Norwegian University of Life Sciences
Faculty of Science and Technology

Philosophiae Doctor (PhD)
Thesis 2019:99

Efficient model selection in the Tikhonov Regularization framework and pre-processing of spectroscopic data

Effektiv modellutvelgelse i Tikhonov Regulariseringsrammeverket og preprosessering av spektroskopisk data

Joakim Skogholt

Efficient model selection in the Tikhonov Regularization framework and pre-processing of spectroscopic data

Effektiv modellutvelgelse i Tikhonov Regulariseringsrammeverket og preprosessering av spektroskopisk data

Philosophiae Doctor (PhD) Thesis

Joakim Skogholt

Norwegian University of Life Sciences
Faculty of Science and Technology

Ås (2019)

Thesis number 2019:99
ISSN 1894-6402
ISBN 978-82-575-1663-5

Supervisory team

Ulf Geir Indahl, Associate Professor (main supervisor)
Faculty of Science and Technology
Norwegian University of Life Sciences

Kristian Hovde Liland, Associate Professor (co-supervisor)
Faculty of Science and Technology
Norwegian University of Life Sciences

Tormod Næs, Professor (co-supervisor)
Nofima
Ås, Norway

Arkadi Ponossov, Professor (co-supervisor)
Faculty of Science and Technology
Norwegian University of Life Sciences

Evaluation committee

Ole Christian Lingjærde, Professor (1st opponent)
Department of Informatics, Research Group for Biomedical Informatics
University of Oslo

Åsmund Rinnan, Associate Professor (2nd opponent)
Faculty of Science
University of Copenhagen, Denmark

Oliver Tomic, Associate Professor (committee coordinator)
Faculty of Science and Technology
Norwegian University of Life Sciences

Acknowledgement

This thesis is submitted to attain the doctoral degree Philosophiae Doctor (PhD) at the Norwegian University of Life Sciences (NMBU). The work has been carried out in the period 2014-2019 at the Faculty of Science and Technology (Realtek) with Ulf Geir Indahl as the main supervisor. This thesis consists of an introduction and 4 enclosed papers.

I would like to thank the people who made this thesis possible. I would like to thank my main supervisor Ulf Geir Indahl, as well as my co-supervisors Kristian Hovde Liland, Tormod Næs, and Arkadi Ponossov. The patience, guidance, ideas, and discussions have been invaluable. My PhD years has been some of the most interesting years of my life. I have learned so much, and I feel like a different person now compared to when I started. I am very grateful for the guidance I have received from my supervisors that made this possible.

I would also like to thank the section for teaching and teacher education for first hiring me at NMBU.

Finally, I would also like to thank my family for support and love. My parents Harald and Gro, my brother Kasper, my wife Karina, and my beautiful daughter Eva.

Joakim Skogholt
Ås, October 2019

Abstract

Machine learning is a hot topic in today's society. Data sets of varying sizes show up in a number of contexts, and learning from data sets is important for answering many questions. There is a plethora of methods that can be used to extract information from data, and in this thesis we consider primarily the Tikhonov Regularization (TR) framework for regularized linear least squares modeling. TR is a very flexible modeling framework, in the sense that it is easy to adjust the type of regularization used as well as including a priori information about the regression coefficients.

The main topic of this thesis is efficient model selection in the TR framework. When using TR regularization for modeling it is necessary to specify one or more model parameters, often called regularization parameters. The regularization parameter can have a significant effect on the quality of the final model, and choosing an appropriate regularization parameter is therefore an important part of the modeling. For large data sets model selection can be time consuming, and it is therefore of interest to obtain efficient methods for selecting between different models. In Paper I it is shown how generalized cross validation can be used for efficient model selection in the TR framework. This discussion continues in Paper III where it is shown how leave-one-out cross validation can be done efficiently in the TR framework. Paper III also suggests a heuristic that can be used for efficient model selection when dealing with data sets with repeated measurements of the same physical sample.

Raw data often needs to pre-processed before useful models can be created. Papers I and II deal with pre-processing and modeling of vibrational spectroscopic data in the extended multiplicative signal correction (EMSC) framework. In the EMSC framework unwanted effects in the data are modeled as multiplicative and additive effects. In Paper I it is shown how the correction of additive effects can be done while creating a regression model in the TR framework and why this can in some cases be advantageous. The multiplicative correction in EMSC is based on a single reference spectrum, but for data sets with very different spectra a single reference spectrum might not be sufficient to accurately correct for multiplicative effects in the measured spectra. Paper II discusses how to extend the EMSC framework to include multiple reference spectra as well as how appropriate reference spectra can be obtained automatically.

Paper IV considers classification using regularized linear discriminant analysis (RLDA). The link between RLDA and regularized regression is used to argue that the efficient validation criteria discussed in papers I and III also can be used for model validation in RLDA. This is tested empirically and the results indicate that good choices of the regularization parameter can be obtained efficiently using a regression-based criterion.

Sammendrag

Maskinlæring er et populært tema i dagens samfunn. Datasett med varierende størrelse dukker opp i mange ulike sammenhenger, og det er av interesse å hente ut informasjon fra datasett. Det er utviklet et bredt utvalg med metoder som kan brukes til å lære fra data, og i denne avhandlingen så fokuserer vi på Tikhonov regulariseringsrammeverket (TR) for regularisert lineær minste kvadraters modellering. TR-rammeverket er veldig fleksibelt i den forstand at det er enkelt å endre typen regularisering, og det er også mulig å inkludere a priori informasjon om regresjonskoeffisientene.

Det gjennomgående temaet i denne avhandlingen er effektiv modelseleksjon i TR-rammeverket. Når man bruker TR-rammeverket så må man spesifisere et eller flere modellparametere, som i denne sammenhengen ofte kalles for regulariseringsparametere. Modellparameteret har betydelig påvirkning på kvaliteten til den endelige modellen, og det er mange ulike metoder som kan brukes for å estimere en god parameterverdi. For store datasett så kan dette være veldig tidskrevende, og det er derfor av interesse å undersøke effektive metoder for å velge blant ulike modeller. I artikkel I så vises det hvordan generalisert kryssvalidering (GCV) kan gjøres effektivt i TR-rammeverket. Denne diskusjonen fortsetter i artikkel III, der det blir vist hvordan også leave-one-out kryssvalidering kan gjøres effektivt i TR-rammeverket. I artikkel III så foreslås det også en heuristikk som kan brukes til effektiv modellutvelgelse for datasett med gjentatte målinger av den samme fysiske prøven.

Rådata må ofte bearbeides og preprosesseres før modellbygging. Artikkel I og artikkel II tar for seg preprosessering av spektroskopiske data i 'extended multiplicative scatter correction' (EMSC) rammeverket. I EMSC-rammeverket så modelleres uønskede effekter i dataene som en kombinasjon av additive og multiplikative effekter. I artikkel I så vises det hvordan korrigering av additiv støy kan gjøres i TR-rammeverket i modelleringsstadiet, og det drøftes når dette kan være hensiktsmessig. Skaleringen i EMSC rammeverket er basert på ett enkelt referansespekter. I datasett der det er stor variasjon mellom de ulike spektrene så er ikke ett referansespekter alltid nok. I artikkel II så diskuteres det hvordan EMSC-rammeverket kan utvides slik at flere referansespektre kan brukes, og det diskuteres også hvordan man kan finne slike referansespektre.

Temaet i artikkel IV er klassifikasjonsproblemer. I artikkelen så brukes sammenhengen mellom regularisert lineær diskriminantanalyse (RLDA) og regularisert regresjon for å argumentere for at modelseleksjonskriteriene fra artikkel I og III også kan brukes til modelseleksjon i RLDA. Dette testes empirisk, og resultatene tyder på at man kan få et godt parametervalg i RLDA ved å bruke et regresjonsbasert kriterie.

List of papers

I. Baseline and interferent correction by the Tikhonov regularization framework for linear least squares modeling

II. Preprocessing of spectral data in the extended multiplicative signal correction framework using multiple reference spectra

III. Model selection by Fast virtual Cross Validation in Ridge Regression and the Tikhonov Regularization framework

IV. Fast identification of good Regularization Parameter Values for Regularized Linear Discriminant Analysis by Cross-validated Ridge Regression

Contents

Acknowledgement ii
Abstract iv
Sammendrag vi
List of papers viii

1 Introduction **1**
1.1 Background and overview 1
1.2 Vibrational spectroscopic data 1
1.3 Regression modeling 4
1.4 Classification with regularized linear discriminant analysis 6
1.5 Validation 8

2 Summaries of papers **11**
2.1 Paper I - Baseline and interferent correction by the Tikhonov regularization framework for linear least squares modeling 11
2.2 Paper II - Preprocessing of spectral data in the extended multiplicative signal correction framework using multiple reference spectra . . 11
2.3 Paper III - Model selection by Fast virtual Cross Validation in Ridge Regression and the Tikhonov Regularization framework 12
2.4 Paper IV - Fast identification of good Regularization Parameter Values for Regularized Linear Discriminant Analysis by Cross-validated Ridge Regression 12

3 Discussion **13**
3.1 Contribution 13
3.2 Future perspectives 14

References **18**

Paper I **19**

Paper II **39**

Paper III **52**

Paper IV **101**

1 Introduction

1.1 Background and overview

Technological advancements allow us to more easily generate large data sets at low costs in many fields of science. Data types such as RNA sequencing data, spectroscopic data, and data arising from nuclear magnetic resonance spectroscopy can consist of thousands (or even tens of thousands) of variables[14, 1, 34]. Large data sets not only arise from methods in the natural sciences. Automatic recognition of handwritten digits is an important problem for automatic sorting of post. More recent problems include spam detection in e-mail and object recognition in images. These problems can naturally be divided into regression problems (where the objective is to predict some numerical quantity from a sample, for example the percentage of fat in a sample from a NIR spectrum) and classification problems (where the objective is to determine class membership for a sample, for example which digit a handwritten digit is). For data sets where the number of features exceed the number of samples it is often not possible to use methods from classical statistics[14]. This is often referred to as the '$p > n$ problem', and there is a need for more methods for analyzing such data[14, 31]. One method for dealing with this problem is to use some form of regularization.

In this thesis we consider modeling primarily using the Tikhonov Regularization (TR) framework for linear least squares modeling. The TR framework is very flexible in the sense that it is straightforward to incorporate additional information and restrictions on the regression coefficients. This makes it possible to include domain knowledge in the model building. Depending on the regression problem considered it is necessary to select a value for one or more regularization parameters. Many methods for choosing a regularization parameter exists (see e.g. [14, 31, 16, 24]), and the computational cost varies between the different methods. For large data sets the computational cost associated with selecting models can limit the number of models one realistically can choose between.

We will show how model selection can be done highly efficiently in the TR framework, allowing for efficient experimentation with a wide variety of models. The flexibility of the TR framework is illustrated mostly on vibrational spectroscopic data in this thesis. Although the TR framework is a regression framework, the results may also be helpful in classification problems when using regularized discriminant analysis. It is well-known that there is a close relationship between linear discriminant analysis and linear regression, and that a similar relationship holds for regularized linear regression and regularized linear discriminant analysis[19, 18]. Because of this relationship it is worth asking whether the efficient model selection methods for the TR framework can also be used for efficient model selection in regularized linear discriminant analysis, and we will argue that this is indeed the case.

1.2 Vibrational spectroscopic data

As most of the data sets considered in this thesis comes from spectroscopy a short discussion of spectroscopic data is included. Spectroscopy deals with the interaction of electromagnetic radiation and matter[6]. Different types of spectroscopy are classified according to which part of the electromagnetic spectrum is used as well as the underlying physical effect that give rise to the spectra. The data sets used in this thesis are mainly Raman spectroscopic data, but also near infrared (NIR) spectroscopic data is used. Both types of spectroscopy study the vibration of molecules but the underlying physical effects are different.

For NIR spectroscopy a light beam is sent towards a sample and the transmitted or reflected light is measured. By varying the wavelength of the light the transmitted or reflected light for different frequencies can be measured, which results in a NIR spectrum. A vibration in a molecule is said to be infrared active if it causes a dipole change in the molecule[6]. Raman spectroscopy relies on a scattering effect. In Raman spectroscopy monochromatic light is beamed towards a sample. When light hits the sample some of the light will scatter. Most of the scattered light will scatter at the same frequency as the incident light and this is referred to as Rayleigh scattering. A small amount of the scattered light will change frequency and this is referred to as Raman scattering. The Raman scattering is further divided into Stokes scattering and anti-Stokes scattering. A decrease in energy in the scattered photon is referred to as Stokes scattering, and an increase in energy of the scattered photon is called anti-Stokes scattering. As most molecules will be in a ground state the Stokes scattering is more intense than the anti Stokes scattering. See Figure 1 for an illustration of the involved energy transitions. The photons scattered at different wavelengths are counted, and the result is a Raman spectrum. The criterion for a vibration to be Raman active is that it changes the polarizability of the molecule[6].

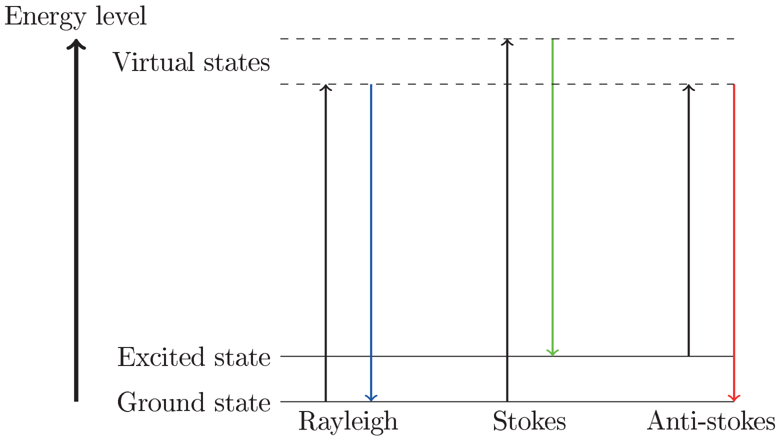


Figure 1: Illustration of the energy changes involved with different types of light scattering in Raman spectroscopy. For Rayleigh scattering the molecule returns to its original state and the scattered light has the same frequency as the incident light. For Stokes scattering the molecule returns to an excited state, decreasing the energy (and therefore frequency) of the scattered photon. For anti-stokes scattering the molecule is initially in an excited state and returns to the ground state, resulting in the energy (and therefore frequency) of the scattered photon increasing.

The basic idea for the applications we consider is that different molecular bonds vibrate at different frequencies, and so the NIR and/or Raman spectrum of a sample gives information about the chemical contents of the sample. By measuring individual spectra for a collection of samples and computing a quantity of interest using, for example, methods from wet chemistry, one can then use multivariate analysis to construct models that can estimate the same quantity from only a spectrum. This is useful because acquiring a spectrum can often be done cheaper, faster, non-invasive, and often with little to none sample preparation compared with other more direct methods[2]. It is often assumed that the quantity of interest depends linearly on the intensity of the spectrum so that linear modeling can be used. For NIR this can be justified using the Beer-Lambert Law[1], which states that the absorption of light in a sample will be proportional to the product of the concentration of what is absorbing the light and the path length of light through the sample.

In practice spectroscopic data also contains unwanted effects that make it difficult to analyze raw data[29, 1, 32]. This could be as simple as random noise or different signals due to inhomogeneous samples, but there are several physical effects that can make modeling challenging. For NIR-spectroscopy it follows from the Beer-Lambert law that variations in path length of light through a sample as well as sample thickness will change the measured transmittance. These effects affect the absorbance multiplicatively and make it difficult to interpret if the apparent difference in absorption between different spectra are due to chemical differences between the samples[1]. Light scattering can also affect spectra[32]. There could also be variations in the fraction of transmitted light collected by the detector[27]. In Raman spectroscopy fluorescence can cause a large baseline in the spectra making it hard to determine what part of the signal comes from Raman scattering[29, 1]. Raman spectra can also be affected by cosmic spikes[11] resulting in large spikes in the spectra, and instrument detector shifts[29] which results in a small ’jump’ in the spectra. Considerable research has been conducted to find pre-processing methods that remove these unwanted effects without removing useful information from the data (see e.g. [33, 15, 7, 30, 1]). Mathematically we can model unwanted effects as a combination of additive and multiplicative effects. Multiplicative effects are corrected for by some sort of scaling procedure, and for additive effects many methods essentially fit a low-degree polynomial to each spectrum and subtract this polynomial. In this thesis we pre-process data primarily using the extended multiplicative signal correction (EMSC) framework[30, 1]. Let $\boldsymbol{X}$ denote a $n \times p$ matrix of spectra consisting of n samples with measurements at p wavenumbers. In the basic EMSC model[1] each spectrum $\boldsymbol{x}$ is decomposed as a sum of the form

$$\boldsymbol{x} = a \cdot \mathbf{1} + b \cdot \boldsymbol{x}_{ref} + c_1 \cdot \boldsymbol{v}_1 + c_2 \cdot \boldsymbol{v}_2 + \boldsymbol{e},$$

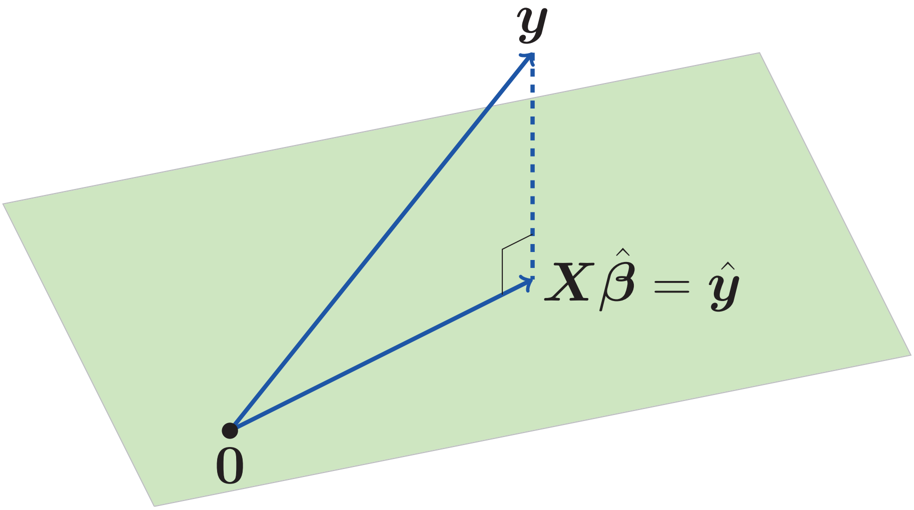
where $\boldsymbol{x}_{ref}$ is a reference spectrum, the vectors $\mathbf{1}, \boldsymbol{v}_1, \boldsymbol{v_2}$ are a basis of the linear space of polynomials of degree 2 and $\boldsymbol{e}$ is the residual. The scalars a, b, c_1, c_2 are obtained using ordinary linear least squares (OLS) regression. The corrected spectrum is given by

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

From the formula for the correction we see that the projection onto the reference spectrum is used to scale the data, and that the polynomials model additive noise. As the reference spectrum will be common to all corrected spectra the chemical information of interest for prediction will be contained in the residual vector $\boldsymbol{e}$[1]. By changing the vectors we project the spectra onto we can adjust the pre-processing from data set to data set. For Raman spectra it can be useful to include higher degree polynomials in the model[1]. If a data set contains interferents this can also be included and corrected for in the model[29]. The EMSC framework has also been extended to correct for Mie scattering[28] as well as replicate correction[26]. In Paper II we discuss the EMSC framework in more detail, and suggest how it can be extended to account for multiple reference spectra.

1.3 Regression modeling

When the data set of interest has been preprocessed (if necessary) methods from multivariate statistics can be used to obtain regression models. Let $\boldsymbol{X}$ be a centred data matrix and $\boldsymbol{y}$ be the associated centred response vector. We consider linear models of the form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{X}$ is an $n \times p$ matrix consisting of n samples and p features, $\boldsymbol{y}$ is the response vector, $\boldsymbol{\beta}$ is the vector of regression coefficients, and $\boldsymbol{\epsilon}$ is the residual. Problems with multiple responses can be solved by considering them as a collection of single-response problems. In ordinary least squares regression we solve the problem $\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2$, which geometrically corresponds to finding the projection of the response vector onto the column space of the data matrix. See Figure 2 for an illustration. In the discussion below we assume that $p > n$, that is that the number of features exceeds the number of samples.



$\boldsymbol{y}$
$\boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{y}}$
$\boldsymbol{0}$

Figure 2: In the ordinary linear least squares problem the vector $\boldsymbol{y}$ of responses is projected onto the column space of the data matrix.

For data sets where the number of features is larger than the number of samples the OLS approach cannot be used directly. Even if OLS could be applied directly this is not always desirable. By using a more flexible model family it is often possible to obtain a model with better predictive performance than the OLS model. An example of a situation where the application of OLS is not desirable (even if it would be possible) can be seen in spectroscopic data which is often high-dimensional, but the chemical information of interest normally lies in a low-dimensional subspace. Some form of dimension reduction is therefore often an important part of modeling. A common method for reducing the dimension of a data set is principal component analysis (PCA)[14]. Principal component regression (PCR) can then be used for model building. In PCR one reduces the dimension of the data to obtain a regression model by finding directions in the data set explaining the most variance. The subspace spanned by the directions explaining the most variance is used for creating the regression model while the remaining directions are discarded. This can be implemented using the singular value decomposition (SVD). To make this more precise let the SVD of the data matrix be given by $\boldsymbol{X} = \boldsymbol{USV}'$. The data can be reduced to a lower dimension by truncating the SVD to only include some of the components. The regression coefficients can then be found by projecting the response vector onto the selected subspace. Mathematically the formula for the regression coefficients are given by the following expression[14]:

$$\boldsymbol{\beta}_k = \sum_{i=1}^{k} \frac{\boldsymbol{u}_i'\boldsymbol{y}}{s_i}\boldsymbol{v}_i,$$

where k is the number of dimensions/components included in the model. In PCR one chooses basis vectors maximizing the explained variance in the data set without considering the response vector. In regression problems we are interested in predicting some quantity, and it makes sense to instead find directions explaining most of the covariance between the data matrix and the response, rather than directions explaining only the variance in the data matrix. This is the idea behind partial least squares (PLS) regression. Mathematically PLS can be viewed as a Krylov subspace method[9, 17, 12]. We use the notation $\mathcal{K}_k(\boldsymbol{A}, \boldsymbol{b})$ for the Krylov subspace spanned by the vectors $\{\boldsymbol{b}, \boldsymbol{Ab}, \boldsymbol{A}^2\boldsymbol{b}, \ldots, \boldsymbol{A}^{k-1}\boldsymbol{b}\}$. PLS solves the following problem[9]:

$$\min_{\boldsymbol{\beta}_k \in \mathbb{R}^p} \|\boldsymbol{X}\boldsymbol{\beta}_k - \boldsymbol{y}\|^2 \text{ subject to } \boldsymbol{\beta}_k \in \mathcal{K}_k(\boldsymbol{X}'\boldsymbol{X}, \boldsymbol{X}'\boldsymbol{y}),\ k = 1, 2, \ldots,$$

where again k refers to the number of components included in the model. There are several algorithms that can be used to solve the PLS optimization problem[5] (notably including Householder bidiagonalization as well as Golub-Kahan-Lanczos bidiagonalization[9]), but not all the algorithms in the literature have good numerical properties[10].

This thesis primarily considers regression modeling in the Tikhonov Regularization (TR) framework[17, 24]. There are many different least squares problem that can be formulated in this framework[24], but a fairly general version of the optimization problem is finding the least squares solution of the following system of linear equations:

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

This is an augmented version of the standard linear least squares problem $\boldsymbol{X\beta} = \boldsymbol{y}$ and the least squares solution minimizes the expression

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{X\beta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{L\beta}\|^2,$$

where the matrix $\boldsymbol{L}$ above is some regularization operator, and the scalar $\lambda > 0$ is the regularization parameter. Note that the regularization parameter controls the trade-off between the two terms in the optimization criterion. Choosing $\boldsymbol{L} = \boldsymbol{I}$ results in L_2-regularization which is also referred to as Ridge Regression[20]. For $\boldsymbol{L} = \boldsymbol{I}$ it is straightforward[14, 17] to show using the SVD of $\boldsymbol{X}$ that the regression coefficients are given by

$$\boldsymbol{\beta} = \sum_{i=1}^{n} \frac{s_i^2}{s_i^2 + \lambda^2} \cdot \frac{\boldsymbol{u}_i'\boldsymbol{y}}{s_i}\boldsymbol{v}_i.$$

From the formula we see that this can be viewed as the regression coefficients from PCR where each term in the sum is multiplied by a scalar less than 1. Per Christian Hansen[17] refers to these scalars as 'filter factors'. This multiplication has a regularizing effect, and for TR dimension reduction is essentially obtained by increasing the regularization parameter. With TR we therefore do not directly project the data onto a lower dimensional space, and as all directions in the sample space contribute to the regression coefficients the dimension reduction can be said to be more 'soft' compared to PCR[17]. The regression problem given above can be modified by adding new rows to the data matrix. This can be used to add additional types of regularization, but also other restrictions to the regression coefficients[24].

1.4 Classification with regularized linear discriminant analysis

In classification problems the goal is to assign a sample to one of several given classes rather than predicting some numerical quantity. If a sample is given by a vector in $\mathbb{R}^p$ and the classes are numbered $1, 2, \ldots, g$ the classification problem can be formulated mathematically as finding a function $f : \mathbb{R}^p \to \{1, 2, \ldots, g\}$ that takes a sample as input and returns the correct class label as output. One well-studied classification method is linear discriminant analysis[14], where the original formulation is due to Fisher[13]. LDA can be motivated both geometrically and statistically. The geometrical idea is to project the data onto a subspace that is good for classification and use a distance based classifier in this subspace. In Fisher's description of LDA this subspace is defined as a subspace where samples of the same class are mapped close to each other but samples of different classes are mapped far apart. To make this mathematically precise, let $\boldsymbol{X}_s$ denote the mean-centered data matrix, and let $\boldsymbol{X}_g$ denote the group centred data matrix. We can then define the total, within group, and between group scatter matrices as $S_T = 1/n \cdot \boldsymbol{X}_s'\boldsymbol{X}_s$, $\boldsymbol{S}_W = 1/n \cdot \boldsymbol{X}_g'\boldsymbol{X}_g$, and $\boldsymbol{S}_B = 1/n \cdot \sum_{k=1}^{g} n_k(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{u})$, where μ is the global mean, μ_k is the mean of class k, and n_k is the number of samples belonging to class k. It can be shown that $\boldsymbol{S}_B = \boldsymbol{S}_T - \boldsymbol{S}_W$[18]. We then seek a subspace that maximizes the Rayleigh quotient associated with the between group scatter relative to the within group scatter:

$$\max_{\boldsymbol{v}\in\mathbb{R}^p} \frac{\boldsymbol{v}'\boldsymbol{S}_B\boldsymbol{v}}{\boldsymbol{v}'\boldsymbol{S}_W\boldsymbol{v}}.$$

This problem can be recast in multiple equivalent ways[22]. The optimization criterion amounts to solving the generalized eigenvalue problem $\boldsymbol{S}_B\boldsymbol{v} = \lambda\boldsymbol{S}_W\boldsymbol{v}$ which, if the within group scatter matrix is invertible, becomes the ordinary eigenvalue problem $\boldsymbol{S}_W^{-1}\boldsymbol{S}_B\boldsymbol{v} = \lambda\boldsymbol{v}$. Alternatively, we can maximize $\boldsymbol{v}'\boldsymbol{S}_B\boldsymbol{v}$ subject to the condition that the $\boldsymbol{v}$-vectors have unit norm in the metric induced by the within group scatter matrix[18]. Because $\boldsymbol{S}_B = \boldsymbol{S}_T - \boldsymbol{S}_W$ we can replace $\boldsymbol{S}_W$ by $\boldsymbol{S}_T$ in the above eigenvalue problem. As $\boldsymbol{S}_B$ and $\boldsymbol{S}_W$ are symmetric the solution will be a series of at most $g-1$ orthogonal eigenvectors. Sorting the eigenvectors by the size of the eigenvalues we essentially get a list of directions sorted by how good they are for discrimination (as measured by the Fisher criterion). Classification can then be done by projecting onto all the eigenvectors, or a dimension reduction can be obtained by projecting onto the space spanned by a selection of the eigenvectors[18].

The alternative probabilistic approach to LDA is to assume that when conditioning on class membership, samples from all classes follow a normal distribution where the mean is different for each class, but the covariance is the same for all classes[14]. Classification here amounts to classifying a sample to the nearest class mean in the metric induced by the within class covariance matrix (normally called the Mahalanobis metric). It can be shown that LDA gives linear decision boundaries[14].

As with OLS there are situations where the application of LDA as described above is not desirable, or not even possible. When the number of features exceeds the number of samples the within group scatter matrix will not be invertible and so LDA cannot be applied directly[22]. There are several possible solutions to this problem. One can first use PCA as a dimension reduction tool and then apply LDA to the projected data[8]. In this thesis we consider regularization by adding a regularization matrix to the within group scatter matrix[18]. That is, we replace the within group scatter matrix with a matrix of the form $\boldsymbol{S}_W + \lambda\boldsymbol{L}$ where $\boldsymbol{L}$ is some regularization matrix (typically the identity matrix). We refer to this modification as regularized LDA (RLDA). This can be done for both the Fisher formulation of LDA and the Mahalanobis formulation, and Hastie et al[19, 18] proved that the two approaches are equivalent for classification (by modifying the Fisher version to account for prior probabilities if applicable). Hastie et al[18] also proved that an equivalent classification can be done based on solving a regression problem similar to Ridge regression. More precisely, they define a penalized optimal scoring problem as minimizing

$$\frac{1}{n}\cdot\left(\|\boldsymbol{Y\Theta} - \boldsymbol{X}_s\boldsymbol{B}\|^2 + \lambda\|\boldsymbol{LB}\|^2\right),$$

subject to $\frac{1}{n}\|\boldsymbol{Y\theta}\|^2 = 1$. Here $\boldsymbol{Y}$ is the $n\times g$ matrix with $0-1$ dummy-coded group membership, and the $\boldsymbol{\Theta}$ is a matrix of scores. The $\boldsymbol{\Theta}$ matrix can be found by solving an eigenvalue problem. This can be viewed as a dummy-regression problem where we right-multiply the dummy-coded responses by a post-processor matrix $\boldsymbol{\Theta}$. The regression coefficients will be right-multiplied by the same post-processor, and so we can view this as a change of basis. The equivalence of these different approaches to linear discriminant analysis shows that there is a close relationship between RLDA and regularized regression.

1.5 Validation

PCR, TR, and PLS all require the selection of a model parameter to obtain a model. For PCR and PLS this parameter is the dimension of the subspace we project onto, and for TR it is the regularization parameter. Choosing an appropriate model parameter is crucial to obtaining an appropriate model[17, 14], and it is therefore necessary to have some method for selecting the model parameter. For large data sets it may be computationally expensive to validate models for a large number of candidate parameter values, and it is therefore of interest to develop computationally efficient methods for model selection. In this thesis we primarily use cross validation (and variations of cross validation) for model selection. In k-fold cross validation the data set is first divided into k folds. A model is then created using $(k-1)$-folds of data while excluding one fold from the modeling. The model is then validated by calculating the mean squared error (MSE) on the fold that was held out during modeling. This is repeated k times so that all folds are held out exactly once. The sum of all these mean squared errors provide a measure of model quality. We can then repeat this process for a selection of model parameters (the number of dimensions for PCR and PLS or the regularization parameter for TR) and select the model parameter that gives the smallest MSE under cross validation (or select the simplest model among the models that have low MSE under cross validation). Let $\hat{y}_{(i)}$ be the estimate obtained by a regression model for the ith sample y_i when it is held out during model training. The mean squared error under cross-validation can then be written as

$$\sum_{i=1}^{n}(y_i - \hat{y}_{(i)})^2.$$

The case where $k = n$ (the number of folds equal the number of samples) is referred to as leave-one-out cross validation (LooCV) and in this case the model statistic is Allens PRESS-statistic[3, 4]. Another criterion that can be used for model selection is the generalized cross-validation (GCV)[16]. The GCV is meant to be '*a rotation invariant version of Allen's PRESS*'[16], and it can be motivated using statistical arguments.

Papers I and III discuss efficient model selection in the TR framework. In Papers I and III it is shown how calculating a single SVD of the centred data matrix allows for very efficient computation of the GCV (Papers I and III) and PRESS-statistic (Paper III) for any choice of regularization parameter value. Slightly more precise, when the SVD has been calculated the cost of calculating the PRESS-statistic (or GCV) for an additional regularization parameter value is roughly two matrix-vector multiplications. This allows us to efficiently compare models for a large number of regularization parameter values. To illustrate this, we consider an example from Paper III. The data set used is NIR spectra of gasoline and consists of $n = 60$ samples and $p = 401$ features[23]. The response variable is the octane number of the sample. We used 40 samples for model training, and considered TR-models with

L_2, first derivative, and second derivative regularization (see the paper for more details). In Figure 3 the data is plotted together with the regression coefficients minimizing the PRESS-statistic for each type of regularization.

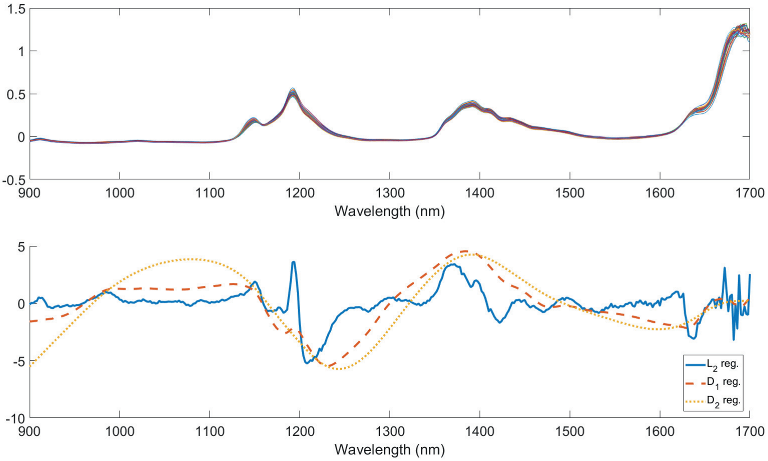


Figure 3: Top: NIR spectra of the gasoline. Bottom: Regression-coefficients minimizing the PRESS-statistic for L_2, first derivative and second derivative regularization.

Computing the GCV and PRESS-statistic for 10000 values of the regularization parameter for the three types of regularization considered took a total of about 0.4 seconds on a home computer. This means that a wide variety of model types can be tested without worrying about the time needed for model selection, and we can get very high-resolution PRESS- and GCV-curves. The PRESS- and GCV-curves for this example are shown in Figure 4. We see from Figure 4 that the PRESS- and GCV-curves are very flat for this example. This indicates that it is possible to choose a simpler model (larger regularization parameter) without sacrificing much predictive power in the model. In paper III we consider two methods for choosing a simpler model than the one minimizing the GCV or the PRESS-statistic, and the resulting regularization parameter is also shown in Figure 4. The two methods are the '1 standard error rule'[14] (select the simplest model with a PRESS-statistic within one standard error of the minimum) and the χ^2-rule [21] (select the simplest model that is not statistically significant from the minimum with respect to the chosen significance level).

For data sets where there are multiple measurements of the same sample, the modeling selection strategies discussed above cannot be applied directly due to data leakage. In Paper III a heuristic is suggested to deal with this problem. The idea is to collect all measurements of a single physical sample in one matrix. The rows are then made orthogonal by finding the SVD of this matrix and left-multiplying by the transpose of the matrix of the left singular vectors. When this is done for all the repeated measurements we can apply the above modeling strategies to this transformed data set. Empirical results indicate that the regularization parameter

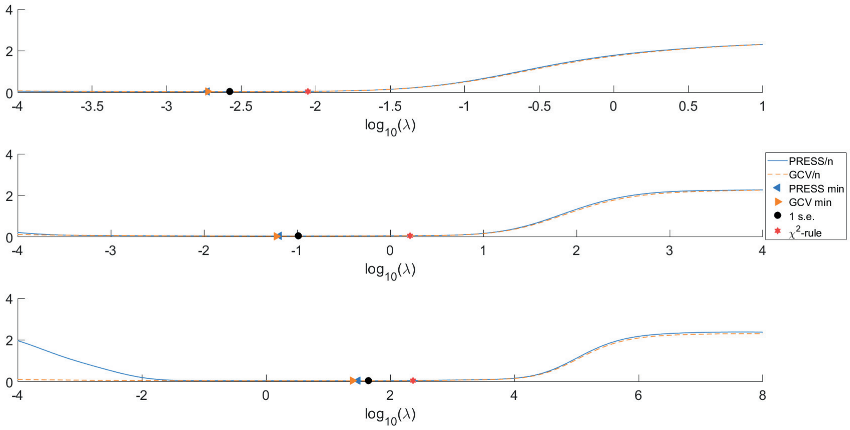


Figure 4: Plot of the PRESS-statistic and GCV divided by the number of samples for the octane data for L_2 regularization (top), first derivative regularization (middle), and second derivative regularization (bottom). Different possible choices of the regularization parameter value is also shown. 1 S.E. refers to the regularization parameter chosen by the ’1 standard error rule’, and ’χ^2-rule’ refers to the regularization parameter chosen by the χ^2-test.

minimizing the PRESS-statistic of the modified data set is approximately the same as the value of the regularization parameter minimizing the cross validation error when applying a segmented cross-validation approach.

Due to the relationship between RLDA and regularized regression it is worth investigating whether the efficient model selection criteria for regression can also be applied to find an appropriate regularization parameter value for LDA. Paper IV is an experimental paper were we tried this, and it appears that choosing a regularization parameter for RLDA using the PRESS-statistic on a $0-1$ coded dummy regression gives a similar regularization parameter as LooCV on classification performance based on the Mahalanobis metric. As the PRESS-statistic can be calculated extremely efficiently this means that a regularization parameter for RLDA can be chosen efficiently.

2 Summaries of papers

2.1 Paper I - Baseline and interferent correction by the Tikhonov regularization framework for linear least squares modeling

Spectroscopic data should generally be pre-processed prior to modeling due to unwanted physical effects and noise in the data. Paper I focuses on the multiplicative signal correction and the extended multiplicative signal correction framework for pre-processing. The paper includes a theoretical discussion of the two preprocessing methods in terms of linear algebra. Further, the paper discusses how parts of the pre-processing can be implemented in the model-building stage when using Tikhonov Regularization and compares the two approaches to pre-processing when using different types of regularization. The effects of derivative regularization on the regression coefficients is discussed, and it is illustrated that requiring global derivative regularization for the regression coefficients can negatively affect model quality. It is also shown how model validation can be done very efficiently using Generalized Cross-Validation once a single SVD of a (possibly modified) centred data matrix has been calculated.

2.2 Paper II - Preprocessing of spectral data in the extended multiplicative signal correction framework using multiple reference spectra

When collecting spectroscopic data the different spectra frequently have very different scales. This can be due to several physical factors, such as path length through a sample or fluorescence (depending on the type of spectroscopy and the measurement). When applying linear modeling for obtaining estimates for some quantity we typically assume that the response is proportional to the intensity of the signal at relevant wavenumbers. Differences in scaling between different spectra that are not due to sample differences can therefore have a big effect on model quality, and an appropriate scaling of the spectra is therefore an important part of pre-processing. In the extended multiplicative signal correction (EMSC) framework this is done by selecting a reference spectrum (typically the mean spectrum) and by normalizing all spectra with respect to the chosen reference spectrum. When there is big variation within the spectra a single reference spectrum may not be appropriate for normalization of all spectra. Paper II shows how the EMSC framework can be extended to include multiple reference spectra by normalizing each spectrum in the subspace spanned by the selected reference spectra. The paper also suggests how the SVD can be used to automatically obtain multiple reference spectra, and discuss when the use of multiple reference spectra is required.

2.3 Paper III - Model selection by Fast virtual Cross Validation in Ridge Regression and the Tikhonov Regularization framework

The underlying topic of Paper III is efficient model selection in the Tikhonov Regularization framework. It is shown that by computing the SVD of the (centred) data matrix once it is possible to perform LooCV for a regularization parameter at the computational cost of approximately two matrix-vector multiplications. This allows for very efficient model selection when considering a large selection of regularization parameters. The paper introduces a heuristic called virtual cross validation for data data sets with repeated measurements of the same physical sample. When dealing with data sets with repeated measurements a LooCV approach cannot be applied directly. This is because holding out one measurement is essentially not reducing the information in the data set due to the other measurements of the same (physical) sample. This results in overfitting to the training data. The problem can be solved by a segmented cross validation approach where each block of repeated measurements is held out in cross validation, but this can be computationally expensive. In Paper III virtual cross validation is presented as a computationally efficient method for model selection with these types of data sets. The idea is to consider each set of repeated measurements as a block of data. The samples within each block is then made orthogonal by finding a (reduced) SVD for each data block and left multiplying by the transpose of the left-singular vectors. This makes the rows within each block orthogonal, and thus for the transformed data set a segmented cross validation approach is equivalent to LooCV. We refer to this method as 'virtual cross validation'. It is shown that in the case where all samples within a block are equal the virtual cross validation is equivalent to segmented cross validation. In general the two methods are not equivalent, but empirically the two methods of model selection appear to produce similar results, while the virtual segmented cross validation is much more computationally efficient. The reason for the computational efficiency of the virtual cross validation compared to the segmented cross validation is that it replaces the computation of the SVDs of a small number of large matrices with the computation of the SVDs of a large number of small matrices.

2.4 Paper IV - Fast identification of good Regularization Parameter Values for Regularized Linear Discriminant Analysis by Cross-validated Ridge Regression

The topic of Paper IV is the application of the fast model selection for regression problems in the TR framework to finding a good value of the regularization parameter for regularized linear discriminant analysis. It is well-known that there is a close link between regularized regression and regularized linear discriminant analysis. It might therefore be possible to use the fast LooCV for regression to compute an appropriate regularization parameter value for RLDA. The paper investigates this idea, and empirical results indicate that a regression based criterion can provide a good choice of regularization parameter. Due to the efficiency of the model selection criterion used for the regression problem this means that a good regularization parameter value for regularized discriminant analysis can be obtained quickly.

3 Discussion

3.1 Contribution

The aim of the thesis was to investigate the use of the TR framework for linear least squares modeling focusing on efficient model selection. This included an investigation of pre-processing methods for spectroscopic data, as well as considering the application of the efficient model selection criteria to classification problems. For spectroscopic data a pre-processing step is often necessary prior to modeling. Paper I discusses the MSC and EMSC pre-processing methods. In terms of linear algebra these methods can be explained as projections onto the subspace spanned by the reference spectra and the polynomial trends and interferents (if any) that are included in the model. Due to the flexibility of the TR framework the removal of unwanted additive effects in the data can be done as a part of the modeling rather than in a pre-processing stage. In Paper I it is illustrated that this can be advantageous when applying derivative regularization in the modeling. Scaling of the spectra cannot be incorporated into the TR framework and must be done prior to modeling. Paper II extends the EMSC framework by showing how multiple reference spectra can be handled. The scaling in EMSC is done by projecting all samples onto a selected subspace and normalizing all samples within a one-dimensional subspace. When dealing with data sets containing outlier spectra or large variation within spectra it may not be sufficient to normalize the data set by projecting each spectrum onto a single reference spectrum. In Paper II we suggest to solve this problem by projecting onto a subspace spanned by a set of reference spectra and normalizing all spectra within this subspace. The reference spectra may be chosen manually, but using the first few right singular vectors of the SVD of the matrix of spectra appears to work well. This is also a reasonable choice of reference spectra as the first few right singular spectra normally will explain most of the variation in spectroscopic data. In the case where almost all the variation in the spectra is explained by the first right singular vector (which will typically be almost equal to the mean spectrum) the addition of multiple reference spectra will have little effect on the pre-processing as the projection onto the other reference spectra will be negligible.

When modeling in the TR framework it is necessary to have some procedure for validating and selecting models. Many methods for validating and selecting regularization parameters are available. Papers I and III discuss how model selection can be done highly efficiently based on a computationally fast version of the LooCV. Paper I shows how this can be done for the GCV criterion, and in Paper III it is shown how this can also be done for the LooCV. The formulae derived in Papers I and III allows for very efficient model selection for a large number of candidate regularization parameter values. This can be done either by sampling a large number of regularization parameter values on a log scale, or by applying a numerical optimizer. The fast model selection formulae can be used for L_2 regularization as well as other types of regularization. This allows for efficient experimentation with a wide variety of regression models.

In classification problems the aim is to classify a sample as belonging to one of several given classes. Paper IV considers the use of regularized discriminant analysis for classification. In classification problems a commonly used criterion for model selection is the number of samples correctly classified using cross validation. It is known that there is a link between RLDA and regularized regression. As Papers I and III discuss efficient methods for model validation for regularized regression it is interesting to investigate whether these efficient parameter selection methods also can be applied to RLDA and classification problems. The empirical results obtained in the Paper suggests that a regression based criterion can be used to obtain a suitable regularization parameter for RLDA.

3.2 Future perspectives

The Tikhonov regularization framework can be extended further, and we have several works in progress which will briefly be discussed here. When modeling data it may be necessary to update models regularly due to, for example, the collection of new samples, change in physical conditions, or new samples with different properties[25]. In this scenario the original model may be useful, but in need of some adjustment. This can be achieved by a modification of the optimization criterion used in the TR framework[25]. With standard L_2-regularized regression the optimization criterion is the weighted sum of the squared errors for the prediction and the L_2 norm of the regression coefficients where the weight is given by the regularization parameter. Instead of requiring the L_2 norm of the regression coefficients to be small, we can require the L_2 norm of the difference between the new and the old regression coefficients to be small. The new regression coefficients will then be regression coefficients that predict the new data well, while at the same time not being to different from the original model. By using arguments similar to the ones used in Paper III, one can develop formulae for fast LooCV of this modified regression problem. This allows for efficiently updating a regression model and finding an appropriate trade-off between keeping the old model while at the same time accounting for new data.

The work in Paper IV discusses how the fast LooCV for regression can be used to obtain a good regularization parameter value for RLDA. The work provides empirical results that an appropriate regularization parameter for RLDA can be obtained from fast LooCV for an associated regression problem. Further study is warranted to establish a more rigorous justification for when this approach appears to parameter selection works, and to establish cases where the regression approach will not yield a good regularization parameter value.

The LooCV discussed in Paper III allows for efficient selection of regression models for different choices of the regularization parameter after computing the (reduced) SVD of the data matrix. The calculation of the SVD can be a computational bottleneck, and this is especially the case for very large data sets. It can be shown that the PRESS-statistic can be computed efficiently from a single bidiagonalization of the data matrix. More precisely, there exists recursion formulae that allows for efficient calculation of the bidiagonalization of the augmented data matrix for any choice of regularization parameter. These recursion formulae can be used to obtain formulae for efficiently computing the PRESS-statistic. Empirical testing indicate that these recursion formulae are sufficiently numerically stable for practical applications. The model update discussed above can also be done using only the bidiagonalization of a data matrix. This will therefore allow for efficient model selection for TR models when considering data sets that are too large for even a single SVD to be computationally feasible. Further, it is well known that PLS can be implemented using the Golub-Kahan-Lanczos bidiagonalization algorithm, and the projection can be obtained by truncating a bidiagonalization of a matrix. This makes the fast LooCV for bidiagonalization interesting not only in terms of the computational savings compared to the SVD, but also because of the additional models we can obtain by combining a truncation of the bidiagonalization with L_2 regularization. The resulting models will be a mix of PLSR and TR models, where we have the full projection onto subspaces provided by PLSR together with the 'soft-tresholding' obtained by TR. Per Christian Hansen[17] refers to this combination as 'the best of both worlds'. These type of models will have two regularization parameters (the parameter from TR and the dimension of the subspace we project onto), and so it will be necessary to investigate efficient parameter selection methods in this case.

References

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

[2] N. K. Afseth, J. P. Wold, and V. H. Segtnan. The potential of raman spectroscopy for characterisation of the fatty acid unsaturation of salmon. *Analytica Chimica Acta*, 572(1):85 – 92, 2006.

[3] D. M. Allen. Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13(3):469–475, 1971.

[4] D. M. Allen. The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1):125–127, 1974.

[5] M. Andersson. A comparison of nine pls1 algorithms. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 23(10):518–529, 2009.

[6] C. N. Banwell, E. M. McCash, et al. *Fundamentals of molecular spectroscopy*, volume 851. McGraw-Hill New York, 1994.

[7] R. J. Barnes, M. S. Dhanoa, and S. J. Lister. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.*, 43(5):772–777, May 1989.

[8] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 19(7):711–720, 1997.

[9] Å. Björck. Stability of two direct methods for bidiagonalization and partial least squares. *SIAM Journal on Matrix Analysis and Applications*, 35(1):279–291, 2014.

[10] Å. Björck and U. G. Indahl. Fast and stable partial least squares modelling: A benchmark study with theoretical comments. *Journal of Chemometrics*, 31(8):e2898, 2017.

[11] F. Ehrentreich and L. Sümmchen. Spike removal and denoising of raman spectra by wavelet transform methods. *Analytical chemistry*, 73(17):4364–4373, 2001.

[12] L. Eldén. Partial least-squares vs. lanczos bidiagonalization-i analysis of a projection method for multiple regression. *Computational Statistics & Data Analysis*, 46(1):11–31, 2004.

[13] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[14] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2009.

[15] P. Geladi, D. MacDougall, and H. Martens. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Appl. Spectrosc.*, 39(3):491–500, May 1985.

[16] G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

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

[18] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, pages 73–102, 1995.

[19] T. Hastie, R. Tibshirani, and A. Buja. Flexible discriminant analysis by optimal scoring. *Journal of the American statistical association*, 89(428):1255–1270, 1994.

[20] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[21] U. Indahl. A twist to partial least squares regression. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 19(1):32–44, 2005.

[22] S. Ji and J. Ye. Generalized linear discriminant analysis: a unified framework and efficient model selection. *IEEE Transactions on Neural Networks*, 19(10):1768–1782, 2008.

[23] J. H. Kalivas. Two data sets of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 37(2):255–259, 1997.

[24] J. H. Kalivas. 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*, 26(6):218–230, 2012.

[25] J. H. Kalivas, G. G. Siano, E. Andries, and H. C. Goicoechea. Calibration maintenance and transfer using tikhonov regularization approaches. *Applied spectroscopy*, 63(7):800–809, 2009.

[26] A. Kohler, U. Böcker, J. Warringer, A. Blomberg, S. Omholt, E. Stark, and H. Martens. Reducing inter-replicate variation in fourier transform infrared spectroscopy by extended multiplicative signal correction. *Applied spectroscopy*, 63(3):296–305, 2009.

[27] A. Kohler, C. Kirschner, A. Oust, and H. Martens. Extended multiplicative signal correction as a tool for separation and characterization of physical and chemical information in fourier transform infrared microscopy images of cryo-sections of beef loin. *Applied spectroscopy*, 59(6):707–716, 2005.

[28] T. Konevskikh, R. Lukacs, R. Blümel, A. Ponossov, and A. Kohler. Mie scatter corrections in single cell infrared microspectroscopy. *Faraday discussions*, 187:235–257, 2016.

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

[30] H. Martens and E. Stark. Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*, 9(8):625 – 635, 1991.

[31] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

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

[33] A. Savitzky and M. J. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[34] F. Savorani, G. Tomasi, and S. B. Engelsen. icoshift: A versatile tool for the rapid alignment of 1d nmr spectra. *Journal of magnetic resonance*, 202(2):190–202, 2010.

PAPER I

Received: 11 January 2017 | Revised: 28 August 2017 | Accepted: 13 September 2017

DOI: 10.1002/cem.2962

RESEARCH ARTICLE

WILEY Journal of CHEMOMETRICS

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

Joakim Skogholt[1] | Kristian Hovde Liland[1,2] | Ulf Geir Indahl[1]

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

[2] Rawmaterial and process optimisation, Nofima Ås, Ås, Norway

Correspondence
Joakim Skogholt, Norwegian University of Life Sciences, Ås, Norway.
Email: joakim.skogholt@nmbu.no

Funding information
Research Council of Norway, Grant/Award Number: 239070

Abstract

Spectroscopic data are usually perturbed by noise from various sources that should be removed prior to model calibration. After conducting a preprocessing 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 noninformative 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, 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. Because of some favorable properties of the singular value decomposition, it is possible to explore the models (including their generalized cross-validation error estimates) associated with a large number of regularization parameter values at low computational cost.

KEYWORDS

multivariate calibration, preprocessing, Tikhonov regularization

1 | INTRODUCTION

Spectroscopic data are often contaminated by various sources of noise and disturbances making analysis and/or interpretations challenging. Preprocessing 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 using 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 preprocessing 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 preprocessing 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 TR 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 TR 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 contain scattering effects that affect the spectra multiplicatively. These effects should be corrected in a preprocessing step prior to model building. Here, we discuss using extended multiplicative signal correction (EMSC) and standard normal variate (SNV) to preprocess data prior to model building. In the examples, we will use EMSC to preprocess the spectra.

For regularization in the TR problem, we will discuss 3 different types of regularizations: (1) L_2 regularization, (2) discrete first derivative regularization, and (3) discrete second derivative regularization. For L_2 regularization without any wavelength selection, we will show that polynomial trends can be corrected for when preprocessing 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 preprocessed spectra, we obtain regression models comparable to Partial Least Squares (PLS) models with EMSC preprocessed spectra.

In the following sections, we give a short review of some common preprocessing 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 2 different data sets of Raman spectra.

2 | PREPROCESSING OF SPECTRAL DATA

2.1 | Preprocessing

Preprocessing 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 EMSC

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 autoscaling 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 follows:

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

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 follows:

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

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 preprocessing). 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, p. 495] 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

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

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 EMSC.[16] The most basic version of the EMSC has the representation

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

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 preprocessing 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 preprocessing):

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

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 preprocessing. This is also pointed out in Rinnan et al,[14] and more details will be given below.

By using the EMSC preprocessing, 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 nonzero. 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 EMSC preprocessing is used and the reference spectrum is taken as the mean spectrum of the training set, then strictly speaking a new EMSC model should be recalculated for each choice of training set, whereas this challenge does not occur when the SNV method is used.

There are also other preprocessing methods that can be used to estimate and correct for scatter effects. One example is the optical path-length estimation and correction,[2] which allows for estimating scatter when the concentration of the components in a sample is known. When using optical path-length estimation and correction, there is also a polynomial correction by projection.

3 | TIKHONOV REGULARIZATION

3.1 | A brief overview of TR for linear least squares modeling

In this section, we briefly review the TR framework for linear least squares modeling. We assume that 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 that 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 as follows:

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

where the regularization parameter λ is considered as fixed. The purpose of the regularization matrix $\boldsymbol{L}$ in (6) and (7) is to impose additional constraints on the regression coefficients and to overcome problems with multicollinearity present in the ordinary least squares 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 $\boldsymbol{XL}^{-1}$ in the place of the original $\boldsymbol{X}$ (see, eg, 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, eg, Hansen[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 SVD of $\boldsymbol{X} = \boldsymbol{USV}'$ (here, $\boldsymbol{S}$ is the diagonal matrix of nonzero singular values, $\boldsymbol{U}$ and $\boldsymbol{V}$ represent the corresponding left and right singular vectors), the solution $\boldsymbol{\beta}$ to (8) simplifies to

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

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 leave-one-out cross-validation (LOOCV) for both TR and PLS. In our experience, minimization 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 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) follows:

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

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 follows:

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

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 Appendix A. The code in Appendix A 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, eg, 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.[23] 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 Appendix A 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 preprocessing 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 Appendix ?? for the required details.

4 | COMPARISON WITH EMSC

4.1 | MSC and EMSC explained by linear algebra

The EMSC preprocessing is used for both eliminating polynomial trends and correcting for scatter effects in spectroscopic data. By using the EMSC preprocessing 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 preprocessing.

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 $W_{MSC} = \{\boldsymbol{1}, \boldsymbol{x}_{ref}\} \subset \mathbb{R}^{p}$ and $W_{EMSC} = \{\boldsymbol{1}, \boldsymbol{x}_{ref}, \boldsymbol{v}_1, \boldsymbol{v}_2\} \subset \mathbb{R}^{p}$, respectively. According to Section 2.2, the formulae for MSC and EMSC preprocessing are given by Equations 3 and 5.

For both types of preprocessing, 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 $W_r = \{\boldsymbol{r}_1, \dots, \boldsymbol{r}_{p-4}\} \subset \mathbb{R}^{p}$ that spans the orthogonal complement of $span(W_{EMSC})$, ie, $span(W_r) = span(W_{EMSC})^{\perp}$ and $\mathbb{R}^{p} = span(W_{EMSC}) \oplus span(W_r)$.

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

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

and

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

for MSC and EMSC, respectively. The first of these equations is obtained by applying the MSC preprocessing 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 preprocessed spectra are clearly located in the subspace spanned by the vectors $\{\boldsymbol{1}, \boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{x}_{ref}\}$.

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:

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

and

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

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 preprocessed spectra, the solution vector will be orthogonal to the trends being corrected for in the EMSC preprocessing. 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 not in general be orthogonal to the trends corrected for in the preprocessing. In this case, adding the extra polynomial block $\sqrt{\mu}\boldsymbol{P}$ to (16) corresponding to the polynomial trends removed in the preprocessing 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 preprocessing 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, eg, EMSC and SNV with trend correction.

The regression coefficients (ie, the model parameters) obtained when using EMSC preprocessing may sometimes represent information considered to be useful for interpretations.[24] When using both MSC preprocessing 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 parameters 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 preprocess the spectra. When using EMSC to correct Raman spectra, it is common to use polynomials up to degree 6 or 7. 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 preprocess the spectra and correct for polynomial trends up to and including degree 6, and we refer to this as EMSC(6) preprocessing.

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. In the process of optimizing the nonnegative regularization parameter, we used a relatively wide interval ranging from 0 to 10^{20} (the upper limit of this interval corresponds to choosing a model that essentially predicts the average response value). For some models we experienced that the minimization process could fail by proposing the right end point value. 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 Appendix A.

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 790 and 3050cm^{-1} (after truncation).

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 preprocess the spectra. The raw spectra and the EMSC(6) preprocessed 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, ie, the mean of the training spectra.

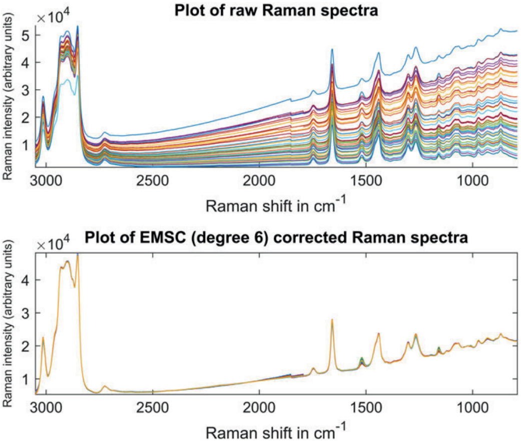


FIGURE 1 Top: raw Raman spectra of salmon oil. Notice in particular the nonlinearities in the baseline. Bottom: EMSC(6) preprocessed Raman spectra of salmon oil. EMSC, extended multiplicative signal correction

Following the steps given at the beginning of Section 5, we generated models for EMSC preprocessed 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) preprocessed data for the results in Table 1, and using MSC preprocessing for the results in Table 2. For each PLS model, the RMSECV was calculated using LOOCV. 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) preprocessed spectra, and a model with 3 components for the MSC preprocessed spectra.

TABLE 1 Fish oil data with EMSC(6) preprocessing

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

Comparison of properties of the regression coefficients. The orthogonality column refers to which polynomials (if any) are added as an additional criterion to the Tikhonov regularization (TR) problem. EMSC, extended multiplicative signal correction; GCV, generalized cross-validation; LOOCV, leave-one-out cross-validation.

TABLE 2 Fish oil data with MSC preprocessing

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

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. GCV, generalized cross-validation; LOOCV, leave-one-out cross-validation; MSC, multiplicative signal correction; TR, Tikhonov regularization.

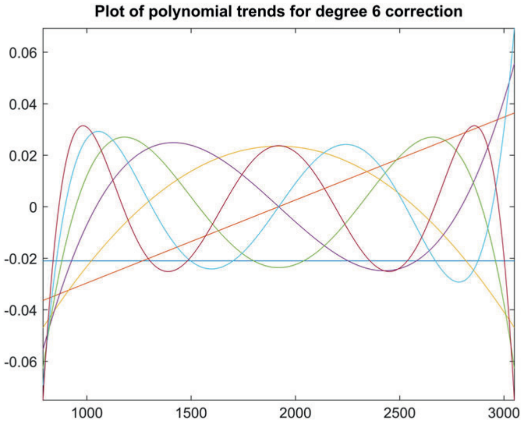


FIGURE 2 Plot of the rows in the matrix $\boldsymbol{P}$ appended to the Tikhonov regularization problem. There are 7 curves as we are correcting for polynomial trends up to and including degree 6

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 ***P*** with the polynomial trends used in this example are plotted in Figure 2.

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 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 preprocessing 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 do 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.

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 3 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, eg, 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.

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 nonzero regression coefficients in the area corresponding to roughly 1800 to 2600cm^{-1}. Comparing this to the coefficients in Figure 4, we see that the coefficients with derivative regularization have nonzero 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 that 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 this results in regression coefficients that are zero for wavenumbers 1800 to 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 second derivative regularization is 1.73. For comparison, PLS coefficients with the same wavelength selection were also calculated. The calculation for PLS was done by excluding the columns of the data matrix corresponding to the wavenumbers that we want to exclude from the regression problem, and afterwards, inserting an appropriately sized zero

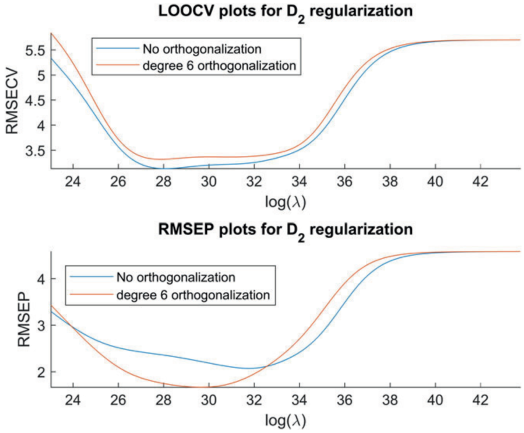


FIGURE 3 Fish oil data with EMSC(6) preprocessing. LOOCV and RMSEP plots for models with D_2 regularization. EMSC, extended multiplicative signal correction; LOOCV, leave-one-out cross-validation

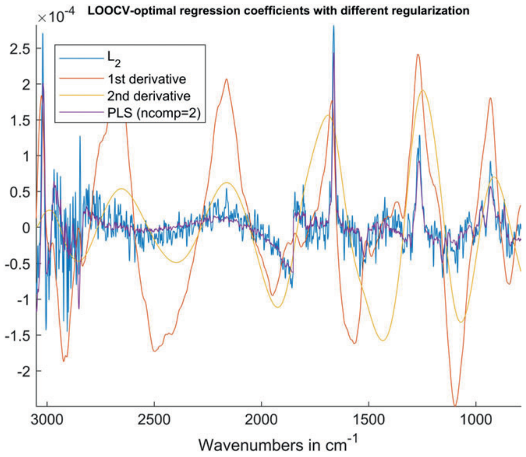


FIGURE 4 Fish oil data with EMSC(6) preprocessing. LOOCV-optimal regression coefficients for different regularizations and an additional orthogonalization criterion in the TR problem (constant term omitted). See Table 1. EMSC, extended multiplicative signal correction; LOOCV, leave-one-out cross-validation; TR, Tikhonov regularization

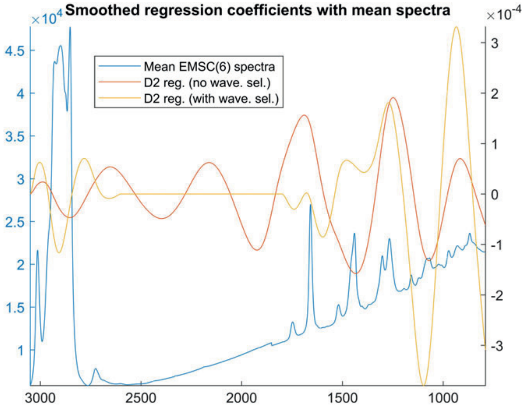


FIGURE 5 Fish oil data with EMSC(6) preprocessing. Plot of mean EMSC(6) preprocessed spectra and regression coefficients with second derivative smoothing (with an extra orthogonalization criterion in the TR problem) with and without wavelength selection (constant term omitted). We can make the regression coefficients zero in a region where we do not expect any chemical information by appending an extra criterion to the TR problem. EMSC, extended multiplicative signal correction; TR, Tikhonov regularization

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 suboptimal 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]

Finally, we consider using MSC to preprocess 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 preprocessing.

The difference can mostly be explained by the different estimates of the multiplicative scalars. If we use MSC to preprocess the spectra and do TR with L_2 regularization, but replace the estimates of the multiplicative scalars with the ones

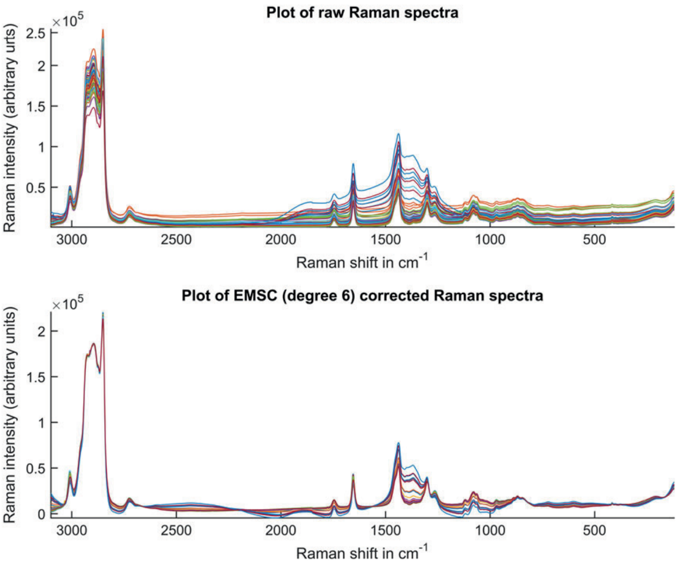


FIGURE 6 Top: raw Raman spectra of adipose tissue. Bottom: EMSC(6) processed Raman spectra of adipose tissue

obtained from the EMSC preprocessed 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 preprocessing 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 120 to 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 preprocess the data. The raw spectra and the corrected spectra for one partition of the data set are plotted in Figure 6. After preprocessing 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 1310 to 1420cm^{-1}. This variation could be removed from the spectra by adding a term representing this interferent to the EMSC preprocessing (this is done in Liland et al[24]), or from only the model by adding an interferent term to the TR problem.

As with the previous data set, we see that there is a large region (again corresponding roughly to the wavenumbers 1800 to 2600 cm^{-1}) in Figure 6 where we do not expect nonzero regression coefficients, but we will have nonzero 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) preprocessing 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 Tables 3 and 4, and LOOCV optimal regression coefficients for one particular train/test set split are plotted in Figures 7 and 8.

TABLE 3 Predicting MUFA from Adipose data with EMSC(6) preprocessing

	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

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.

TABLE 4 Predicting iodine value from Adipose data with EMSC(6) preprocessing

	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

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.

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 second derivative regularization results in a more than 30% improvement on RMSEP, making the models created using second derivative regularization comparable to the other models.

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

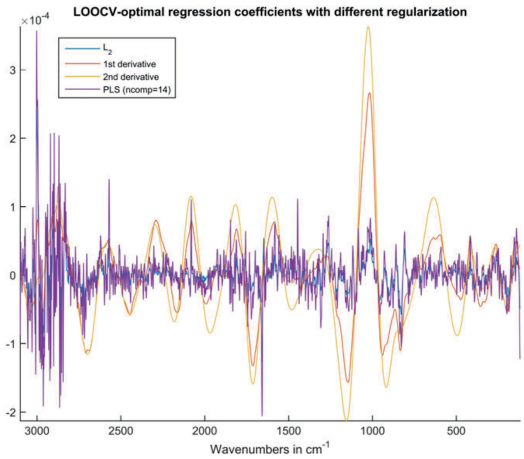


FIGURE 7 Adipose data with EMSC(6) pre-processing. LOOCV-optimal regression coefficients for MUFA prediction with different regularizations (constant term omitted). See Table 3

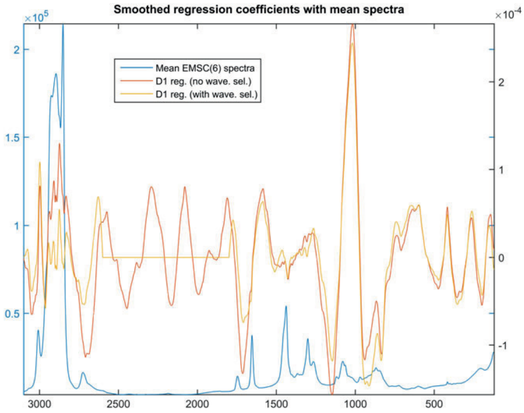


FIGURE 8 Plot of mean EMSC(6) preprocessed spectra and regression coefficients with 1st derivative smoothing with and without wavelength selection (constant term omitted) for predicting MUFA. See Table 3

6 | CONCLUSIONS

Using the SVD TR 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 TR framework cannot correct for them directly. This can be done for example using EMSC or SNV. With TR, 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.

ACKNOWLEDGEMENTS

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

ORCID

Joakim Skogholt http://orcid.org/0000-0001-8511-993X
Kristian Hovde Liland http://orcid.org/0000-0001-6468-9423
Ulf Geir Indahl http://orcid.org/0000-0002-3236-463X

REFERENCES

1. Rinnan Å. Pre-processing in vibrational spectroscopy—when, why and how. *Anal Methods*. 2014;6:7124-7129. https://doi.org/10.1039/C3AY42270D.
2. Chen ZP, Morris J, Martin E. Extracting chemical information from spectral data with multiplicative light scattering effects by optical path-length estimation and correction. *Anal Chem*. 2006;78(22):7674-7681. PMID: 17105158.
3. Andries E, Kalivas JH. Interrelationships between generalized Tikhonov regularization, generalized net analyte signal, and generalized least squares for desensitizing a multivariate calibration to interferences. *J Chemometrics*. 2013;27(5):126-140. https://doi.org/10.1002/cem.2501.
4. Kalivas JH. Overview of two-norm (l2) and one-norm (l1) Tikhonov regularization variants for full wavelength or sparse spectral multivariate calibration models or maintenance. *J Chemometrics*. 2012;26(6):218-230. https://doi.org/10.1002/cem.2429.
5. Stout F, Kalivas JH. Tikhonov regularization in standardized and general form for multivariate calibration with application towards removing unwanted spectral artifacts. *J Chemometrics*. 2006;20(1-2):22-33. https://doi.org/10.1002/cem.975.
6. Vogt F, Steiner H, Booksh K, Mizaikoff B. Chemometric correction of drift effects in optical spectra. *Appl Spectrosc*. 2004;58(6):683-692. http://as.osa.org/abstract.cfm?URI=as-58-6-683.
7. Engel J, Gerretzen J, Szymańska E, et al. Breaking with trends in pre-processing? *Trends in Analytical Chemistry*. 2013;50:96-106. http://www.sciencedirect.com/science/article/pii/S0165993613001465.
8. Liland KH, Rukke EO, Olsen EF, Isaksson T. Customized baseline correction. *Chemom Intell Lab Syst*. 2011;109(1):51-56. http://www.sciencedirect.com/science/article/pii/S0169743911001535.
9. Afseth NK, Kohler A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemom Intell Lab Syst*. 2012;117:92-99. http://www.sciencedirect.com/science/article/pii/S0169743912000494.
10. Liland KH, Almøy T, Mevik BH. Optimal choice of baseline correction for multivariate calibration of spectra. *Appl Spectrosc*. 2010Sep;64(9):1007-1016. http://as.osa.org/abstract.cfm?URI=as-64-9-1007.
11. Barnes RJ, Dhanoa MS, Lister SJ. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl Spectrosc*. 1989;43(5):772-777. http://as.osa.org/abstract.cfm?URI=as-43-5-772.
12. Geladi P, MacDougall D, Martens H. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Appl Spectrosc*. 1985May;39(3):491-500. http://as.osa.org/abstract.cfm?URI=as-39-3-491.
13. Gautam R, Vanga S, Ariese F, Umapathy S. Review of multidimensional data processing approaches for raman and infrared spectroscopy. *EPJ Tech Instrum*. 2015;2(1):1-38. https://doi.org/10.1140/epjti/s40485-015-0018-6.
14. Rinnan Å, van den Berg F, Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry*. 2009;28(10):1201-1222. http://www.sciencedirect.com/science/article/pii/S0165993609001629.
15. Ref. 121985Geladi et al.Geladi, MacDougall, and Martens.
16. Martens H, Stark E. Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy. *J Pharm Biomed Anal*. 1991;9(8):625-635. http://www.sciencedirect.com/science/article/pii/073170859180188F.
17. Fearn T, Riccioli C, Garrido-Varo A, Guerrero-Ginel JE. On the geometry of SNV and MSC. *Chemom Intell Lab Syst*. 2009;96(1):22-26. http://www.sciencedirect.com/science/article/pii/S0169743908002098.
18. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. 1970;12(1):55-67.
19. Hansen P. *Discrete inverse problems*. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2010. http://epubs.siam.org/doi/abs/10.1137/1.9780898718836.
20. Friedman J, Hastie T, Tibshirani R. *The Elements of Statistical Learning*, Vol. 1. Berlin: Springer Series in Statistics Springer; 2009.
21. Forrester JB, Kalivas JH. Ridge regression optimization using a harmonious approach. *J Chemometrics*. 2004;18(7-8):372-384. https://doi.org/10.1002/cem.883.

22. Golub GH, Heath M, Wahba G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*. 1979;21(2):215-223.
23. Kreyszig E. *Introductory Functional Analysis with Applications*, Vol. 81, New York, NY: wiley New York; 1989.
24. Liland KH, Kohler A, Afseth NK. Model-based pre-processing in raman spectroscopy of biological samples. *J Raman Spectroscopy*. 2016;47:643-650. https://doi.org/10.1002/jrs.4886.
25. Afseth NK, Wold JP, Segtnan VH. The potential of raman spectroscopy for characterisation of the fatty acid unsaturation of salmon. *Analytica Chimica Acta*. 2006;572(1):85-92. http://www.sciencedirect.com/science/article/pii/S0003267006009457.
26. Afseth NK, Segtnan VH, Wold JP. Raman spectra of biological samples: a study of preprocessing methods. *Appl Spectrosc*. 2006;60(12):1358-1367. http://as.osa.org/abstract.cfm?URI=as-60-12-1358.
27. Brown CD, Green RL. Critical factors limiting the interpretation of regression vectors in multivariate calibration. *TrAC Trends in Analytical Chemistry*. 2009;28(4):506-514. http://www.sciencedirect.com/science/article/pii/S0165993609000363.
28. Olsen EF, Rukke EO, 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(4):628-634. http://www.sciencedirect.com/science/article/pii/S0309174007000484.

How to cite this article: Skogholt J, Liland KH, Indahl U. Baseline and interferent correction by the Tikhonov regularization framework for linear least squares modeling. *Journal of Chemometrics*. 2018;32:e2962. https://doi.org/10.1002/cem.2962

APPENDIX A: 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 = (—1:2/(l—1):1)';
for k = 1:d
    P(:,k+1) = x.^k;
end
[Q,R] = qr(P,0);

end
```

PAPER II

Received: 9 July 2018 | Revised: 11 October 2018 | Accepted: 28 October 2018

DOI: 10.1002/jrs.5520

RESEARCH ARTICLE

WILEY Journal of RAMAN SPECTROSCOPY

Preprocessing of spectral data in the extended multiplicative signal correction framework using multiple reference spectra

Joakim Skogholt | Kristian Hovde Liland | Ulf Geir Indahl

Faculty of Science and Technology, Norwegian University of Life Sciences, Ås, Norway

Correspondence
Joakim Skogholt, Faculty of Science and Technology, Norges miljo- og biovitenskapelige universitet, Norway.
Email: joakim.skogholt@nmbu.no

Funding information
Research Council of Norway, Grant/Award Number: 239070

Abstract

Extended multiplicative signal correction (EMSC) is a widely used framework for preprocessing 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 preprocessing 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 quares 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), modelling, preprocessing, Raman spectroscopy

1 | INTRODUCTION

Because raw spectral data often contain unwanted artefacts and noise that make modelling and interpretation difficult, some kind of preprocessing is often required.[1–4] The goal of preprocessing spectral data is to transform the raw data into a form that is more suitable for modelling or interpretation. A vast amount of preprocessing methods for spectral data are available. The most widely used preprocessing 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 preprocessing 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 preprocessing 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 data sets containing one or several outlier spectra. By including additional reference spectra that better accounts for the chemical profiles of the outlier spectra, the preprocessing 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 preprocessing 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 PREPROCESSING

When modelling by the traditional EMSC preprocessing framework, the spectra are scaled according to a prespecified 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 $\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_d$. In the traditional EMSC framework, a spectrum $\boldsymbol{x}$ is projected onto the subspace spanned by the vectors in the basis

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

Note that the exact choice of basis vectors in Equation (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 as follows:

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

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 zero, 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 the MSC.

Several extensions of the traditional EMSC model have been proposed in the literature. If any known interferents are also 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 Equation (1) to include the vectors representing the interferents. This results in the following extended set of basis vectors:

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

The correction of a spectrum $\boldsymbol{x}$ is obtained by subtracting its projection onto the subspace spanned by the interferents in Equation (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 Equations (4) and (5) as our starting point. Because 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 Fourier-transform infrared spectroscopy.[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 PREPROCESSING WITH MULTIPLE REFERENCE SPECTRA

The purpose of the reference spectrum in the EMSC preprocessing 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 (a) a strategy for deriving more than one reference spectrum, and (b) a generalization of the EMSC-correction given in Equation (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 preprocessing result that is quite similar to the result obtained by using the mean spectrum as the reference.

In the traditional EMSC preprocessing, a nonorthogonal 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 Equation (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 nonorthogonal 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-factorization. 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 ith 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-factorization used to obtain an orthonormal basis will then produce the associated Legendre polynomials.[13] To distinguish between the traditional nonorthogonal 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-factorization as described above. Let n_{ref} be the total number of reference spectra (identified by the SVD or some other insights), and denote the ith reference spectrum by $\boldsymbol{r}_i$. For the orthonormal basis of the suggested modified EMSC-framework, we use the notation:

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

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, that is, $\alpha_i = (\boldsymbol{v}^o_i)^t \boldsymbol{x}$, $\delta_j = (\boldsymbol{w}^o_j)^t \boldsymbol{x}$ and $\gamma_k = (\boldsymbol{r}^o_k)^t \boldsymbol{x}$. Expressing a spectrum $\boldsymbol{x}$ with respect to this basis therefore yields as follows:

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

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 $\boldsymbol{r}_k^o$'s), that is

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

where the reference combination $\boldsymbol{r_x} = \frac{1}{\sqrt{\sum_{k=1}^{n_{ref}} \gamma_i^2}} \cdot \sum_{i=1}^{n_{ref}} \gamma_i \boldsymbol{r}_i^o$ 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

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

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

Note that for the traditional EMSC preprocessing 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 nonorthogonal to the $\boldsymbol{r_x}$'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

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

The vectors $\boldsymbol{r_x}$ 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 preprocessing. 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.

Preprocessing 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 Equation (6), the spectra preprocessed according to Equation (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 preprocessing, 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 preprocessing is included in the Appendix.

4 | EXAMPLES

In this section, we will compare using the traditional EMSC preprocessing 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 sixth degree is included for all the preprocessing alternatives. No interferents will be added to the preprocessing models. The traditional EMSC framework using the mean spectrum as the reference spectrum will be referred to as simply (standard) EMSC preprocessing. 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 analyzed.[7] After truncation, there are $p = 2263$ wave numbers. The response is the associated measured iodine values. The raw spectra are shown in Figure 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 (PUFAs) quantified as percentage of total weight, and PUFA as percentage of total fat content. The spectra are truncated to the wave numbers $675cm^{-1} - 1770cm^{-1}$. This truncation has been used before when the data set has been analyzed.[19,20] There are a total of $n = 69$ measured samples in the data set, and after truncation there are $p = 1096$ wave numbers. The raw truncated spectra are shown in Figure 2.

For modelling, the following procedure was used: A nested cross-validation strategy was employed to separate preprocessing and parameter optimization from model validation. The outer validation loop was a repeated two-fold (50:50) shuffle-split, whereas the inner optimization loop was a leave-one-out cross-validation (LooCV). For each outer split, the first half of the samples were used to create preprocessing models and subsequently estimate model parameters (using LooCV) for TR and PLS on the preprocessed data. The second half of the outer split was preprocessed 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

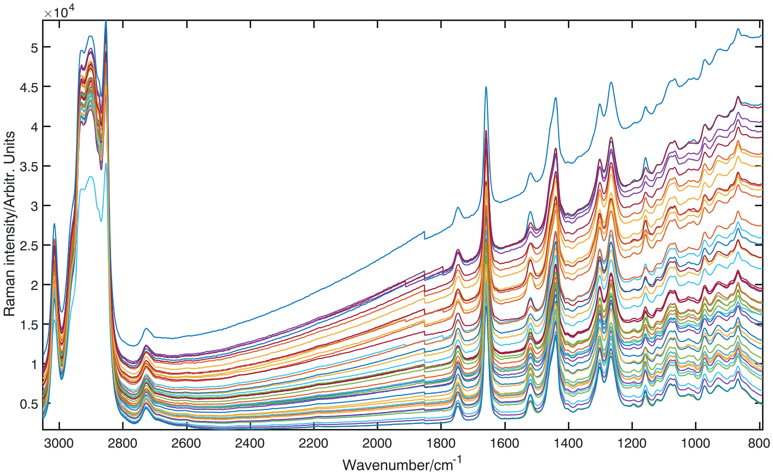


FIGURE 1 Fish oil data: Raw Raman spectra [Colour figure can be viewed at wileyonlinelibrary.com]

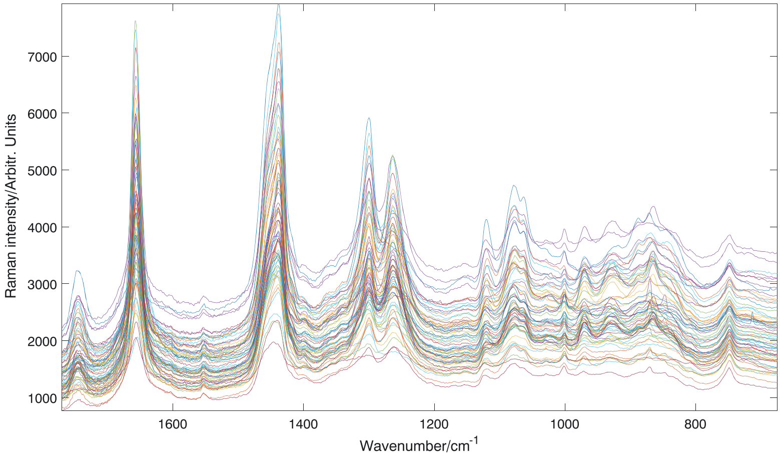


FIGURE 2 Emulsion data: Raw Raman spectra. The spectra have been truncated to the range $675cm^{-1} - 1770cm^{-1}$ [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Fish oil data: Average RMSEP over 500 random data splits

Model \ Preprocessing	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

Note. EMSC: extended multiplicative signal correction; PLS: partial least squares; RMSEP: root mean squared errors of prediction; TR: Tikhonov regularization.

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

Model \ Preprocessing	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

Note. EMSC: extended multiplicative signal correction; PLS: partial least squares; RMSEP: root mean squared errors of prediction; TR: Tikhonov regularization.

derivative regularization were used.[21] 1,000 values of the regularization parameter were selected uniformly on a log scale, and the parameter value minimizing the RMSECV was selected. Note that there is some data leakage for the LooCV in the inner loop as the data was preprocessed 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 Figure 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 reference spectrum to be sufficient to obtain an appropriate scaling. Inspecting

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

Model \ Preprocessing	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

Note. EMSC: extended multiplicative signal correction; PLS: partial least squares; PUFA: polyunsaturated fatty acid; RMSEP: root mean squared errors of prediction; TR: Tikhonov regularization.

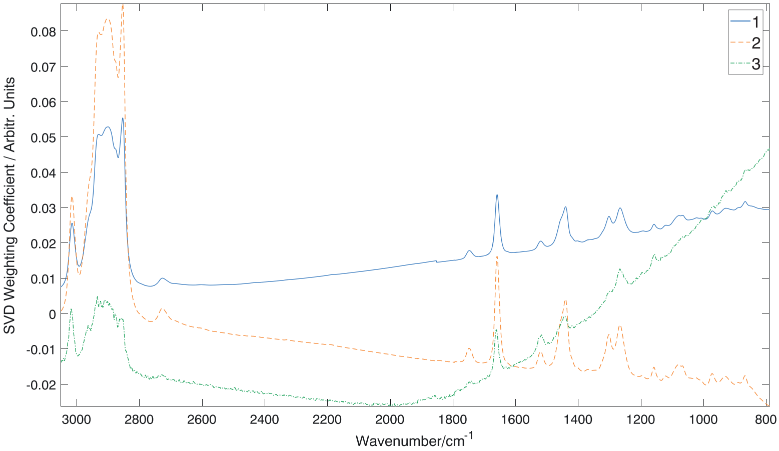


FIGURE 3 Fish oil data: The first three right singular vectors [Colour figure can be viewed at wileyonlinelibrary.com]

the first three right singular vectors of the fish oil data plotted in Figure 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 preprocessing 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 are larger than 10° (corresponding to a correlation lower than 0.9848). For six 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 Figure 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 preprocessed emulsion spectra are plotted in Figure 5 and Figure S2 (Supporting Information). In Figure 5, there is no apparent visual difference between the two preprocessing alternatives, except for the scale difference between the standard EMSC preprocessed spectra and the modified EMSC preprocessed spectra. The similarities between the standard EMSC and EMSC(1) is supported by Figure S1 (Supporting Information), from

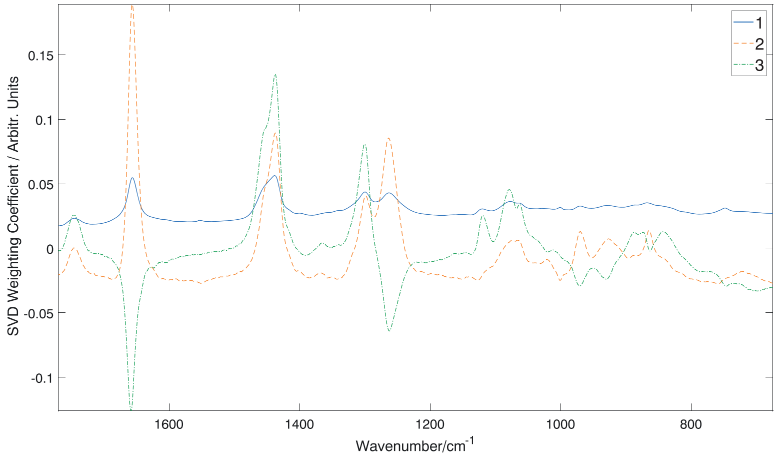


FIGURE 4 Emulsion data: The first three right singular vectors [Colour figure can be viewed at wileyonlinelibrary.com]

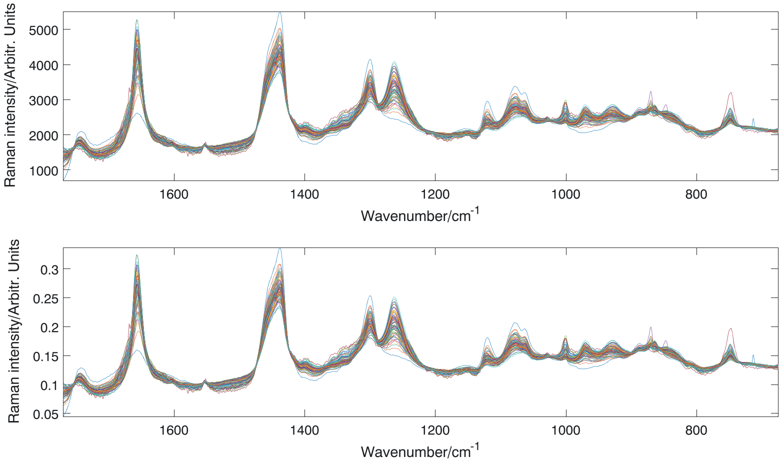


FIGURE 5 Emulsion data: Preprocessed Raman spectra using different preprocessing methods. Top: standard EMSC. Bottom: EMSC(1) [Colour figure can be viewed at wileyonlinelibrary.com]

which it is clear that the mean spectrum and the first right singular vector are very similar. Because the two spectra are that similar, we expect the standard EMSC and EMSC(1) preprocessed 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 Figure S2 (Supporting Information) that this does not result in a huge visual impact on the spectra, with the notable exception of one spectrum (see in particular the peak at about 1445cm^{-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 preprocessing 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 preprocessing is approximately the same as the RMSEP obtained from the EMSC(1) preprocessed 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 Figure 6 and Figure 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

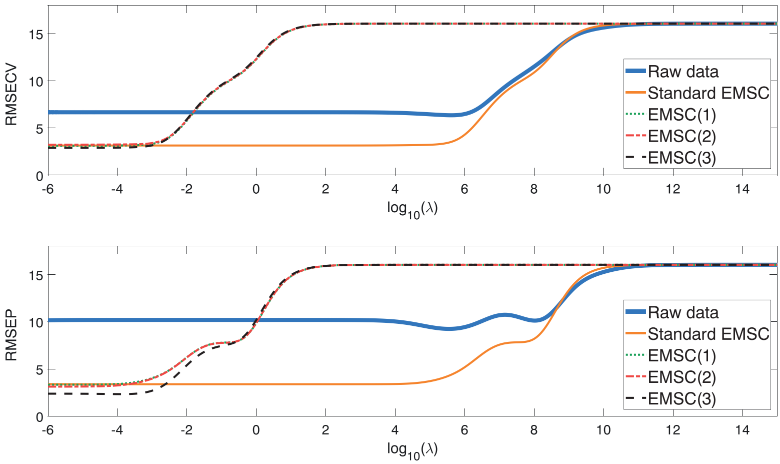


FIGURE 6 Emulsion data: TR modelling (L_2 regularization) for the response PUFA as % of total fat content for a particular split of the data. Top: RMSECV. Bottom: RMSEP. In the top plot we see that the RMSECV curves for the modified EMSC preprocessing are overlapping. In the bottom plot we see that the RMSEP curves for the modified EMSC using 1 and 2 reference spectra are overlapping [Colour figure can be viewed at wileyonlinelibrary.com]

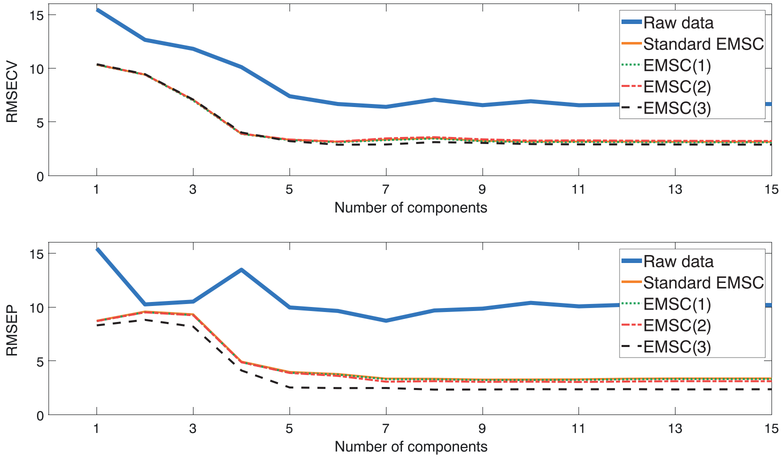


FIGURE 7 Emulsion data: PLS modelling for the response PUFA as % of total fat content for a particular split of the data. Top: RMSECV. Bottom: RMSEP. In the top plot the RMSECV curves for all preprocessing alternatives are overlapping. In the bottom plot the RMSEP curves are overlapping for all preprocessing alternatives except for EMSC(3) preprocessing [Colour figure can be viewed at wileyonlinelibrary.com]

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.

The prediction errors for the response PUFA as percent 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 three reference spectra compared with using one 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 Figure 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 preprocessing, much better scaling estimates are obtained for these spectra.

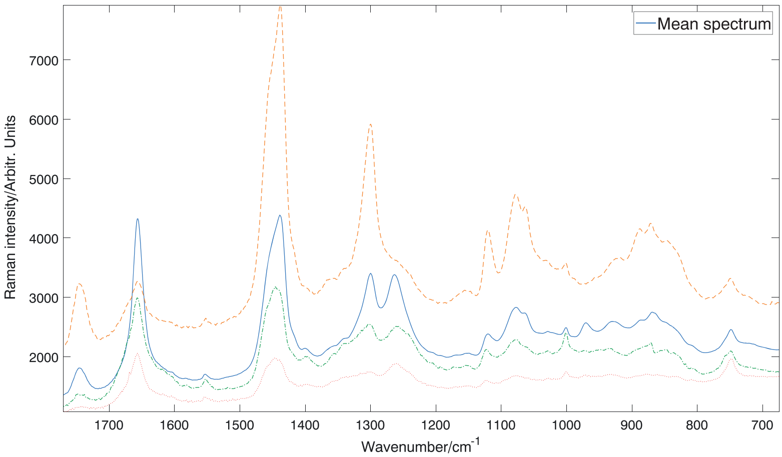


FIGURE 8 Emulsion data: Mean spectrum together with the three spectra with worst cross-validated prediction errors when using standard EMSC preprocessing [Colour figure can be viewed at wileyonlinelibrary.com]

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 utilize 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 preprocessing 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 preprocessing, 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 preprocessing 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, whereas more or less pure interferent spectra should be exploited as such in the EMSC.

ACKNOWLEDGEMENTS

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

ORCID

Joakim Skogholt https://orcid.org/0000-0001-8511-993X
Kristian Hovde Liland https://orcid.org/0000-0001-6468-9423

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*, *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, Philadelphia, PA **2010**.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Skogholt J, Liland KH, Indahl UG. Preprocessing of spectral data in the extended multiplicative signal correction framework using multiple reference spectra. *J Raman Spectrosc*. 2019;50:407–417. https://doi.org/10.1002/jrs.5520

APPENDIX 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);
```

PAPER III

Model selection by Fast virtual Cross–validation in Ridge Regression and the Tikhonov Regularization framework[*]

Ulf G. Indahl[†], Kristian H. Liland[†] and Joakim Skogholt[†]

September 27, 2019

†) *Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, N-1432 Ås, Norway* (ulf.indahl@nmbu.no).

[*] Preprint submitted to Computational Statistics and Data Analysis

Abstract

Ridge regression (RR) is an important special case in the Tikhonov regularization (TR) framework for biased linear regression modelling. The shrinking properties of RR and TR models often yield more attractive linear regression models than those obtained by the ordinary least squares (OLS), in particular for situations with highly correlated predictors and when the number of predictors exceeds the number of observed data points. The model selection task (i.e. the problem of choosing an appropriate value of the regularization parameter) in such cases is traditionally approached by either considering the so-called ridge traces of the regression coefficients, or by some choice of cross–validation strategy.

The purpose of this paper is to draw attention to a computationally efficient model selection strategy for the TR framework. The proposed strategy is derived by considering the orthogonal projections and the associated singular values obtained by the compact singular value decomposition (SVD). The resulting formulas provide highly efficient calculations for the exact leave-one-out cross–validation and associated predicted residual error sum of squares (PRESS) statistic for a continuous range of non-negative regularization parameter values. The closely related generalized cross–validation (GCV) measures and model degrees of freedom (df) are obtained simultaneously at insignificant additional computational costs.

By proposing an approach based on orthogonal transformations for situations with repeated or highly dependent measurements in the observations, we advocate a computationally fast method that approximates the PRESS statistic for the associated segmented cross–validation approach.

The capability of our theoretical findings and heuristic arguments are demonstrated to provide computationally efficient model selection tools for RR/TR in several practical applications. In particular, the proposed approach for approximating the PRESS–values obtained from the segmented cross–validation is demonstrated to provide precision levels that are similar to the PRESS–approximations by GCV of exact leave-one-out cross–validation PRESS–values.

Keywords: Cross–validation, SVD, Tikhonov regularization, Ridge regression, GCV, PRESS statistic.

1 Introduction

Model-/parameter selection in statistical modelling is frequently justified from the maximum likelihood (ML) principle in combination with some measure of model quality (such as the AIC, BIC, Mallows C_p, the PRESS statistic etc.) that estimates the expected predictive performance for some candidate model(s), see Friedman et al. (2009).

According to Hjorth (1993) the application of cross–validation measures as a methodology for model-/parameter selection in statistical applications was introduced by Stone (1974). Stones ideas motivated the invention of the generalized cross–validation (GCV) method by Golub et al. (1979). The GCV is a computationally efficient method for choosing a good ridge parameter in ridge regression (RR) modelling.

The RR method was introduced to the statistics community by Hoerl and Kennard (1970), and it is considered to be the most important special case in the Tikhonov (1963) regularization (TR) framework of linear regression methods. Originally, the TR framework was introduced to the community of numerical mathematics for solving linear discrete ill–posed problems in the context of inverse modelling. A good elementary introduction to the field can be found in Hansen (2010).

The GCV is not only a computationally efficient approximation to the leave-one-out cross–validation (LooCV) method. It is also invariant under orthogonal transformations of the data set. The *Predicted Residual Sum of Squares* (PRESS) statistic associated with LooCV was shown by Allen (1971, 1974) to be available for the ordinary least squares (OLS) regression by direct calculations avoiding the explicit and tedious remodelling usually associated with cross–validation schemes. From the Sherman–Morrison–Woodbury updating formula for calculating matrix inverses, see Householder (1965), it is possible to derive the individual scaling factors for the fitted model residuals to obtain the exact PRESS statistic associated with the LooCV method without explicit re–modelling. The required scaling factors for adjusting the residuals correctly are derived directly from the diagonal elements of the projection matrix associated with the regression problem. These diagonal elements (often referred to as the *leverage values*, see Best and Wolf (2014)) can easily be calculated from any orthogonal basis for the subspace spanned by the columns of the data matrix.

The purpose of the present paper is to present a prediction based framework for computationally efficient model selection in the TR framework for biased linear regression modelling. This is obtained as follows:

i) First we derive the simple and fast LooCV calculations utilizing the compact singular value decomposition (SVD) of our data matrix to quickly obtain PRESS values associated with any choice of the regularization parameter for a TR–problem. In particular this enables fast graphing of the PRESS–values as a function of the regularization parameter at any desired level of detail.

ii) Then we propose an approximation of the segmented (K–fold) cross–validation strategy by invoking the computationally inexpensive LooCV strategy after conducting an appropriate orthogonal transformation of the data matrix. The particular orthogonal transformation is constructed from the left singular vectors of the K local SVDs associated with the K distinct data segments. In situations where repeated re–modelling by leaving out one segment at a time is the most appropriate alternative to obtain a realistic PRESS–estimate, the suggested strategy provides a useful approximation of the PRESS–statistic at substantial computational savings – in particular for large data sets containing many segments (large K) of either identical, or highly related measurement values.

2 Linear regression preliminaries

2.1 Model estimation in ordinary least squares and ridge regression

In ordinary least squares (OLS) regression (Friedman et al. (2009)) one minimizes the *residual sum of squares*

$$RSS(\mathbf{b}) = \|\mathbf{X}\mathbf{b} - \mathbf{y}\|^2, \tag{1}$$

to identify the least squares solution(s) of (1) with respect to the regression coefficients $\mathbf{b}$. A least squares solution $\mathbf{b}_{OLS}$ of (1) corresponds to an exact solution of the associated *normal equations*

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}, \tag{2}$$

where $\mathbf{b}_{OLS}$ is unique when $\mathbf{X}'\mathbf{X}$ is non-singular. If otherwise not stated we assume that $\mathbf{X}$ is a centered $(n \times p)$ data matrix ($\mathbf{X}'$ denotes the transpose of $\mathbf{X}$) and that the corresponding $(n \times 1)$ vector $\mathbf{y}$ of responses is also centered.

For later predictions of uncentered data, the associated vector of fitted values is given by

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}_{OLS} + b_0, \tag{3}$$

where the constant term (intercept) $b_0 = \bar{y} - \bar{\mathbf{x}}\mathbf{b}_{OLS}$. Here, $\bar{y}$ and $\bar{\mathbf{x}}$ denotes the (column) averages of $\mathbf{y}$ and $\mathbf{X}$ before centering, respectively.

For various reasons ($\mathbf{X}'\mathbf{X}$ may be singular or poorly conditioned, the solution of (2) is not unique or inappropriate etc.) a minimizer $\mathbf{b}_{OLS}$ of $RSS(\mathbf{b})$ in equation (1) is not always the most attractive choice from a predictive point of view, see Friedman et al. (2009); Hansen (2010); Kalivas (2012). An alternative and quite useful solution was independently recognized by Tikhonov (1963), Phillips (1962) and Hoerl and Kennard (1970). Instead of directly minimizing the $RSS(\mathbf{b})$, their alternative proposal was to minimize the weighted bi–objective least squares problem

$$RSS_\lambda(\mathbf{b}) = \|\mathbf{X}\mathbf{b} - \mathbf{y}\|^2 + \lambda\|\mathbf{I}\mathbf{b} - \mathbf{0}\|^2 = \|\mathbf{X}\mathbf{b} - \mathbf{y}\|^2 + \lambda\|\mathbf{b}\|^2, \tag{4}$$

where the scalar $\lambda > 0$ represents a fixed *regularization parameter value* (of appropriate magnitude), the matrix $\mathbf{I}$ is the $(p \times p)$ identity matrix and $\mathbf{0}$ is a $(p \times 1)$ vector of zeros. This formulation explicitly represents a penalization with respect to the Euclidean (L_2) norm $\|\mathbf{b}\|$ of the regression coefficients, and for each fixed λ–value the unique minimizer of (4) is given by $\mathbf{b}_\lambda$ of equation (7) below. The rightmost part of equation (4) is often referred to as a TR–problem in *standard form*, see Hansen (2010).

The minimization of equation (4) with respect to the vector $\mathbf{b}$ is equivalent to solving the OLS problem associated with the augmented data matrix and response vector:

$$\mathbf{X}_\lambda = \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I} \end{bmatrix}, \quad \mathbf{y}_0 = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}. \tag{5}$$

Note that linear independence of the $\mathbf{X}_\lambda$–columns trivially follows from linear independence of the included $\mathbf{I}$–columns. The matrix product $\mathbf{X}_\lambda'\mathbf{X}_\lambda$ in the associated normal equations

$$\mathbf{X}_\lambda'\mathbf{X}_\lambda\mathbf{b} = \mathbf{X}_\lambda'\mathbf{y}_0 \tag{6}$$

is therefore non–singular, and the corresponding least squares solution

$$\mathbf{b}_\lambda = (\mathbf{X}_\lambda'\mathbf{X}_\lambda)^{-1}\mathbf{X}_\lambda'\mathbf{y}_0 \tag{7}$$

of the augmented problem associated with (5) becomes unique. Trivial algebraic simplifications of (6) result in the the familiar normal equations associated with the RR–problem

$$(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})\mathbf{b} = \mathbf{X}'\mathbf{y}, \tag{8}$$

and the solution in (7) simplifies to

$$\mathbf{b}_\lambda = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}. \tag{9}$$

For subsequent applications of the λ–regularized model to uncentered $\mathbf{X}$–data, the appropriate constant term in the resulting regression model is

$$b_{0,\lambda} = \bar{y} - \bar{\mathbf{x}}\mathbf{b}_\lambda, \tag{10}$$

and the associated vector of fitted values $\hat{\mathbf{y}}_\lambda$ is given by

$$\hat{\mathbf{y}}_\lambda = \mathbf{X}\mathbf{b}_\lambda + b_{0,\lambda}. \tag{11}$$

2.2 The Tikhonov L_2-regularization framework

Tikhonov (1963) noted that it is straight forward to generalize the above L_2 regularization of $\mathbf{b}$ to more specialized solution alternatives through a corresponding regularization matrix $\mathbf{L}$. These cases are expressed in terms of identifying the minimizing solution of the bi–objective least squares problem

$$RSS_{\mathbf{L},\lambda}(\mathbf{b}) = \|\mathbf{Xb} - \mathbf{y}\|^2 + \lambda\|\mathbf{Lb} - \mathbf{0}\|^2 = \|\mathbf{Xb} - \mathbf{y}\|^2 + \lambda\|\mathbf{Lb}\|^2, \tag{12}$$

for some fixed $\lambda > 0$. The minimization of equation (12) with respect to $\mathbf{b}$ can be obtained by considering the augmented data $\mathbf{X}_{\mathbf{L},\lambda} = \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{L} \end{bmatrix}$ and $\mathbf{y}_0 = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}$, and solving the normal equations

$$\mathbf{X}'_{\mathbf{L},\lambda}\mathbf{X}_{\mathbf{L},\lambda}\mathbf{b} = \mathbf{X}'_{\mathbf{L},\lambda}\mathbf{y}_0 \Rightarrow (\mathbf{X}'\mathbf{X} + \lambda\mathbf{L}'\mathbf{L})\mathbf{b} = \mathbf{X}'\mathbf{y} \tag{13}$$

associated with the OLS problem $\mathbf{X}_{\mathbf{L},\lambda}\mathbf{b} = \mathbf{y}_0$.

To avoid technical distractions we will in the following restrict our attention to the cases of square and non-singular regularization matrices $\mathbf{L}$ (even for situations where a non–square regularization matrix is the immediate choice to obtain solutions with particular characteristics, a non–singular $(p \times p)$–alternative that serves the same purpose is often available). By defining $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{L}^{-1}$, the solution of the OLS problem in (13) is equivalent to finding the unique OLS–solution $\boldsymbol{\beta}_\lambda$ of the transformed problem $\tilde{\mathbf{X}}_\lambda\boldsymbol{\beta} = \mathbf{y}_0$, where $\tilde{\mathbf{X}}_\lambda = \mathbf{X}_{\mathbf{L},\lambda}\mathbf{L}^{-1} = \begin{bmatrix} \tilde{\mathbf{X}} \\ \sqrt{\lambda}\mathbf{I} \end{bmatrix}$ and $\boldsymbol{\beta} = \mathbf{Lb}$. The associated expression minimized by $\boldsymbol{\beta}_\lambda$ is

$$\|\tilde{\mathbf{X}}\boldsymbol{\beta} - \mathbf{y}\|^2 + \lambda\|\boldsymbol{\beta}\|^2, \tag{14}$$

i.e. in the standard form (4), and the minimizing solution $\mathbf{b}_\lambda$ of the original problem (12) is obtained by

$$\mathbf{b}_\lambda = \mathbf{L}^{-1}\boldsymbol{\beta}_\lambda. \tag{15}$$

Among all the possible choices for the regularization matrix $\mathbf{L}$ we describe a few that are particularly useful:

1. **diagonal scaling** (e.g. the standardization of variables often advised for RR applications):

$$\mathbf{L}_{std} = \begin{bmatrix} \hat{\sigma}_1 & & & \\ & \hat{\sigma}_2 & & \\ & & \ddots & \\ & & & \hat{\sigma}_p \end{bmatrix},$$

   where $\hat{\sigma}_i$ is an estimate of the standard deviation of the i-th variable $(1 \leq i \leq p)$.

2. a (full) rank p discrete **1. derivative approximation**:

$$\mathbf{L}_1 = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \\ \sqrt{\epsilon}c_1 & \sqrt{\epsilon}c_1 & \ldots & \sqrt{\epsilon}c_1 & \sqrt{\epsilon}c_1 \end{bmatrix}.$$

3. a (full) rank p discrete **2. derivative approximation**:

$$\mathbf{L}_2 = \begin{bmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ \sqrt{\epsilon}c_1 & \sqrt{\epsilon}c_1 & \ldots & \sqrt{\epsilon}c_1 & \sqrt{\epsilon}c_1 & \sqrt{\epsilon}c_1 \\ -\sqrt{\epsilon}c_2 p/2 & -\sqrt{\epsilon}c_2(p-1)/2 & \ldots & \ldots & \sqrt{\epsilon}c_2(p-1)/2 & \sqrt{\epsilon}c_2 p/2 \end{bmatrix}.$$

The alternatives $\mathbf{L}_1$ and $\mathbf{L}_2$ are appropriate for problems where the $\mathbf{X}$–data are associated with discretized (uniform) sampling of continuous signals, so that some smoothness in the solution candidates $\mathbf{b}_\lambda$ is a reasonable expectation. The two last rows in $\mathbf{L}_2$ (and the last row in $\mathbf{L}_1$) above are scaled versions of the discretized and normalized Legendre polynomials (Kreyszig

(1978)) of order 0 and 1, respectively (c_1 and c_2 represent the normalization constants, and $\epsilon > 0$ is a scaling factor to be commented on below). It should be noted that these rows (considered as vectors) are orthogonal to the above rows in the discrete derivative matrices where they appear. The main purpose of the included Legendre vectors is to ensure full rank of the regularization matrices that is required to obtain the attractive computational advantages described below.

Appropriate regularization of the solutions $\mathbf{b}_\lambda$ may be obtained by choosing the fixed scaling factor $\epsilon > 0$ to be

- either sufficiently large to make $\mathbf{b}_\lambda$ practically orthogonal to the subspace of polynomial trends spanned by the included Legendre vectors, or
- sufficiently small to inhibit any notable penalization effect with respect to the same polynomial trends.

The choice of ϵ in the last case can therefore not be made arbitrary small in practice, but must be chosen large enough to avoid numerical difficulties in the computations of $\tilde{\mathbf{X}}$ and $\mathbf{b}_\lambda$. Additional (non–invertible) differentiation matrix candidates taking various boundary condition requirements into account are discussed in Hansen (2010).

2.3 Calculating the $\mathbf{b}_\lambda$–solutions effectively from the SVD

The full SVD of $\mathbf{X} = \mathbf{USV}'$ yields $\mathbf{VV}' = \mathbf{I}_p$ and $\mathbf{X}'\mathbf{X} = \mathbf{VS}'\mathbf{SV}'$. The right singular vectors $\mathbf{V}$ of $\mathbf{X}$ are obviously eigenvectors for both $\mathbf{X}'\mathbf{X}$ and

$$\mathbf{X}_\lambda'\mathbf{X}_\lambda = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p) = \mathbf{V}(\mathbf{S}'\mathbf{S} + \lambda\mathbf{I}_p)\mathbf{V}', \tag{16}$$

and their corresponding eigenvalues are given by the diagonals of $\mathbf{S}'\mathbf{S}$ and $\mathbf{S}'\mathbf{S} + \lambda\mathbf{I}_p$, respectively. The inverse matrix $(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1} = \mathbf{V}(\mathbf{S}'\mathbf{S} + \lambda\mathbf{I}_p)^{-1}\mathbf{V}'$, and the expression (9) for the TR-regression coefficients of a problem on standard form therefore simplifies (Friedman et al. (2009)) to

$$\mathbf{b}_\lambda = \mathbf{V}(\mathbf{S}'\mathbf{S} + \lambda\mathbf{I}_p)^{-1}\mathbf{V}'\mathbf{VSU}'\mathbf{y} = \mathbf{V}(\mathbf{S}'\mathbf{S} + \lambda\mathbf{I}_p)^{-1}\mathbf{SU}'\mathbf{y}. \tag{17}$$

In the following we assume that $\mathbf{X}$ has full rank, i.e. $r = rank(\mathbf{X}) = \min(n, p)$. Then there will be exactly r non-zero rows in the $\mathbf{S}$-factor of $\mathbf{b}_\lambda$, and the zero rows of $\mathbf{S}$ cancel both the associated columns in $\mathbf{V}(\mathbf{S}'\mathbf{S} + \lambda\mathbf{I}_p)^{-1}$ and rows in $\mathbf{U}'$. By considering the compact SVD of $\mathbf{X} = \mathbf{U}_r\mathbf{S}_r\mathbf{V}_r'$ (the vanishing dimensions associated with the singular value 0 are omitted from the factorization), the expression (17) for the regression coefficients $\mathbf{b}_\lambda$ simplifies to

$$\mathbf{b}_\lambda = \mathbf{V}_r(\mathbf{S}_r^2 + \lambda\mathbf{I}_r)^{-1}\mathbf{S}_r\mathbf{U}_r'\mathbf{y} = \mathbf{V}_r(\mathbf{S}_r + \lambda\mathbf{S}_r^{-1})^{-1}\mathbf{U}_r'\mathbf{y} = \mathbf{V}_r\mathbf{c}_\lambda, \tag{18}$$

where the coordinate vectors $\mathbf{c}_\lambda = (\mathbf{S}_r + \lambda\mathbf{S}_r^{-1})^{-1}\mathbf{U}_r'\mathbf{y} = [c_{\lambda,1} \,...\, c_{\lambda,r}]' \in \mathbb{R}^r$ has entries

$$c_{\lambda,k} = \frac{\mathbf{u}_k'\mathbf{y}}{s_k + \lambda/s_k}, \text{ for } 1 \leq k \leq r. \tag{19}$$

Compared to the relatively large computational costs associated with calculating the (compact) SVD of $\mathbf{X}$, calculation of the regression coefficient candidates (even for a large number of candidate λ-values) just requires computing the vectors $\mathbf{c}_\lambda$ according to (19) and the matrix-vector multiplications $\mathbf{b}_\lambda = \mathbf{V}_r\mathbf{c}_\lambda$ as derived in equation (18).

For the regularized multivariate regression with several (q) responses $\mathbf{Y} \in \mathbb{R}^{n\times q}$, the associated matrix of regression coefficients is

$$[\mathbf{b}_{1,\lambda} \,...\, \mathbf{b}_{q,\lambda}] = \mathbf{V}_r(\mathbf{S}_r + \lambda\mathbf{S}_r^{-1})^{-1}\mathbf{U}_r'\mathbf{Y} = \mathbf{V}_r\mathbf{C}_\lambda, \tag{20}$$

where $\mathbf{C}_\lambda = (\mathbf{S}_r + \lambda\mathbf{S}_r^{-1})^{-1}\mathbf{U}_r'\mathbf{Y}$ is the obvious multivariate generalization of the vector $\mathbf{c}_\lambda$ described above.

3 The computationally fast leave–one–out cross–validation for TR–problems

3.1 The OLS case

With linearly independent columns in the data matrix $\mathbf{X}$, the associated OLS–solution $\mathbf{b}_{OLS}$ of the normal equations (2) is unique and a computationally fast version of the LooCV can be derived from the Sherman–Morrison–Woodbury formula for updating matrix inverses, see Householder (1965).

Let $\hat{y}_{k,-1}$ denote the prediction of the k–th sample after deleting it from the regression problem in (1). Then the PRESS–statistic proposed by Allen (1971, 1974), is given by

$$PRESS = \sum_{k=1}^{n}(y_k - \hat{y}_{k,-1})^2 = \sum_{k=1}^{n}\left(\frac{y_k - \hat{y}_k}{1 - h_k - 1/n}\right)^2. \tag{21}$$

In (21) $\hat{y}_k$ is the k–th entry in the fitted values $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}_{OLS} + b_0$, and h_k is the k–th diagonal element of the projection matrix $\mathbf{H}$ defined in (22) below. The denominator $(1 - h_k - 1/n)$ scaling the k–th model residual $(y_k - \hat{y}_k)$ yields precisely the corresponding LooCV prediction residual $(y_k - \hat{y}_{k,-1})$. The term $1/n$ in this denominator accounts for the centering of the $\mathbf{X}$–columns and the inclusion of a constant term (b_0) in the regression model (3). Note that the projection matrix can be expressed as follows

$$\mathbf{H} \overset{\text{def}}{=} \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{T}\mathbf{T}', \tag{22}$$

where $\mathbf{T}$ may be any orthogonal $(n \times r)$–matrix spanning the column space of the centered $\mathbf{X}$–data. The diagonal element h_k is often referred to as the *leverage value* associated with the k–th sample (row) in $\mathbf{X}$.

From the last identity of equation (22) it is clear that the entries of the n–vector $\mathbf{h} = [h_1\ h_2\ ...\ h_n]'$ representing the diagonal elements of $\mathbf{H}$ is identical to the vector containing the squared norms of the $\mathbf{T}$–rows, i.e.

$$\mathbf{h} = (\mathbf{T} \odot \mathbf{T})\mathbf{1}, \tag{23}$$

where $\mathbf{T} \odot \mathbf{T}$ denotes the Hadamard (element-wise) product of $\mathbf{T}$ with itself and $\mathbf{1} \in \mathbb{R}^r$ is the constant vector with 1's in all entries. Appropriate choices of the matrix $\mathbf{T}$ can be obtained by various strategies including both the SVD, the QR-factorization or some alternative Gram–Schmidt process based on the columns of $\mathbf{X}$. One should note that calculating the matrix inverse $(\mathbf{X}'\mathbf{X})^{-1}$ in the process for finding the diagonal $\mathbf{h}$ of $\mathbf{H}$ in (22) is neither required nor recommended in practice. In general, the explicit calculation of matrix inverses (for non-diagonal matrices) should be avoided whenever possible due to various unfavourable computational aspects, see (Björck, 2016, Section 1.2.6).

3.1.1 The generalized cross–validation

The GCV was proposed by Golub et al. (1979) as a fast method for choosing good regularization parameter values in RR. The GCV is explained as a rotation invariant alternative to the LooCV that provides an approximation of the $PRESS$–statistic when considering it as a function of the regularization parameter λ. Here, we prefer using the particular definition

$$GCV(\lambda) \overset{\text{def}}{=\!=} \sum_{k=1}^{n} \left(\frac{y_k - \hat{y}_{\lambda,k}}{1 - \bar{h}_\lambda - 1/n} \right)^2 = (1 - df(\lambda)/n)^{-2} \|\mathbf{y} - \mathbf{X}\mathbf{b}_\lambda\|^2, \tag{24}$$

where $(y_k - \hat{y}_{\lambda,k})$ is the k-th entry of the residual vector $\mathbf{r}_\lambda = \mathbf{y} - \hat{\mathbf{y}}_\lambda$, $\bar{h}_\lambda \overset{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{r} \frac{s_k}{s_k + \lambda/s_k}$ and the effective degrees of freedom $df(\lambda) \overset{\text{def}}{=\!=} n\bar{h}_\lambda + 1$. This definition of $GCV(\lambda)$ is proportional (by the sample size n) to the definition given in (Golub et al., 1979, page 216).

From the elementary matrix–vector multiplication formula (18) for computing the regression coefficients $\mathbf{b}_\lambda$, it is clear that $GCV(\lambda)$ can be calculated very effectively for a large number of different λ–values once the non-zero singular values of $\mathbf{X}$ are available.

In their justification of $GCV(\lambda)$ as the preferable choice over the exact LooCV–based $PRESS(\lambda)$, Golub and co–workers stressed the unsatisfactory properties of the $PRESS$–function in situations where the rows of $\mathbf{X}$ are orthogonal or nearly orthogonal. In such situations the estimated regression coefficient $\mathbf{b}_\lambda^{(k)}$ (obtained by excluding the k-th row $\mathbf{x}_k$ of $\mathbf{X}$) must be correspondingly orthogonal (or nearly orthogonal) to the excluded sample $\mathbf{x}_k$. Consequently, the associated leave–one–out prediction $\hat{y}_{k,-1}(= \mathbf{x}_k\mathbf{b}_\lambda^{(k)})$ becomes a poor estimate of the corresponding k-th response value y_k.

In situations such as the one just described, it hardly makes any sense to think of the $\mathbf{X}$–data as a collection of independent random samples, and the statistical motivation for considering the LooCV idea becomes correspondingly inferior. The claim in Golub et al. (1979) that any parameter selection procedure should be invariant under orthogonal transformations of the $(\mathbf{X}, \mathbf{y})$–data will be discussed below (our scepticism to this requirement as an inexpedient restriction, relates to the context of approximating the PRESS-statistic for situations where a segmented/folded cross–validation approach is appropriate).

From the matrix and vector augmentation (5) in the above preliminaries and equation (21), it is immediately clear that the computationally fast version of the LooCV and the associated $PRESS$–statistic is also valid for TR–problems when the regularization parameter λ is treated as a fixed quantity. Below we will derive an equation assuring fast calculations of the regularized leverage vectors $\mathbf{h}_\lambda$. These calculations are surprisingly similar to a computationally efficient alternative for obtaining the fitted values $\hat{\mathbf{y}}_\lambda$ and closely related to corresponding regularized regression coefficients $\mathbf{b}_\lambda$ in (18). Both $\mathbf{h}_\lambda$, $\hat{\mathbf{y}}_\lambda$ (and $\mathbf{b}_\lambda$) can be calculated efficiently by utilizing the SVD of the centered data matrix $\mathbf{X}$. This makes the computations of the exact LooCV–based $PRESS(\lambda)$–function defined in (28) below about as efficient as the approximation obtained by the $GCV(\lambda)$ in (24).

3.2 The exact LooCV–based $PRESS(\lambda)$–function for TR–problems

We assume that the centered $\mathbf{X}$ has full rank r and that $\mathbf{X} = \mathbf{U}_r\mathbf{S}_r\mathbf{V}_r'$ is the associated compact SVD. By defining $\mathbf{S}_{\lambda,r}$ to be the diagonal $r \times r$ matrix with non-zero diagonal entries $\sqrt{s_k^2 + \lambda}$, $k = 1, ..., r$, the r most dominant singular values of the augmented matrix $\mathbf{X}_\lambda$ in (5) are given by the diagonal elements of $\mathbf{S}_{\lambda,r}$. From equation (16) in Section 2.3, the right singular vectors $\mathbf{V}_r$ of $\mathbf{X}$ are also the right singular vectors of the augmented matrix $\mathbf{X}_\lambda$, and the associated r left singular vectors are

$$\mathbf{T}_{\lambda,r} = \mathbf{X}_\lambda\mathbf{V}_r\mathbf{S}_{\lambda,r}^{-1} = \begin{bmatrix} \mathbf{X}\mathbf{V}_r\mathbf{S}_{\lambda,r}^{-1} \\ \sqrt{\lambda}\mathbf{I}\mathbf{V}_r\mathbf{S}_{\lambda,r}^{-1} \end{bmatrix} = \begin{bmatrix} \mathbf{U}_r\mathbf{S}_r\mathbf{S}_{\lambda,r}^{-1} \\ \sqrt{\lambda}\mathbf{V}_r\mathbf{S}_{\lambda,r}^{-1} \end{bmatrix} = \begin{bmatrix} \mathbf{U}_{\lambda,r} \\ \sqrt{\lambda}\mathbf{V}_r\mathbf{S}_{\lambda,r}^{-1} \end{bmatrix}, \tag{25}$$

where the matrix $\mathbf{U}_{\lambda,r} \overset{\text{def}}{=} \mathbf{U}_r\mathbf{S}_r\mathbf{S}_{\lambda,r}^{-1}$ denoting the upper n rows of $\mathbf{T}_{\lambda,r}$ is the part of actual interest (the additional left singular vectors not included in (25) are all zeros in the upper n entries). Because $\mathbf{S}_r\mathbf{S}_{\lambda,r}^{-1}$ is $(r \times r)$ diagonal, $\mathbf{U}_{\lambda,r}$ is obtained by scaling the k–th column $(1 \leq k \leq r)$ of $\mathbf{U}_r$ with the factor $\sqrt{s_k/(s_k + \lambda/s_k)}$.

From the above definition of $\mathbf{U}_{\lambda,r}$, calculation of the PRESS–residuals associated with the

n original $(\mathbf{X}, \mathbf{y})$ data points in the augmented least squares problem $\mathbf{X}_\lambda \mathbf{b} = \mathbf{y}_0$ is straight forward. According to (23), the leverage values $\mathbf{h}_\lambda = [h_{\lambda,1} \; ... \; h_{\lambda,n}]'$ corresponding to the n samples in the regularized version of our data set are given by the matrix-vector multiplication

$$\mathbf{h}_\lambda = (\mathbf{U}_{\lambda,r} \odot \mathbf{U}_{\lambda,r})\mathbf{1} = (\mathbf{U}_r \odot \mathbf{U}_r)\mathbf{d}_\lambda, \tag{26}$$

where the coefficient vector $\mathbf{d}_\lambda = [d_{1,\lambda} \; ... \; d_{r,\lambda}]' = (\mathbf{S}_r \mathbf{S}_{\lambda,r}^{-1})^2 \mathbf{1} \in \mathbb{R}^r$ has the entries

$$d_{k,\lambda} = \frac{s_k^2}{s_k^2 + \lambda} = \frac{s_k}{s_k + \lambda/s_k}, \text{ for } 1 \leq k \leq r. \tag{27}$$

For each choice of the regularization parameter $\lambda > 0$, the fitted values are $\hat{\mathbf{y}}_\lambda = \mathbf{X}\mathbf{b}_\lambda + b_{0,\lambda}$. Hence, the PRESS–values

$$PRESS(\lambda) \overset{\text{def}}{=\!=} \sum_{k=1}^{n} \left(\frac{y_k - \hat{y}_{\lambda,k}}{1 - h_{\lambda,k} - 1/n} \right)^2, \tag{28}$$

where $y_k - \hat{y}_{\lambda,k}$ is the k–th entry of the residual vector $\mathbf{r}_\lambda = \mathbf{y} - \hat{\mathbf{y}}_\lambda$ and the leverage $h_{\lambda,k}$ is the corresponding k-th entry of $\mathbf{h}_\lambda$. Note that $\bar{h}_\lambda$ in the denominator of equation (24) defining $GCV(\lambda)$ is identical to the mean value of the $\mathbf{h}_\lambda$–entries, i.e. $\bar{h}_\lambda = (1/n)\sum_{k=1}^{n} h_{\lambda,k}$, due to the fact that $\mathbf{U}_r$ is an orthogonal matrix.

Based on the compact SVD of $\mathbf{X}$, the expression for the regression coefficients in (18) and the identity $\mathbf{S}_r \mathbf{c}_\lambda = \mathbf{U}_r'\mathbf{y} \odot \mathbf{d}_\lambda$ we obtain the fitted values as

$$\hat{\mathbf{y}}_\lambda = \mathbf{X}\mathbf{b}_\lambda = \mathbf{U}_r \mathbf{S}_r \mathbf{c}_\lambda = \mathbf{U}_r(\mathbf{U}_r'\mathbf{y} \odot \mathbf{d}_\lambda). \tag{29}$$

Consequently, the evaluation of the $PRESS(\lambda)$–function defined in (28) is essentially available at the additional computational cost of two matrix–vector multiplications (equations 26 and 29) for each choice of λ. The associated coefficient vectors $\mathbf{d}_\lambda$ and $\mathbf{U}_r'\mathbf{y} \odot \mathbf{d}_\lambda$ are obtained by elementary arithmetic operations where everything except for the regularization parameter λ is fixed. A note on the number of floating point operations (flops) required for the fast calculation of the LooCV–based $PRESS(\lambda)$–function is included in Appendix C. An efficient prototype MATLAB–routine for computing the PRESS–statistic and regression coefficients is available in Appendix A. A corresponding implementation in R code will be made available upon publication at https://cran.r-project.org/web/packages/TR.

3.3 The segmented virtual cross–validation

There are obviously situations where a direct application of the LooCV approach may be inappropriate for both model validation and –selection. Most typical are the situations where some repeated or closely related measurements (based on an experimental design or some other type of rigorous framework) leads to subsets of highly similar rows in the data matrix $\mathbf{X}$. A leave–one–out cross–validation strategy is then usually not reliable but rather likely to produce overoptimistic PRESS–values.

In such situations it is more appropriate to handle a data set according to the present sample segment structure, and to calculate the PRESS statistic according to a segmented cross–validation (SCV) strategy of repeated remodelling by successively holding out the entire sample segments. However, for large data sets (containing either a large number of samples and/or variables), the SCV strategy may be computationally slow or at the worst practically infeasible. We therefore propose a considerably faster alternative that approximates the SCV approach for the type of situations just described. In the following we assume (without loss of generality) that the uncentered data matrix

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_K \end{bmatrix} \text{ together with the uncentered response vector } \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_K \end{bmatrix} \quad (K \geq 2) \qquad (30)$$

is composed by K distinct sample segments. For $1 \leq k \leq K$, we assume that $\mathbf{U}_k\mathbf{S}_k\mathbf{V}_k' = \mathbf{X}_k$ denotes the compact SVD of segment number k, and that n_k is the number of rows in segment $\mathbf{X}_k$ so that the total number of samples included in $\mathbf{X}$ is $n = \sum_{k=1}^{K} n_k$.

From the above SVD for the k–th segment, the identity $\mathbf{U}_k'\mathbf{X}_k = \mathbf{S}_k\mathbf{V}_k$ immediately follows. Consequently, the orthogonal transformation performed by left multiplication with the $(n_k \times n_k)$ matrix $\mathbf{U}_k'$ transforms the segment $\mathbf{X}_k$ into a matrix of strictly orthogonal rows. Now we can define the two block diagonal matrices

$$\mathbf{T} = \begin{bmatrix} \mathbf{U}_1 & & & \\ & \mathbf{U}_2 & & \\ & & \ddots & \\ & & & \mathbf{U}_K \end{bmatrix} \text{ and } \tilde{\mathbf{T}} = \begin{bmatrix} \mathbf{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \tag{31}$$

with the properties $\mathbf{T}'\mathbf{T} = \mathbf{T}\mathbf{T}' = \mathbf{I}$ and $\tilde{\mathbf{T}}'\tilde{\mathbf{T}} = \tilde{\mathbf{T}}\tilde{\mathbf{T}}' = \mathbf{I}$, i.e. both $\mathbf{T}$ and $\tilde{\mathbf{T}}$ are orthogonal.

The formulation of TR–modelling for uncentered $\mathbf{X}$ and explicit inclusion of the constant term corresponds to finding the least squares solution of the linear system

$$\begin{bmatrix} \mathbf{1} & \mathbf{X} \\ \mathbf{0} & \sqrt{\lambda}\mathbf{L} \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}, \tag{32}$$

and left multiplication of (32) by the orthogonal matrix $\tilde{\mathbf{T}}'$ yields the system

$$\begin{bmatrix} \mathbf{T}'\mathbf{1} & \mathbf{T}'\mathbf{X} \\ \mathbf{0} & \sqrt{\lambda}\mathbf{L} \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \mathbf{T}'\mathbf{y} \\ \mathbf{0} \end{bmatrix}. \tag{33}$$

Note that the associated normal equations of the systems in (32) and (33) are identical. Hence, their least squares solutions are also identical.

3.3.1 Definition of the segmented virtual cross–validation

The *segmented virtual cross–validation (SvCV)* strategy is defined as the process of applying the LooCV strategy to the transformed system in equation (33). As is noted above, multiplication by $\mathbf{T}'$ has the effect of orthogonalizing the rows within each of the K segments in the $\mathbf{X}$ matrix.

The heuristic argument for justifying the SvCV approach as an approximation of a SCV approach is that the rows within each transformed data segment are unsupportive of each other under the LooCV strategy (due to the internal "decoupling" of each segment into a set of mutually orthogonal row vectors). However, because the complete dataset is used to derive the transformation $\mathbf{T}$, it can be observed that in practical situations the accuracy of this approximation depends on the level of similarity between the original samples within each segment of data points.

Note that contrary to the LooCV, the GCV is not useful in combination with the SvCV strategy. The obvious reason for this is that the singular values of $\mathbf{X}$ are invariant under orthogonal transformations. From equation (24) and the definition of $\bar{h}_\lambda$ it follows that $GCV(\lambda)$ is also invariant under orthogonal transformations, i.e. the systems in (32) and (33) lead to the exact same $GCV(\lambda)$–function.

3.3.2 Segment decomposition in three different situations

In the following we will discuss the proposed SvCV strategy more closely for three different situations:

a) Segments of identical rows.

b) Segments of collinear rows.

c) The general case (segments with no particular structure in the rows).

Identical rows:

Let us assume that all the rows of a segment $\mathbf{X}_i$, $(1 \leq i \leq K)$ are identical. In this particular case the $PRESS$–function associated with the SvCV is identical to the $PRESS$–function obtained by the SCV.

The alleged identity can be derived by noting that the left-multiplication of the left- and right hand sides of a linear system by an orthogonal matrix affects neither the least squares solution nor the norm of the associated residual vector. Consequently, the SCV strategy applied to the two systems (32) and (33) will result in identical $PRESS$-functions. With all rows within each segment $\mathbf{X}_k \in \mathbb{R}^{n_k \times p}$ being identical to its first row (denoted $\mathbf{x}_{k,1}$) of the segment, it is straight forward to verify that $\mathbf{X}_k$ has only one non-zero singular value

$s_{k,1} = \sqrt{\mathbf{x}_{k,1}\mathbf{x}'_{k,1}n_k}$ and the corresponding left– and right singular vectors are

$$\mathbf{u}_{k,1} = \frac{1}{\sqrt{n_k}}\begin{bmatrix} 1 \\ 1 \\ : \\ 1 \end{bmatrix} \in \mathbb{R}^{n_k} \text{ and } \mathbf{v}_{k,1} = \frac{1}{\sqrt{\mathbf{x}_{k,1}\mathbf{x}'_{k,1}}}\mathbf{x}'_{k,1} \in \mathbb{R}^p. \tag{34}$$

By the orthogonality requirements of the SVD, any other left singular vector $\mathbf{u}$ must satisfy $\mathbf{u}'\mathbf{u}_{k,1} = 0$. Consequently

$$\mathbf{U}'_k\mathbf{X}_k = \begin{bmatrix} \sqrt{n_k}\mathbf{x}_{k,1} \\ \mathbf{0} \\ : \\ \mathbf{0} \end{bmatrix} \text{ and } \mathbf{U}'_k\mathbf{1} = \begin{bmatrix} \sqrt{n_k} \\ 0 \\ : \\ 0 \end{bmatrix}, \tag{35}$$

meaning that there will be only one non-zero row in each segment on the left hand side of the $\tilde{\mathbf{T}}$-transformed system (33). It is therefore sufficient to demonstrate that the *PRESS*–functions obtained from applying the SCV and the LooCV to the system in (33) are equal: Clearly, for any row containing just zeros in the left hand side of (33) the prediction based on it is trivially identical to 0 (zero) for either of the cross–validation strategies (regardless of the regression coefficients). Because such zero rows do not contribute in the calculation of the regression coefficients, we are forced to conclude that the regression coefficients obtained by holding out the (only) non–zero row of a segment must be equal to the regression coefficients obtained from holding out the entire segment. Thus the predicted values for the non-zero row in each segment must also be identical for both cross–validation strategies, and we can conclude that the *PRESS* functions obtained by the SCV– and the SvCV strategies must be identical.

Collinear (proportional) rows:

One might suspect the same result to hold when the rows within a segment are proportional. This is however not the case with the modelling strategy described above. The reason for this is that the inclusion of a constant term will make each of the K segments become a rank 2 – rather than a rank 1 submatrix. With more than one non-zero row on the left hand side in each segment the argument of the previous situation fails, and doing LooCV on the transformed data is no longer equivalent to doing SCV on the original data. However, if omitting the constant term from the modelling, each of the K segments has rank 1, and the SCV and SvCV approaches will result in identical $PRESS(\lambda)$–functions. The rigorous explanation is similar to the argument given for the above situation with identical rows.

The general case:

In general, the goodness of the approximation provided by the SvCV of the exact SCV is related to the similarity of the rows within each segment. With the SvCV we are clearly cross–validating on the orthogonal phenomena caused by the samples within each segment. As all the samples in a segment contribute to identifying these directions, the SvCV cannot be expected to provide exactly the same results as the SCV. One may, however, expect that when the different segments are carefully arranged to contain highly similar samples only (which is a reasonable assumption to make for most organized studies with such data segments), then the SvCV provides a useful approximation to the SCV. This will be demonstrated in the applications described below.

Computational aspects in the leverage corrections for the SvCV

As is noted in association with (30), the SvCV procedure requires an initial calculation of the transformation $\mathbf{T}$ from the segments of the uncentered $\mathbf{X}$–data. For a successful and correct implementation of the computational shortcuts similar to those of the LooCV, it is necessary to mean center the data matrix $\mathbf{X}$ prior to applying the $\mathbf{T}$–transformation and doing the least squares modelling. In practice, one must therefore mean center the data prior to the multiplication with $\mathbf{T}'$ (or, equivalently, one can multiply by $\mathbf{T}'$ and subtract the projection of the transformed data onto the transformed vector $\mathbf{T}'\mathbf{1}$ of ones). As $\mathbf{T}$ is an orthogonal transformation the angles and in particular the orthogonality between vectors will be preserved. For the transformed data, modelling by including a constant term is therefore associated with the transformed vector $\mathbf{T}'\mathbf{1}$ of ones. With $\mathbf{X}_c$ and $\mathbf{y}_c$ denoting the centred data matrix and the associated centred response vector, respectively, the vector $\mathbf{T}'\mathbf{1}$ is orthogonal to the columns of the transformed centred data $\mathbf{T}'\mathbf{X}_c$ and $\|\mathbf{T}'\mathbf{1}\| = \|\mathbf{1}\| = \sqrt{n}$. The justification for the leverage correction described earlier therefore still holds, but the particular correction terms $(1/n)$ changes.

With the transformed predictors $\tilde{\mathbf{X}} = \mathbf{T}'\mathbf{X}_c$ and responses $\tilde{\mathbf{y}} = \mathbf{T}'\mathbf{y}_c$ in (33), the associated fitted (centred) values as $\hat{\tilde{\mathbf{y}}}_\lambda = \tilde{\mathbf{X}}\mathbf{b}_\lambda$, the PRESS-function for the SvCV is given by

$$PRESS_{SvCV}(\lambda) = \sum_{k=1}^{n}(\tilde{y}_k - \hat{\tilde{y}}_{\lambda,k,-1})^2 = \sum_{k=1}^{n}\left(\frac{\tilde{y}_k - \hat{\tilde{y}}_{\lambda,k}}{1 - h_{\lambda,k} - m_k/n}\right)^2. \qquad (36)$$

Here the leverages $h_{\lambda,k}$ are calculated as in (26) based on the transformed version $\tilde{\mathbf{X}}$ of the centered data, and the enumerator of the correction terms are the entries of the vector $\mathbf{m} = \mathbf{T}'\mathbf{1} \odot \mathbf{T}'\mathbf{1} \in \mathbb{R}^n$. This means that the correction terms $1/n$ in the denominator of (28) must be replaced by m_k/n in (36), where m_k denotes the k-th entry of the vector $\mathbf{m}$ (to be consistent with the orthogonal transformation of regularized least squares problem).

A comparison of the number of flops required for the SvCV compared to the SCV is included in Appendix D. An efficient prototype MATLAB-routine for computing the SvCV is available in Appendix B. Corresponding R-code will be made available upon publication at https://cran.r-project.org/web/packages/TR.

3.4 A short note on model selection heuristics

The key formulas derived above allow for efficient model selection procedures by minimizing the $PRESS(\lambda)$– or the $GCV(\lambda)$–functions with respect to the regularization parameter λ. However, the minima of these functions will not necessarily assure the selection of an optimal model in terms of future predictions. This is particularly the case when the $PRESS(\lambda)$– and $GCV(\lambda)$–functions are relatively flat for some large interval of λ–values containing the minimum value. In such situations it is often useful to invoke the heuristic principles of *Occam's razor* for identifying a simpler model (in terms of the norm of the regression coefficients) at a small additional cost in terms of the PRESS (or the GCV):

The '**1 standard error rule**' described in Friedman et al. (2009) obtains a simpler (more regularized) alternative by selecting a model where the PRESS–statistic is within one standard error of the PRESS-minimal model. More precisely, we first identify the minimum PRESS value and calculate the standard error of the squared cross–validation errors associated with this model. Then the largest regularization parameter value where the associated model has a PRESS–statistic within one standard error of the PRESS–minimum is selected.

The 'χ^2 **model selection rule**' to determine the regularization parameter was originally introduced for model selection with Partial Least Squares regression modelling, see Indahl (2005). By assuming that the residuals associated with the minimum value $PRESS_{min}$ of $PRESS(\lambda)$ are randomly drawn from a normal distribution, the statistic $n \cdot PRESS_{min}/\sigma^2$, where σ^2 is the associated (unknown) variance, follows a χ^2_n distribution (where n is the degrees of freedom). By fixing a particular significance level α, the selection rule says: "Choose the largest possible value of λ so that $n \cdot PRESS_{min}/PRESS(\lambda) \geq \chi^2_{n,\alpha}$", where $\chi^2_{n,\alpha}$ is the lower α–quantile of the χ^2_n distribution and $PRESS(\lambda)$ is a substitute for σ^2.

Based on the efficient formulas for calculating $PRESS(\lambda)$, both these model selection alternatives can be implemented without significant increases of the total computational costs.

4 Applications

In the following we will present some applications of our fast cross–validation approaches for model selection within the TR framework based on several real world data sets. We will consider situations where both the LooCV and the SvCV are appropriate. The required algorithms were implemented and executed in MATLAB, and prototype code is given in the appendices. We used a computer running Windows 10 and MATLAB R2017, with 16GB ram, an Intel i7–4790k processor and a NVIDIA GTX–970 graphics card. For the discrete derivative regularization matrices we use the full rank approximations described in Section 2.2 with the scaling coefficient set to $\epsilon = 10^{-10}$ in the appended rows. This is done to alleviate the numerical impact from these rows in the resulting regression coefficients.

4.1 The fast leave-one-out cross–validation

4.1.1 Data sets

The following data sets will be considered in the examples presented below:

1. *Octane data*, see Kalivas (1997). This data set consists of near infrared (NIR) spectra of gasoline. There are 60 samples and 401 features (wavelengths in the range $900nm - 1700nm$). The response value is the octane number measured for each sample.

2. *Pork fat data*, see Lyndgaard et al. (2012). This data set consists of Raman spectra measured on pork fat tissue. There are 105 samples, 5567 features (wavenumbers in the range $200.1cm^{-1} - 1889.9cm^{-1}$), and 19 different responses. For modelling and prediction we only consider the response consisting of saturated fatty acids as percentage of total fatty acids, hereafter referred to as SFA.

3. *Prostate gene data*, see Singh et al. (2002). The data set is a microarray gene expression data set. There are 102 samples, and the gene expression of 12600 different genes were measured. The response is binary (cancer/not cancer), and we consider the "dummy–regression" approach to the underlying classification problem. For this data set we standardize the data prior to modelling. The standardization will introduce a small bias in the model selection that will be discussed later.

For all datasets we have used approximately 2/3 of the available samples for model building and –selection. The remaining 1/3 of samples were used for testing the selected models. We considered the following model selection alternatives identifying good regularization parameter candidates: (i) $PRESS_{min}$ – the minimum $PRESS(\lambda)$–value, (ii) GCV_{min} – the minimum $GCV(\lambda)$–value, (iii) the 1 standard error rule for $PRESS(\lambda)$, (iv) the χ^2–rule for $PRESS(\lambda)$ using the significance level $\alpha = 0.2$.

4.1.2 Model selection and prediction

For each data set the modelling was based on 1000 regularization parameter candidate values spaced uniformly on a log–scale. For the octane data the displayed parameter values were from the range 10^{-4} to 10^5, for the Pork fat data from the range 10^2 to 10^{25}, and for the Prostate data from the range 10^{-1} to 10^8. Different ranges were chosen for each data set to avoid irrelevant levels of regularization, and to obtain a good display of the PRESS– and GCV curves including the located minima. In Figures 1–3 the PRESS/n and GCV/n are plotted as functions of the regularization parameter for the different data sets and the different choices of the regularization matrix. Such plots are useful for model selection as they allow for a direct comparison of the model quality for different values of the regularization parameter. Division of the PRESS– and GCV values by the samples size n makes the model selection statistics directly comparable to the prediction results as measured by the *mean squared error* (MSE) obtained for the test sets. The test set results are shown in the Tables 1–3.

For the prostate data, the percentage correctly classified on the training set using cross–validation (classifying each sample to the largest of the fitted target values when using 0/1 dummy–coding for the group memberships) is 91.2% for all the parameter selection methods (it should be noted that this number happens to be identical to the test set result for most of the parameter selection methods).

It should be noted that that most of the displayed PRESS– (and GCV–) curves are relatively flat without a very distinct minimum point. Therefore it may be advantageous to employ either the 1 S.E. rule or the χ^2–rule to assure the selection of a simpler model. For the Prostate data, in particular, we note that the smallest available candidate regularization parameter value provides the minimum PRESS–value. The effect in terms of prediction when using the 1 S.E. rule or the χ^2–rule to obtain a simpler model varies between the data sets. For the Pork fat data the χ^2–rule gives better prediction than the other parameter selection methods for the SFA response, while the χ^2–rule selects a poorer model than the other parameter selection methods on the Prostate data.

For the most precise identification of the PRESS– and GCV–minima a numerical optimizer should be used. However, in most practical situations the suggested strategy of considering just a relatively dense subset of candidate regularization parameter values is usually enough for a good approximation of the minima before doing the subsequent identification of parsimonious models (based on the principle of Occam's razor) that predicts well.

Regularization type Parameter selection method	L_2	First derivative	Second derivative
Minimum PRESS value	0.057	0.047	0.038
Minimum GCV value	0.057	0.047	0.039
PRESS and 1 standard error rule	0.059	0.045	0.036
PRESS and χ^2–rule	0.073	0.047	0.039

Table 1: *Octane data. MSE (for the test data) using various regularization types and parameter selection methods.*

Regularization type Parameter selection method	L_2	First derivative	Second derivative
Minimum PRESS value	4.46	5.39	5.56
Minimum GCV value	4.36	5.45	5.58
PRESS and 1 standard error rule	4.58	5.56	5.72
PRESS and χ^2-rule	4.11	4.32	4.20

Table 2: *Pork fat data. MSE (for the test data) for the SFA response using various regularization types and parameter selection methods.*

Parameter selection method	PCC test set
Minimum PRESS value	91.2
Minimum GCV value	91.2
PRESS and 1 standard error rule	91.2
PRESS and χ^2-rule	88.2

Table 3: *Prostate data. Percentage of correctly classified (PCC) samples using the test set predictions of the selected $0-1$ dummy regression model based on L_2 regularization.*

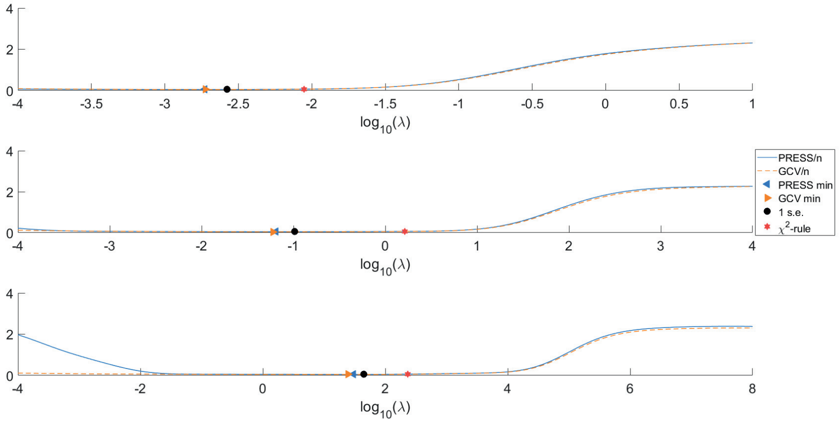


Figure 1: *Octane data. PRESS/n and GCV/n for a range of regularization parameter values and different regularization matrices. Top: L_2 regularization. Middle: 1st derivative regularization. Bottom: 2nd derivative regularization. The minimum PRESS and GCV values has been marked, as well as the regularization parameter value selected by the* 1 *S.E. rule and the χ^2-rule.*

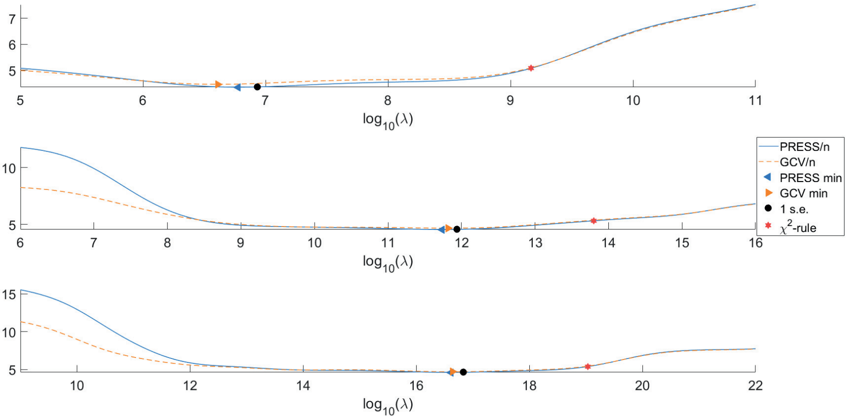


Figure 2: *Pork fat data and SFA response. PRESS/n and GCV/n for a range of regularization parameter values and different regularization matrices. Top: L_2 regularization. Middle: 1st derivative regularization. Bottom: 2nd derivative regularization. The minimum PRESS and GCV values has been marked, as well as the regularization parameter value selected by the* 1 *S.E. rule and the χ^2-rule.*

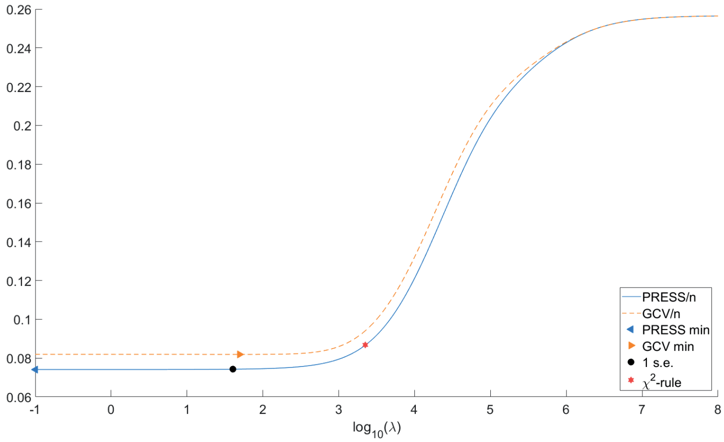


Figure 3: *Prostate data. PRESS/n and GCV/n for a range of regularization parameter values using L_2 regularization. The minimum PRESS and GCV values has been marked, as well as the regularization parameter value selected by the* 1 *S.E. rule and the χ^2–rule.*

4.1.3 Regression coefficients

Figure 4 shows the octane data together with the PRESS–minimal regression coefficients using the L_2–, the first derivative–, and the second derivative regularizations. Note that the choice of regularization matrix heavily influences the appearance of the regression coefficients without causing notable differences in the minimum PRESS– or GCV values. Table 1 confirms that the predictive powers are relatively similar for all these models. Doing consistent model interpretations solely based on the regression coefficients in figure 4 is clearly a challenging (if not impossible) task. Similar issues are discussed in Brown and Green (2009).

4.1.4 Computational speed

Table 4 shows the computational times for model selection with the different data sets and different types of regularization when varying the number of regularization parameter candidate values. The computing times in Table 4 also includes calculation of the regression coefficients corresponding to the minimal GCV and PRESS values for all responses. The main differences in computational time between finding the SVD in the case of L_2 regularization and in the

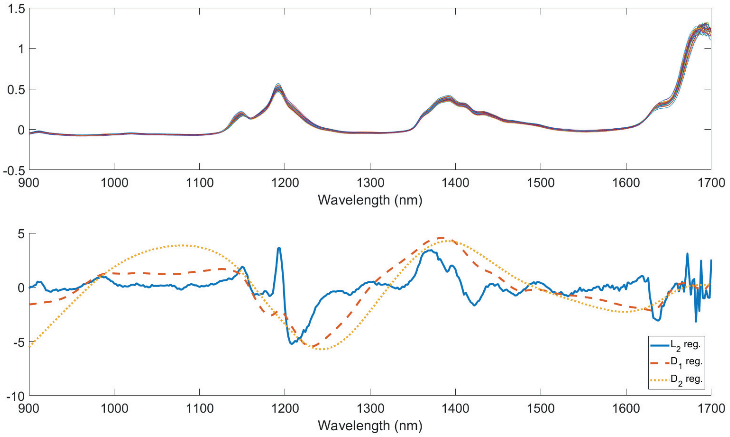


Figure 4: *Octane data. Top: Plot of the NIR spectra of octane. Bottom: PRESS–minimal regression coefficients based on different regularization matrices.*

cases of first- and second derivative regularization is due to the initial calculations of the transformed data $\tilde{\mathbf{X}}$, see Section 2.2. Similarly, the required transformation of the regression coefficients (see (15)) explains the increase in computational time from calculating the SVD only to finding PRESS, GCV and regression coefficients for a single regularization parameter value for first and second derivative regularization.

Number of λ–values Data (reg. type)	0 (SVD only)	1	10	100	1000	10000
Octane (L_2)	0.0014	0.0014	0.0014	0.0016	0.0024	0.013
Octane (1st derivative)	0.0034	0.0046	0.0051	0.0052	0.0055	0.017
Octane (2nd derivative)	0.0048	0.0074	0.0082	0.0082	0.0087	0.020
Pork fat (L_2)	0.018	0.023	0.023	0.026	0.040	0.26
Pork fat (1st derivative)	0.096	0.22	0.22	0.22	0.24	0.46
Pork fat (2nd derivative)	0.23	0.59	0.60	0.62	0.64	0.85
Prostate (L_2)	0.038	0.072	0.077	0.078	0.078	0.11

Table 4: *Computing time (in seconds) for model selection including finding the PRESS– and GCV–minimal regression coefficients when varying the number of candidate regularization parameter values. The times are the averages of* 50 *repeated runs rounded to the two most significant digits.*

4.2 The fast segmented virtual cross–validation

4.2.1 Datasets

In the following we will demonstrate the use of segmented virtual cross–validation with L_2 regularization for two datasets:

1. Raman spectra of fish oil, see Afseth et al. (2006). The data set consists of 42 sample segments including 3 replicate spectra of each unique sample giving a total of 126 rows and 2801 wavenumbers in the range $400cm^{-1}$ to $3200cm^{-1}$. The response variable was the iodine value (the response values were identical across each segment), which is frequently used as an indicator of the degree of unsaturation of fat, see Afseth et al. (2006). The spectra of this data set are plotted in Figure 5.

2. Raman milk spectra, see Afseth et al. (2010); Randby et al. (2012); Liland et al. (2016). The data set consists of 232 sample segments including between 6 and 12 replicate measurements of the associated unique sample giving a total of 2682 rows and 2981 wavenumbers in the range $120cm^{-1}$ to $3100cm^{-1}$. The response variables were the iodine value and the concentration of conjugated linoleic acid (CLA). Also for this dataset the response values were identical across each segment. The spectra of this data set are plotted in Figure 6.

For both datasets we have excluded the endpoint regions of the original spectra due to nosy and poor quality of the measurements. The wave numbers reported above are those included after this truncation. Approximately 2/3 of the replicate segments were used for model building and –selection, and the remaining 1/3 of segments were used as a test set.

The following four model selection strategies were considered: (i) $PRESS_{min}$ – the minimum $PRESS(\lambda)$–value from LooCV (ignoring the presence of sample segments), (ii) GCV_{min} – the minimum $GCV(\lambda)$–value, (iii) the $PRESS_{min}$ from the SCV (successively holding out the entire sample segments), and (iv) the $PRESS_{min}$ from the SvCV. We have chosen to focus only on the parameter selections associated with the minima of the various error curves in this part of our study (neither the χ^2-rule nor the 1 S.E. rule turned out to affect the model selections much).

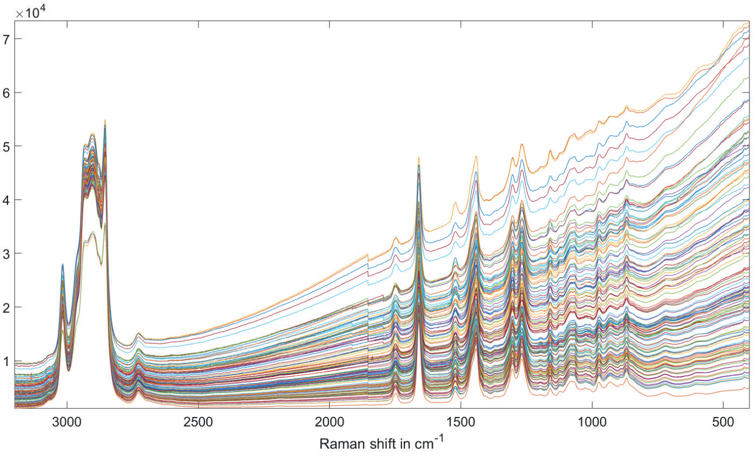


Figure 5: *Plot of the fish oil spectra.*

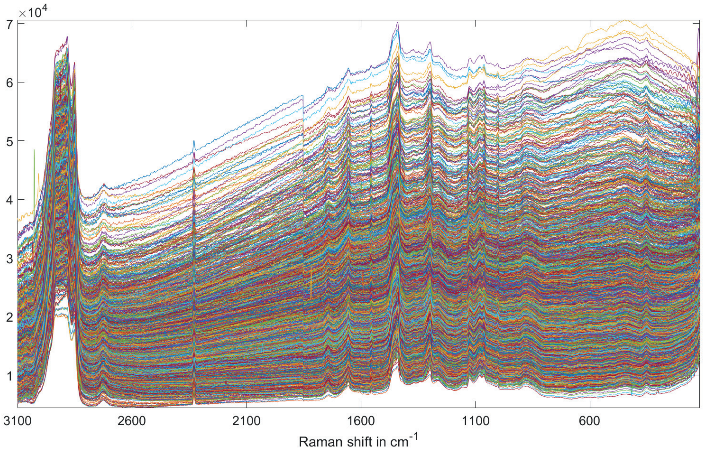


Figure 6: *Plot of the milk spectra.*

4.2.2 Model selection and prediction with raw data

For the fish oil data, the different error curves for model selection are shown in Figure 7. Note that:

1. The less relevant PRESS (and GCV–) curves based on the LooCV show considerably smaller values than the corresponding PRESS–values based on the SCV and the SvCV.

2. The regularization parameter values minimizing the PRESS (and GCV) are approximately a factor 10 smaller than the regularization parameter values minimizing the SCV and SvCV.

3. Although the PRESS–values based on the SvCV are clearly smaller than the PRESS-values based on the SCV, the shapes of the SvCV and SCV curves are quite similar, and the regularization parameter values defining the minima and selected models are not very different. Figure 8 shows the corresponding and highly similar $PRESS(\lambda)$–minimal regression coefficients based the SCV and the SvCV.

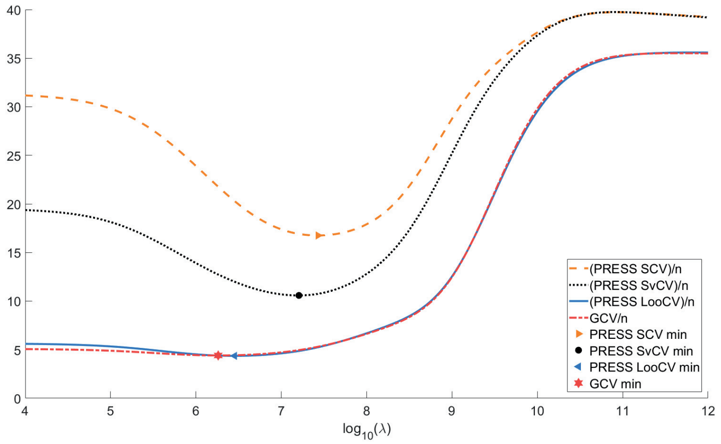


Figure 7: *Fish oil data (no pre-processing). Different model selection strategies for a range of regularization parameter values using* L_2 *regularization.*

Figure 9 shows the GCV and PRESS curves for the two available response alternatives (CLA and Iodine) in the milk data set. In this case all the associated minima indicate high agreement between the methods in the selection regularization parameter value for both responses. Near their minima the PRESS values based on the SvCV and the SCV are quite similar in magnitude, and consistently (but not much) larger than the corresponding less relevant GCV–

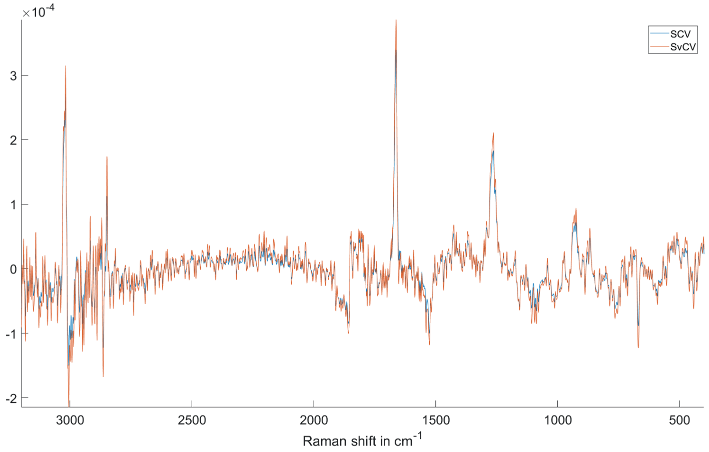


Figure 8: *Fish oil data (no pre-processing). $PRESS(\lambda)$–minimal regression coefficients (constant term omitted) for the SCV and the SvCV using L_2 regularization.*

and PRESS values based on the LooCV.

Selection curves / Data set	LooCV	GCV	SvCV	SCV
Fish oil data	20.3	21.5	12.3	9.7
Milk data (CLA)	0.0093	0.0093	0.0093	0.0093
Milk data (iodine)	2.59	2.58	2.58	2.58

Table 5: *MSE (from test data) for the data sets without any pre-processing according to the various model selection criteria.*

The associated prediction errors for the test data are shown in Table 5. The fish oil data indicate that the better models are obtained by considering either the SCV or the SvCV. In the view of Table 5, Figure 7 indicates that model selections based on the LooCV– and the GCV errors will lead to poorer predictions (the selected regularization parameters seems to be too small). For the milk data set there are no such clear distinction. Table 5 shows that the prediction results are almost identical for all the parameter selection alternatives as one should expect from the various error minima shown in Figure 9.

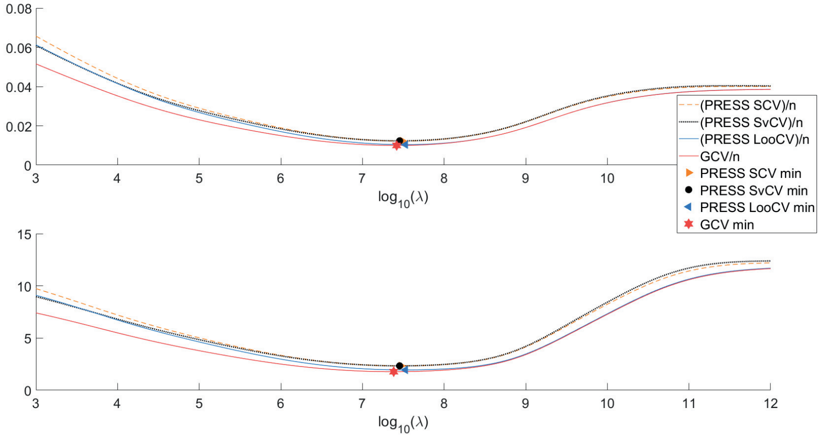


Figure 9: *Milk data (no pre-processing). Different model selection strategies for a range of regularization parameter values using L_2 regularization. Top: CLA. Bottom: Iodine value.*

4.2.3 Model selection and –predictions with pre–processed data

Spectroscopic measurements may be corrupted by both additive and multiplicative types of noise. Pre–processing of such data prior to modelling is therefore usually required. It is therefore of particular interest also to investigate how the model selection strategies considered above compare for pre–processed data. In particular we will consider the Extended Multiplicative Signal Correction (EMSC), see Afseth and Kohler (2012), with replicate corrections as described in Kohler et al. (2009).

The goal of the EMSC pre–processing is to adjust all the measured spectra to a common scale and to eliminate the eventual effects of additive noise. This includes the estimation of an individual scaling constant for each spectrum and an orthogonalization step that de–trends the spectra with respect to some set of lower order polynomial trends (the reader is referred to the provided references for the technical details). In the present examples with Raman spectra, the samples were orthogonalized with respect to the subspace including all polynomial trends up to the 6-th degree.

For datasets including segments of replicated measurements, a replicate correction step is often considered to alleviate the presence of inter-replicate variance. Such correction can be done by an initial EMSC–based pre–processing of the spectra in each sample segment. Thereafter, the corrected sample segments can be individually mean-centered, and organized into a full data matrix.

As we expect the dominant right singular vectors of the full matrix to account for the most dominant inter–replicate variance, orthogonalization of the data with respect to one or more of the associated dimensions contributes to making the replicates more similar, see Kohler et al. (2009) about the details. Because every sample in the training data set is included in the pre-processing, some bias affecting the subsequent PRESS–calculations and model selection must be expected.

Figure 10 shows the model selection for pre–processed fish oil data based on the pure EMSC and for the EMSC where 30% of the inter-replicate variance is removed. It is evident that the SCV and the SvCV become considerably more similar in the latter case. As one should expect, the GCV– and PRESS curves based on the LooCV seems to provide unrealistically low error values and the selection of lesser regularized models. This phenomenon does not occur with the SCV where an entire segment of replicates is held out in each cross–validation step. The SvCV seems remarkably robust against the removal of inter–replicate variance.

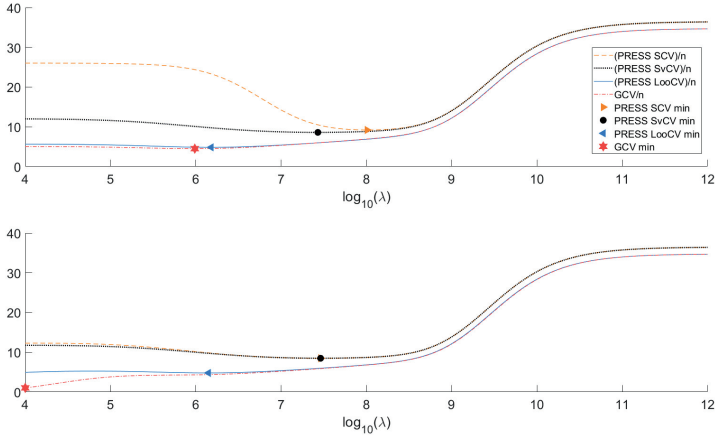


Figure 10: *Fish oil data. Model selection for data pre-processed with the EMSC both with and without replicate correction. Top: Standard EMSC pre-processing. Bottom: EMSC with 30% of the inter-replicate variance removed.*

Selection curves / Pre proc.	LooCV	GCV	SvCV	SCV
Raw data	20.3	21.5	12.3	9.7
EMSC	14.4	15.1	6.9	4.5
EMSC + 30% inter-replicate variance removed	14.4	15.9	6.7	6.7

Table 6: *Fish oil data. MSE (from test data) for different model selection strategies and different pre-processing alternatives.*

The prediction results for the test set of the fish oil data with the various pre-processing alternatives are presented in Table 6, and shows that the best predictions are obtained with the ordinary EMSC pre–processing and model selection based on the SCV. By simultaneously considering Figure 10, it is clear that the more heavily regularized among the selected models (those based on the largest regularization parameter values) perform better on the test set. With standard EMSC pre-processing the minima of the SvCV is located at a smaller regularization parameter value than for the SCV, suggesting an explanation of the difference in predictive performance.

For the milk data, the prediction error estimates obtained after pre–processing the data are similar for all the parameter selection methods (table omitted), as was also the case with the raw data.

4.2.4 Computational speed

Table 7 shows the computational times for the different model selection strategies. Both the PRESS– and the GCV values are included as computing only one of them takes approximately the same time as computing both. Because the size of the replicate segments are relatively small for these data sets (3 replicate measurements for the fish oil data and 6 to 12 replicate measurements for the milk data), the SVDs required for the internal orthogonalizations of the segments contribute insignificantly to the total computational load. The amount of computations required for model selection based on the SvCV is therefore quite comparable to the computations required for the LooCV version of PRESS (and for the GCV).

Data set	PRESS+GCV time	SCV time	SvCV+GCV time	SCV time / SvCV time ratio
Fish oil	0.017	0.98	0.023	42
Milk	2.8	416	3.2	130

Table 7: *Computational time for different model selection strategies for the Fish oil data and Milk data when considering* 500 *candidate regulariation parameter values. The time is given in seconds, rounded to two significant digits, and is the average of* 50 *repeated runs.*

5 Discussion and conclusions

The essence of the TR–framework described in the present work is that from a single SVD–calculation (of either the original data matrix $\mathbf{X}$ or a transformed version of it $\tilde{\mathbf{X}}$) it is possible to explore the entire regularized regression problem of interest. Our most notable finding is that the PRESS– and GCV values required for model selection(s) based on the LooCV or the GCV can be obtained at the computational cost of two matrix–vector multiplications for each choice of the regularization parameter value λ.

The applications in Section 4 confirm that our framework scales well when increasing the number of candidate regularization parameter values, as well as when considering multiple responses and in the case of 'small n with large p' problems. For smaller and medium sized data as well as for other situations where the required SVD can be calculated (or approximated) reasonably fast, the acquired computational efficiency allows for the exploration of a large number of candidate models in a very short amount of time. In many situations we have observed that such explorations can lead to relatively wide ranges of regularization parameter values corresponding to models of almost identical predictive performance.

For datasets having segments of similar samples, we have seen that the proposed SvCV gives a computationally efficient approximation of the traditional SCV. In the applications (Section 4) we correspondingly observed that the SvCV approximation of the SCV appears to work particularly well for model selection in the case of highly similar samples within each segment. Model selection based on the LooCV or GCV in such situations is not recommended as these tend to favour insufficient regularization resulting in models that predict poorer.

It is important to note that when the data set is pre–processed and/or transformed by a data dependent method, some bias both in the LooCV– and SvCV based PRESS values must be expected. The standardization of variables commonly used in RR is a typical example. The EMSC pre–processing that was used in Section 4.2.3 with or without replicate corrections is another. However, as the main purpose of the LooCV– and SvCV based PRESS values in the proposed framework is model selection rather than error estimation, the bias introduced by such pre–processing methods is not disruptive as long as the (training) data does not contain any serious outliers.

Although leverage correction of the model residuals for obtaining fast calculation of the LooCV in linear least squares regression problems is a well established result, there are some misleading assertions in the literature regarding both the properties and the accuracy of PRESS–values that requires clarification: i) In (Hansen, 2010, page 96) it is claimed that the leverage values are not invariant under row permutations of the $\mathbf{X}$–data making the PRESS–values dependent of the ordering of the data. This is not correct. When the rows of the data matrix are permuted it is actually simple to verify that the leverage values are invariant, and undergoes precisely the same permutation. Consequently, the correct leverage values will match up with the corresponding model residuals in the calculation of the $PRESS(\lambda)$ calculations assuring its invariance under any permutation of rows in the $(\mathbf{X}, \mathbf{y})$–data matrix. ii) In (Myers, 1990, page 399) it is claimed that the expression for fast calculation of $PRESS(\lambda)$ is only an approximation when performing centering and scaling of the data. This is, however, only true when the scaling factors are calculated from the data to be used in the model building process. The data centering, as such, does not corrupt the leverage– and $PRESS(\lambda)$–values as long as the $1/n$ terms are included in the associated leverage corrections of the model residuals. iii) The version of Ridge regression implemented in the MASS package by Venables and Ripley (2002) for the R programming language includes a fast calculation of the $GCV(\lambda)$–values for a desired vector of corresponding λ–values. The $1/n$ term is, however ignored when correcting the model residuals by the required averaged leverage value. Consequently, the resulting GCV–values are incorrect when centering of the data is included as a part of the Ridge Regression modelling.

Aren't there already existing fast algorithms for regularized regression, with CV to chose the regularization parameter? For sure the *glmnet* (a widely used R–package, see Friedman et al. (2010)) uses a clever numerical optimization and computational tricks to fit the models and to choose the regularization parameter at the same time via SCV, for either linear or logistic regression models, including the possibilities of doing both L_1 and L_2 regularization. This a reasonable objection, and we have therefore done a small comparison between an R–implementation of our method and the glmnet based on the R version 3.6.1. The comparison was executed based on 100 λ–values with an Intel® Xeon® E5-2630 v4 CPU at 2.2 GHz with 10 physical processor cores having access to 128 GB RAM, giving the following results:

	gasoline	fish	pork	prostate	milk
t$_{glmnet}$	5.04	57.96	134.62	60.46	713371.32
t$_{TR}$	0.01	0.13	0.16	0.34	66.69
t$_{ratio}$	458.45	452.84	820.82	179.93	10696.51
n (samples)	60	126	105	102	2682
p (predictors)	401	2801	5667	12600	2981

Table 8: *Comparison of LooCV for gmlnet and TR (measured in seconds) for differently sized datasets using 100 regularization parameter values.*

Table 8 shows that glmnet spent 8 days and 6 hours cross–validating the milk–data in comparison to the TR using only 67 seconds. Currently we cannot report any algorithm of similar performance for L_1–regularization problems. However, we think that the huge computational advantages available for the L_2-case may generate a genuine motivation to search for similar results with the regularization used in the LASSO, Elastic Net etc., see Friedman et al. (2009).

In conclusion, we believe in the presented work as a useful reference for future statistical texts and software dealing with parameter selection issues for Ridge Regression (and Tikhonov Regularization). The fast calculation of the $PRESS(\lambda)$ in (28), heavily relying on the SVD, represents yet another simple but powerful application of linear algebra to the benefit of multivariate data analysis.

A TR Prototype MATLAB code

```

function [press, bcoefs, b, lambda, H, U, s, V, GCV, L, idmin, rescv] = TregsLooCV(X, y, lambdas, type)
% ————————————————————————————————————————————————————————————————
% INPUTS:
% X       — Data matrix
% y       — Response vector
% lambdas — Vector of regularization parameter values
% type    — Regularization type (—1 for standardization, 0 for L2, 1 for 1st derivative regularization, etc ...)
% ————————————————————————————————————————————————————————————————
% OUTPUTS:
% press   — PRESS—statistic for input lambdas
% bcoefs  — Regression coefficients for selected lambda (no constant term)
% b       — Regression coefficients for PRESS—minimal lambda (with constant term)
% lambda  — Value of lambda minimising the PRESS—statistic
% H       — Vector of leverage values for all values of lambda
% U, s, V — SVD of matrix
% GCV     — GCV—statistic for input lambdas
% L       — Regularization matrix (empty for L2 regularizatoin)
% idmin   — Index of lambda value minimising the PRESS—statistic
% rescv   — LooCV—residuals
% ————————————————————————————————————————————————————————————————

[n,p] = size(X);
mX = mean(X); my = mean(y);
X = bsxfun(@minus,X,mX); y = y—my;

L = [];
if type > 0 % Create full rank discrete derivative matrix of order 'type'.
    epsilon = 1e—14;
    L = diff([speye(p);sparse(type,p)],type);
    L(end—type+1:end,:) = sqrt(epsilon)*Plegendre(type—1,p);
elseif type < 0 % Create variable standardization matrix.
    L = spdiags(std(X)',0,p,p);
end
if type ∼= 0, X = X/L; end

[U, S, V] = svd(X,'econ'); s = diag(S);
denom      = bsxfun(@plus,s,bsxfun(@rdivide,lambdas,s));
bcoefs     = V*bsxfun(@rdivide,(U'*y),denom);
H          = (U.^2)*bsxfun(@rdivide,s,denom)+1/n;
resid      = bsxfun(@minus,y,U*bsxfun(@rdivide,s.*(U'*y),denom));
rescv      = bsxfun(@rdivide,resid,(1—H));
press      = sum(rescv.^2)';
GCV        = (sum(resid.^2)./mean(1—H).^2)';

% Finding press—minimal model and corresponding regression coefficients:
[∼,idmin] = min(press); lambda = lambdas(idmin); h = H(:,idmin);
if type  ∼= 0, bcoefs = L\bcoefs; end
b          = [my—mX*bcoefs(:,idmin); bcoefs(:,idmin)]; % Constant term

end
```

```

function Q = Plegendre(d,p)
P = ones(p,d+1);
x = (-1:2/(p-1):1)';
for k = 1:d
    P(:,k+1) = x.^k;
end
[Q,~] = qr(P,0);
Q = Q';
end
```

B SvCV Prototype MATLAB code

```
function [press, bcoefs, b, lambda, H, U, s, V, GCV, L, idmin, rescv, Usegments] = TregsSvCV(X, y, lambdas, type, ...
     segments)
% ————————————————————————————————————————————————————————————————
% INPUTS:
% X          — Data matrix
% y          — Response vector
% lambdas    — Vector of regularization parameter values
% type       — Regularization type (—1 for standardization, 0 for L2, 1 for 1st derivative regularization, etc ...)
% segments   — List of integers identifying cross——validation segments
% ————————————————————————————————————————————————————————————————
% OUTPUTS:
% press      — PRESS—statistic for input lambdas
% bcoefs     — Regression coefficients for selected lambda (no constant term)
% b          — Regression coefficients for PRESS—minimal lambda (with constant term)
% lambda     — Value of lambda minimising the PRESS—statistic
% H          — Vector of leverage values for all values of lambda
% U, s, V    — SVD of matrix
% GCV        — GCV—statistic for input lambdas
% L          — Regularization matrix (empty for L2 regularizatoin)
% idmin      — Index of lambda value minimising the PRESS—statistic
% rescv      — LooCV—residuals
% Usegments — Sparse matrix representing the orthogonal transformations used in the SvCV
% ————————————————————————————————————————————————————————————————

% Finding orthogonal transformation and the modification to the leverage correction:
Usegments = segmentORTH(X, segments);
bs = (sum(Usegments,1).^2)';

[n,p] = size(X);
mX = mean(X); my = mean(y);
X = bsxfun(@minus,X,mX); y = y—my;

% Transforming data:
X = Usegments'*X; y = Usegments'*y;

L = [];
if type > 0
    epsilon = 1e—14;
    L = diff([speye(p);sparse(type,p)],type);
    P = Plegendre(type—1,p);
    L(end—type+1:end,:) = sqrt(epsilon)*P;
elseif type < 0
    L = spdiags(std(X)',0,p,p);
end

if type ∼= 0, X = X/L; end

[U, S, V]              = svd(X,'econ'); s = diag(S);
s_plus_lambdas_over_s = bsxfun(@plus,s,bsxfun(@rdivide,lambdas,s));

H      = bsxfun(@plus, (U.^2)*bsxfun(@ldivide,s_plus_lambdas_over_s, s), bs/n);
```

```
bcoefs = V*bsxfun(@ldivide,s_plus_lambdas_over_s,(U'*y));
res    = bsxfun(@minus,y,X*bcoefs);
rescv  = bsxfun(@rdivide,res,(1-H));
press  = sum(rescv.^2)';
GCV    = sum(bsxfun(@rdivide,res,mean(1-H)).^2)';

if type ~= 0, bcoefs = L\bcoefs; end

% Finding press-minimal model and corresponding regression coefficients:
[~,idmin] = min(press); lambda = lambdas(idmin); h = H(:,idmin);
if type  ~= 0, bcoefs = L\bcoefs; end
b         = [my-mX*bcoefs(:,idmin); bcoefs(:,idmin)]; % Constant term

end

function U = segmentORTH(X, segments)
n         = size(X,1);
nsegments = max(segments);
U         = sparse(n,n);
for k = 1:nsegments
    ind             = find(segments==k);
    [U(ind, ind),~] = svd(X(ind,:),'econ');
end
end

function Q = Plegendre(d,p)
P = ones(p,d+1);
x = (-1:2/(p-1):1)';
for k = 1:d
    P(:,k+1) = x.^k;
end
[Q,~] = qr(P,0);
Q = Q';
end
```

C Computational complexity of the fast LooCV

For a more precise description of the computational complexity involved in calculating the fast LooCV, an approximate count of the floating point operations (flop) is required. According to Björck Björck (2016), an approximate flop count for finding the reduced SVD (using a QR–SVD algorithm with Golub–Kahan–Householder bidiagonalisation) of a $(n \times p)$–matrix is $12pn^2 + (16/3)n^3$ when assuming $p \geq n$. The remaining computations consist of centering, calculating the PRESS values, and calculating the PRESS–minimal regression coefficients for every response. With q different responses, the approximate flop count for these computations is given by:

$$(3np + 3nq + nr + 2nrq - q + 2prq + pq) + n_\lambda(3r + 2nr + 2nrq + qr + 4nq), \tag{37}$$

where n_λ denotes the number of different candidate regularization parameter values. For $p \geq n$, the computations needed to evaluate the $PRESS(\lambda)$–function for one additional regularization parameter is of the order $\mathcal{O}(qn^2)$, and in particular the additional computations are independent of the number (p) of measured features. This makes the fast LooCV highly useful also for problems where the number of features are even larger than the number of samples. To calculate the cost of finding the corresponding $GCV(\lambda)$–values as well as GCV–minimal regression coefficients one should add $5nn_\lambda q - q + q(2pr + p)$ to the above flop count. Note that the choice of regularization matrix $\mathbf{L}$ matters here, and for $\mathbf{L} \neq \mathbf{I}$ there are additional calculations (see Section 2.2) that must be taken into account. The exact number of flops associated with these additional calculations will depend on the sparsity structure of $\mathbf{L}$ and to what extent that sparsity can be utilized in the required calculations.

D Computational savings of the SvCV compared to the SCV

To assess the computational savings of the SvCV over the SCV, flop count approximations for the associated $PRESS$-values must be compared. (We only consider the situation involving L_2 regularization, i.e. the identity matrix $\mathbf{I}$ acting as the regularization matrix.) Let K denote the number of segments , and assume for simplicity that the various segments sizes are all bouded from above by the constant B_{ss}. The approximate number of flops required for the SVDs for the different parameter selection methods when using the entire data set for training are given by the formulas in Table 9 (using the approximate flop count for the SVD given in Björck (2016)). The Table shows that the size of (all but one of) the required SVDs for the SvCV are much smaller than for the SCV (assuming the size of each segment is much smaller than the total number of samples, which is obviously the case in most real applications). This is primarily what makes the SvCV superior to the SCV in terms of computational efficiency.

If the block diagonal structure of the transformation matrix T is utilized, the matrix multiplication part of the orthogonal transformation (33) for the SvCV requires approximately

$$2B_{ss}(B_{ss}-1) + K \cdot B_{ss} \cdot p(2B_{ss}-1) + q \cdot B_{ss}(2B_{ss}-1) \tag{38}$$

flops. For keeping track of the remaining computations needed for the SvCV we can use the flop count approximations in Section C, as the flop count for the SvCV and the LooCV will be identical after applying the orthogonal transformation required for the SvCV. The approximate flop count of the remaining computations for the SCV is given by

$$2K \cdot B_{ss}(q+p) - q + r_{train} \cdot q(2B_{ss}-1) + q \cdot n_{\lambda} \cdot K[3r_{train} + 2p \cdot r_{train} + p + 2p \cdot n_{test} + 3n_{test}] \tag{39}$$

where $r_{train} = \min(n_{train}, p)$ and n_{train} is the number of samples in the training set.

Although the main computational cost with model validation is with the initial SVD(s) there will also be an additional computational cost for each candidate regularization parameter value for which we want to validate the model. Consider the case $p > n$ of most interest for the present work (the number of features is greater than the number of samples). From the above reasoning we observe that when considering additional regularization parameter values, the SCV flop count depends on the number of features p for each candidate value. The above flop count for the SvCV and the LooCV flop count in appendix C shows that this is not the case for the SvCV. When p is very large it might therefore be computationally inefficient (or even infeasible) to validate models for a large number of regularization parameter values based on the SCV. Clearly, the SvCV is the method of choice among the two in such cases.

Par. sel. method	Approx. flops for SVD(s)	Approx. for Raman milk dataset
LooCV/GCV	$12pn^2 + \frac{16}{3} \cdot n^3$	$3.602 \cdot 10^{11}$
SvCV	$12pn^2 + \frac{16}{3} \cdot n^3 + K \cdot \big(12p \cdot B_{ss}^2 + \frac{16}{3} \cdot B_{ss}^3\big)$	$3.614 \cdot 10^{11}$
SCV	$K \cdot \big(12p \cdot (n - B_{ss})^2 + \frac{16}{3} \cdot (n - B_{ss})^3\big)$	$8.272 \cdot 10^{13}$

Table 9: *Approximate flop counts for the required SVD(s) in the different parameter selection methods when assuming* $p \geq n$. *Figures for the Raman milk data set (with* $n = 2682$, $p = 2981$, $K = 232$ *and* $B_{SS} = 12$, *see Section 4.2.1) are shown in the last column.*

E Acknowledgements

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

References

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

N. K. Afseth, J. P. Wold, and V. H. Segtnan. The potential of raman spectroscopy for characterisation of the fatty acid unsaturation of salmon. Analytica chimica acta, 572(1): 85–92, 2006.

N. K. Afseth, H. Martens, Å. Randby, L. Gidskehaug, B. Narum, K. Jørgensen, S. Lien, and A. Kohler. Predicting the fatty acid composition of milk: A comparison of two fourier transform infrared sampling techniques. Applied spectroscopy, 64(7):700–707, 2010.

D. M. Allen. Mean square error of prediction as a criterion for selecting variables. Technometrics, 13(3):469–475, 1971.

D. M. Allen. The relationship between variable selection and data agumentation and a method for prediction. Technometrics, 16(1):125–127, 1974.

H. Best and C. Wolf. The Sage handbook of regression analysis and causal inference. Sage, 2014.

Å. Björck. Numerical methods in matrix computations. Springer, 2016.

C. D. Brown and R. L. Green. Critical factors limiting the interpretation of regression vectors in multivariate calibration. TrAC Trends in Analytical Chemistry, 28(4):506 – 514, 2009. ISSN 0165-9936. doi: http://dx.doi.org/10.1016/j.trac.2009.02.003. URL http://www.sciencedirect.com/science/article/pii/S0165993609000363.

J. Friedman, T. Hastie, and R. Tibshirani. The elements of statistical learning, volume 1. Springer series in statistics Springer, Berlin, 2009.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1):1–22, 2010.

G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics, 21(2):215–223, 1979.

P. C. Hansen. Discrete Inverse Problems. Society for Industrial and Applied Mathematics, 2010. doi: 10.1137/1.9780898718836. URL `http://epubs.siam.org/doi/abs/10.1137/1.9780898718836`.

J. S. U. Hjorth. Computer Intensive Statistical Methods: Validation, Model Selection and Bootstrap. Chapman and Hall/CRC, 1993.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1):55–67, 1970.

A. S. Householder. The theory of matrices in numerical analysis. Blaisdell book in the pure and applied sciences. Blaisdell Pub. Co., 1965.

U. Indahl. A twist to partial least squares regression. Journal of Chemometrics, 19(1):32–44, 2005.

J. H. Kalivas. Two data sets of near infrared spectra. Chemometrics and Intelligent Laboratory Systems, 37(2):255 – 259, 1997. ISSN 0169-7439. doi: https://doi.org/10.1016/S0169-7439(97)00038-5. URL `http://www.sciencedirect.com/science/article/pii/S0169743997000385`.

J. H. Kalivas. 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, 26(6):218–230, 2012.

A. Kohler, U. Böcker, J. Warringer, A. Blomberg, S. Omholt, E. Stark, and H. Martens. Reducing inter-replicate variation in fourier transform infrared spectroscopy by extended multiplicative signal correction. Applied spectroscopy, 63(3):296–305, 2009.

E. Kreyszig. Introductory functional analysis with applications, volume 1. Wiley New York, 1978.

K. H. Liland, A. Kohler, and N. K. Afseth. Model-based pre-processing in raman spectroscopy of biological samples. Journal of Raman Spectroscopy, 47(6):643–650, 2016.

L. B. Lyndgaard, K. M. Sørensen, F. Berg, and S. B. Engelsen. Depth profiling of porcine adipose tissue by raman spectroscopy. Journal of Raman Spectroscopy, 43(4):482–489, 2012.

R. H. Myers. Classical and modern regression with applications, volume 1. Duxbury Press, 1990.

D. L. Phillips. A technique for the numerical solution of certain integral equations of the first kind. Journal of the ACM (JACM), 9(1):84–97, 1962.

Å. Randby, M. R. Weisbjerg, P. Nørgaard, and B. Heringstad. Early lactation feed intake and milk yield responses of dairy cows offered grass silages harvested at early maturity stages. Journal of Dairy science, 95(1):304–317, 2012.

D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer cell, 1(2):203–209, 2002.

M. Stone. Cross-validatory choice and assessment of statistical predictions. Journal of the royal statistical society. Series B (Methodological), pages 111–147, 1974.

A. N. Tikhonov. On the solution of ill-posed problems and the method of regularization. In Doklady Akademii Nauk, volume 151, number 3, pages 501–504. Russian Academy of Sciences, 1963.

W. N. Venables and B. D. Ripley. Modern Applied Statistics with S. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0. R MASS package Available at: https://cran.r-project.org/web/packages/MASS/index.html. Version 7.3.50.

PAPER IV

Fast identification of good Regularization Parameter Values for Regularized Linear Discriminant Analysis by Cross-validated Ridge Regression

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

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

Abstract

Regularization in Linear Discriminant Analysis (LDA) is required in many practical situations. When using Regularized LDA (RLDA) it is necessary to determine an appropriate regularization parameter for obtaining a robust model giving accurate predictions. Many methods have been developed for choosing a regularization parameter, and cross-validation in particular is frequently used. For big data sets the computations needed to validate many different candidate values for the regularization parameter can be significant. In this work we suggest a computationally efficient regression-based heuristic for selecting a value for the regularization parameter for RLDA. The motivation for the regression-based criterion comes from the link between Ridge regression (RR) and RLDA. By using fast model selection strategies for RR we are able to select a good regularization parameter value for RLDA from a large number of candidate regularization parameter values very efficiently. The heuristic is tested on several data sets, and empirical results indicate that the predictive power of the models obtained with this criterion appears to be comparable to the predictive power of models obtained by cross-validation on the percentage of samples correctly classified.

1 Introduction

In the present work we are particularly interested in procedures for fast identification of good regularization parameter values for Regularized Linear Discriminant Analysis (RLDA). RLDA is a well-studied generalization of Linear Discriminant Analysis (LDA), and is particularly useful when LDA cannot be applied directly. This is for example the case when the number of predictor variables is greater than the number of samples, or when the LDA problem is ill-conditioned[9, 16, 25]. In such cases the total and within-groups scatter matrices derived from the data set will not be invertible[16]. Several useful approaches have been proposed to overcome this and similar problems[3, 4, 1, 8, 25]. Our focus is on the application of regularization in the form of adding a scaled identity matrix to the within group scatter matrix. To obtain a good model it is necessary to select an appropriate regularization parameter value[16], and it is therefore necessary to employ some criteria for choosing a good value for this parameter. Selection of the regularization parameter is often done by cross-validation on the percentage of samples correctly classified, but for large data sets this can be computationally expensive and time consuming.

There is a close relationship between RLDA and regularized regression. It is well-known that solving the LDA problem for a 2-class problem is equivalent to solving the associated linear regression problem using a 0/1 dummy coding of the group memberships as the response vector, in the sense that the LDA discriminant function coefficients are proportional to the regression coefficients obtained by the dummy regression[3]. Considerable efforts have been put into extending this relationship to classification problems with more than 2 classes as well as the inclusion of regularization [11, 9, 24, 25, 18]. By utilizing the link between Ridge regression (RR) and RLDA, we suggest an efficient heuristic for selecting an appropriate regularization parameter value based on the PRESS-statistic obtained from a 0/1 dummy coded regression problem. Using a regression based criterion to determine the regularization in RLDA has been mentioned as a possibility in the earlier literature[9], but to our knowledge no detailed investigations are available.

The structure of the present work is as follows. We start by reviewing LDA, its regularized version, and RR. Then we discuss the computational relationship between RLDA and RR, and its implications regarding the selection of good regularization parameter values for RLDA. Finally the suggested heuristic will be tested on several data sets and compared to regularization parameter selection by cross-validation on the percentage of samples correctly classified.

2 Regularized Linear Discriminant Analysis

In the following we use $\boldsymbol{X}$ to denote an $n \times p$ data matrix of measurements (n samples and p features), and the p-vector $\boldsymbol{\mu}$ denotes the (global) mean of the $\boldsymbol{X}$-columns. The sample group memberships are represented by both an n-dimensional vector $\boldsymbol{G}$ containing the labels $\{1, 2, 3, \ldots, g\}$ (where g is the number of different groups) and an associated $n \times g$ matrix $\boldsymbol{Y}$ of $0/1$ dummy-coded group memberships. The number of samples in the k-th group is denoted by n_k for $1 \leq k \leq g$, and the feature means associated with the individual groups are represented by the p-vectors $\boldsymbol{\mu}_k$ for $1 \leq k \leq g$. We use $\mathbf{1}_m$ and $\mathbf{0}_m$ to denote the m-dimensional vectors of ones and zeros, respectively. All vectors are assumed to be column vectors unless otherwise stated. We denote the globally centred data matrix by $\boldsymbol{X}_S$, and the group-centred data matrix by $\boldsymbol{X}_G$. Finally we define the total-, the within groups, and between groups scatter matrices by

$$\boldsymbol{S}_T = \frac{1}{n} \cdot \boldsymbol{X}_S' \boldsymbol{X}_S,$$

$$\boldsymbol{S}_W = \frac{1}{n} \cdot \boldsymbol{X}_G' \boldsymbol{X}_G,$$

$$\boldsymbol{S}_B = \frac{1}{n} \cdot \sum_{k=1}^{K} n_k (\boldsymbol{\mu}_k - \boldsymbol{\mu})' (\boldsymbol{\mu}_k - \boldsymbol{\mu}),$$

respectively. It can be shown[3] that $\boldsymbol{S}_B = \boldsymbol{S}_T - \boldsymbol{S}_W$.

In Fisher's[2] description of linear discriminant analysis the idea is to identify linear combinations $\boldsymbol{t} = \boldsymbol{X}\boldsymbol{v}$ of the original features that are particularly useful for linear discrimination between two or more groups. This is made more precise by finding directions maximizing the between group scatter relative to the within group scatter[3]. That is, we want to maximize the following ratio:

$$\frac{\boldsymbol{v}' \boldsymbol{S}_B \boldsymbol{v}}{\boldsymbol{v}' \boldsymbol{S}_W \boldsymbol{v}}. \tag{1}$$

The intuition behind the above optimization problem is that we want to find directions in the sample space that maximizes the distance between samples of different groups while at the same time making the distances between samples corresponding to the same group small. Maximization of the ratio in (1) can be obtained by solving the generalized eigenvalue/-vector problem:

$$\boldsymbol{S}_B \boldsymbol{v} = \alpha \boldsymbol{S}_W \boldsymbol{v}, \tag{2}$$

and since $\boldsymbol{S}_B = \boldsymbol{S}_T - \boldsymbol{S}_W$ it follows[25] that a solution $\boldsymbol{v}$ of (2) alternatively can be obtained by solving the generalized eigenvalue/-vector problem:

$$\boldsymbol{S}_B\boldsymbol{v} = \alpha\boldsymbol{S}_T\boldsymbol{v}. \tag{3}$$

If the matrix $\boldsymbol{S}_T$ is invertible, this problem corresponds to the ordinary eigenvalue/-vector problem:

$$\boldsymbol{S}_T^{-1}\boldsymbol{S}_B\boldsymbol{v} = \alpha\boldsymbol{v}. \tag{4}$$

If $\boldsymbol{S}_T$ is not invertible (which is obviously the case when the number of features $p > n$) or poorly conditioned, a useful solution to Fisher's LDA problem cannot be found directly and some sort of stabilization or regularization is required. One method of regularization (see e.g. [9], [16] or [25]) is to add a scaled version of the identity matrix to $\boldsymbol{S}_W$ or $\boldsymbol{S}_T$. Let $\lambda \geq 0$ be the scaling factor of the added identity matrix. The resulting regularized RLDA eigenvalue/-vector problem is:

$$\boldsymbol{S}_B\boldsymbol{v} = \alpha(\boldsymbol{S}_W + \lambda\boldsymbol{I})\boldsymbol{v}, \tag{5}$$

alternatively with $\boldsymbol{S}_W$ replaced by $\boldsymbol{S}_T$ if the formulation in terms of the total scatter matrix is used[25]. As eigenvectors are only unique up to scaling factor this eigenvalue problem has infinitely many solutions, and it is therefore common to add a constraint to the eigenvectors to obtain a unique solution. One such constraint[9] is:

$$\boldsymbol{v}'(\boldsymbol{S}_W + \lambda \cdot \boldsymbol{I})\boldsymbol{v} = 1. \tag{6}$$

Another possible constraint is to similarly require unit norm of the eigenvectors in the metric induced by the regularized total scatter matrix[25]. The matrix $\boldsymbol{V}$ whose columns consist of all vectors satisfying (5) and (6) is referred to as a solution to the RLDA problem. In the present work samples will be classified as belonging to the class of the nearest group centre in the Euclidean metric in the projected space. That is, the classification of a sample $\boldsymbol{x}$ is given by:

$$\underset{i=1,\ldots,g}{arg\ min}\ \|(\boldsymbol{x} - \boldsymbol{\mu}_i)\boldsymbol{V}\|_2. \tag{7}$$

We note that it is also possible to use, for example, a nearest neighbour classifier in projected space[16].

An alternative approach to LDA is statistically motivated and based on Bayes rule together with the assumption of multivariate normally distributed data for each group with a common covariance structure[3]. This leads to classification based on the Mahalanobis metric defined by

$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)'\boldsymbol{S}_W^{-1}(\boldsymbol{x}_1 - \boldsymbol{x}_2)}$. If we denote the prior probability of a sample belonging to class i by p_i, the classification of a sample $\boldsymbol{x}$ is given by

$$\underset{i=1,\ldots,g}{arg\ min}\ (d(\boldsymbol{x}, \boldsymbol{\mu}_i) - 2\log(p_i))\,. \tag{8}$$

If the prior probabilities are identical for all classes, then the classification formula reduces to classifying a sample to the closest group center in the Mahalanobis metric. The corresponding regularized version is obtained by replacing $\boldsymbol{S}_W$ with $\boldsymbol{S}_W + \lambda\boldsymbol{I}$ for an appropriate choice of the regularization parameter $\lambda > 0$ in the formula for the Mahalanobis distance. When used for classification the two approaches to (R)LDA are equivalent (as long as one projects onto the full subspace for RLDA and use the same priors if applicable)[9].

3 Ridge Regression

The motivation for introducing Ridge regression (RR)[13] is similar to the motivation for introducing regularization in LDA. Consider the ordinary least squares problem $\boldsymbol{X}_S\boldsymbol{b} = \boldsymbol{y}_S$, where $\boldsymbol{y}_S$ is a centered vector of responses. If the number of features is greater that the number of samples, the solution of the corresponding normal equations $(\boldsymbol{X}_S'\boldsymbol{X}_S)\boldsymbol{b} = \boldsymbol{X}_S'\boldsymbol{y}$ with respect to $\boldsymbol{b}$ is not unique as the matrix $\boldsymbol{X}_S'\boldsymbol{X}_S$ is singular. Attempting to apply ordinary least squares directly in this case will result in overfitting to the data. If the matrix $\boldsymbol{X}_S'\boldsymbol{X}_S$ is ill-conditioned the solution vector may exhibit unwanted behavior such as neighboring regression coefficients being large in absolute value with different signs[3] which may give solutions that do not make physical sense for practical problems[13]. A better solution can in this case be obtained by adding a regularization term to the least squares problem. The RR problem can then be formulated as the following modified least squares problem:

$$\begin{bmatrix} \boldsymbol{X}_S \\ \sqrt{\lambda}\boldsymbol{I} \end{bmatrix} \cdot \boldsymbol{b} = \begin{bmatrix} \boldsymbol{y}_S \\ \boldsymbol{0}_p \end{bmatrix}. \tag{9}$$

The least squares solution of the above set of linear equations minimizes the sum

$$||\boldsymbol{X}_S\boldsymbol{b} - \boldsymbol{y}_S||^2 + \lambda||\boldsymbol{b}||^2. \tag{10}$$

The magnitude of λ regulates the strength of this requirement, i.e. by increasing λ the L_2-norm of $\boldsymbol{b}$ is forced to decrease. From a Bayesian viewpoint, the solution to (9) is equivalent to the solution of the standard linear regression problem when using a Gaussian prior with mean zero[19]. From the statistical viewpoint the regularization parameter λ can be interpreted as being related to the variance of the regression coefficients.

Denoting the reduced singular value decomposition (SVD) of $\boldsymbol{X}_S$ by $\boldsymbol{X}_S = \boldsymbol{USV}'$, the solution to (9) and (10) can be formulated as (see [3] and [10]):

$$\boldsymbol{b} = \boldsymbol{V}(\boldsymbol{S}'\boldsymbol{S} + \lambda\boldsymbol{I})^{-1}\boldsymbol{S}\boldsymbol{U}'\boldsymbol{y}_S. \tag{11}$$

The above expression is useful because it involves only the inversion of a diagonal matrix and a few matrix-vector multiplications. Thus, having calculated the reduced SVD of the data matrix once, the computational costs of calculating the regression coefficients for any choice of the parameter λ are low[8]. The constant term b_0 is given by $b_0 = \overline{\boldsymbol{y}} - \boldsymbol{X}_S\boldsymbol{b}$.

Next we outline how the predicted residual error sum of squares (PRESS) statistic can be calculated efficiently for any value of λ when the reduced SVD of the centred data matrix has been calculated based on Indahl et al.[15]. The idea is to use the Sherman-Morrison-Woodbury formula[7] to perform rank 1 updates. More precisely, it can be shown[3] that a Leave-one-out (Loo) residual for a standard linear regression model can be calculated as

$$r_j^* = \frac{r_j}{1 - h_{\lambda,j}} \tag{12}$$

where r_j^* is the LOO residual for the j-th sample, r_j is the residual for the j-th sample for the full model, and $h_{\lambda,j}$ is the leverage of the j-th sample (the j-th diagonal element of the projection matrix projecting onto the solution space of the least squares problem). We have shown that regression coefficients can be calculated efficiently which means that regression residuals can also be calculated efficiently. To deal with the leverage values, let $\boldsymbol{S}_\lambda$ denote the diagonal matrix with elements $\frac{s_i}{\sqrt{s_i^2+\lambda}}$, where s_i is the i-th singular value of the matrix $\boldsymbol{X}_S$. It can be shown[15] that the upper n rows of the projection matrix for the RR–problem is given by

$$\boldsymbol{U}_\lambda = \boldsymbol{USS}_\lambda^{-1}. \tag{13}$$

As the leverage values are the diagonal elements of the projection matrix, which is given by $\boldsymbol{U}_\lambda\boldsymbol{U}_\lambda'$, we can calculate the leverages by squaring the elements of $\boldsymbol{U}_\lambda$ and adding the elements row-wise. This shows that the PRESS-statistic associated with (9) can be calculated at very little computational cost for a large range of λ-values, making model selection based on the PRESS-statistic highly efficient. See appendix B for a prototype MATLAB implementation of the above method for calculating the PRESS-statistic.

The extension of the RR problem to a multivariate response is straightforward. For a $n \times g$ response matrix $\boldsymbol{Y}$, the multi response RR problem is given by

$$\begin{bmatrix} \boldsymbol{X}_S \\ \sqrt{\lambda} \cdot \boldsymbol{I} \end{bmatrix} \boldsymbol{B} = \begin{bmatrix} \boldsymbol{Y}_S \\ \boldsymbol{0}_{p,g} \end{bmatrix}, \tag{14}$$

where $\boldsymbol{B}_0 \in \mathbb{R}(1, g)$ and $\boldsymbol{B} \in \mathbb{R}(p, g)$. Here we wish to minimize the Frobenius norm of the residual. Note that solving (14) is equivalent to solving the univariate problems obtained by considering each column of $\boldsymbol{Y}$ as a single response vector. In principle we can therefore calculate a PRESS-minimal regularization parameter for each univariate problem. As we want to use the regularization parameter obtained from regression in an RLDA model we require a single regularization parameter. We will therefore select the regularization parameter minimizing the overall PRESS-statistic for the multivariate problem.

Hastie et al.[9] established a link between regularized regression (in the form of an optimal scoring problem) and regularized linear discriminant analysis. One of their results can be formulated as saying that for any regularization parameter $\lambda \geq 0$, there exists a matrix $\boldsymbol{Z}_\lambda \in \mathbb{R}(g, g-1)$ such that for any $\boldsymbol{x} \in \mathbb{R}^p$

$$\underset{i=1,\dots,g}{arg\ min} \|(\boldsymbol{x} - \boldsymbol{\mu}_i)' \boldsymbol{B} \boldsymbol{Z}_\lambda\|^2 = \underset{i=1,\dots,g}{arg\ min} (\boldsymbol{x} - \boldsymbol{\mu}_i)' (\boldsymbol{S}_W + \lambda \boldsymbol{I})^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i), \tag{15}$$

where $\boldsymbol{B}$ is the least squares solution of (14). The $\boldsymbol{Z}_\lambda$ matrix can be viewed as a post-processor[9] that changes basis to a space where nearest centroid classification with the Euclidean metric gives the same classification as the one obtained using RLDA. The existence of such a matrix provides a clear relationship between regularized regression and RLDA, and motivates the idea of using the PRESS-statistic from a regression problem for selecting the regularization parameter for RLDA.

4 Parameter selection in RLDA

4.1 Cross-validation on percentage of samples correctly classified

A common way of selecting the regularization parameter for RLDA is using cross-validation. In this approach one calculates the percentage of samples correctly classified under cross-validation for a selection of regularization parameter values, and then select the regularization parameter value maximizing the percentage of samples correctly classified. In the case of a tie between different regularization parameter values one can, for example, select the largest parameter value (i.e. simplest model) from the parameters giving the best prediction. The naive approach to cross-validation would be a complete refit of the model for each choice of regularization parameter value. This can be computationally expensive, and there are some tricks that can be used for significant computational savings. In the examples we will use 5-fold cross-validation based on the algorithm given in [16] as one method for parameter selection. This algorithm achieves considerable computational savings compared to the naive approach. The algorithm is based on a similar idea to the one allowing for efficient calculation of ridge regression coefficients for a large number of regularization parameters once a single SVD of the data matrix has been calculated. Note that the regularization in RLDA only affects the non-zero eigenvalues of the within group (or total) scatter matrix, and that the eigenvectors of the within group (or total) scatter matrix does not change when adding a scalar multiple of the identity matrix. This implies that a single SVD of the data matrix (globally or group centred depending on whether $\boldsymbol{S}_T$ or $\boldsymbol{S}_W$ is used) is sufficient to efficiently invert the scatter matrix used for any choice of regularization parameter The algorithm still requires the computation of a SVD for each choice of regularization parameter because of the maximization of the between group scatter, but on a much smaller matrix making the algorithm quite efficient. See [16] for the details.

For the special case of LooCV there exists formulae for calculating the LooCV distances without refitting the model. These formulae are based on rank 1 updates and calculates adjustment factors to convert the training set distances to LOO distances. The formulae are given in Hjort[12] and Fukunaga[5], and are reproduced in Ripley[22]. As noted in an errata[21] to Ripley's book, there are some mistakes in the formulae given in [22], and the formulae reproduced below are the corrected ones from the errata. Let Δ^2_{jk} be the squared training set distance between sample j and the group center of group k in the Mahalanobis metric associated with the regularized within group covariance matrix. If sample j belongs to group c, then the squared LOO Mahalanobis distance between sample j and group c is given by

$$\tilde{\Delta}^2_{jc} = \Delta^2_{jc} \cdot \left(\frac{n_c}{n_c - 1}\right)^2 / \left(1 - \frac{n_c}{(n_c - 1)(n - g)}\Delta^2_{jc}\right). \tag{16}$$

If sample j does not belong to group c, then squared LOO Mahalanobis distance between sample j and group c is given by

$$\tilde{\Delta}^2_{jc} = \Delta^2_{jc} \cdot \left(1 + \frac{((\boldsymbol{x}_j - \boldsymbol{\mu}_c)'(\boldsymbol{S}_W + \lambda \boldsymbol{I})^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_c))^2}{((n-g)(n_c - 1)/n_c - \Delta^2_{jc})\Delta^2_{jk}}\right). \tag{17}$$

A prototype MATLAB implementation of using the above formulae for LOOCV based on prediction performance is included in appendix C.

One disadvantage of the above approach is that the number of samples correctly classified is not an ideal metric for model selection as there will generally be a range of regularization parameter values that apparently produce the same classification rate. We tried an alternative criterion motivated by Fisher's idea behind LDA. We used a cross-validation approach, and for each fold we calculated distances from each sample to all group centers. We then calculate the ratio of the distance between a sample and the correct group center to the distance between the sample and the closest incorrect group center. We then sum these ratios over all samples, and select the regularization parameter value minimizing this sum. The motivation of this criterion is then to find a choice of regularization parameter value mapping samples as close to the correct group center as possible and as far away from other group centers as possible. In experiments the criterion appeared to be a bit too sensitive to outliers and require too much computation, and we therefore abandoned this approach. The sensitivity to outliers can be adjusted for by setting a maximum number each ratio of distances are allowed to contribute to the total, but the computational inefficiency makes other parameter selection methods more favorable.

4.2 Dummy regression

An alternative parameter selection method from RLDA can be obtained from a regression problem. Consider the multivariate regression problem

$$\begin{bmatrix} \boldsymbol{X}_S \\ \sqrt{\lambda} \cdot \boldsymbol{I} \end{bmatrix} \boldsymbol{B} = \begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{0}_{p,g} \end{bmatrix}, \tag{18}$$

where the $\boldsymbol{Y}$-matrix consists of $0/1$ dummy-coded group memberships (so $\boldsymbol{Y}_{ij} = 1$ if and only if the sample in row i of the $\boldsymbol{X}$ matrix belongs to group j). As shown in Section 3 the PRESS-statistic for this problem can be calculated very efficiently for a range of regularization parameter values. We then select the regularization parameter minimizing the total PRESS-statistic for the multivariate regression problem (since we need a single regularization parameter for RLDA). This criterion is a heuristic, but in Section 5 we show empirically that this criterion appears to work quite well. Below we attempt to motivate this criterion by appealing to the link between regularized regression and RLDA given in [9].

From the link between RR and RLDA we know that there exists a linear transformation mapping the regression coefficients and the $0/1$ dummy-coded responses to a space where classifying to the nearest group centre in the Euclidean metric gives the same classification as when using RLDA. Now, think of the rows of $\boldsymbol{Y}$ as group centers in g-dimensional space, and think of the regression coefficients as a linear map that maps each sample to the same g-dimensional space.

With this viewpoint solving the regression problem amounts to finding a linear transformation (the regression coefficients) that maps the samples as close to the correct group center as possible[18]. In the multivariate regression problem (18) we are minimizing the Frobenius norm of the residual matrix, so minimizing the square norm of the columns of the residual matrix is equivalent to minimizing the square of the norm of the rows. Finding the λ-value minimizing the PRESS-statistic can then be viewed as finding the λ-value that maps the samples as close to the correct group center as possible. Even though this space is not necessarily good for classification, we know that there is a linear transformation mapping the samples and group centers to a space where the Euclidean distance can be used for classification. As the same linear transformation is applied to both the dummy-coded responses and to the regression coefficients, it is not unreasonable that the regularization parameter value minimizing the distance between samples and their corresponding group centers will also provide good discriminatory ability after a linear transformation. This is, of course, a heuristic criterion, but the above justification together with the results in Section 5 provide some justification for why this criterion is sensible.

5 Examples and discussion

5.1 Selection of data sets and the procedure used

In this section we illustrate the parameter selection methods we have discussed with several examples. We consider the following data sets: tumor14[20], Yale32B[6], ORL[23], and MNIST[17]. The tumor14 data set is a gene expression data set where the responses are different types of tumors. The yale32B and ORL data sets are face recognition data sets. The MNIST data set is a handwritten digit recognition data set. For more details about the data sets see the given references. Summaries of the sizes of the data sets are given in Table 1.

For model performance we consider the percentage of samples correctly classified. With the MNIST data set the standard training/test set split was used. For the three other data sets 50 random splits into training and test sets were used, where 1/2 of the data set was used for training. For each split, it was verified that there was at least two samples from each group in the training set and that each group would be represented in each training set during CV. If this criteria was not met, new random splits were generated until a split satisfying this criteria was found. When using CV to estimate the predictive accuracy of a model there will typically be a range of values of the regularization parameter giving the best cross-validated predictive performance. In this case we select the largest value of λ (simplest model) among the values giving the best cross-validated predictive performance. For the PRESS-statistic we consider both the value of λ giving

Properties Data set name	n	p	g	Reference
tumor14	308	15009	26	[20]
yale32B	2414	1024	15	[6]
MNIST	60000+10000	5000[†]	10	[17]
ORL	400	10304	40	[23]

Table 1: *Overview of the data sets used as examples in this Section. n is the number of samples, p is the number of features, and g is the number of groups.*

the minimum PRESS-statistic, as well as the value of λ obtained by a χ^2-test[14] with $\alpha = 0.1$. The idea behind the χ^2-test is to select a larger regularization parameter value than the one obtained by the PRESS-minimal value as long as this can be done without affecting the PRESS-statistic of the selected model too much. See [14, 15] for details. For all data sets we sampled 500 values of λ on a log scale. In the implementations of the RLDA-functions based on the Mahalanobis metric and the Fisher formulation of the RLDA problem we omitted the $1/n$ factor in the covariance matrices. This was done to make the regularization parameter values comparable with the ones obtained from the PRESS-statistic for the $0/1$ dummy-coded regression problem. In the tables and text below we refer to the regularization parameter obtained by 5-fold CV based on the pseudocode given in [16] simply as '5-fold CV', we refer to the regularization parameter obtained by the fast Loo update formulae for the Mahalanobis distances as 'LooCV', and we refer to the regularization parameter values obtained by PRESS and PRESS together with the χ^2-test as 'PRESS' and 'PRESS+χ^2', respectively.

Using RLDA on the MNIST data set without any additional pre-processing gives a classification error on the test set of about 12%. The classification performance can be significantly improved by using randomly generated features. More precisely, we generate features based on contrast differences between images of different classes. The idea is to sample two images from different classes, consider their difference (as vectors), and using the pixels/elements where this difference is large as a feature. See appendix A for details about the feature generation. By using 784 random features (this gives us the same number of features as the raw MNIST data set) we obtain an error rate on the test set of about 4%. The feature generation and matrix multiplication prior to modeling for this number of features takes less than 5 seconds. The rest of the computational time is unaffected compared to using RLDA on the original MNIST data set, as the new data set has the same size as the original data set. In the results presented in the tables we have used 5000 random features. For 5000 random features the feature generation and matrix multiplication needed to obtain the new training and test sets took about 30 seconds, and this time is not included when comparing the time used for the different parameter selection methods.

[†] *See discussion in the text about feature generation for MNIST.*

5.2 Computational time and overview of prediction results

In Table 2 the time used for parameter selection for the various data sets and parameter selection methods are shown. For all data sets except MNIST the results reported are the averages over the 50 data splits considered. For the MNIST data set there is a standard training/test set split and so model selection was done only once for this data set. From Table 2 we see that the slowest method is the Mahalanobis LOOCV. This is due to the LOO update formulae depending on the group membership of the removed sample, but the implementation used could also be a factor. In Table 3 we illustrate how the different parameter selection methods scale with respect to the number of regularization parameter values considered. It is clear that the fastest method is the PRESS-statistic. As the main calculation when using the PRESS-statistic is the initial SVD we see that increasing the number of regularization parameter considered has little effect on the computational time needed for model selection.

Validation method Data set	5-fold CV	LOOCV	PRESS-statistic
tumor14	3.87	4.09	0.20
yale32B	12.0	293	1.28
ORL	4.3	11.6	0.23
MNIST	700	22000	160

Table 2: *Time needed for model selection with* 500 *candidate regularization parameter values for the different data sets and parameter selection methods. For tumor14, yale32B and the ORL data set the given time is the average time over the* 50 *different data splits. For MNIST the time given is for a single model selection.*

Number of λ values Data set	10	50	100	500	1000	5000
LooCV	6.4	31.2	60.1	313	575	2800
PRESS	0.43	0.50	0.57	1.23	1.70	6.66
5-fold CV	2.1	3.0	3.91	12.4	19.54	91.72

Table 3: *The Table shows the average time needed for model selection for different number of regularization parameter values for the Yale32B data set. This gives an indication of how the different parameter selection methods scale with an increasing number of candidate regularization parameter values.*

In Table 4 the percentage correctly classified (PCC) samples for the test sets are shown for the different parameter selection methods considered. We see that both PCC under cross validation (PCCCV) and the PRESS-statistic from dummy regression give similar predictive performance. This is partly due to the values of λ selected by the different methods for many of the splits being similar, but also due to the fact that the test set performance is not always very sensitive to the choice of λ. This is illustrated in Figures 1 and 3. In Figure 1 we see that the PRESS curve correlates well with PCCCV, and the PRESS-curve gives a very good indication of where the

Data set \ selection method	5-fold CV	LOOCV	PRESS	PRESS+χ^2
tumor14	64.8	64.5	64.5	60.8
yale32B	94.5	94.7	94.6	94.1
ORL	92.0	92.6	92.3	92.8
MNIST	98.1	98.1	98.0	98.0

Table 4: *Average PCC on test set for the different model selection methods λ.*

optimal value of the regularization parameter is. The value of the regularization parameter chosen by LOOCV and the PRESS-statistic are very similar. In Figure 3 we see that there is a large interval of λ-values that give approximately the same classification accuracy, so that the test set result is in this case not very sensitive to the exact value of the chosen regularization parameter.

From Table 4 we can see that using a χ^2-test can both improve and worsen predictive performance. From Figures 1 and Figure 3 it appears that when the PRESS-curve has a clear minimum the PRESS-minimal λ provides a good regularization parameter value. In this case it seems that adding more regularization can result in a worse model, as shown in the results in Table 4. For a flat PRESS-curve the position of the PRESS-minimal λ can be slightly arbitrary, and in this situation the χ^2-test can be useful. From Figure 2 we see that the PRESS-curve is very flat, and selecting the PRESS-minimal regularization parameter results in a smaller regularization parameter than the one obtained from the other parameter selection methods. By applying the PRESS+χ^2 parameter selection method here we obtain a value of the regularization parameter more similar to the ones obtained from 5-fold CV and LooCV.

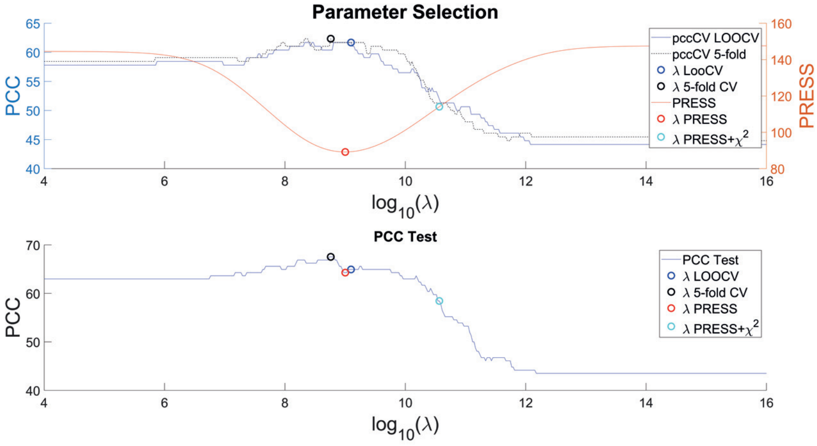


Figure 1: *Parameter selection and corresponding test set results for one particular split of the Tumor 14 data set. Top: Model selection for CV and PRESS. Bottom: Results on the test set.*

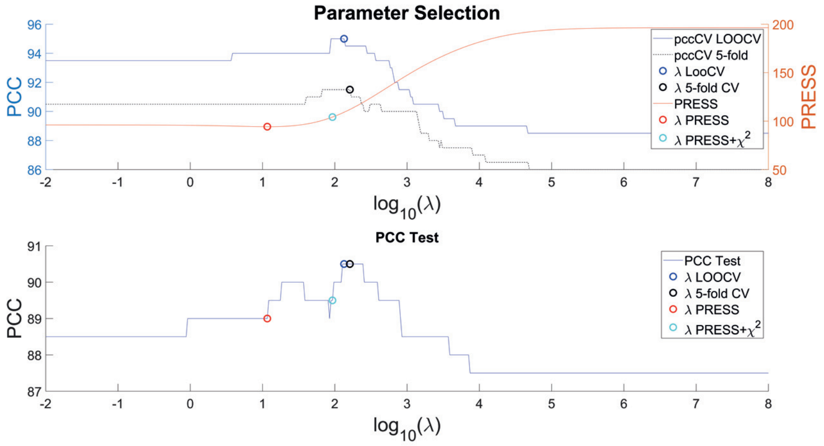


Figure 2: *Parameter selection and corresponding test set results for one particular split of the ORL data set. Top: Model selection for CV and PRESS. Bottom: Results on the test set.*

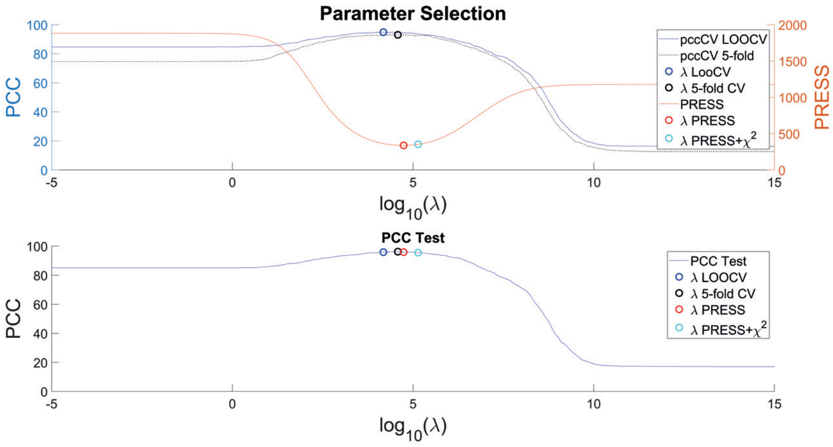


Figure 3: *Parameter selection and corresponding test set results for one particular split of the Yale32B data set. Top: Model selection for CV and PRESS. Bottom: Results on the test set.*

In Table 5 we show the relative difference between the largest and smallest value of λ selected by the different methods over all 50 iterations with different training and test sets. For the data sets considered in this work, the PRESS criterion appears to be much less sensitive to the choice of partition of the training and test set than the other criteria. A large difference in the chosen regularization parameter value is not necessarily problematic as there is often a large range of regularization parameter values giving similar predictions, and manual inspection showed that this was indeed the case here. It is, however, interesting how the PRESS-statistic appears to not be

Validation method Data set	5-fold CV	LOOCV	PRESS	PRESS+χ^2
tumor14	$5 \cdot 10^3$	$2 \cdot 10^3$	1.43	$2 \cdot 10^6$
yale32B	24.5	7.7	0.20	0.35
ORL	$2.4 \cdot 10^7$	192	1.5	1.8

Table 5: *Relative difference of the largest and smallest values of the regularization parameter value selected by the different criteria over all the training and test set splits. Calculated as* $\frac{\lambda_{max}-\lambda_{min}}{\lambda_{min}}$.

very sensitive to the training/test set split.

6 Conclusion

In the present work we have argued that the PRESS-statistic obtained from the dummy regression problem (18) can be used as an efficient criterion for selecting a value of the regularization parameter for RLDA. The motivation for this idea comes from viewing the regression residuals as distances to the group centers. In the examples we showed that the computational savings can in some cases be significant by using the PRESS-statistic for selecting the regularization parameter. The PRESS-statistic can also be useful as an explorative tool, as it allows one to quickly investigate the data set and find an approximation to the optimal value of the regularization parameter, or at the very least a neighborhood in which to look for a good regularization parameter value. In the case where the PRESS-curve has a clear minimum it seems that the PRESS-minimal regularization parameter provides a good regularization parameter value for RLDA, but in the case of a flat PRESS–curve additional regularization may be needed. From the regression problem alone it is not clear that the regularization parameter value chosen by the PRESS-statistic should be so close to the value chosen by CV on predictive performance. We believe the similarity is explained by the visual interpretation of RR as clustering samples close to the group centres. When choosing among several values of the regularization parameter, it is intuitive that the value of the regularization parameter providing the best clustering of the groups will also provide the best (or close to best) prediction. A more theoretical explanation would be useful as it would help explain exactly when the PRESS-statistic gives us a good regularization parameter value for RLDA.

A Improving MNIST classification by using random features

In this appendix we describe the random feature generation used for the MNIST data set in detail. The features were generated in the following way:

1. Randomly pick two samples $\boldsymbol{x}_1, \boldsymbol{x}_2$ from the training set from different groups.
2. Set a threshold value (for each feature we chose a random threshold value in the interval $[30, 100]$). For the sampled vectors set the pixels with values greater than the threshold value equal to 1 and the other pixels to 0. Label the vectors obtained by $\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2$.
3. Define a new feature $\boldsymbol{f} = abs(\tilde{\boldsymbol{x}}_1 - \tilde{\boldsymbol{x}}_2)$.
4. Repeat the above steps until the desired number of features have been generated.

After the features have been generated and collected as columns in the matrix $\boldsymbol{F} \in \mathbb{R}(p, n_{features})$, the new training and test sets are obtained by evaluating the matrix products $\boldsymbol{X}_{train}\boldsymbol{F}$ and $\boldsymbol{X}_{test}\boldsymbol{F}$.

The idea of the above feature generation is to generate features such that each individual feature provides discriminatory information between two classes. Each sampled image is considered to be prototype for its group, and the difference between the two sampled images (after thresholding) gives information about pixels that are of high intensity in one of the images, but not in the other. This feature generation method works for the MNIST data set because the samples are centred and scaled to have the same size. A consequence of this pre-processing is that we can expect that samples from the same class should have high pixel intensity in roughly the same pixels. An exception to this is digits that can be drawn in multiple ways (such as 4), but this is not a problem in practice as long as enough features are generated. The threshold in step (2) is to exclude pixels of very low intensity as these are often not useful for classification. The subtraction in step (3) removes pixels that are of high intensity in both sampled images as these pixels do not provide discriminatory information between the two classes sampled.

B Prototype MATLAB code for PRESS calculation

```matlab
function [press,lambda,lambdaIndex,U,s,V] = TregPRESS(X,Y,lambdas,type) % 'type' is an integer. type > 0 indicate the desired ...
       regularization. type < 0 to standardize the dataset (ordinary ridge regression).
%% Gives PRESS—statistic from dummy regression with minimal computations
[n,p] = size(X); mX = mean(X); mY = mean(Y); X = bsxfun(@minus,X,mX); Y = bsxfun(@minus,Y,mY); % Size and centering of data matrix X ...
       and response vector y.
L = [];
if type > 0
    L = diff([speye(p);sparse(type,p)],type);    %L(end—type+1:end,:) = sqrt(eps)*L(end—type+1:end,:);
    X = X/L;                                                          % For penalizing lack of smoothness wrt indicated derivative ...
           (type).
elseif type < 0
    L = spdiags(std(X)',0,p,p);  X = X/L;
end        % For standardization and ordinary ridge regression.
g = size(Y,2);
press = zeros(length(lambdas),g);
[U, S, V] = svd(X,'econ'); s = diag(S);                           % SVD (PCA) of centered (and scaled) X—data & extraction of the ...
       singular values.
denom     =  bsxfun(@plus,s,bsxfun(@rdivide,lambdas,s));          % Denominator factors for both bcoefs and PRESS.
H         =  (U.^2)*bsxfun(@rdivide,s,denom)+1/n;            % The leverages for all lambdas.
for i=1:g
    resid     =  bsxfun(@minus,Y(:,i),U*bsxfun(@rdivide,s.*(U'*Y(:,i)),denom));
    press(:,i)     =  sum(bsxfun(@rdivide,resid,(1—H)).^2)';
end

[lambda, lambdaIndex] = min(sum(press,2));
```

C Prototype MATLAB code for fast Mahalanobis LOOCV

```
function [pccCV, pcc, SpCV, nG, muG0, s0, V, GhatCV, d2CV, Ghat, d2] = RLDA(X, G, lambdas)
%% Declare LDA-parameters & calculate basic stuff
nlambdas = length(lambdas);  % Number of lambda-values to be investigated
n = size(X,1);   g = max(G); % # samples, # X-variables and # groups
d2 = zeros(n,g); d2CV = zeros(n,g,nlambdas);  % Squared Mahalanobis distances (Fitted values and LooCV)
pccCV = zeros(nlambdas,1);   % The first function output (LooCV percent correct classification for each of the lambdas.)
pcc = zeros(nlambdas,1);     % Percent correct classification by resubstitution
Yd = dummyvar(G); nG = sum(Yd)'; muG0 = (Yd'*Yd)\Yd'*X; % Dummy coding of the groups, % Groups sizes, % Group means
SpCV = zeros(nlambdas,1);    % Summned CV-probabilities for the correct classifications
Ghat = zeros(n,nlambdas);
GhatCV = zeros(n,nlambdas);

[s0, V] = rsvd(X,G);         % Ridge-adapted SVD according to groups in G
X = X*V; muG = muG0*V;
for i = 1:nlambdas
    isr = 1./sqrt(s0.^2+lambdas(i))'; % Inverse of the singular values corresponding to ridge data: [Xs;sqrt(lambda(i))*eye(p)].
    for j = 1:g % Calculation of regularized squared Mahalanobis distances to the various group means.
        d2(:,j) = sum(bsxfun(@times,bsxfun(@minus,X,muG(j,:)),isr).^2,2);
    end
    [~, Ghat(:,i)] = min(d2,[],2); pcc(i) = 100*sum(G == Ghat(:,i))/n;
    %% Fast Calculation of LooCV Mahalanobis distances to the various group means:
    for j = 1:n
        c = G(j);    % c is the true group-membership of sample j
        nc = nG(c);  % nc is the size of group c
        Xc = X(j,:)-muG(c,:); % j-th sample (row) group centered wrt the correct group.
        for k = 1:g  % Compute the adjusted squared Mahalanobis distance from from xj to center of group k:
            if k==c, d2CV(j,c,i) = d2(j,c) * (nc/(nc-1)).^2 / (1-(nc/(nc-1))*d2(j,c)); else
                Xk = X(j,:)-muG(k,:); % j-th sample (row) group centered wrt uncorrect group.
                d2CV(j,k,i) = d2(j,k) * (1 + [(Xc.*isr)*(Xk.*isr)']^2 / ([(nc-1)/nc - d2(j,c)]*d2(j,k)) );
            end
        end
    end
    [~, GhatCV(:,i)] = min(d2CV(:,:,i),[],2); pccCV(i) = 100*sum(G == GhatCV(:,i))/n;
end

%% LDA ridge-adaption of SVD:
function [s, V] = rsvd(X,G)
[n, p] = size(X);
if nargin==1, Yd = ones(n,1); else Yd = dummyvar(G); end
g = size(Yd,2);  muG  = (Yd'*Yd)\Yd'*X;
[~, S, V] = svd((X-muG(G,:)),'econ');
k = min(p,n-g); m = min(n,p); h = m-k;
s = [diag(S(1:k,1:k)); zeros(h,1)];
if h > 0, [V, ~] = qr([V(:,1:k) muG'],0); end
```

References

[1] P. J. Di Pillo. The application of bias to discriminant analysis. *Communications in Statistics-Theory and Methods*, 5(9):843–854, 1976.

[2] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of human genetics*, 7(2):179–188, 1936.

[3] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2009.

[4] J. H. Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.

[5] K. Fukunaga and D. L. Kessell. Estimation of classification error. *IEEE Transactions on Computers*, 100(12):1521–1527, 1971.

[6] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE transactions on pattern analysis and machine intelligence*, 23(6):643–660, 2001.

[7] G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

[8] Y. Guo, T. Hastie, and R. Tibshirani. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86–100, 2006.

[9] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, pages 73–102, 1995.

[10] T. Hastie and R. Tibshirani. Efficient quadratic regularization for expression arrays. *Biostatistics*, 5(3):329–340, 2004.

[11] T. Hastie, R. Tibshirani, and A. Buja. Flexible discriminant analysis by optimal scoring. *Journal of the American statistical association*, 89(428):1255–1270, 1994.

[12] N. Hjort. Notes on the theory of statistical symbol recognition. *Norwegian Computing Center report no. 778*, 1986.

[13] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[14] U. Indahl. A twist to partial least squares regression. *Journal of Chemometrics*, 19(1):32–44, 2005.

[15] U. G. Indahl, K. H. Liland, and J. Skogholt. Model selection by fast virtual cross-validation in ridge regression and the tikhonov regularization framework. *Submitted for publication.*

[16] S. Ji and J. Ye. Generalized linear discriminant analysis: a unified framework and efficient model selection. *IEEE Transactions on Neural Networks*, 19(10):1768–1782, 2008.

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[18] K. Lee and J. Kim. On the equivalence of linear discriminant analysis and least squares. In *AAAI*, pages 2736–2742, 2015.

[19] K. P. Murphy. *Machine learning: a probabilistic perspective.* MIT press, 2012.

[20] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98(26):15149–15154, 2001.

[21] B. D. Ripley. Errata for 'Pattern Recognition and Neural Networks', 1998. [Online; accessed 26/01/2018].

[22] B. D. Ripley. *Pattern recognition and neural networks.* Cambridge university press, 2007.

[23] F. S. Samaria and A. C. Harter. Parameterisation of a stochastic model for human face identification. In *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, pages 138–142. IEEE, 1994.

[24] J. Ye. Least squares linear discriminant analysis. In *Proceedings of the 24th international conference on Machine learning*, pages 1087–1093. ACM, 2007.

[25] Z. Zhang, G. Dai, C. Xu, and M. I. Jordan. Regularized discriminant analysis, ridge regression and beyond. *Journal of Machine Learning Research*, 11(Aug):2199–2228, 2010.

ISBN: 978-82-575-1663-5
ISSN: 1894-6402

NMBU

Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no

105452 / ANDVORDGRAFISK.NO