of the mulitplicative effects.

5.2 Adipose data

We will now investigate a data set of Raman spectra of fat from pork adipose tissue.[28] This data set was also analyzed in Liland et al.[24] The data set consists of 77 samples, with 50 samples being used for training. From the data we made 500 random partitions into training and test sets. We will perform a similar analysis as for the previous data set, but we will primarily report the mean results from these 500 different partitions. There are 4967 wavenumbers evenly distributed in the range $120cm^{-1} - 3099.6cm^{-1}$ after trimming. The response variables are monounsaturated fatty acids (MUFA), polyunsaturated fatty acids (PUFA), iodine values, and saturated fatty acids (SFA). Here we only look at the responses MUFA and iodine value as the results for PUFA and SFA are similar to the results for MUFA and iodine value. As with the fish oil data, we use EMSC with a degree 6 polynomial correction to pre-process the data. The raw spectra and the corrected spectra for one partition of the data set are plotted in Figure 6. After pre-processing the data, much of the variation between the spectra is removed. We note, however, that there is still large variation in the spectra in particular in the region $1310cm^{-1} - 1420cm^{-1}$. This variation could be removed from the spectra by adding a term representing this interferent to the EMSC pre-processing (this is done in Liland et al.[24]), or from only the model by adding an interferent term to the TR problem.

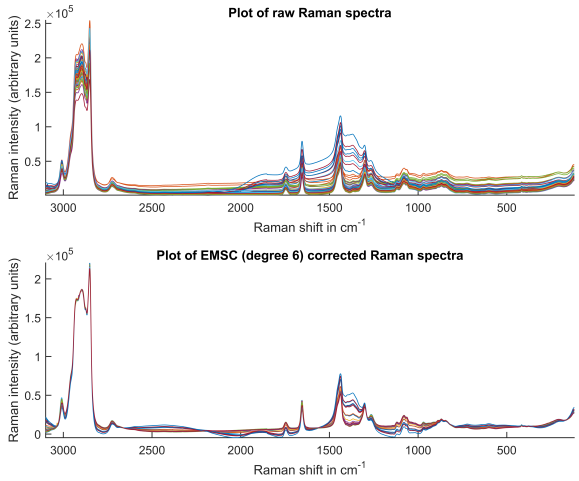

Figure 6: *Top: Raw Raman spectra of adipose tissue. Bottom: EMSC(6) processed Raman spectra of adipose tissue.*

As with the previous data set we see that there is a large region (again corresponding roughly to the wavenumbers $1800cm^{-1} - 2600cm^{-1}$) in Figure 6 where we do not expect non-zero regression coefficients, but we will have non-zero regression coefficients for D_1 and D_2 regularization as a consequence of the smooth derivative criterion. We will therefore also create regression models where we have excluded these wavelengths. We note that these are roughly the same wavelengths that are excluded in Olsen et al.[28]

We will perform the same analysis as on the previous data set: We create TR models using L_2, D_1 and D_2 regularization, and also create a PLS model for comparison (using EMSC(6) pre-processing for all methods). The number of components in the PLS model was chosen using LOOCV. We begin by considering the MUFA response. The mean results from the 500 train/test set splits are summarized in Table 3 and Table 4, and LOOCV optimal regression coefficients for one particular train/test set split are plotted in Figures 7 and 8.

For PLS the mode number of components is 8 both with and without wavelength selection. From Table 3 we see that the inclusion of an extra orthogonalization criterion in the TR problem generally improves prediction. Including wavelength selection also improves prediction for all models. The effects of the extra orthogonalization criterion in the TR problem and wavelength selection is most apparent for 2nd derivative regularization. Including both the extra orthogonality criterion and wavelength selection for 2nd derivative regularization results in a more than 30% improvement on RMSEP, making the models created using 2nd derivative regularization comparable to the other models.

Consider next the iodine response and the results given in Table 4. In this case the mode

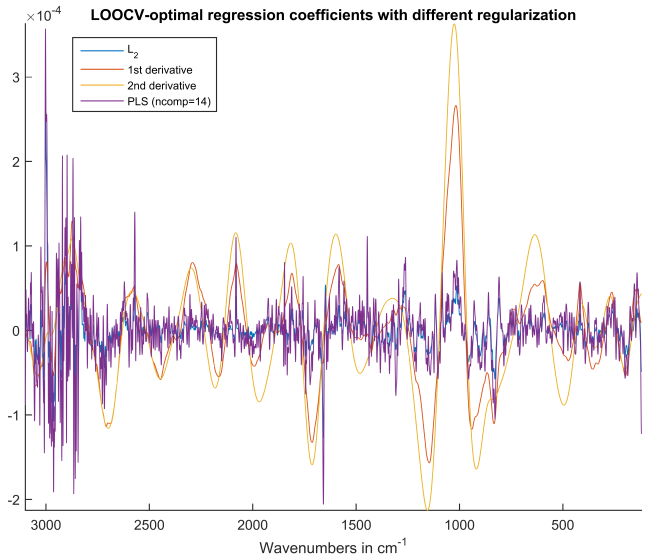

Figure 7: *Adipose data with ESMC(6) pre-processing. LOOCV-optimal regression coefficients for MUFA prediction with different regularizations (constant term omitted). See Table 3.*

number of PLS components is 5 without wavelength selection and 4 components with wavelength selection. For this response the extra orthogonality criterion has very little effect on both RMSECV and RMSEP. For the iodine response we also see that wavelength selection has a large negative effect on the models using 2nd derivative regularization. The bad results here are partly explained by roughly 5 of the training/test splits giving a very large RMSEP, but even removing these splits the 2nd derivative models still perform worse than the other models. This shows that incorporating wavelength selection can also worsen model performance. We also note that although the RMSECV is reasonably close to the RMSEP for most models, this only holds because we are calculating average values over many different splits of the data set. On a single split of the data set the RMSECV is not necessarily a good indicator of model performance.

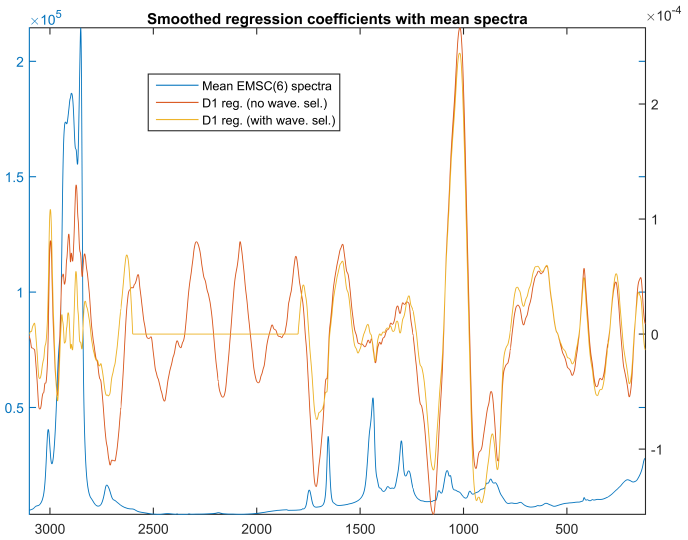

Figure 8: *Plot of mean EMSC(6) pre-processed spectra and regression coefficients with 1st derivative smoothing with and without wavelength selection (constant term omitted) for predicting MUFA. See Table 3.*

	Orthogonalization	Reg.	Optimal λ (LOOCV)	Min. RMSECV (LOOCV)	RMSEP (LOOCV)	RMSEP (GCV)
No wave. sel.	TR (No orth.)	L_2	$7.87 \cdot 10^6$	0.98	1.04	1.07
	TR (Degree 6)	L_2	$7.87 \cdot 10^6$	0.98	1.04	1.07
	TR (No orth.)	D_1	$3.14 \cdot 10^{13}$	1.38	1.42	1.21
	TR (Degree 6)	D_1	$1.45 \cdot 10^{13}$	1.23	1.26	1.19
	TR (No orth.)	D_2	$1.02 \cdot 10^{18}$	1.78	1.85	1.68
	TR (Degree 6)	D_2	$6.96 \cdot 10^{17}$	1.62	1.70	1.45
	PLS	NA	NA	0.97	1.06	NA
With wave. sel.	TR (No orth.)	L_2	$8.19 \cdot 10^6$	0.97	1.03	1.05
	TR (Degree 6)	L_2	$8.00 \cdot 10^6$	0.97	1.02	1.04
	TR (No orth.)	D_1	$3.50 \cdot 10^{13}$	1.38	1.40	1.14
	TR (Degree 6)	D_1	$6.18 \cdot 10^{11}$	1.00	1.03	1.02
	TR (No orth.)	D_2	$1.12 \cdot 10^{13}$	1.18	1.26	1.24
	TR (Degree 6)	D_2	$6.83 \cdot 10^{12}$	1.09	1.14	1.13
	PLS	NA	NA	0.97	1.05	NA

Table 3: *Predicting MUFA from Adipose data with EMSC(6) pre-processing. Above thick line: Without wavelength selection. Below thick line: With wavelength selection. All numbers are mean values for 500 randomized splits of the data into training and test sets.*

	Orthogonalization	Reg.	Optimal λ (LOOCV)	Min. RMSECV (LOOCV)	RMSEP (LOOCV)	RMSEP (GCV)
No wave. sel.	TR (No orth.)	L_2	$8.42 \cdot 10^7$	1.01	1.01	1.00
	TR (Degree 6)	L_2	$8.42 \cdot 10^7$	1.01	1.01	1.00
	TR (No orth.)	D_1	$2.10 \cdot 10^{11}$	1.02	1.04	1.04
	TR (Degree 6)	D_1	$2.25 \cdot 10^{11}$	1.02	1.04	1.04
	TR (No orth.)	D_2	$6.85 \cdot 10^{14}$	1.06	1.10	1.12
	TR (Degree 6)	D_2	$1.10 \cdot 10^{15}$	1.07	1.13	1.14
	PLS	NA	NA	1.02	1.04	NA
With wave. sel.	TR (No orth.)	L_2	$7.18 \cdot 10^7$	1.00	1.00	0.99
	TR (Degree 6)	L_2	$7.08 \cdot 10^7$	1.00	0.99	0.98
	TR (No orth.)	D_1	$1.71 \cdot 10^{11}$	1.02	1.03	1.02
	TR (Degree 6)	D_1	$1.50 \cdot 10^{11}$	1.01	1.01	1.00
	TR (No orth.)	D_2	$3.26 \cdot 10^{14}$	1.86	1.52	1.66
	TR (Degree 6)	D_2	$3.71 \cdot 10^{14}$	2.04	1.40	1.62
	PLS	NA	NA	1.02	1.05	NA

Table 4: *Predicting iodine value from Adipose data with EMSC(6) pre-processing. Above thick line: Without wavelength selection. Below thick line: With wavelength selection. All numbers are mean values for 500 randomized splits of the data into training and test sets.*

6 Conclusions

Using the SVD Tikhonov regularization with GCV for model selection can be implemented very efficiently. The examples considered here demonstrates that the GCV performs very similar to using LOOCV for selecting the regularization parameter in TR. As the GCV statistic can be calculated very efficiently we recommend using GCV for selecting the regularization parameter in TR. For data where multiplicative effects are present, these effects should be corrected prior to model building as the Tikhonov Regularization framework cannot correct for them directly. This can be done for example using (E)MSC or SNV. With Tikhonov regularization we can also easily impose extra criteria on our regression coefficients. Here domain knowledge is important, as for example knowing which wavenumbers of spectra contain useful chemical information can be incorporated into the model to give better predictions. Smooth regression coefficients can be obtained by using derivative regularization and can in some cases improve the predictive power of the models. We have shown that using derivative regularization can impose structure on the regression coefficients that are not supported by the data, so that some form of wavelength selection can be useful for derivative regularization. The addition of polynomial corrections as an extra criterion to the TR problem is not necessary for L_2 regularization if the correction is made for the training set, but for derivative regularization a polynomial criterion in the TR problem is in general necessary to obtain regression coefficients orthogonal to unwanted polynomial trends. For the examples included in this paper, the models created using TR were comparable to the models created using PLS. As the model generation in TR is done quickly one can quickly generate optimal models from several model families and afterwards make a decision about which model to use.

7 Acknowledgements

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

8 Appendix - Prototype MATLAB code

```
function [b, lambda, gcv, bcoefs,  U, s, V] = TregGCV(X,y,lambdas, dtype, otype, fminbndMax)
% dtype — Degree of derivative regularization, dtype=0 gives L2 regularization
% otype — polynomial trend to correct for in TR problem

if nargin < 6
    fminbndMax = 1e20;
end

function gcv = gcvValue(lambda)
    D = bsxfun(@plus,s2,lambda);
    b = V * bsxfun(@rdivide, (U'*[y;zeros(otype+1,1)]).*s, D);
    H = (U.^2) * bsxfun(@rdivide, s2, D) + 1/n; H = H(1:n,:);
    gcv = sum(bsxfun(@rdivide,bsxfun(@minus, y, X(1:n,:)*b),(1—repmat(mean(H,1),n,1))).^2)';
end

[n,p] = size(X); mX = mean(X); my = mean(y);
X = X—ones(n,1)*mX; y = y—my;
mu = 1e24;

if otype >= 0, P = Plegendre(otype, p); X = [X; sqrt(mu)*P']; end
if dtype > 0, L = diff([speye(p);sparse(dtype,p)],dtype); X = X/L; end % Standardizing if using derivative regularization

[U, S, V] = svd(X,'econ'); s = diag(S); s2 = s.^2;
D = bsxfun(@plus,s2,lambdas); % Factor in the bcoefs & H calculations below
bcoefs = V*bsxfun(@rdivide,(U'*[y;zeros(otype+1,1)]).*s,D);
H = (U.^2)*bsxfun(@rdivide,s2,D)+1/n; H = H(1:n,:); % Matrix of leverage—values (one column per lambda—value)
% The following three lines calculates the GCV statistic for lambda values given as input and find the lambda with minimum GCV statistic
gcv = sum(bsxfun(@rdivide,bsxfun(@minus, y, X(1:n,:)*bcoefs),(1—repmat(mean(H,1),n,1))).^2)';
[~,id] = min(gcv);
lambda = lambdas(id);
% The line below uses fminbnd to numerically find an optimal lambda value
[lambda, gcv] = fminbnd(@(x) gcvValue(x),0,fminbndMax);

if dtype > 0, bcoefs = L\bcoefs; end % Transform regression coeffs to match original X—data.
if dtype > 0, b = L\b; end

b = [my—mX*b; b];                % Regression coeffs with constant term of minimum GCV—model.
bcoefs = [my—mX*bcoefs; bcoefs]; % GCV—optimal regression coefficients

end

function [Q, R] = Plegendre(d,l)
% The function generates vectors representing the polynomial trends we correct for in the TR problem
% Generate 'd' 'l'—dimensional orthonormal vectors corresponding to the
% Legendre—polynomials up to degree 'd':
P = ones(l,d+1);
x = (—l:2/(l—1):l)';
for k = 1:d
    P(:,k+1) = x.^k;
end
[Q,R] = qr(P,0);

end
```

References

[1] Rinnan, Å. Pre-processing in vibrational spectroscopy - when, why and how. *Anal. Methods* **2014**, *6*, 7124–7129.

[2] Chen, Z.-P.; Morris, J.; Martin, E. Extracting Chemical Information from Spectral Data with Multiplicative Light Scattering Effects by Optical Path-Length Estimation and Correction. *Analytical Chemistry* **2006**, *78*, 7674–7681, PMID: 17105158.

[3] Andries, E.; Kalivas, J. H. Interrelationships between generalized Tikhonov regularization, generalized net analyte signal, and generalized least squares for desensitizing a multivariate calibration to interferences. *Journal of Chemometrics* **2013**, *27*, 126–140.

[4] Kalivas, J. H. Overview of two-norm (L2) and one-norm (L1) Tikhonov regularization variants for full wavelength or sparse spectral multivariate calibration models or maintenance. *Journal of Chemometrics* **2012**, *26*, 218–230.

[5] Stout, F.; Kalivas, J. H. Tikhonov regularization in standardized and general form for multivariate calibration with application towards removing unwanted spectral artifacts. *Journal of Chemometrics* **2006**, *20*, 22–33.

[6] Vogt, F.; Steiner, H.; Booksh, K.; Mizaikoff, B. Chemometric Correction of Drift Effects in Optical Spectra. *Appl. Spectrosc.* **2004**, *58*, 683–692.

[7] Engel, J.; Gerretzen, J.; Szymańska, E.; Jansen, J. J.; Downey, G.; Blanchet, L.; Buydens, L. M. Breaking with trends in pre-processing? *Trends in Analytical Chemistry* **2013**, *50*, 96 – 106.

[8] Liland, K. H.; Rukke, E.-O.; Olsen, E. F.; Isaksson, T. Customized baseline correction. *Chemometrics and Intelligent Laboratory Systems* **2011**, *109*, 51 – 56.

[9] Afseth, N. K.; Kohler, A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemometrics and Intelligent Laboratory Systems* **2012**, *117*, 92 – 99.

[10] Liland, K. H.; Almøy, T.; Mevik, B.-H. Optimal Choice of Baseline Correction for Multivariate Calibration of Spectra. *Appl. Spectrosc.* **2010**, *64*, 1007–1016.

[11] Barnes, R. J.; Dhanoa, M. S.; Lister, S. J. Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. *Appl. Spectrosc.* **1989**, *43*, 772–777.

[12] Geladi, P.; MacDougall, D.; Martens, H. Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat. *Appl. Spectrosc.* **1985**, *39*, 491–500.

[13] Gautam, R.; Vanga, S.; Ariese, F.; Umapathy, S. Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Techniques and Instrumentation* **2015**, *2*, 8.

[14] Rinnan, Å.; van den Berg, F.; Engelsen, S. B. Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry* **2009**, *28*, 1201 – 1222.

[15] Ref. 121985Geladi et al.Geladi, MacDougall, and Martens, p. 495.

[16] Martens, H.; Stark, E. Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis* **1991**, *9*, 625 – 635.

[17] Fearn, T.; Riccioli, C.; Garrido-Varo, A.; Guerrero-Ginel, J. E. On the geometry of SNV and MSC. *Chemometrics and Intelligent Laboratory Systems* **2009**, *96*, 22 – 26.

[18] Hoerl, A. E.; Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67.

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

[20] Friedman, J.; Hastie, T.; Tibshirani, R. *The elements of statistical learning*; Springer series in statistics Springer, Berlin, 2009; Vol. 1.

[21] Forrester, J. B.; Kalivas, J. H. Ridge regression optimization using a harmonious approach. *Journal of Chemometrics* **2004**, *18*, 372–384.

[22] Golub, G. H.; Heath, M.; Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **1979**, *21*, 215–223.

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

[24] Liland, K. H.; Kohler, A.; Afseth, N. K. Model-based pre-processing in Raman spectroscopy of biological samples. *Journal of Raman Spectroscopy* **2016**, *47*, 643–650.

[25] Afseth, N. K.; Wold, J. P.; Segtnan, V. H. The potential of Raman spectroscopy for characterisation of the fatty acid unsaturation of salmon. *Analytica Chimica Acta* **2006**, *572*, 85 – 92.

[26] Afseth, N. K.; Segtnan, V. H.; Wold, J. P. Raman Spectra of Biological Samples: A Study of Preprocessing Methods. *Appl. Spectrosc.* **2006**, *60*, 1358–1367.

[27] Brown, C. D.; Green, R. L. Critical factors limiting the interpretation of regression vectors in multivariate calibration. *TrAC Trends in Analytical Chemistry* **2009**, *28*, 506 – 514.

[28] Olsen, E. F.; Rukke, E.-O.; Flåtten, A.; Isaksson, T. Quantitative determination of saturated-, monounsaturated- and polyunsaturated fatty acids in pork adipose tissue with non-destructive Raman spectroscopy. *Meat Science* **2007**, *76*, 628 – 634.