GENERALIZED METHODS FOR CALIBRATION

Michael Rasmussen

LYNGBY 2001
EKSAMENSPROJEKT
NR. 2001/03

IMM

Trykt af IMM, DTU

Abstract

In chemometrics traditional calibration in case of spectral measurements express a quantity of interest (e.g. a concentration) as a linear combination of the spectral measurements at a number of wavelengths. Often the spectral measurements are performed at a large number of wavelengths and in this case the number of coefficients in the linear combination is magnitudes larger than the number of observations. Traditional approaches to handling this problem includes Principal Components Regression, (PCR), Partial Least Squares regression, (PLS), Ridge Regression, (RR) and variable selection. They are all presented with theory and application. Least Absolute Shrinkage and Selection Operator, (LASSO), which has recently been improved to handle sigular design matrices, is also presented here. Furthermore a new approach that combines these methods with B-spline basis functions is presented.

The empirical work is done using NIR-spectra for gasoline and wheat.

Keywords

Calibration, NIR spectroscopy, linear regression, cross-validation, singular value decomposition, regularization, minimum length least squares, principal components regression, partial least squares regression, cyclic subspace regression, ridge regression, subset selection, forward selection, LASSO, adaptive ridge regression, B-spline basis, basis function regression, Matlab.

Foreword

This thesis constitutes the main part of my work for the Master of Science degree during the period September 2000 to February 2001 at The Institute of Informatics and Mathematical Modelling, The Technical University of Denmark. Parts of the thesis have already been published in one conference paper, [47].

I would like to thank the entire staff and my fellow students at the institute for creating a friendly and inspiring environment. Specifically I would like to thank Assist. Prof. Henrik Aalborg Nielsen that made the work in Sections 12.2 and 12.3 possible. Furthermore I would like to thank my supervisor Prof. Henrik Madsen and PhD. Henrik Öjelund for valuable contributions and librarian Finn Kuno Christensen for always being very helpfull.

The thesis is written in LaTeX, while the plots are created using Matlab.

Lyngby, 1. February 2001

Michael Rasmussen

Contents

1 Introduction **9**

2 Gasoline example **11**

2.1 Data 11

3 Calibration and the Least Squares method **13**

3.1 Introduction 13

3.1.1 Properties of the OLS estimator 14

3.2 Minimum Length Least Squares 15

3.2.1 Range and Null spaces 16

3.2.2 Rank-retaining factorizations 18

3.2.3 The complete orthogonal factorization 20

3.2.4 The Singular Value Decomposition 22

3.2.5 Moore-Penrose generalized inverse 23

4 Regularization methods **25**

4.1 Introduction 25

4.1.1 Model selection by cross-validation 26

4.2 MLLS applied to gasoline example 28

5 Ridge Regression **31**

5.1 Introduction 31

5.2 Theory 31

5.3 Bayesian Motivation 33

5.4 Ridge applied to gasoline example 34

6 Principal Components Regression **39**

6.1 Introduction 39

6.2 Theory 39

6.3 Selection principles 40

6.4 PCR applied to gasoline example 41

6.4.1 Comments to the PCR methods 44

7 Partial Least Squares Regression **45**

7.1 Introduction 45

7.2 Theory 46

7.3 PLS-algorithm 46

7.4 PLS applied to gasoline example 48

7.5 Summary of Ridge, PCR and PLS 50

8 Cyclic Subspace Regression **51**

8.1 Introduction 51

8.2 Theory 52

8.3 CSR applied to gasoline example 53

9 Subset Selection **55**

9.1 Introduction 55

9.2 Theory 55

9.3 Forward Selection applied to gasoline example 57

10 LASSO Regression **61**

10.1 Introduction 61

10.2 Definition 61

10.3 Convex duality and the LASSO 62

10.4 LASSO as Bayes estimate 65

10.5 Algorithm 65

10.6 LASSO applied to gasoline example 67

11 Adaptive Ridge Regression **71**

11.1 Introduction 71

11.2 Theory 71

11.3 The equivalence to LASSO 72

11.4 Adaptive Ridge Regression applied to gasoline example . . 75

12 Basis-Function Regression **79**

12.1 Introduction 79

12.2 Model 80

12.3 Approximations 81

12.4 Basis Function Regression applied to gasoline example . . . 83

12.5 Comments 87

12.6 Range selection using the BFR estimates 92

12.7 Summary for the gasoline example 95

13 Wheat example **99**

13.1 Data 99

13.1.1 Pretreatment of data 100

13.1.2 Setup for 5-fold cross-validation 102

13.2 Results 103

13.2.1 MLLS applied to wheat example 103
13.2.2 Ridge applied to wheat example 106
13.2.3 PCR applied to wheat example 109
13.2.4 PLS applied to wheat example 112
13.2.5 CSR applied to wheat example 116
13.2.6 FSR applied to wheat example 116
13.2.7 LASSO applied to wheat example 119
13.2.8 BFR applied to wheat example 121
13.2.9 Range selection using the BFR estimates 130
13.3 Summary for the wheat example 136

14 Summary **139**
14.1 A calibration strategy 141
14.2 Further enhancements 142

15 Conclusion **143**

A Gasoline **145**

B Wheat **149**

C Some Matlab functions **157**
C.1 Ridge Regression 157
C.2 Principal Components Regression 158
C.3 Partial Least Squares Regression 159
C.4 Cyclic Subspace Regression 160
C.5 Forward Selection Regression 161
C.6 Adaptive Ridge Regression 163

Chapter 1
Introduction

The developments of measurement instrument technology in many areas of science has made it easier to generate large data sets. The computer revolution makes it possible to handle these data sets numerically. The tool to extract quantitative information i.e. transforming data sets into information and knowledge is statistics.

Calibration can be described by the following situation:

- There are typically two types of measurements or observations for each item.
  1. An expensive or laborious characteristic y.
  2. A quick and cheap measurement x.

  x is used as an indirect measurement of y, that is estimate or predict the corresponding unknown y when x has been measured.
- The model for the relationship between y and x; is often assumed to be a linear regression of x on y or y on x, where in the latter case a joint distribution for (x, y) is assumed.
- A calibration (training) sample of complete pairs (x, y) is needed to build a model.

In the typical chemical situation the concentration y of one or several substances jointly should be determined. The "true" concentrations are known for a number of calibration samples obtained either by specially prepared reference samples or from other traditional chemical methods. The x-values

are measurements of absorption, reflection or transmission of light, that are easily and quickly obtained by an instrument. The multivariate character appears when the light is measured at several different wavelengths jointly.

Chapter 2
Gasoline example

2.1 Data

This data set contains 60 gasoline samples with specified octane numbers, see figure 2.1. Samples were measured using diffuse reflectance (R) as $\log(1/R)$ from 900 to 1700 nm in 2 nm intervals. So we have $n = 60$ and $p = 401$. The results for this example will be presented along with the introduction of the different methods. The data set can be obtained from ftp://ftp.clarkson.edu/pub/hopkepk/chemdata/kalivas.

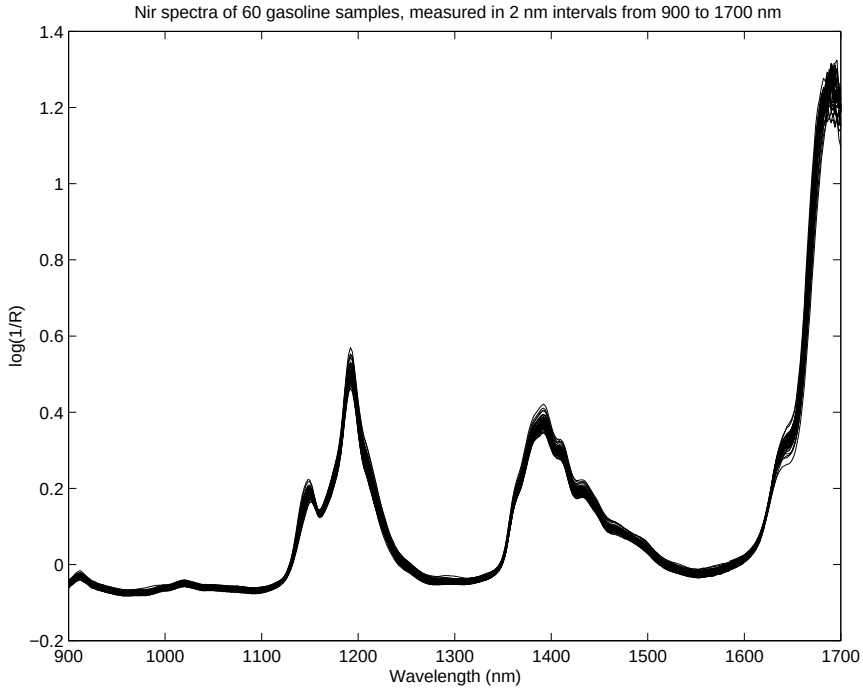

Figure 2.1: 60 NIR spectra of gasoline

Chapter 3
Calibration and the Least Squares method

3.1 Introduction

The n-observation linear regression model is

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{3.1}$$

where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^T)^T$ is a $(p+1)$ vector of parameters to be estimated, $\mathbf{y}$ is an $(n \times 1)$ vector of dependent variables, $\mathbf{Z} = [\mathbf{1}\ \mathbf{X}]$ is an $n \times (p+1)$ matrix with ith row $(1, x_i^T)$, $i = 1, \ldots, n$, containing the independent variables, where each row of the matrix $\mathbf{X}$ contains the measurements for a given sample and each column the measurements for a given variable and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector containing error terms assumed to follow an n-dimensional multivariate normal distribution with $E[\boldsymbol{\epsilon}] = \mathbf{0}$ and $V[\boldsymbol{\epsilon}] = \mathbf{I}\sigma^2$. For the estimation of $\boldsymbol{\theta}$ in equation (3.1) the full rank case requires that the $n \times (p+1)$ matrix $\mathbf{Z} = [\mathbf{1}\ \mathbf{X}]$ is such that $n \geq (p+1)$ and is of full rank $(p+1)$.

An ordinary least squares, (OLS), estimator is expressed by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta}[(\mathbf{y} - \mathbf{Z}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{Z}\boldsymbol{\theta})] \tag{3.2}$$

The solution is given by the "normal" equations,

$$\mathbf{Z}^T\mathbf{Z}\boldsymbol{\theta} = \mathbf{Z}^T\mathbf{y} \tag{3.3}$$

If $\mathbf{Z}$ is of full rank then all $(p+1)$ parameters are uniquely estimated by

$$\hat{\boldsymbol{\theta}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y} \tag{3.4}$$

see [11] and [7]

Proof:

$$\begin{aligned} L &= [(\mathbf{y} - \mathbf{Z}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{Z}\boldsymbol{\theta})] \\ \frac{dL}{d\boldsymbol{\theta}} &= -2\mathbf{Z}^T\mathbf{y} + 2\mathbf{Z}^T\mathbf{Z}\boldsymbol{\theta} \\ \frac{dL}{d\boldsymbol{\theta}} &= 0 \Rightarrow \\ 0 &= -\mathbf{Z}^T\mathbf{y} + \mathbf{Z}^T\mathbf{Z}\boldsymbol{\theta} \Leftrightarrow \\ \mathbf{Z}^T\mathbf{Z}\boldsymbol{\theta} &= \mathbf{Z}^T\mathbf{y} \Leftrightarrow \\ \boldsymbol{\theta} &= (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y} \end{aligned}$$

see [63] p. 26

3.1.1 Properties of the OLS estimator

The OLS estimator given by equation (3.4) has the following properties

- It is an unbiased estimate (central), $E[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$ since

$$\begin{aligned} E[\hat{\boldsymbol{\theta}}] &= E[(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}] \\ &= E[(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T(\mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\epsilon})] \\ &= (\mathbf{Z}^T\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{Z})\boldsymbol{\theta} \\ &= \boldsymbol{\theta} \end{aligned} \tag{3.5}$$

- $V[\hat{\boldsymbol{\theta}}] = E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T] = \sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1}$ since

$$\begin{aligned} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} &= (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y} - \boldsymbol{\theta} \\ &= (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T(\mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\epsilon}) - \boldsymbol{\theta} \\ &= (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\epsilon} \Rightarrow \\ V[\hat{\boldsymbol{\theta}}] &= E[(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}] \\ &= (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1} \\ &= \sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1} \end{aligned} \tag{3.6}$$

- It provides unbiased estimates of the elements of $\boldsymbol{\theta}$ which have the minimum variance of *any* linear function of the observations (Gauss-Markov). An estimator with this property is also called BLUE (Best, amongst Linear, Unbiased, Estimators). See [37] p. 33 and [11] p. 87.

It is good numerical procedure to center all variables, that is

$$\mathbf{y} = \mathbf{y} - \mathbf{1}\bar{\mathbf{y}}\ , \quad \sum_{i=1}^{n} Z_{ij} = 0, \quad j = 1, \ldots, p.$$

where $\bar{\mathbf{y}} = \sum_{i=1}^{n}(y_i)/n$ is the mean of $\mathbf{y}$. The model may then be presented in a slightly different parametrization

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.7}$$

From now on $\mathbf{X}$, $(n \times p)$, and $\mathbf{y}$, $(n \times 1)$, are assumed to be centered. Often the explanatory variables are standardized to give the different variables the same influence to the modeling. The general opinion is that if the variables are on different scales it is important to do it, otherwise it is not recommended, [56].

3.2 Minimum Length Least Squares

When there are more variables than observations the estimator, $\hat{\boldsymbol{\beta}}$, cannot be determined uniquely due to dependency in $(\mathbf{X}^T\mathbf{X})$, which causes $(\mathbf{X}^T\mathbf{X})$ to be singular and then the inverse of $(\mathbf{X}^T\mathbf{X})$ does not exist. When working with the experimental setup described in the introduction we are often

dealing with *near-collinearities* in $\mathbf{X}$. Near-collinearities are approximate linear dependencies between the columns of $\mathbf{X}$ and are essentially constant for all responses y. They are a consequence of duplication of information that is provided by the variables. More formally, if there exist k linearly independent nonzero vectors $w_j = (w_{1j}, \ldots, w_{nj})^T$ such that

$$\sum_{q=1}^{n} w_{qj} x_q = 0 \quad j = 1, \ldots, k \tag{3.8}$$

then the columns of $\mathbf{X}$ are said to be collinear. The closer the linear combinations in equation (3.8) are to zero, the stronger are the near-collinearities and the more damaging are their effects on the OLS estimator, [58]. As shown in [10] pp. 4.57-58, [28] p. 56 and [63] p. 29 this will lead to an estimate with a large variance

$$\sum_{j=1}^{p} V[\hat{\boldsymbol{\beta}}_j] = \sigma^2 trace(\mathbf{X}^T\mathbf{X})^{-1}$$

This has prompted research into *biased* regression estimators. The estimators attempt to introduce a small bias into the regression estimator while greatly reducing the variance appearing in the OLS estimator. The first to be introduced here is known as the *Minimum Length Least Squares*, which is the minimum 2-norm solution.

The observed x-vectors span only an n-dimensional Euclidian subspace of the potential p-dimensional. The idea is now that the x-variables do not vary independently, but are very much correlated, and that all essential sources of variation for x should show up in the n-dimensional subspace, hopefully including those connected with the explainable part of the variation in $\mathbf{y}$, [7] p. 52 and [11] p. 258. The key to produce the MLLS solution is rank-retaining factorizations of $\mathbf{X}$, [18] p. 230.

3.2.1 Range and Null spaces

Let S be a set containing vectors of dimension n, then S is a *subspace* of $\mathbb{R}^n$ if, for any scalars α_1 and α_2:

$$\mathbf{x}, \mathbf{y} \in S \quad \text{implies} \quad \alpha_1\mathbf{x} + \alpha_2\mathbf{y} \in S \tag{3.9}$$

This property implies that every subspace must contain the zero vector, ($\alpha_1 = \alpha_2 = 0$). Given a subspace S, a set of k vectors $\{a_j\}, j = 1, \ldots, k$ is said to *span* S if every vector $\mathbf{x}$ in S can be written as a linear combination of the set of vectors. The set of all n-vectors that are linear combinations of the columns of the $n \times p$ matrix $\mathbf{X}$ is called the *range space* of $\mathbf{X}$, and will be denoted by $\mathcal{R}(\mathbf{X})$:

$$\mathbf{y} \in \mathcal{R}(\mathbf{X}) \quad \text{if and only if there exists} \quad \boldsymbol{\beta} \in \mathbb{R}^p \quad \text{such that} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta}$$

$\mathcal{R}(\mathbf{X})$ is a subspace of $\mathbb{R}^n$. Consider any two vectors $\mathbf{y_1}, \mathbf{y_2} \in \mathcal{R}(\mathbf{X})$, by definition of $\mathcal{R}(\mathbf{X})$ it must hold that $\mathbf{y_1} = \mathbf{X}\boldsymbol{\beta}_1$ and $\mathbf{y_2} = \mathbf{X}\boldsymbol{\beta}_2$ for some $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Because $\mathbf{X}$ is a linear transformation for any scalars α_1, α_2 we have:

$$\alpha_1\mathbf{y_1} + \alpha_2\mathbf{y_2} = \alpha_1\mathbf{X}\boldsymbol{\beta}_1 + \alpha_2\mathbf{X}\boldsymbol{\beta}_2 = \mathbf{X}(\alpha_1\boldsymbol{\beta}_1 + \alpha_2\boldsymbol{\beta}_2) \in \mathcal{R}(\mathbf{X})$$

Which means that $\mathcal{R}(\mathbf{X})$ satisfies (3.9). The set of all p-vectors that are linear combinations of the (transposed) rows of $\mathbf{X}$ defines a subspace of vectors expressible as $\mathbf{X}^T\mathbf{v}$ for some n-vector v. This subspace is denoted by $\mathcal{R}(\mathbf{X}^T)$ and is defined as follows:

$$\boldsymbol{\beta} \in \mathcal{R}(\mathbf{X}^T) \quad \text{if and only if there exists} \quad \mathbf{v} \in \mathbb{R}^n \quad \text{such that} \quad \boldsymbol{\beta} = \mathbf{X}^T\mathbf{v}$$

The *column rank* of a matrix $\mathbf{X}$ (the dimension of $\mathcal{R}(\mathbf{X})$) is the maximum number of linearly independent columns of $\mathbf{X}$. Similarly, the *row rank* of a matrix is the maximum number of linearly independent rows. The row and column ranks of a matrix must be equal and their common value is called the *rank* of the matrix, r. Any matrix $\mathbf{X}$ defines two other important subspaces apart from $\mathcal{R}(\mathbf{X})$ and $\mathcal{R}(\mathbf{X}^T)$. For an $n \times p$ matrix $\mathbf{X}$, the set of all n-vectors orthogonal to vectors in $\mathcal{R}(\mathbf{X})$ is called the *null-space* of $\mathbf{X}^T$, denoted by $\mathcal{N}(\mathbf{X}^T)$. Elements of $\mathcal{N}(\mathbf{X}^T)$ is defined as follows

$$\mathbf{z} \in \mathcal{N}(\mathbf{X}^T) \quad \text{if and only if} \quad \mathbf{X}^T\mathbf{z} = 0$$

The null-space of $\mathbf{X}$ itself, $\mathcal{N}(\mathbf{X})$, is the subspace consisting of all p-vectors $\mathbf{q}$ such that $\mathbf{X}\mathbf{q} = 0$. Because $\mathcal{N}(\mathbf{X}^T)$ and $\mathcal{R}(\mathbf{X})$ contain only the zero vector in common the expression of any non-zero n-vector $\mathbf{y}$ in the following form is unique:

$$\mathbf{y} = \mathbf{y}_R + \mathbf{y}_N \quad \text{, with} \quad \mathbf{y}_R \in \mathcal{R}(\mathbf{X}), \mathbf{y}_N \in \mathcal{N}(\mathbf{X}^T)$$

and the vectors $\mathbf{y}_R$ and $\mathbf{y}_N$ satisfy:

$$\mathbf{y}_R^T\mathbf{y}_N = 0 \quad \text{and} \quad \mathbf{y}^T\mathbf{y} = \mathbf{y}_R^T\mathbf{y}_R + \mathbf{y}_N^T\mathbf{y}_N$$

Turning to $\mathcal{R}(\mathbf{X}^T)$, any nonzero p-vector, $\boldsymbol{\beta}$, similarly has a unique representation of the form

$$\boldsymbol{\beta} = \boldsymbol{\beta}_R + \boldsymbol{\beta}_N$$

Where $\boldsymbol{\beta}_R \in \mathcal{R}(\mathbf{X}^T)$ and $\boldsymbol{\beta}_N \in \mathcal{N}(\mathbf{X})$. A summary of the four fundamental subspaces from a rank-retaining factorization, $\mathbf{X} = \mathbf{GH}$, is given in Table 3.1

Subspace	Basis	Dimension	Specification
$\mathcal{R}(\mathbf{X})$	$\mathbf{G}$	$n \times r$	
$\mathcal{N}(\mathbf{X}^T)$	$\mathbf{K}$	$n \times (n-r)$	$\mathbf{G}^T\mathbf{K} = 0$
$\mathcal{R}(\mathbf{X}^T)$	$\mathbf{L}$	$p \times r$	$\mathbf{L}^T\mathbf{Z} = 0$
$\mathcal{N}(\mathbf{X})$	$\mathbf{Z}$	$p \times (p-r)$	$\mathbf{H}^T\mathbf{Z} = 0$

Table 3.1: The four subspaces

The theory above is from [18] pp. 17-20, p. 241.

3.2.2 Rank-retaining factorizations

Let $\mathbf{X}$ be a nonzero $n \times p$ matrix of rank r. When $\mathbf{X}$ has full column rank ($r = p$), any solution of $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ is unique. In contrast there are infinitely many solutions if $r < p$. It can be desirable to compute the solution of minimum Euclidian length (MLLS). The procedures to be described are required only when $r < p$, see [18] pp. 187-196. Any $n \times p$ matrix $\mathbf{X}$ can be written as $\mathbf{X} = \mathbf{GH}$, where $\mathbf{G}$ is $n \times r$ and $\mathbf{H}$ is $r \times p$ both with rank r, [18] p. 187. This rank-retaining form of writing $\mathbf{X}$ is used to derive a representation of the MLLS solution. $\mathbf{G}$ has linearly independent columns and $\mathbf{H}$ has linearly independent rows, which means that the matrices $\mathbf{G}^T\mathbf{G}$ and $\mathbf{HH}^T$ are $r \times r$ and nonsingular. Assume the existence of a vector $\boldsymbol{\beta}$ (now representing the MLLS solution) satisfying $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$ that lies in $\mathcal{R}(\mathbf{X}^T)$, i.e., such that $\boldsymbol{\beta} = \mathbf{X}^T\mathbf{v}$ for some vector $\mathbf{v}$. Since $\mathbf{y}$ lies in $\mathcal{R}(\mathbf{X})$ it also lies in $\mathcal{R}(\mathbf{G})$, so we have $\mathbf{y} = \mathbf{Gs}$ for a unique r-vector $\mathbf{s}$. Substituting the expressions $\mathbf{X} = \mathbf{GH}$, $\boldsymbol{\beta} = \mathbf{X}^T\mathbf{v}$ and $\mathbf{y} = \mathbf{Gs}$ into the relation $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$, we have

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{XX}^T\mathbf{v} = \mathbf{GHH}^T\mathbf{G}^T\mathbf{v} = \mathbf{Gs} = \mathbf{y} \tag{3.10}$$

Because $\mathbf{G}$ has linearly independent columns it can be cancelled from both sides of (3.10), which gives

$$\mathbf{H}\mathbf{H}^T\mathbf{G}^T\mathbf{v} = \mathbf{s} \quad \text{or} \quad \mathbf{G}^T\mathbf{v} = (\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{s} \tag{3.11}$$

Recall the assumption $\boldsymbol{\beta} = \mathbf{X}^T\mathbf{v}$, and $\mathbf{X}^T = \mathbf{H}^T\mathbf{G}^T$, then a new expression for $\boldsymbol{\beta}$ is obtained by multiplying (3.11) with $\mathbf{H}^T$

$$\boldsymbol{\beta} = \mathbf{H}^T\mathbf{G}^T\mathbf{v} = \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{s} \quad \text{or} \tag{3.12}$$

$$\boldsymbol{\beta} = \mathbf{H}^T\mathbf{z} \quad \text{, where} \quad \mathbf{H}\mathbf{H}^T\mathbf{z} = \mathbf{s} \quad \text{and} \quad \mathbf{G}\mathbf{s} = \mathbf{y} \tag{3.13}$$

For the Least Squares problem,(3.2), that determines the smallest possible Euclidian norm of the residuals, $\boldsymbol{\rho} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$, $\boldsymbol{\rho}$ and $\mathbf{y}$ are both n-vectors and can according to 3.2.1 be written as:

$$\mathbf{y} = \mathbf{y}_R + \mathbf{y}_N \quad \text{, where} \quad \mathbf{y}_R = \mathbf{X}\mathbf{y}_X \tag{3.14}$$

$$\boldsymbol{\rho} = \boldsymbol{\rho}_R + \boldsymbol{\rho}_N \quad \text{, where} \quad \boldsymbol{\rho}_R = \mathbf{X}\boldsymbol{\rho}_X \tag{3.15}$$

here the notation $\mathbf{c}_X$ refers to any p-vector satisfying $\mathbf{c}_R = \mathbf{X}\mathbf{c}_X$. Combining the definition of $\boldsymbol{\rho}$ as $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ with (3.14) and (3.15) we have

$$\begin{aligned} \boldsymbol{\rho} = \boldsymbol{\rho}_R + \boldsymbol{\rho}_N &= \mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{y}_R + \mathbf{y}_N - \mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}_R - \mathbf{X}\boldsymbol{\beta} + \mathbf{y}_N \end{aligned}$$

According to the definition of range and null-space, 3.2.1, the range-space and null-space components of the residual must satisfy

$$\boldsymbol{\rho}_R = \mathbf{y}_R - \mathbf{X}\boldsymbol{\beta} \quad \text{and} \quad \boldsymbol{\rho}_N = \mathbf{y}_N$$

Since $\mathbf{y}_N$ is retained in its entirety in the residual, $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2$ will be minimized when the two-norm of its range-space component, $\|\mathbf{y}_R - \mathbf{X}\boldsymbol{\beta}\|_2$ is as small as possible. Since $\mathbf{y}_R \in \mathcal{R}(\mathbf{X})$ by definition, a vector $\boldsymbol{\beta}$ must exist such that $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}_R$. For this $\boldsymbol{\beta}$ the entire range-space component of $\mathbf{y}$ is removed by subtraction of $\mathbf{X}\boldsymbol{\beta}$, which means that $\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{y}_N$, and the Euclidian norm of the residual is equal to its lower bound $\|\mathbf{y}_N\|_2$. This leads to two equivalent characterizations of the optimal Least Squares solution

$$\begin{aligned} \boldsymbol{\beta} \quad & \text{minimizes} \quad \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \\ & \text{if and only if} \quad \mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \end{aligned} \tag{3.16}$$

and

$$\text{if and only if} \quad \mathbf{X}\boldsymbol{\beta} = \mathbf{y}_R \quad \text{and} \quad \mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{y}_N \tag{3.17}$$

see [18] pp. 218-220. (3.17) shows that the MLLS solution must be the solution to the system $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}_R$. In the expressions (3.12) and (3.13), one only needs to replace $\mathbf{y}$ with $\mathbf{y}_R$. These expressions cannot be used directly to find the MLLS solution since the Least Squares problem, (3.2), is stated in terms of the vector $\mathbf{y}$ itself, and not its range-space component $\mathbf{y}_R$. Therefore we need to obtain the vector $\mathbf{s}$ satisfying $\mathbf{Gs} = \mathbf{y}_R$ from the original vector $\mathbf{y}$. $\mathcal{N}(\mathbf{X}^T)$ and $\mathcal{N}(\mathbf{G}^T)$ are the same, so $\mathbf{G}^T\mathbf{v} = 0$ for any vector in $\mathcal{N}(\mathbf{X}^T)$. Since $\mathbf{y}_N \in \mathcal{N}(\mathbf{X}^T)$ it follows that

$$\mathbf{G}^T\mathbf{y} = \mathbf{G}^T(\mathbf{y}_R + \mathbf{y}_N) = \mathbf{G}^T\mathbf{y}_R \tag{3.18}$$

With the substitution of $\mathbf{y}$ with $\mathbf{y}_R$ in (3.12) and (3.13), $\mathbf{y}_R$ is defined as $\mathbf{Gs}$

$$\mathbf{G}^T\mathbf{Gs} = \mathbf{G}^T\mathbf{y}_R = \mathbf{G}^T\mathbf{y} \tag{3.19}$$

Since $\mathbf{G}^T\mathbf{G}$ is nonsingular, it follows that $\mathbf{s}$ must be the unique solution of

$$\mathbf{G}^T\mathbf{Gs} = \mathbf{G}^T\mathbf{y} \quad \text{namely} \quad \mathbf{s} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{y} \tag{3.20}$$

Now $\mathbf{y}_R$ can be defined as

$$\mathbf{y}_R = \mathbf{Gs} = \mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{y} \tag{3.21}$$

Substituting the expression for $\mathbf{s}$ in (3.12) we have obtained a representation for the MLLS solution expressed only in terms of $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{y}$:

$$\boldsymbol{\beta} = \mathbf{H}^T(\mathbf{HH}^T)^{-1}\mathbf{s} = \mathbf{H}^T(\mathbf{HH}^T)^{-1}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{y} \tag{3.22}$$

Any ill-conditioning in $\mathbf{X}$ will be reflected in $\mathbf{G}$ and $\mathbf{H}$, and the occurrence of $\mathbf{G}^T\mathbf{G}$ and $\mathbf{HH}^T$ can cause numerical difficulties. There are several rank-retaining forms of $\mathbf{X}$ that can be used to avoid this problem.

3.2.3 The complete orthogonal factorization

The rank-retaining form, $\mathbf{X} = \mathbf{GH}$, of the QR factorization is

$$\mathbf{X} = \mathbf{Q}_r\mathbf{RP}^T \quad \text{, with} \quad \mathbf{G} = \mathbf{Q}_r \quad \text{and} \quad \mathbf{H} = \mathbf{RP}^T$$

where $\mathbf{Q}_r$ (the first r columns of the orthogonal matrix $\mathbf{Q}$) has orthonormal columns, $\mathbf{R}$ is an $r \times p$ upper triangle, and $\mathbf{P}$ is a permutation. The orthogonality of the columns of $\mathbf{Q}_r$ means that the product $\mathbf{G}^T\mathbf{G}$ is simply

$\mathbf{Q}_r^T\mathbf{Q}_r = \mathbf{I}_r$, where $\mathbf{I}_r$ is the r-dimensional identity matrix. The expression $\mathbf{G}^T\mathbf{G}^{-1}\mathbf{G}^T$ appearing in (3.20) simplifies, and $\mathbf{s}$ and $\mathbf{y}_R$, (3.21), are given by

$$\mathbf{s} = \mathbf{Q}_r^T\mathbf{y} \quad \text{and} \quad \mathbf{y}_R = \mathbf{Q}_r\mathbf{Q}_r^T\mathbf{y} \tag{3.23}$$

The matrix $\mathbf{H}\mathbf{H}^T$ is $\mathbf{R}\mathbf{P}^T\mathbf{P}\mathbf{R}^T = \mathbf{R}\mathbf{R}^T$, so the identical expression for the MLLS solution, (3.22), is

$$\boldsymbol{\beta} = \mathbf{P}\mathbf{R}^T(\mathbf{R}\mathbf{R}^T)^{-1}\mathbf{Q}_r^T\mathbf{y} \tag{3.24}$$

If $\mathbf{X}$ is ill-conditioned this result will also be numerically unstable due to the matrix $\mathbf{R}\mathbf{R}^T$. The way to deal with this problem is to split $\mathbf{R}$ into $\mathbf{R}_{11}$ and $\mathbf{R}_{12}$, where $\mathbf{R}_{12}$ has $p-r$ columns. If $\mathbf{R}$ happened to be non-singular, (i.e., if $\mathbf{R}_{12}$ were of dimension zero), then $\mathbf{R}\mathbf{R}^T$ would disappear:

$$(\mathbf{R}\mathbf{R}^T)^{-1} = \mathbf{R}^{-T}\mathbf{R}^{-1} \quad \text{, so that} \quad \mathbf{R}^T(\mathbf{R}\mathbf{R}^T)^{-1} = \mathbf{R}^T\mathbf{R}^{-T}\mathbf{R}^{-1} = \mathbf{R}^{-1}$$

An orthogonal matrix $\mathbf{V}$ that annihilates $\mathbf{R}_{12}$ when applied to $\mathbf{R}$ on the right, can be constructed by a sequence of Householder transformations, (see [18] pp. 121-122)

$$\mathbf{R}\mathbf{V} = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \end{pmatrix}\mathbf{V} = \begin{pmatrix} \bar{\mathbf{R}} & \mathbf{0} \end{pmatrix}$$

where $\bar{\mathbf{R}}$ is an $r \times r$ non-singular upper triangle matrix. Since $\mathbf{V}$ is orthogonal the following holds

$$\mathbf{R} = \begin{pmatrix} \bar{\mathbf{R}} & \mathbf{0} \end{pmatrix}\mathbf{V}^T$$

and $\mathbf{R}\mathbf{R}^T$ can be expressed as

$$\mathbf{R}\mathbf{R}^T = \begin{pmatrix} \bar{\mathbf{R}} & \mathbf{0} \end{pmatrix}\mathbf{V}^T\mathbf{V}\begin{pmatrix} \bar{\mathbf{R}}^T \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{R}} & \mathbf{0} \end{pmatrix}\begin{pmatrix} \bar{\mathbf{R}}^T \\ \mathbf{0} \end{pmatrix} = \bar{\mathbf{R}}\bar{\mathbf{R}}^T \tag{3.25}$$

The expression $\mathbf{R}^T(\mathbf{R}\mathbf{R}^T)^{-1}$, from (3.24), can be written as:

$$\mathbf{R}^T(\mathbf{R}\mathbf{R}^T)^{-1} = \mathbf{V}\begin{pmatrix} \bar{\mathbf{R}}^T \\ \mathbf{0} \end{pmatrix}(\bar{\mathbf{R}}\bar{\mathbf{R}}^T)^{-1} = \mathbf{V}\begin{pmatrix} \bar{\mathbf{R}}^{-1} \\ \mathbf{0} \end{pmatrix} \tag{3.26}$$

The *undesirable* $\mathbf{R}\mathbf{R}^T$ appears no longer and the MLLS solution in (3.24) can now be written as:

$$\boldsymbol{\beta} = \mathbf{P}\mathbf{V}\begin{pmatrix} \bar{\mathbf{R}}^{-1} \\ \mathbf{0} \end{pmatrix}\mathbf{Q}_r^T\mathbf{y} = \mathbf{P}\mathbf{V}\begin{pmatrix} \bar{\mathbf{R}}^{-1}\mathbf{Q}_r^T\mathbf{y} \\ \mathbf{0} \end{pmatrix} \tag{3.27}$$

So by first computing the QR factorization of $\mathbf{X}$ and then reducing $\mathbf{R}$ it is possible to compute the MLLS solution without squaring the condition number. The complete orthogonal factorization of $\mathbf{X}$ leads to the rank-retaining factorization

$$\mathbf{X} = \mathbf{GH} \quad \text{, with} \quad \mathbf{G} = \mathbf{Q}_r \quad \text{and} \quad \mathbf{H} = \begin{pmatrix} \bar{\mathbf{R}} & \mathbf{0} \end{pmatrix} \mathbf{V}^T \mathbf{P}^T$$

3.2.4 The Singular Value Decomposition

The Singular Value Decomposition, SVD, is also a complete orthogonal decomposition. It is another way to handle the MLLS problem with an ill-conditioned data-matrix $\mathbf{X}$. if $\mathbf{X}$ is a real $n \times p$ matrix, then there exist orthonormal matrices $\mathbf{U} = [u_1, \ldots, u_n] \in \mathbb{R}^{n \times n}$ and $\mathbf{V} = [v_1, \ldots, v_p] \in \mathbb{R}^{p \times p}$ such that

$$\mathbf{U}^T \mathbf{X} \mathbf{V} = \mathbf{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_t) \in \mathbb{R}^{n \times p}, \quad \sigma_1 \geq \cdots \geq \sigma_t \geq 0 \tag{3.28}$$

where $t = \min(n, p)$. The σ_j are the *singular* values of $\mathbf{X}$. The vectors u_j and v_j are the jth *left singular vector* and the jth *right singular vector*, respectively. For proof see [19] Thm. 2.5.1.

The SVD reveals a great deal about the structure of a matrix. If the SVD of $\mathbf{X}$ is given as above, and r is defined as

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_t = 0$$

then

$$\begin{aligned} \text{rank}(\mathbf{X}) &= r, \\ \mathcal{R}(\mathbf{X}) &= \mathcal{R}([u_1, \cdots, u_r]), \\ \mathcal{N}(\mathbf{X}) &= \mathcal{R}([v_{r+1}, \cdots, v_p]), \\ \mathcal{R}_r(\mathbf{X}) = \mathcal{R}(\mathbf{X}^T) &= \mathcal{R}([v_1, \cdots, v_r]), \\ \mathcal{N}_r(\mathbf{X}) = \mathcal{N}(\mathbf{X}^T) &= \mathcal{R}([u_{r+1}, \cdots, u_n]) \end{aligned}$$

which are the four fundamental subspaces from Table 3.1. Moreover if $\mathbf{U}_r = [u_1, \ldots, u_r]$, $\mathbf{\Sigma}_r = \text{diag}(\sigma_1, \ldots, \sigma_r)$ and $\mathbf{V}_r = [v_1, \ldots, v_r]$ then the SVD expansion of $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^T = \sum_{j=1}^{r} \sigma_j u_j v_j^T \tag{3.29}$$

see [19] p. 72. From (3.28) it follows that

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\mathbf{V}^T \tag{3.30}$$
$$\mathbf{X}\mathbf{X}^T = \mathbf{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\mathbf{U}^T. \tag{3.31}$$

Thus $\sigma_j^2 = 1, \ldots, t$ are eigenvalues, λ, of the symmetric matrices $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$, and v_j and u_j are the corresponding eigenvectors.

The rank-retaining form $\mathbf{X} = \mathbf{G}\mathbf{H}$ of the SVD is

$$\mathbf{X} = \mathbf{G}\mathbf{H} \quad \text{, with} \quad \mathbf{G} = \mathbf{U}_r \quad \text{and} \quad \mathbf{H} = \boldsymbol{\Sigma}_r\mathbf{V}_r^T \tag{3.32}$$

Due to the orthogonality of $\mathbf{U}$ and $\mathbf{V}$, $\mathbf{G}^T\mathbf{G} = \mathbf{I}_r$ and $\mathbf{H}\mathbf{H}^T = \boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_r^T$. By inserting this into (3.22) a new expression for the MLLS solution is derived

$$\boldsymbol{\beta} = \mathbf{V}_r\boldsymbol{\Sigma}_r^{-1}\mathbf{U}^T\mathbf{y} = \sum_{j=1}^{r}\frac{u_j^T\mathbf{y}}{\sigma_j}v_j \tag{3.33}$$

which minimizes $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2$ and has the smallest 2-norm of all minimizers. For the proof see [19] Thm. 5.5.1.

3.2.5 Moore-Penrose generalized inverse

The SVD can be used to define a matrix $\mathbf{X}^+ \in \mathbb{R}^{p\times n}$ which will be referred to as the *Moore-Penrose generalized inverse* or the *pseudo-inverse* of $\mathbf{X}$:

$$\mathbf{X}^+ = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^T \tag{3.34}$$

where

$$\boldsymbol{\Sigma}^{-1} = \operatorname{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}, 0, \ldots, 0\right) \in \mathbb{R}^{p\times n}$$

(3.33) can now be written as

$$\boldsymbol{\beta} = \mathbf{X}^+\mathbf{y} \tag{3.35}$$

Typically, $\mathbf{X}^+$ is defined to be the unique matrix $\mathbf{A} \in \mathbb{R}^{p\times n}$ that satisfies the four *Moore-Penrose conditions*

$$\begin{array}{llll} \text{(i)} & \mathbf{X}\mathbf{A}\mathbf{X} = \mathbf{X} & \text{(iii)} & (\mathbf{X}\mathbf{A})^T = \mathbf{X}\mathbf{A} \\ \text{(ii)} & \mathbf{A}\mathbf{X}\mathbf{A} = \mathbf{A} & \text{(iv)} & (\mathbf{A}\mathbf{X})^T = \mathbf{A}\mathbf{X} \end{array} \tag{3.36}$$

see [10] p. 1.35. Combining (ii) and (iii) with (3.34), one gets

$$\begin{aligned} \boldsymbol{\beta} &= \mathbf{X}^{+}\mathbf{y} = \mathbf{X}^{+}(\mathbf{X}^{T})^{+}\mathbf{X}^{T}\mathbf{y} && (3.37)\\ &= \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^{T}\mathbf{U}(\boldsymbol{\Sigma}^{T})^{-1}\mathbf{V}^{T}\mathbf{X}^{T}\mathbf{y} && (3.38)\\ &= \mathbf{V}(\boldsymbol{\Sigma}^{T}\boldsymbol{\Sigma})^{-1}\mathbf{V}^{T}\mathbf{X}^{T}\mathbf{y} && (3.39) \end{aligned}$$

which resembles more the OLS expression from (3.4), [62].

Chapter 4
Regularization methods

4.1 Introduction

In the case of more explanatory variables than observations, ($n < p$), the MLLS is known as a regularization method or shrinking method. These methods produce biased estimates but with smaller variance. It is still Least Squares estimation where we minimize the residual sum of squares, but now with a limit on the squared length of $\hat{\boldsymbol{\beta}}$. The new cost function can be written as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta}[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})] \quad \text{subject to} \sum_{j=1}^{p} \beta_j^2 \leq t. \tag{4.1}$$

where the constraint $\sum_{j=1}^{p} \beta_j^2 \leq t$ is equivalent to the addition of a penalty term $\lambda \sum_{j=1}^{p} \beta_j^2$, where λ is the Lagrange multiplier varying with the bound t on the norm of the parameters, see [60].

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2\right] \tag{4.2}$$

See [57] and [64]. The most popular regularization methods are Ridge Regression, (RR), Principal Component Regression, (PCR), and Partial

Least Squares regression, (PLS). In the notation of Frank and Friedman, [15], the estimator, $\hat{\boldsymbol{\beta}}$, for these three methods can be expressed in a general form by using (3.30) to express $\mathbf{X}^T\mathbf{X}$ as

$$\mathbf{X}^T\mathbf{X} = \sum_{j=1}^{r} \lambda_j v_j v_j^T \tag{4.3}$$

where λ_j are the positive eigenvalues of $\mathbf{X}^T\mathbf{X}$, then

$$\hat{\boldsymbol{\beta}} = \sum_{j=1}^{r} f(\lambda_j)\hat{\alpha}_j v_j \qquad ; \hat{\alpha}_j = \frac{1}{\lambda_j} z_j \tag{4.4}$$

where $f(\lambda_j)$ are shrinkage factors, the $\hat{\alpha}_j$ are the coefficients of the OLS estimator in the principal directions v_j and the $z_j = v_j^T\mathbf{X}^T\mathbf{y}$ are the canonical covariances, [8]. $f(\lambda_j)$ will be derived along with the introduction for the three aforementioned methods.

4.1.1 Model selection by cross-validation

For comparison of models intended for prediction it is highly inadequate to look just at model fit. There exist many methods for selecting a model based on some validation criterion. Cross-validation is a method for model selection according to the predictive ability of the models. The data set is split into two parts, where the first part of the data contains n_c observations which are used for fitting a model (calibration), and the second part that consists of $n_v = n - n_c$ observations is reserved for assessing the predictive ability of the model (validation). There are $\binom{n}{n_v}$ different ways to split the data set. Cross-validation selects the model with the best average predictive ability based on all (or some) different ways of data splitting.

The main focus of researchers' attention has for a long time been on the choice $n_v = 1$. This type of cross-validation is called *leave-one-out* cross-validation. It has been shown[1] that this particular type of cross-validation is asymptotically equivalent to other methods for model selection, such as the Akaike information criterion (AIC) (Akaike 1974) [1], the C_p (Mallows 1973) [39], the jackknife and the bootstrap (Efron 1983) [12]. Furthermore,

[1] Stone 1977a, [55]

the *leave-one-out* cross-validation is known to be too conservative in the sense that it tends to select an unnecessarily large model[2].

Five-fold crossvalidation is one of many ways of partitioning the data and is the method which will be used in the following, it has been suggested by Breiman, [5], and Shao, [54] as an alternative to the methods mentioned above. The data is split into five different sets, the calibration part consists consecutively of 4 different parts, and the validation data is the part left out of the calibration data. When splitting the data it is important to construct the groups in a way that the response-variables span approximately the same levels. For the gasoline example this is achieved by sorting the octane numbers in ascending order and then numbering them successively from 1 to 5 in order to get five sets that cover approximately the same range.

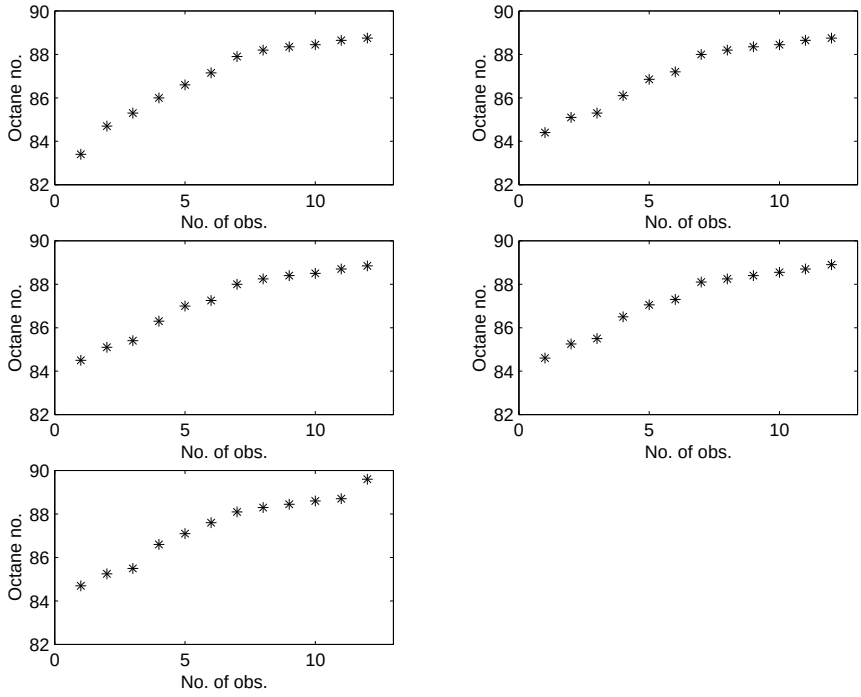

Figure 4.1: Partition of the octane observations in the gasoline data

[2]see [54] p. 486., [5] and [6]

For comparing the results of the methods we need a measure to judge the goodness of prediction. In the literature there are several equivalent measures. Suppose the calibration data is split into s sets then

- RSS = $\sum_{i=1}^{s}(\mathbf{y}_i - \hat{\mathbf{y}}_{\bar{i}})^T(\mathbf{y}_i - \hat{\mathbf{y}}_{\bar{i}})$
- MSEP = RSS$/n$
- RMSEP = $\sqrt{MSEP}$

where $\hat{\mathbf{y}}_{\bar{i}}$ are the fitted values for the ith split in the cross-validation. See [7] p. 54 and [56] chap.3. This procedure will be fair if

- the relation between y and x remains the same.
- future x-vectors are alike the calibration ones.

$$MSEP = E[(\mathbf{y} - \hat{\mathbf{y}})^2] = E[(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^2] \tag{4.5}$$

$$= V[\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}] + (E[\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}])^2 \tag{4.6}$$

see [63]. As seen above it can be an advantage to allow a little bias if this can reduce the variance. The Root Mean Squared Error of Prediction, RMSEP, will be used here.

4.2 MLLS applied to gasoline example

The parameter estimate for the MLLS solution based on the full data set when applied to the gasoline problem is shown in Figure 4.2. The regression coefficients in Figure 4.2 oscillate wildly with the wavelength at a high amplitude. This is due to the presence of near-collinearities. The cross-validated RMSEP-value is shown in Table 4.1

Method	$\|\hat{\boldsymbol{\beta}}\|_2$	RMSEP
MLLS	217.7	0.34

Table 4.1: RMSEP-value for the MLLS solution.

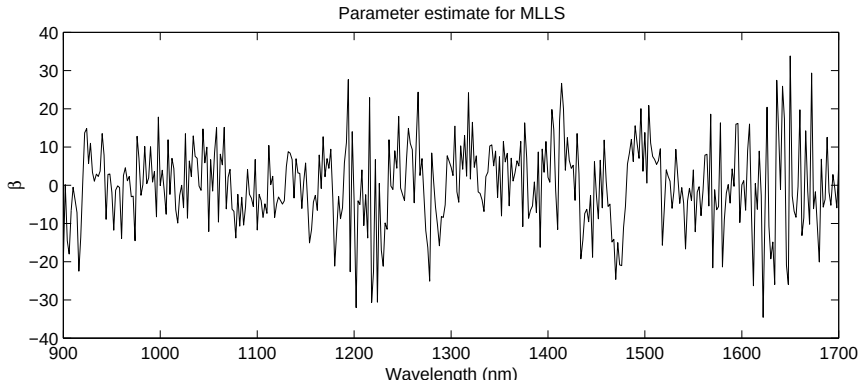

Figure 4.2: Parameter estimates, $\hat{\boldsymbol{\beta}}$, for MLLS

Chapter 5
Ridge Regression

5.1 Introduction

The Ridge regression, (RR), method was introduced as a method for stabilizing estimates in the presence of near-collinearity. RR produces a non-central estimate with smaller variance than the OLS estimate. The idea of adding a constant to the diagonal elements of the $\mathbf{X}^T\mathbf{X}$ matrix became popular when presented by Hoerl & Kennard (1970) in [28].

5.2 Theory

Again the linear regression model from (3.7) is considered, with the usual assumptions about $\mathbf{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$. The Ridge estimate is given as the solution to (4.2), [19] p. 565, which for $k > 0$ must satisfy

$$\hat{\boldsymbol{\beta}}_{RR} = (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{5.1}$$

Adding to $\mathbf{X}^T\mathbf{X}$ a multiple of $\mathbf{I}$ has a stabilizing effect, [15]. For $k = 0$ the Ridge estimate will be identical to the OLS estimate. See [28] p. 57, [7] p. 56, [10] p. 4.61 and [11] chap.6.7. The relationship of a Ridge estimate,

$\hat{\boldsymbol{\beta}}_{RR}$, to an OLS estimate, $\hat{\boldsymbol{\beta}}$, is given by the following linear transformation

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{RR} &= (\mathbf{I} + k(\mathbf{X}^T\mathbf{X})^{-1})^{-1}\hat{\boldsymbol{\beta}} && (5.2)\\ &= \mathbf{Z}\hat{\boldsymbol{\beta}} && (5.3) \end{aligned}$$

$\hat{\boldsymbol{\beta}}_{RR}$ is not unbiased since

$$E[\hat{\boldsymbol{\beta}}_{RR}] = \mathbf{Z}\boldsymbol{\beta}$$

[10] p. 4.61. The covariance matrix of $\hat{\boldsymbol{\beta}}_{RR}$ is

$$V[\hat{\boldsymbol{\beta}}_{RR}] = \sigma^2(\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1} \tag{5.4}$$

[10] p. 4.61. Here it is obvious that for $k \to \infty$ we have $V[\hat{\boldsymbol{\beta}}_{RR}] \to 0$.
To look at $\hat{\boldsymbol{\beta}}_{RR}$ from the point of view of mean square error the following expression is obtained by application of the expectation operator and (5.2)

$$\begin{aligned} E[L_1^2] &= E[(\hat{\boldsymbol{\beta}}_{RR} - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}}_{RR} - \boldsymbol{\beta})] && (5.5)\\ &= E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\mathbf{Z}^T\mathbf{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \\ &\quad + (\mathbf{Z}\boldsymbol{\beta} - \boldsymbol{\beta})^T(\mathbf{Z}\boldsymbol{\beta} - \boldsymbol{\beta}) && (5.6)\\ &= \sigma^2 trace(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Z}^T\mathbf{Z} \\ &\quad + \boldsymbol{\beta}^T(\mathbf{Z} - \mathbf{I})^T(\mathbf{Z} - \mathbf{I})\boldsymbol{\beta} && (5.7)\\ &= \sigma^2[trace(\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1} \\ &\quad - k trace(\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-2}] && (5.8)\\ &\quad + k^2\boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-2}\boldsymbol{\beta} \\ &= \sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i + k)^2} && (5.9)\\ &\quad + k^2\boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-2}\boldsymbol{\beta} \end{aligned}$$

[28] p. 60, where the last part of (5.9) is the squared distance from $\mathbf{Z}\boldsymbol{\beta}$ to $\boldsymbol{\beta}$. It will be zero when $k = 0$, since $\mathbf{Z}$ is then equal to $\mathbf{I}$. Thus $k^2\boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-2}\boldsymbol{\beta}$ can be considered the square of a bias introduced when $\hat{\boldsymbol{\beta}}_{RR}$ is used instead of $\hat{\boldsymbol{\beta}}$. The first term in (5.9) can be shown to be the total variance of the estimate. From (5.2) and (5.4) we get

$$\begin{aligned} V[\hat{\boldsymbol{\beta}}_{RR}] &= \mathbf{Z}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T V[\mathbf{y}]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Z}^T \\ &= \sigma^2\mathbf{Z}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Z}^T \end{aligned} \tag{5.10}$$

The sum of the diagonal elements of (5.10) is the sum of all variances of $\hat{\boldsymbol{\beta}}_{RR_i}$. The total variance decreases as k increases while the squared bias increases with k. By noticing that $\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^2}$ is a decreasing function of k, while $k^2\boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X}+k\mathbf{I})^{-2}\boldsymbol{\beta}$ is increasing, we have that there exist values of k for which the mean square error is less for $\hat{\boldsymbol{\beta}}_{RR}$ than for $\hat{\boldsymbol{\beta}}$, [10] p. 4.61-62, [11] p. 316 and [28] p. 60.

By using the SVD, (3.29), it is clear that (5.1) can be expressed as

$$\hat{\boldsymbol{\beta}}_{RR} = (\mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^T + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \quad ; \, k > 0 \tag{5.11}$$

by adding k to the diagonal elements of $\boldsymbol{\Sigma}^2$ (5.11) can be written as

$$\hat{\boldsymbol{\beta}}_{RR} = \sum_{j=1}^{r} \frac{1}{\lambda_j + k} v_j^T \mathbf{X}^T \mathbf{y} v_j \quad ; \, k > 0 \tag{5.12}$$

this means that the function $f(\lambda)$, in (4.4), determining the shrinking factors must be

$$f(\lambda) = \frac{\lambda}{\lambda + k} \quad ; \, k > 0 \tag{5.13}$$

Relating this to (4.4) where $\hat{\alpha}_j$ represents the unbiased OLS estimate, it is clear that RR is a shrinking method. The method is now clear, by iteratively selecting values of $k > 0$, we introduce a small bias and substantially reduce the variance, thereby improving the RMSEP-value. The optimal value of k is found for the smallest RMSEP-value.

5.3 Bayesian Motivation

For RR the parameters β_i are assumed to be independent normally distributed with mean zero and known variance σ_β^2. The posterior density of $\boldsymbol{\beta}$ is proportional to the likelihood times the prior, that is, proportional to

$$\exp(-1/2)\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/\sigma^2 + \boldsymbol{\beta}^T\boldsymbol{\beta}/\sigma_\beta^2\} \tag{5.14}$$

The exponential argument is quadratic in $\boldsymbol{\beta}$ and thus $\boldsymbol{\beta}$ is normal *a posteriori*. Completing the square in the quadratic form in $\boldsymbol{\beta}$ this multivariate normal posterior has mean $\hat{\boldsymbol{\beta}}_{RR}$ and covariance matrix

$$\sigma^2(\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1} \tag{5.15}$$

where $k = \sigma^2/\sigma_\beta^2$. If σ_β^2 is very large then the prior variances of the β_i are large and k is near zero, whence the posterior mean, $\hat{\boldsymbol{\beta}}_{RR}$, approaches the MLLS solution. The implication of this Bayesian motivation is that the Ridge estimator will tend to do well if these prior assumptions are met. For RR to perform optimal the regression coefficients should be clustered around zero and look like a random sample from a zero mean normal distribution, [7] p. 61.

5.4 Ridge applied to gasoline example

For the Ridge method the optimal model was found by iterating through values of k chosen on an equally spaced grid on the logarithmic scale, see Figure 5.1.

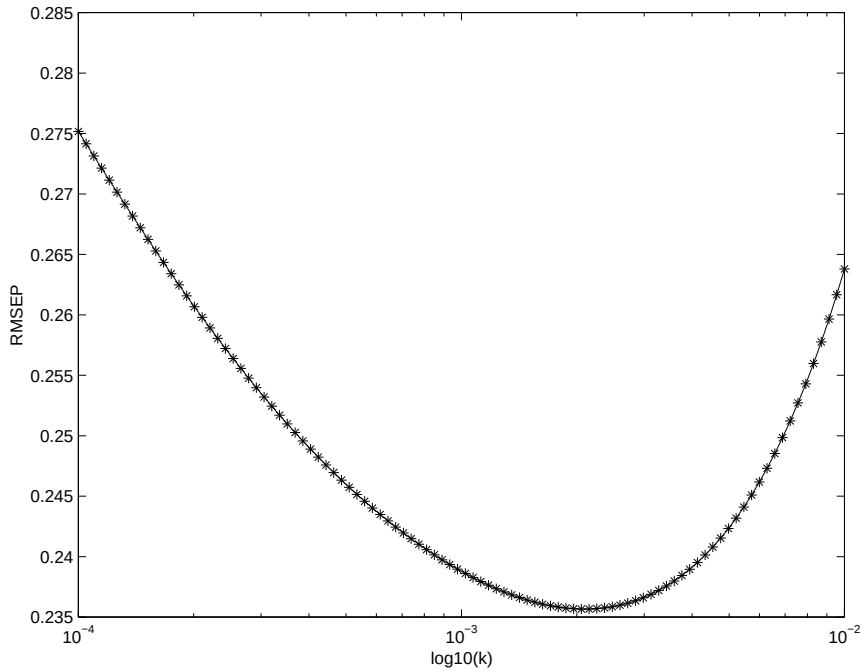

Figure 5.1: RMSEP as a function of $\log_{10}(k)$

Method	Regularization parameter	$\|\hat{\boldsymbol{\beta}}\|_2$	RMSEP
Ridge	$k = 0.002$	$\|\hat{\boldsymbol{\beta}}\|_2 = 27.48$	0.24

Table 5.1: Result for Ridge Regression.

For $k = 0.002$ the Ridge method is applied to the full data set, and the resulting parameter estimate can be seen in Figure 5.2. It is clear to see how the variation in the parameter estimate decreases when k increases. Remember that Ridge regression approximates the MLLS estimate[1] for $k \to 0$.

The Ridge method provides a smoother parameter estimate than MLLS. To see this consider the SVD of $\mathbf{X}$: ($\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$). In the gasoline example the singular values range from 1.61 to 0.002, the rows of $\mathbf{X}$ are weighted averages of the columns of the right singular vectors, $\mathbf{V}$. Figure 5.3 shows the first three and the last three columns of $\mathbf{V}$. Note that the columns corresponding to the larger singular values are much smoother than the columns corresponding to the smaller singular values.

The MLLS estimate which in (3.33) was formulated as $\boldsymbol{\beta} = \mathbf{V}_r\boldsymbol{\Sigma}_r^{-1}\mathbf{U}_r^T\mathbf{y}$ will obviously tend to be rough, since the rougher columns of $\mathbf{V}$ are multiplied by larger elements of $\boldsymbol{\Sigma}^{-1}$. Providing that k is not chosen too small, the Ridge estimate, (5.11), will be smoother since the rough columns of $\mathbf{V}$ have smaller multipliers, see [22].

[1] See Figure 4.2

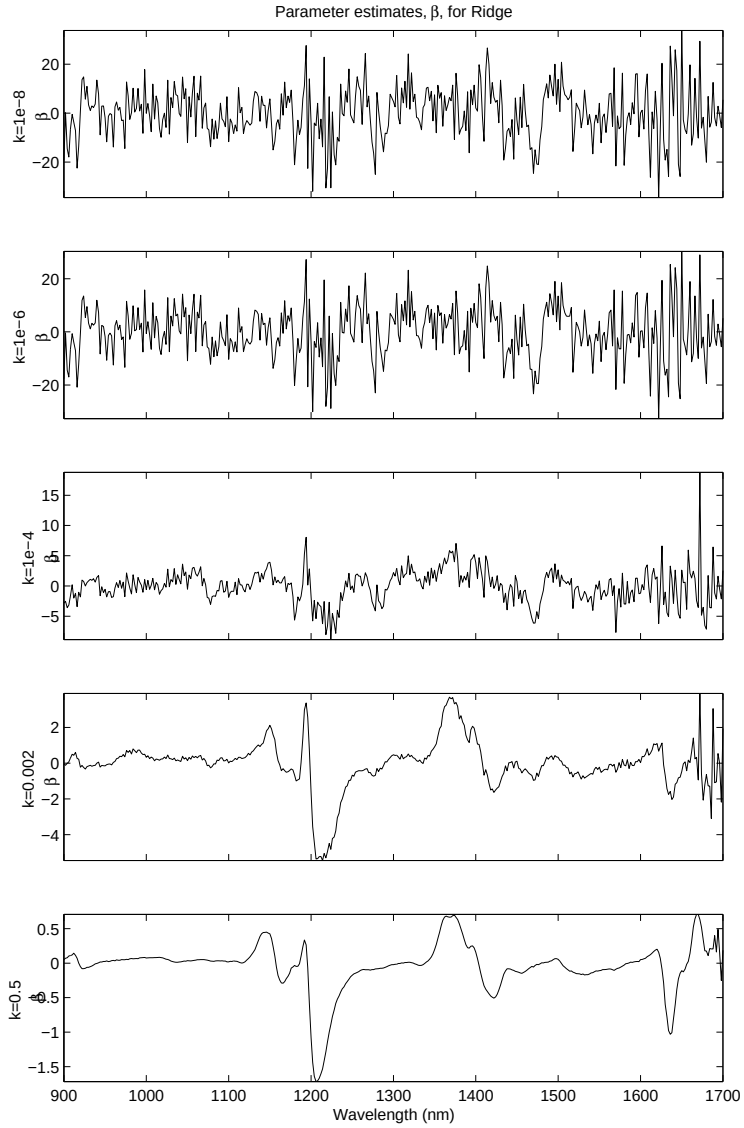

Figure 5.2: The $\hat{\boldsymbol{\beta}}$ estimate plotted for different k. Notice the decrease in the variation of $\hat{\boldsymbol{\beta}}$ when k increases. Also note the scaling of the y-axis.

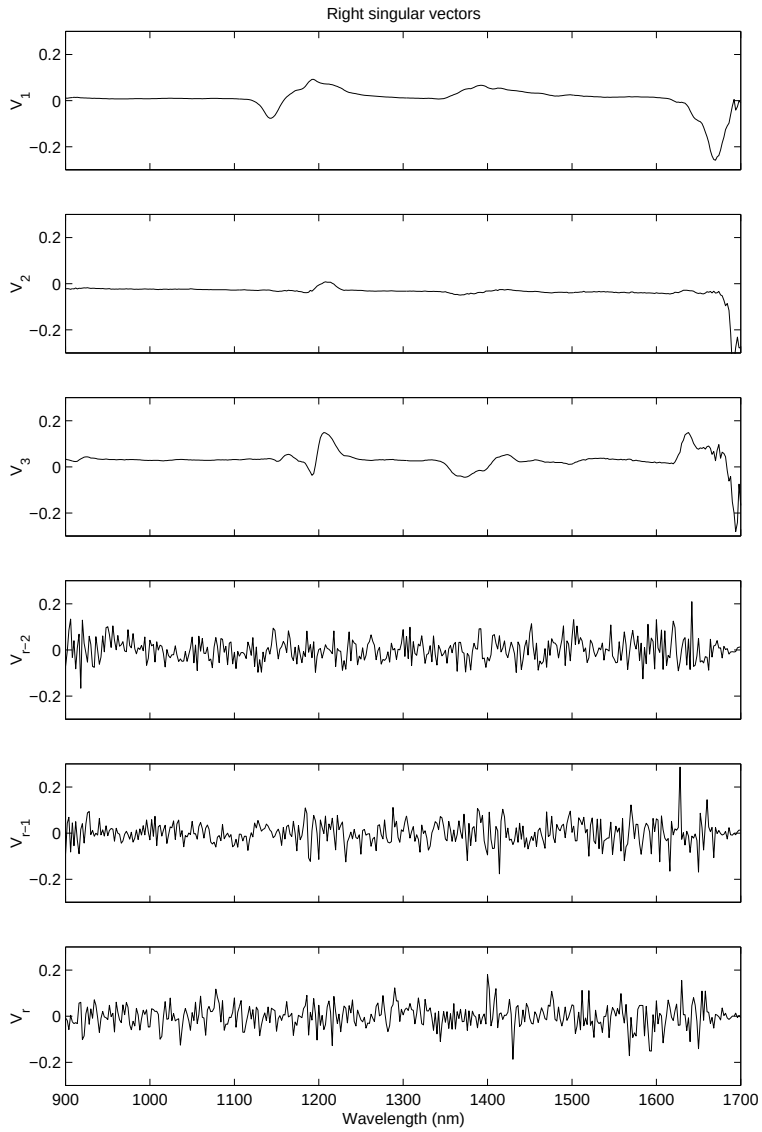

Figure 5.3: The Right singular vectors of $\mathbf{X}$ as a function of the wavelengths

Chapter 6
Principal Components Regression

6.1 Introduction

Principal Components Regression, PCR, forms a new set of canonical variables by canonical variance analysis of $\mathbf{X}^T\mathbf{X}$. The new variables, called the principal components, are used to predict the response variables. As shown earlier the variance of the least squares estimate becomes very large when the x-variables are almost linear dependent. In PCR this is avoided by choosing a smaller amount of the canonical variables.

6.2 Theory

By using the spectral decomposition, SVD, we can define the PCR estimate as:

$$\hat{\boldsymbol{\beta}}_{PCR} = \left(\sum_{j=1}^{K} (1/\lambda_j) \mathbf{v}_j {\mathbf{v}_j}^T \right) \mathbf{X}^T \mathbf{y} \quad ; K = 1, \ldots, r \tag{6.1}$$

where $\mathbf{v}_j$ are the eigenvectors of $\mathbf{X}^T\mathbf{X}$ and λ_j are the corresponding eigenvalues. r is the rank of $\mathbf{X}^T\mathbf{X}$ and corresponds to the number of nonzero

eigenvalues. The singular value decomposition, SVD, leads to a new expression for (3.7)

$$\mathbf{t} = \mathbf{U}^T\mathbf{y} = \mathbf{\Sigma}\mathbf{V}^T\boldsymbol{\beta} + \mathbf{U}^T\boldsymbol{\epsilon} \tag{6.2}$$

Due to the properties of $\mathbf{U}$ the error term is unaffected by this transformation. Using (3.34) and letting $\mathbf{V}^T\boldsymbol{\beta} = \boldsymbol{\alpha}$, the model becomes

$$t_j = \sqrt{\lambda_j}\alpha_j + \epsilon_j, \quad j \leq K \tag{6.3}$$

$$t_j = \epsilon_j, \quad j > K \tag{6.4}$$

[7] p.58. The PCR estimate, $\hat{\alpha}_{PCR}$, of this model is

$$\hat{\alpha}_{j,PCR} = \frac{t_j}{\sqrt{\lambda_j}}, \quad j \leq K \tag{6.5}$$

$$\hat{\alpha}_{j,PCR} = 0, \quad j > K \tag{6.6}$$

For α-coefficients corresponding to zero eigenvalues the least squares estimator is indeterminate, but is zero for the minimum length least squares estimator. Thus the PCR estimate is equal to the MLLS estimate if all the eigenvalues are used, otherwise $\hat{\alpha}_{PCR}$ is shrunken towards zero. The shrinking function $f(\lambda)$ from (4.4) is obviously

$$f(\lambda) = 1 \quad ;\lambda_j \geq \lambda_K \tag{6.7}$$

$$= 0 \quad ;\lambda_j < \lambda_K \tag{6.8}$$

6.3 Selection principles

In most applications of PCR, principal components (PC's) are included in regression models in sequence according to respective variances, i.e. the magnitude of singular values associated with respective PC's. A suitable number of PC's is determined by the predictive ability obtained through cross-validation. This principle of selection is often referred to as *top-down* selection, e.g [36]. In [29] it is concluded that other selection principles are necessary, since there is no guarantee that the low-variance components are unimportant. Brown, [7] briefly mentions another selection strategy that includes those PC's most correlated with the response variables. This strategy has been investigated in [36], it will here be denoted Correlation-PCR, (CPCR). A third selection strategy from [36] is called Forward Selection PCR (FSPCR). The FSPCR can be described as follows:

Step 1: Compute all the PC's of $\mathbf{X}$ using a SVD as done with PCR and CPCR.

Step 2: Determine the first PC producing the minimum RMSEP-value in a cross-validation, or for another chosen criterion.

Step 3: Based on this best single PC subset, the best two PC subset is identified as the one providing the minimum criterion from all possible two PC combinations containing the best single PC.

Step 4: Analogous to step 3, the best third PC to augment the best two PC subset is identified.

Step 5: The process continues until all PC's have been included in the model. The best model should be that with the lowest criterion.

The underlying assumption of FSPCR is that the most predictive components should be included in the model in the early stage of the sequential process and should not be excluded from the model as the number of PC's increases.

It is important to mention that the FSPCR method should only be compared when an external validation set is present. FSPCR based only on the cross-validated RMSEP-values tend to adapt to specific features of the data[1]. For PCR and CPCR the number of components is determined by use of cross-validation whereas for FSPCR the cross-validation is used to determine which components are to be included in the model. This can result in overfitting and therefore a fair comparison of the methods is not possible without an external validation set. Nevertheless the results for the FSPCR method applied to the gasoline example are shown.

6.4 PCR applied to gasoline example

The PCR methods are tested for a number of principal components. The result of this is seen in Figure 6.1.

[1] See [36] p.22

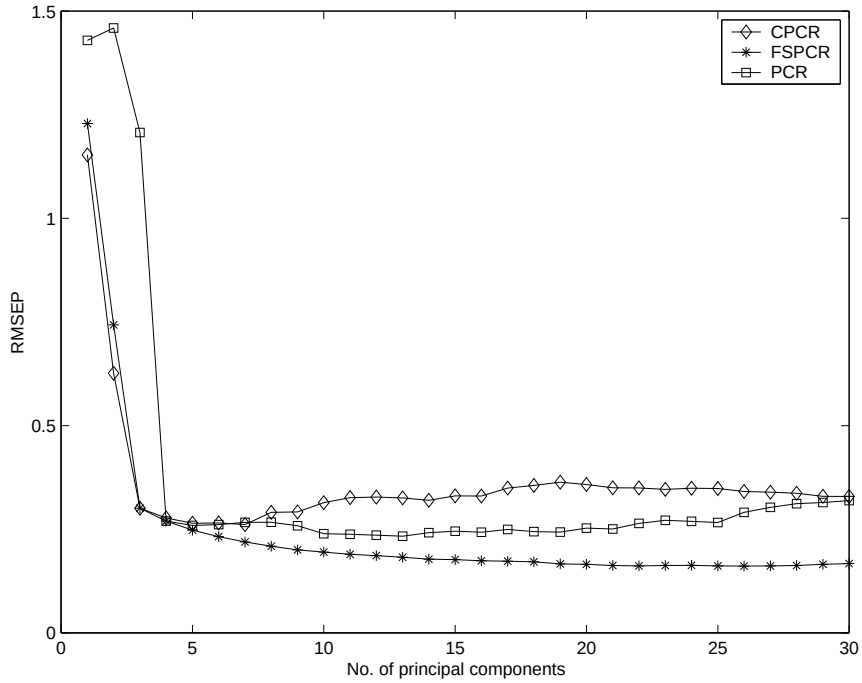

Figure 6.1: The RMSEP value is plotted as a function of the number of principal components. The optimal number of principal components for the PCR-method is 13, for the CPCR-method it is 7, and the FSPCR-method has it's minimum RMSEP-value for 26 components.

Method	Regularization parameter	RMSEP
PCR	No. of comps.= 13	0.23
CPCR	No. of comps.= 7	0.26
FSPCR	No. of comps.= 26	0.17

Table 6.1: Results for PCR, CPCR and FSPCR

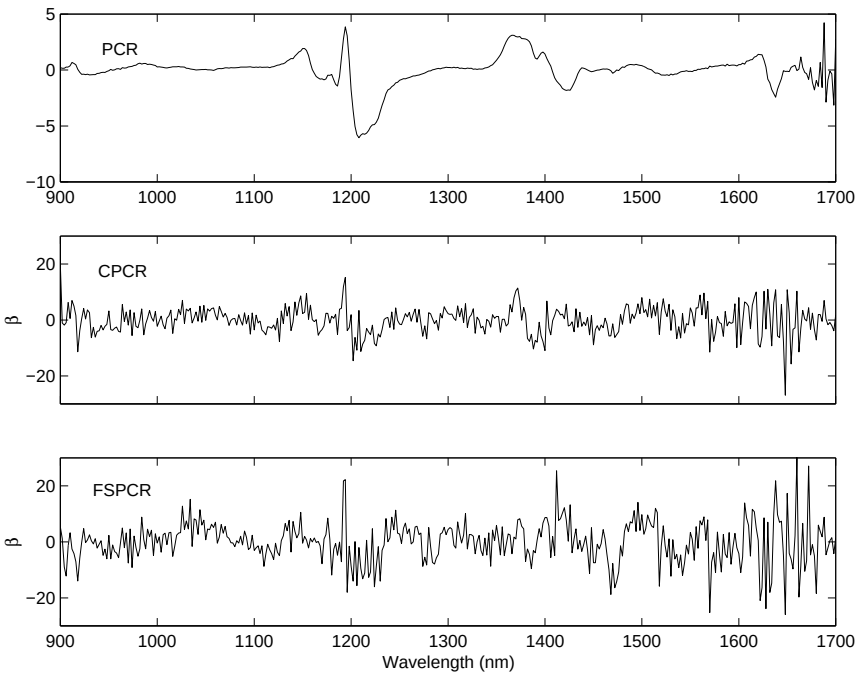

Figure 6.2: The $\hat{\boldsymbol{\beta}}$ estimate plotted for the three methods . Notice the increase in the variation of $\hat{\boldsymbol{\beta}}$ when PC's corresponding to the smaller eigenvalues are chosen.

Table 6.2: Order of the PC's chosen by the methods, (only the first 15 PC's are shown). The absolute correlation is also shown.

FSPCR order	CPCR order	$\lvert r \rvert$
4	4	0.7155
3	3	0.5185
1	1	0.4358
2	2	0.0794
10	9	0.0731
41	11	0.0587
5	42	0.0378
38	12	0.0370
29	17	0.0316

46	25	0.0305
40	5	0.0297
39	38	0.0295
42	27	0.0260
27	30	0.0241
8	37	0.0227

6.4.1 Comments to the PCR methods

FSPCR and CPCR perform better than PCR for the first several PC's. After this, CPCR produces worse results compared to PCR. The first 4 principal components are the same for all three methods. Comparing the PCR and the CPCR method reveals that the top-down selection in this case is the best choice for choosing the principal components. The parameter estimates for Ridge and PCR look very much alike, see Figure 5.2 and 6.2.

Chapter 7
Partial Least Squares Regression

7.1 Introduction

Partial Least Squares regression, (PLS), was first introduced in the 1960s by Herman Wold. PLS uses a number of factors as well as PCR. In PCR only the explanatory variables are used to form these, (principal components), whereas in PLS the calculation of the factors uses both the explanatory and the response variables, (found by canonical covariance analysis on $\mathbf{X}$ and $\mathbf{y}$). PLS is often just presented as an algorithm. The original PLS algorithm , [24], has been shown to be related to the conjugate gradient method[1] for inverting matrices, by S. Wold, [61], this relation has been explored in detail by Manne (1987), [40].

[1] See [19] pp.519-523

7.2 Theory

The matrix $\mathbf{X}$ of n observations on p explanatory variables can be described in a bilinear factor form

$$\mathbf{X} = \mathbf{t}_1\mathbf{v}_1^T + \mathbf{t}_2\mathbf{v}_2^T + \cdots + \mathbf{t}_K\mathbf{v}_K^T + \boldsymbol{\epsilon}_K \tag{7.1}$$

where $\mathbf{t}_i$ is of length n and $\mathbf{v}_i$ is a p-vector. The idea behind PLS is that the relationship between $\mathbf{X}$ and $\mathbf{y}$ is conveyed through the orthogonal factors, thus we have

$$\mathbf{y} = \mathbf{t}_1 b_1 + \mathbf{t}_2 b_2 + \cdots + \mathbf{t}_K b_K + \mathbf{f}_K \tag{7.2}$$

for scalar b_i. Conditions need to be imposed for uniqueness. Forcing $\mathbf{t}_i$ to be mutually orthogonal in $\mathcal{R}^n$ and $\mathbf{v}_i$ to be mutually orthogonal in $\mathcal{R}^p$ will lead to the traditional PCR which bases choice on eigenvalues, since then $\mathbf{t}_i$ would be the eigenvectors of $\mathbf{X}\mathbf{X}^T$ and $\mathbf{v}_i$ the eigenvectors of $\mathbf{X}^T\mathbf{X}$, i.e the factors are based on $\mathbf{X}$ alone. Due to this property there are two different algorithms for PLS depending on whether $\mathbf{t}_i$ or $\mathbf{v}_i$ are determined as orthogonal. They both lead to the same result. The algorithm presented below is based on $\mathbf{t}_i$ being orthogonal in $\mathcal{R}^n$.

7.3 PLS-algorithm

In [7] and [24] PLS is described as a two-stage approach. The first stage of PLS determines the $K \leq r$, $r = min(n, p)$, factors $\mathbf{t}_i$ of length n to be included in the regression. The ith factor $\mathbf{t}_i = \mathbf{X}w_i$ is chosen to maximize $t_i^T\mathbf{y}$ subject to the constraints that $|w_i| = 1$ and that $\mathbf{t}_i$ is orthogonal to the space spanned by the basis $\{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{i-1}\}$. In the second stage ordinary least squares is applied to the regression of $\mathbf{y}$ on the factors $\mathbf{t}_i$, $i = 1, 2, \ldots, K$. Imposing the condition that the scores, $\mathbf{t}_i$, are to be orthogonal in $\mathcal{R}^n$ gives the following algorithm, [24] p.588. By choosing a maximum number of, K, factors, the aim now is to find representations of (7.1) and (7.2). By writing $\boldsymbol{\epsilon}_0 = \mathbf{X}$ and $\mathbf{f}_0 = \mathbf{y}$ one must have

$$\boldsymbol{\epsilon}_i = \boldsymbol{\epsilon}_{i-1} - \mathbf{t}_i\mathbf{v}_i^T \tag{7.3}$$

$$\mathbf{f}_i = \mathbf{f}_{i-1} - \mathbf{t}_i b_i \tag{7.4}$$

for i=1,...,K. $\mathbf{t}_i$, $\mathbf{v}_i$ and b_i are determined by induction. $\mathbf{t}_i$ is determined as a linear combination of the $\mathbf{x}$-residuals from the previous step. In particular for i=1

$$\mathbf{t}_1 = \sum_{j=1}^{p} \mathbf{x}_j \mathbf{w}_{j1} = \mathbf{X}\mathbf{w}_1 \tag{7.5}$$

where $\mathbf{w}_1$ is a p-dimensional weight-vector. To secure that $\mathbf{t}_1$ is highly correlated with $\mathbf{y}$, the choice for each component $\mathbf{w}_{j1}$ is made proportional to the covariance between $\mathbf{x}_j$ and $\mathbf{y}$

$$\mathbf{w}_{j1} = \mathbf{x}_j^T \mathbf{y}, \quad \text{i.e.:} \quad \mathbf{w}_1 = \mathbf{X}^T \mathbf{y} \tag{7.6}$$

so in general

$$\mathbf{t}_i = \boldsymbol{\epsilon}_{i-1} \mathbf{w}_i \tag{7.7}$$

$$\mathbf{w}_i = \boldsymbol{\epsilon}_{i-1}^T \mathbf{f}_{i-1} \tag{7.8}$$

then $\mathbf{v}_i$ and b_i are determined such that a best possible fit in (7.3) and (7.4) is obtained. For i=1, the best possible fit to $\mathbf{y} = \mathbf{t}_1 b_1 + \mathbf{f}_1$ is given by the regression coefficient $b_1 = \mathbf{y}^T \mathbf{t}_1 / \mathbf{t}_1^T \mathbf{t}_1$. So in general

$$\mathbf{v}_i = \boldsymbol{\epsilon}_{i-1}^T \mathbf{t}_i / \mathbf{t}_i^T \mathbf{t}_i \tag{7.9}$$

$$b_i = \mathbf{f}_{i-1}^T \mathbf{t}_i / \mathbf{t}_i^T \mathbf{t}_i \tag{7.10}$$

The new residuals, $\boldsymbol{\epsilon}_i$ and $\mathbf{f}_i$ are then found from (7.3) and (7.4). The second stage of PLS regresses $\mathbf{y}$ on the factors $\mathbf{t}_i$, $i = 1, 2, \ldots, K$. This involves minimizing the sum of squares

$$\left(\mathbf{y} - \sum_{i=1}^{K} b_i \mathbf{t}_i\right)^T \left(\mathbf{y} - \sum_{i=1}^{K} b_i \mathbf{t}_i\right) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{PLS})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{PLS}) \tag{7.11}$$

with respect to b_i, where $\boldsymbol{\beta}_{PLS} = \sum_{i=1}^{K} b_i w_i$ is the PLS parameter vector, [8]. If now $\mathbf{x}_0 = (x_{01}, x_{02}, \ldots, x_{0p})^T$ is a new spectrum, and $\mathbf{e}_0 = \mathbf{x}_0 - \bar{\mathbf{x}}$ with $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_p)^T$, then the factors and residuals are

$$\mathbf{t}_{i0} = \mathbf{e}_{i-1}^T \mathbf{w}_i, \qquad \mathbf{e}_i = \mathbf{e}_{i-1} - \mathbf{t}_{i0} \mathbf{v}_i \tag{7.12}$$

and prediction of the corresponding unknown response-variable is then

$$\hat{\mathbf{y}}_{K0} = \bar{\mathbf{y}} + \sum_{i=1}^{K} \mathbf{t}_{i0} b_i = \bar{\mathbf{y}} + \sum_{i=1}^{K} \mathbf{t}_{i0} (\mathbf{t}_i^T \mathbf{t}_i)^{-1} \mathbf{t}_i^T \mathbf{y} \tag{7.13}$$

where K is based on some cross-validatory method, [24]. From [24] p. 595 we have that

$$\text{span}\{w_1, \ldots, w_K\} = \text{span}\{\mathbf{X}^T\mathbf{y}, (\mathbf{X}^T\mathbf{X})\mathbf{X}^T\mathbf{y}, \ldots, (\mathbf{X}^T\mathbf{X})^{K-1}\mathbf{X}^T\mathbf{y}\} \tag{7.14}$$

which in the numerical litterature is known as a *Krylov sequence*[2]. The dimension of the space spanned by this sequence will be the maximal number of factors that the PLS-algorithm can give, which is equal to the number of eigenvectors in $\mathbf{X}^T\mathbf{X}$ with non-zero components along $\mathbf{X}^T\mathbf{y}$.[3] So an alternative form for $\hat{\boldsymbol{\beta}}_{PLS}$ is

$$\hat{\boldsymbol{\beta}}_{PLS} = \sum_{i=1}^{K} \hat{\gamma}_i (\mathbf{X}^T\mathbf{X})^{i-1}\mathbf{X}^T\mathbf{y} \tag{7.15}$$

where the parameters $\hat{\gamma}_i$ minimize equation (7.11), [8]. Using the SVD of $\mathbf{X}$, it follows from [15], that $\hat{\boldsymbol{\beta}}_{PLS}$ is given by (4.4) with shrinkage funtion

$$f(\lambda) = \sum_{i=1}^{K} \hat{\gamma}_i \lambda^i \tag{7.16}$$

see [15] and [8].

An interesting aspect of the PLS solution is that (unlike RR and PCR) it not only shrinks the OLS solution in some eigendirections ($f(\lambda) \leq 1$) but expands in others ($f(\lambda) > 1$). For a K-component PLS solution the OLS solution is expanded in the subspace defined by the eigendirections associated with the eigenvalues closest to the Kth eigenvalue. Directions associated with somewhat larger eigenvalues tend to be slightly shrunk, and those with smaller eigenvalues are substantially shrunk, see [15] p. 121.

7.4 PLS applied to gasoline example

The RMSEP-values for different number of PLS components is shown in Figure 7.1. The result is shown in TABLE 7.1.

[2] See [19] pp. 476-477
[3] See [24] p. 595.

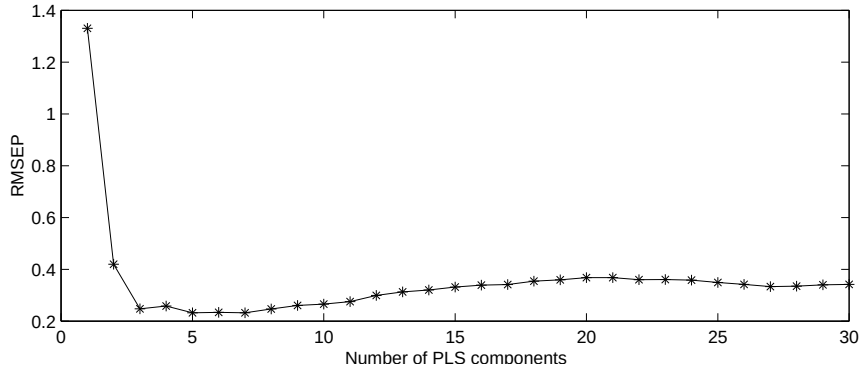

Figure 7.1: The RMSEP value is plotted as a function of the number of PLS factors. The optimal number of PLS factors is 7.

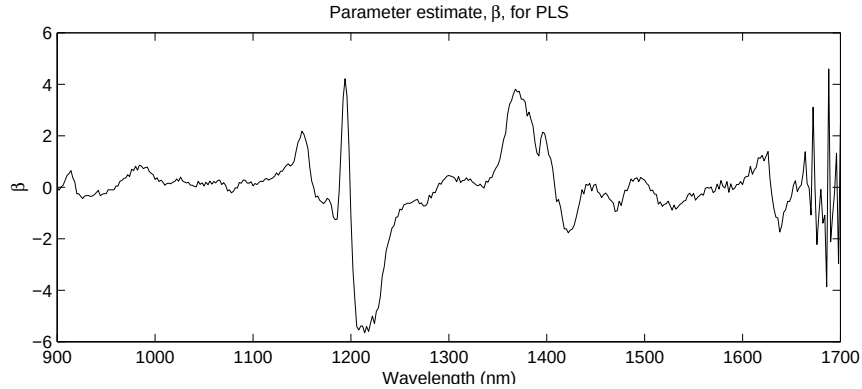

Figure 7.2: Parameter estimate, $\hat{\boldsymbol{\beta}}$, for PLS.

Method	Regularization parameter	RMSEP
PLS	No. of comps.= 7	0.23

Table 7.1: Result for PLS.

PLS obtains the same RMSEP-value as PCR. PLS uses 7 components compared to the 13 principal components which was optimal for PCR.

7.5 Summary of Ridge, PCR and PLS

Ridge regression, PCR and PLS are seen to operate in a similar fashion. Their principal goal is to shrink the parameter estimate away from the MLLS solution. It is shown above how the methods produce biased estimates. The effect of this bias is to pull the parameter estimate away from the MLLS solution toward directions in which the data have larger spread. The degree of this bias is regulated by the value of the model selection parameter. For Ridge regression, setting $k = 0$ yields the MLLS solution, whereas $k > 0$ introduces increasing bias along with increased shrinkage of the length of the parameter estimate. In PCR the degree of bias is controlled by the value of K, that is the number of eigenvectors used in (6.2). If $K = r$, (rank of $\mathbf{X}^T\mathbf{X}$), one obtains the MLLS solution. For $K < r$, bias is introduced. For PLS the situation is similar to that of PCR. The degree of bias is regulated by the number of components used. The MLLS solution is obtained for $K = r$.

Frank & Friedman, [15] and Helland, [24][4], have done a thorough comparative study on the behavior of the three methods. Their general conclusion is that when the methods are applied to problems involving data with high collinearity in which the variance of the estimates tends to dominate the bias, the solutions and hence the performance tends to be quite similar. That PLS uses fewer components than PCR is generally seen when the methods are used for calibration purposes. In any case, either method is free to choose its own number of components (bias-variance trade-off) through cross-validation. Both PCR and PLS span a full (but not the same) spectrum of models from the most biased to the least biased (MLLS solution). The fact that PLS tends to balance this trade-off with fewer components is (in general) neither an advantage nor disadvantage[5].

[4] Helland focuses solely on PCR and PLS.
[5] See [15] p. 122 and [24] p. 602.

Chapter 8
Cyclic Subspace Regression

8.1 Introduction

Cyclic Subspace Regression, CSR [33], makes explicit mathematical connections between PCR, PLS and MLLS. The theory leads to a simple algorithm that provides not only solutions for PCR, PLS and MLLS but also a finite number of other related methods. CSR also shows that the methods differ only in the amount of information used from the calibration data. The motivation for developing this algorithm comes from the work done by Manne, [40] & Helland, [24] where it is shown that PLS is tied to a cyclic subspace of the matrix $\mathbf{X}^T\mathbf{X}$, see (7.14). The main idea of CSR is to start with a complete singular value decomposition of $\mathbf{X}$ and then to generate subspaces of the full row and column spaces by a Krylov procedure.[1]

For methods like PCR and PLS the parameter estimate in (3.7) can be written as

$$\hat{\boldsymbol{\beta}} = \mathbf{R}(\mathbf{R}^T\mathbf{X}^T\mathbf{X}\mathbf{R})^{-1}\mathbf{R}^T\mathbf{X}^T\mathbf{y} \tag{8.1}$$

for some $(p \times K)$ matrix $\mathbf{R}$ defining the subspace onto which $\mathbf{x}$ is projected, [26] p. 239. PCR results from (8.1) by letting the columns of $\mathbf{R}$ be K eigenvectors of $(\mathbf{X}^T\mathbf{X})$. As shown by Helland, [24], PLS fits into (8.1) by taking $\mathbf{R} = (\mathbf{X}^T\mathbf{y}, (\mathbf{X}^T\mathbf{X})\mathbf{X}^T\mathbf{y}, \ldots, (\mathbf{X}^T\mathbf{X})^{K-1}\mathbf{X}^T\mathbf{y})$

[1] See [19] pp. 476-477.

8.2 Theory

In what follows it is assumed that $\mathbf{X}$ is $n \times p$, $\mathbf{y}$ is $n \times 1$, and $\boldsymbol{\beta}$ is $p \times 1$. A linear relationship of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ is believed to hold and $\text{rank}(\mathbf{X}) = r \geq 1$. Then the matrices $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$ have r nonzero eigenvalues, λ. Associated with these eigenvalues are two sets of orthonormal eigenvectors. Let $\mathbf{U}$ and $\mathbf{V}$ denote these, see 3.2.4. Recall from 3.2 that $\mathbf{y}$ can be written as $\mathbf{y} = \mathbf{y}_R + \mathbf{y}_N$. By combining (3.23) and (3.32) $\mathbf{y}_R$ can be written as

$$\mathbf{y}_R = \sum_{i=1}^{r} \mathbf{y}^T\mathbf{u}_i\mathbf{u}_i \tag{8.2}$$

This representation of $\mathbf{y}$ combined with the SVD of $\mathbf{X}$ implies that

$$\mathbf{X}^T\mathbf{y} = \sum_{i=1}^{r} \sigma_i\mathbf{y}^T\mathbf{u}_i\mathbf{v}_i \tag{8.3}$$

where $\sigma_i = \sqrt{\lambda_i}$. Now let l be a fixed integer satisfying $1 \leq l \leq r$ and form the following vector

$$\boldsymbol{\beta}_l = \mathbf{X}^T\mathbf{U}\mathbf{U}^T\mathbf{y} = \sum_{i=1}^{l} \sigma_i\mathbf{y}^T\mathbf{u}_i\mathbf{v}_i \tag{8.4}$$

where l signifies the amount of information extracted from the singular directions $\mathbf{v}_1, \ldots, \mathbf{v}_r$ in (8.4). Associated with $\boldsymbol{\beta}_l$ is the following l-dimensional cyclic $\mathbf{X}^T\mathbf{X}$-invariant subspace of $\mathcal{R}(X^T)$

$$\mathbf{X}_l = \text{span}\{\boldsymbol{\beta}_l, (\mathbf{X}^T\mathbf{X})\boldsymbol{\beta}_l, \ldots, (\mathbf{X}^T\mathbf{X})^{l-1}\boldsymbol{\beta}_l\} \tag{8.5}$$

which is tied to the theory regarding PLS, see (7.6) and (7.14). Let j be a fixed integer satisfying $1 \leq j \leq l \leq r$. Define

$$\mathbf{A}_l^j = (\boldsymbol{\beta}_l, (\mathbf{X}^T\mathbf{X})\boldsymbol{\beta}_l, \ldots, (\mathbf{X}^T\mathbf{X})^{j-1}\boldsymbol{\beta}_l) \tag{8.6}$$

and set

$$\mathbf{B}_l^j = \mathbf{X}\mathbf{A}_l^j \tag{8.7}$$

$\mathbf{A}_l^j$ is a $p \times j$ matrix obtained by using the first j vectors from the above representation of the subspace $\mathbf{X}_l$. The notation on $\mathbf{B}_l^j$ identifies that it results from $\mathbf{A}_l^j$. (8.4) always has a unique solution in $\mathcal{R}(X^T)$. Suppose

$\boldsymbol{\beta}_l$ in (8.4) is sought from a particular subspace $W \in \mathcal{R}(X^T)$. Then (8.4) may be incapable of producing such a solution as the unique solution in $\mathcal{R}(X^T)$ may not live in W. To assure that such a solution can be found it is necessary to form the matrix

$$\mathbf{A}_l^j \big((\mathbf{A}_l^j)^T \mathbf{A}_l^j\big)^{-1} (\mathbf{A}_l^j)^T \tag{8.8}$$

which is a projection of $\mathbb{R}^p$ onto $W \in \mathcal{R}(X^T)$. Having set $\mathbf{B}_l^j = \mathbf{X}\mathbf{A}_l^j$ the following matrix

$$\mathbf{B}_l^j \big((\mathbf{B}_l^j)^T \mathbf{B}_l^j\big)^{-1} (\mathbf{B}_l^j)^T \tag{8.9}$$

is a projection of $\mathbb{R}^n$ onto $\mathcal{R}(X\mathbf{A}_l^j \big((\mathbf{A}_l^j)^T \mathbf{A}_l^j\big)^{-1} (\mathbf{A}_l^j)^T)$. The equation

$$\mathbf{B}_l^j \big((\mathbf{B}_l^j)^T \mathbf{B}_l^j\big)^{-1} (\mathbf{B}_l^j)^T \mathbf{y} = \mathbf{X}\mathbf{A}_l^j \big((\mathbf{A}_l^j)^T \mathbf{A}_l^j\big)^{-1} (\mathbf{A}_l^j)^T \boldsymbol{\beta} \tag{8.10}$$

has a unique solution coming from the sought subspace $W \in \mathcal{R}(X^T)$ specified by the factors l and j. The parameter estimate of the CSR, $\boldsymbol{\beta}_{CSR}$, from use of the first j factors taken from the l-dimensional subspace $\mathbf{X}_l$ can now be written as

$$\boldsymbol{\beta}_{CSR} = \mathbf{A}_l^j \big((\mathbf{B}_l^j)^T \mathbf{B}_l^j\big)^{-1} (\mathbf{B}_l^j)^T \mathbf{y} \tag{8.11}$$

see [33]. By varying l and j this procedure can produce $(r^2 + r)/2$ different parameter estimates. As mentioned in the introduction this algorithm produces results for PCR and PLS amongst others. Results generated by CSR corresponds to PCR when $l = j$, PLS when $l = r$, and MLLS when $l = j = r$, see [33].

8.3 CSR applied to gasoline example

The $(r^2 + r)/2$ models were generated, and based on the RMSEP-values from the five-fold cross-validation, the best model was found to be the one consisting of 7 factors based on 19 eigenvectors from the SVD. The result is shown in Table 8.1.

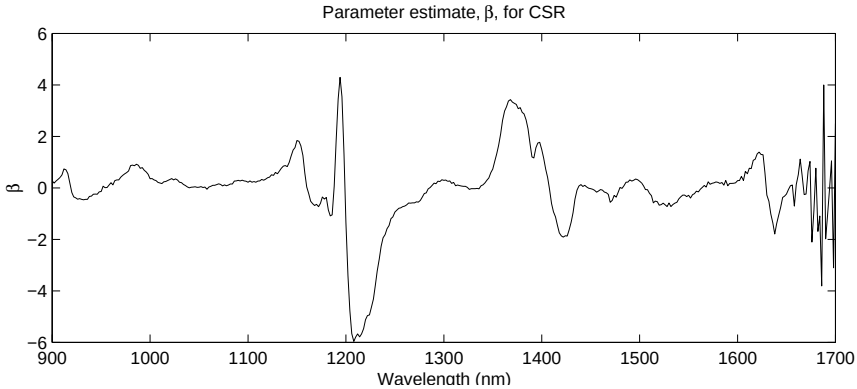

Figure 8.1: Parameter estimate, $\hat{\boldsymbol{\beta}}$, for CSR.

Method	Regularization parameter	RMSEP
CSR	No. of eigenvectors and factors = 19 and 7	0.23

Table 8.1: Results for CSR.

Chapter 9
Subset Selection

9.1 Introduction

Subset Selection, (SS), is a variable selection method which focuses on selecting a number of coefficients and to set others to zero. SS can provide easily interpreted models and aids the reduction in variance of the regression estimator, $\hat{\boldsymbol{\beta}}$.

Miller, [43], has given some reasons for using only a subset of the available variables:

- to estimate or predict at a lower cost by reducing the number of variables on which the data are to be collected;
- to predict more accurately by eliminating uninformative variables; and
- to estimate regression coefficients with smaller standard errors (particularly when some of the variables are highly correlated).

9.2 Theory

The regression problem may be solved using subset selection as follows, [44]. Given a set of explanatory variables $X_1, X_2, \ldots, X_p$, the aim is to find

a subset of variables $X_{(1)}, X_{(2)}, \ldots, X_{(k)}$, with $k < p$ that minimizes

$$S = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{k} \hat{\beta}_{(j)} x_{i(j)} \right)^2 \tag{9.1}$$

where $x_{i(j)}$ is the ith observation of the variable $X_{(j)}$ in the final subset of selected variables. $\hat{\beta}_{(j)}$ is the least squares regression coefficient for the corresponding selected variable. Determining the optimal value of k is an ongoing problem, and more details on defining correct stopping rules can be seen in [44]. The problem of finding the optimal set of explanatory variables is a 2^p problem, where p is the number of measured wavelengths. This is a very computational demanding problem which has spurred the development of a number of more or less heuristic selection methods. One of these methods is forward selection. For this method the first variable selected is the variable X_j for which

$$S_1 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_j x_{ij} \right)^2 \tag{9.2}$$

is minimized, where $\hat{\beta}_j$ minimizes S_1 for variable $X_{(j)}$. $\hat{\beta}_j$ is given by

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} x_{ij} y_i}{\sum_{i=1}^{n} x_{ij}^2} \tag{9.3}$$

This leads to the following value of S_1

$$S_1 = \frac{\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} x_{ij} y_i \right)^2}{\sum_{i=1}^{n} x_{ij}^2} \tag{9.4}$$

and hence the first selected variable is the one that maximizes

$$\frac{\left(\sum_{i=1}^{n} x_{ij} y_i \right)^2}{\sum_{i=1}^{n} x_{ij}^2} = \frac{\|\mathbf{x}_j^T \mathbf{y}\|_1^2}{\|\mathbf{x}_j\|_2^2} \tag{9.5}$$

Dividing this expression by $\|y_i\|_2$ the cosine of the angle between $\mathbf{x}_j$ and $\mathbf{y}$ is obtained, since the mean has been subtracted from each variable this value represents the correlation between the variable X_j and the response $\mathbf{y}$, see [44] p. 45.

This first selected variable, which is forced into all further subsets, is denoted $X_{(1)}$. The residuals $\mathbf{y} - \mathbf{x}_1\hat{\beta}_1$ are orthogonal to $X_{(1)}$. Therefore in order to work out which variable is to be included next, the orthogonal space to $X_{(1)}$ is searched. From each variable, X_j, other than the one already selected the following expression is formed

$$\mathbf{x}_{j.(1)} = \mathbf{x}_j - \hat{\beta}_{j.(1)}\mathbf{x}_1 \tag{9.6}$$

where $\hat{\beta}_{j.(1)}$ is the least squares coefficient of X_j on $X_{(1)}$. Now in expression (9.5) $\mathbf{y}$ is replaced with $\mathbf{y} - \mathbf{x}_{(1)}\hat{\beta}_{(1)}$ and $\mathbf{x}_j$ is replaced with $\mathbf{x}_{j.(1)}$. The variable $X_{j.(1)}$ which maximizes (9.5) is the next to be included. This ensures that the new variable selected is that which has the largest partial correlation in absolute value with $\mathbf{y}$ after $X_{(1)}$ has been fitted, see [44] p. 46. This process is repeated until a subset of variables $X_{(1)}, X_{(2)}, \ldots, X_{(k)}$ has been selected.

This procedure is advantageous because the required sums of squares and vector products can be obtained from previous calculations, making the whole process less computationally expensive.

Other subset selection methods exist, such as Efroymson's algorithm (stepwise regression) and backward elimination of variables, see [44], [11].

Subset selection, however, may not always be satisfactory because it can be variable, since the final subset of selected variables can change dramatically for small changes in the original data $\mathbf{y}$. A stability investigation by Breiman, [5], categorizes the subset selection procedures as unstable whereas a method like Ridge regression is considered to be a stable method. A new method called *Mean Subset* that overcomes the instability issues of subset selection has been proposed in [65].

9.3 Forward Selection applied to gasoline example

Using the 5-fold cross-validation to choose the number of variables, k, results in the following Figure 9.1

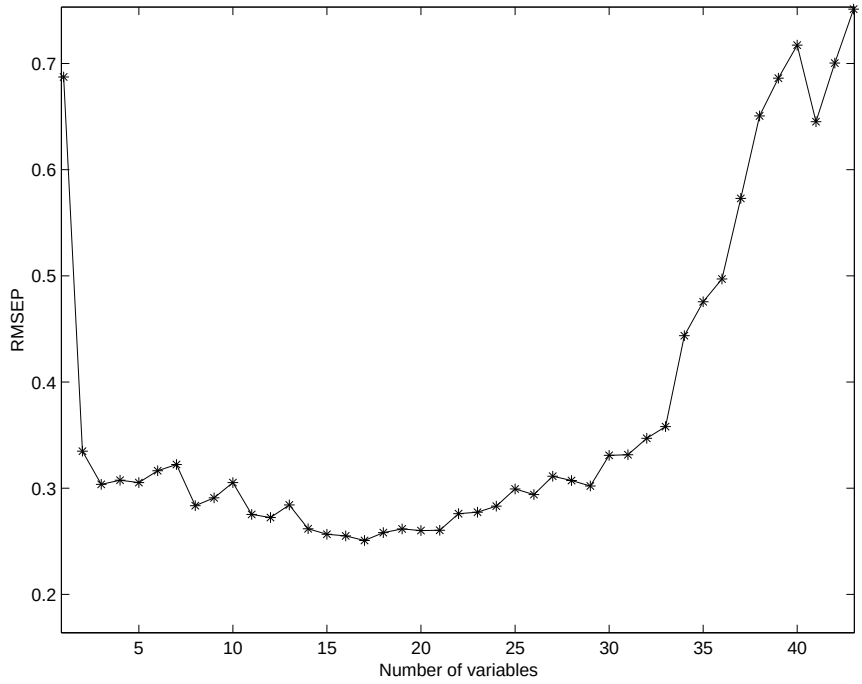

Figure 9.1: RMSEP as a function of the number of variables

Method	Regularization parameter	RMSEP
FSR	No. of variables = 17	0.25

Table 9.1: RMSEP-value for the forward selection method.

The optimal number of variables is 17 for the forward selection method. With just 17 out of 401 variables it is possible to obtain good predictions of octane. The resulting parameter estimate is seen in Figure 9.2.

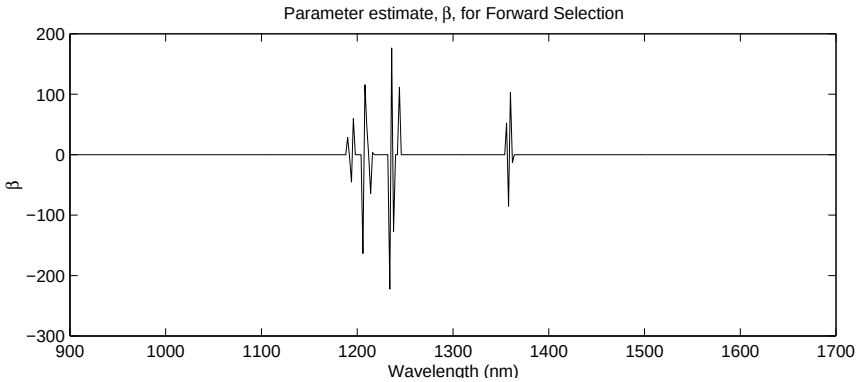

Figure 9.2: Parameter estimate, $\hat{\boldsymbol{\beta}}$, for the forward Selection method.

Chapter 10
LASSO Regression

10.1 Introduction

Least Absolute Shrinkage and Selection Operator, (LASSO), is a method developed by Robert Tibshirani in 1996, [57]. The LASSO minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. The constraints in the model allows for coefficients that are exactly zero[1], so it embodies the advantageous features of both Ridge Regression and Subset Selection.

10.2 Definition

The LASSO estimate, $\hat{\boldsymbol{\beta}}$, is defined as a constrained optimization problem by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2} \left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right] \quad \text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t \qquad (10.1)$$

[1] For a good geometrical interpretation of this see [64]

where $t \geq 0$ is a hyper-parameter. A closely related optimization problem is

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left[\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right] \tag{10.2}$$

where λ is the Lagrange parameter. This problem is the same as (10.1) because for a given λ, $0 \leq \lambda < \infty$, there exists a $t \geq 0$ such that they both share the same solution. LASSO limits the length of the p parameters, $\boldsymbol{\beta}$, with $\sum_{i=1}^{p} |\beta_j| \leq t$, where t is some constant. t controls the amount of shrinkage that is applied to the estimates. Let $\tilde{\boldsymbol{\beta}}$ be the full least squares estimate and let $t_0 = \sum |\tilde{\beta}_j|$. Values of $t < t_0$ will cause shrinkage of the solutions towards zero, and as mentioned earlier some may be exactly zero, see [57] p. 271.

The optimization problem (10.1), that consists in finding a vector that optimizes (i.e. minimizes or maximizes) a linear objective function subject to a finite set of linear constraints is easily stated, solving it numerically though is no trivial exercise. There are 2^p inequality constraints, corresponding to the 2^p different possible signs for the $\beta_j s$. When p is large it is not practical possible to solve this problem by direct application, [57] p. 268. The algorithm proposed by Tibshirani[2] is adequate for moderate values of n but is not the most efficient possible. This effect is greatly magnified when a technique like cross-validation is used to select an appropriate value of t. Moreover, Tibshiranis algorithm is not usable at all when applied to a problem where $p > n$, see [49] and [50].

10.3 Convex duality and the LASSO

In [49] and [50] the problem (10.1) is treated as a convex programming problem and the dual optimization problem is derived.

Problem (10.1) can be written as

$$\text{minimize}_{\boldsymbol{\beta}} \quad f(\boldsymbol{\beta}) \tag{10.3}$$

subject to

$$g(\boldsymbol{\beta}) \geq 0 \tag{10.4}$$

[2] Available in Fortran with an S-plus interface from STATLIB at http://www.stat.unipq.it/pub/stat/stalib/S/lasso

where

$$f(\boldsymbol{\beta}) = \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \frac{1}{2}\mathbf{r}^T\mathbf{r} \tag{10.5}$$

and

$$g(\boldsymbol{\beta}) = t - \sum_{j=1}^{p} |\beta_j| \tag{10.6}$$

Here $\mathbf{r} = \mathbf{r}(\boldsymbol{\beta})$ is the vector of residuals corresponding to $\boldsymbol{\beta}$ and $g(\boldsymbol{\beta})$ is implicitly a function of t, which is treated as fixed in the following. f is continous and the region of feasible β vectors is compact so a solution to (10.3), (10.4) is guaranteed to exist, [49] p. 322.

Treating (10.3), (10.4) as a convex programming problem the Lagrangian is, (see [14] p. 198)

$$\mathcal{L}(\boldsymbol{\beta}, \lambda) = f(\boldsymbol{\beta}) - \lambda g(\boldsymbol{\beta}) \tag{10.7}$$

If

$$\mathcal{L}^*(\boldsymbol{\beta}) = \sup_{\lambda \geq 0} \mathcal{L}(\boldsymbol{\beta}, \lambda) \tag{10.8}$$

is defined, then

$$\mathcal{L}^*(\boldsymbol{\beta}) = \begin{cases} f(\boldsymbol{\beta}) & \text{if} \quad g(\boldsymbol{\beta}) \geq 0, \\ \infty & \text{if} \quad g(\boldsymbol{\beta}) < 0. \end{cases} \tag{10.9}$$

Hence minimizing $\mathcal{L}^*(\boldsymbol{\beta})$ is equivalent to solving (10.3), (10.4). $\mathcal{L}^*(\boldsymbol{\beta})$ is called the primal problem and $f(\boldsymbol{\beta})$ is called the primal objective function. For $\lambda \geq 0$ the dual objective objective function is defined to be

$$\mathcal{L}_*(\lambda) = \inf_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}, \lambda), \tag{10.10}$$

and the dual problem is

$$\underset{\lambda \geq 0}{\text{maximize}}\ \mathcal{L}_*(\lambda), \tag{10.11}$$

If $\lambda \geq 0$ is fixed, then $\mathcal{L}(\boldsymbol{\beta}, \lambda)$ is a convex function in $\boldsymbol{\beta}$ and $\mathcal{L}(\boldsymbol{\beta}, \lambda) \to \infty$ as $\|\boldsymbol{\beta}\|_1 \to \infty$. Hence $\mathcal{L}(\cdot, \lambda)$ has at least one minimum and $\bar{\boldsymbol{\beta}}$ minimizes $\mathcal{L}(\boldsymbol{\beta}, \lambda)$ if and only if the p-dimensional null-vector is an element of [3]

$$\partial_{\boldsymbol{\beta}} \mathcal{L}(\bar{\boldsymbol{\beta}}, \lambda) = -\mathbf{X}^T\mathbf{r} + \lambda\mathbf{v}, \tag{10.12}$$

[3] (10.12) is in the optimization litterature known as the Kuhn-Tucker conditions (1951), see e.g. [38] p. 19 or [14] p. 200

Here $\mathbf{v} = (v_1, \ldots, v_p)^T$ is of the following form:

$$v_i \in [-1, 1] \quad \text{if} \quad \beta_i = 0 \quad \text{or} \quad v_i = \begin{cases} 1 & \text{if} \quad \beta_i > 0, \\ -1 & \text{if} \quad \beta_i < 0 \end{cases}$$

Thus, if $\bar{\boldsymbol{\beta}}$ minimizes $\mathcal{L}(\boldsymbol{\beta}, \lambda)$ for a given value of λ, then

$$\mathbf{0} = -\mathbf{X}^T \bar{\mathbf{r}} + \lambda \mathbf{v}, \tag{10.13}$$

for some $\mathbf{v}$ of the form described above and $\bar{\mathbf{r}} = \mathbf{r}(\bar{\boldsymbol{\beta}}) = \mathbf{y} - \mathbf{X}\bar{\boldsymbol{\beta}}$. The form of $\mathbf{v}$ implies that $\mathbf{v}^T \bar{\boldsymbol{\beta}} = \|\bar{\boldsymbol{\beta}}\|_1$ and thus it follows from (10.13) that if $\bar{\boldsymbol{\beta}}$ minimizes $\mathcal{L}(\boldsymbol{\beta}, \lambda)$, then $\lambda = \bar{\mathbf{r}}^T \mathbf{X}\bar{\boldsymbol{\beta}} / \|\bar{\boldsymbol{\beta}}\|_1$. Alternatively if $\bar{\boldsymbol{\beta}} \neq \mathbf{0}$, then $\|\mathbf{v}\|_\infty = 1$ and from (10.13) it shows that $\lambda = \|\mathbf{X}^T \bar{\mathbf{r}}\|_\infty$. These two expressions for λ will be used below to derive a new expression for the dual function.

$$\begin{aligned} \mathcal{L}_*(\lambda) = \mathcal{L}(\bar{\boldsymbol{\beta}}, \lambda) &= \frac{1}{2}\bar{\mathbf{r}}^T \bar{\mathbf{r}} - \frac{\bar{\mathbf{r}}^T \mathbf{X}\bar{\boldsymbol{\beta}}}{\|\bar{\boldsymbol{\beta}}\|_1}(t - \|\bar{\boldsymbol{\beta}}\|_1) \\ &= \frac{1}{2}\bar{\mathbf{r}}^T \bar{\mathbf{r}} + \bar{\mathbf{r}}^T \mathbf{X}\bar{\boldsymbol{\beta}} - t\frac{\bar{\mathbf{r}}^T \mathbf{X}\bar{\boldsymbol{\beta}}}{\|\bar{\boldsymbol{\beta}}\|_1} \\ &= \frac{1}{2}\mathbf{y}^T \mathbf{y} - \frac{1}{2}\bar{\boldsymbol{\beta}}^T (\mathbf{X}^T \mathbf{X})\bar{\boldsymbol{\beta}} - t\frac{\bar{\mathbf{r}}^T \mathbf{X}\bar{\boldsymbol{\beta}}}{\|\bar{\boldsymbol{\beta}}\|_1} \\ &= \frac{1}{2}\mathbf{y}^T \mathbf{y} - \frac{1}{2}\bar{\boldsymbol{\beta}}^T (\mathbf{X}^T \mathbf{X})\bar{\boldsymbol{\beta}} - t\|\mathbf{X}^T \bar{\mathbf{r}}\|_\infty \end{aligned}$$

and if the following is defined

$$\begin{aligned} \tilde{h}(\boldsymbol{\beta}) &= \frac{1}{2}\mathbf{y}^T \mathbf{y} - \frac{1}{2}\boldsymbol{\beta}^T (\mathbf{X}^T \mathbf{X})\boldsymbol{\beta} - t\frac{\mathbf{r}^T \mathbf{X}\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_1} \\ \bar{h}(\boldsymbol{\beta}) &= \frac{1}{2}\mathbf{y}^T \mathbf{y} - \frac{1}{2}\boldsymbol{\beta}^T (\mathbf{X}^T \mathbf{X})\boldsymbol{\beta} - t\|\mathbf{X}^T \mathbf{r}\|_\infty \end{aligned}$$

then the dual function can be written as

$$\mathcal{L}_*(\lambda) = \mathcal{L}(\bar{\boldsymbol{\beta}}, \lambda) = \bar{h}(\bar{\boldsymbol{\beta}}) = \tilde{h}(\bar{\boldsymbol{\beta}}) \quad \text{for any} \quad \bar{\boldsymbol{\beta}} \quad \text{for which} \quad \mathbf{0} \in \partial_{\boldsymbol{\beta}} \mathcal{L}(\bar{\boldsymbol{\beta}}, \lambda)$$

So $\bar{\boldsymbol{\beta}}$ is a solution to (10.3), (10.4) if and only if it satisfies (10.13), see [49] p. 323. For the equivalent Ridge regression solution $\lambda\bar{\boldsymbol{\beta}}$ replaces $\lambda\mathbf{v}$ in (10.13).

10.4 LASSO as Bayes estimate

The tendency for the LASSO method to produce estimates that are either zero or large can be reflected by deriving the LASSO estimate as the Bayes posterior mode under independent double-exponential priors for the β_js.

$$f(\beta_j) = \frac{1}{2\tau} exp\left(\frac{-|\beta_j|}{\tau}\right) \tag{10.14}$$

Here $\tau = 1/\lambda$, see [11] p. 88. The double-exponential density puts more mass near 0 and in the tails. This reflects the greater tendency of the LASSO to produce estimates that are either large or 0, see [57] p. 277.

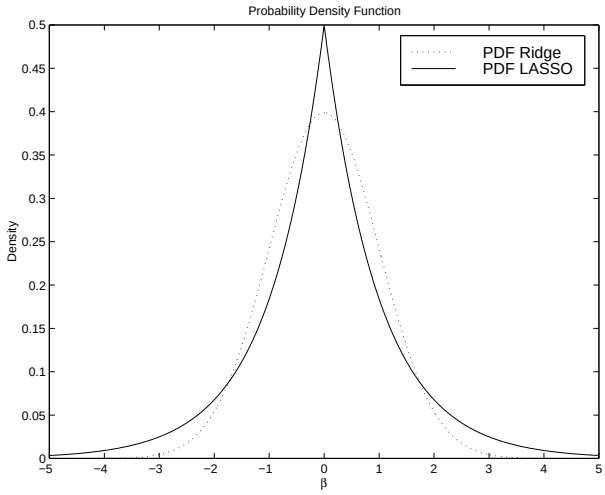

Figure 10.1: The normal density $(\cdots)$ is the prior distribution of β_j used by the Ridge method. The double-exponential density (-) is the prior distribution of β_j used by the LASSO method.

10.5 Algorithm

The algorithm based on the theory above can be sketched as follows: [4]

[4]See [49] and [50] for the detailed description. A C-implementation with an S-plus interface can be found on http://www.stat.unipq.it/pub/stat/stalib/S/lasso2

(1) Setup

- Create $\alpha = \{i : \beta_i \neq 0\}$, which is an index vector indicating which of the coefficients are non-zero.
- $\mathbf{P}$ is a permutation matrix that collects the non-zero components of $\boldsymbol{\beta}$ in the first $|\alpha|$ positions. This is equivalent to moving all the non-zero coefficients in $\boldsymbol{\beta}$ to the start of the $\boldsymbol{\beta}$ vector. In mathematical notation, $\boldsymbol{\beta} = \mathbf{P}^T \binom{\boldsymbol{\beta}_\alpha}{\mathbf{0}}$.
- Let $\boldsymbol{\theta}_\alpha = \text{sign}(\boldsymbol{\beta}_\alpha)$ have entry 1 if the corresponding entry in $\boldsymbol{\beta}_\alpha$ is positive and -1 otherwise.

(2) Main step

To obtain the next iterate from the current $\boldsymbol{\beta}$, solve

$$\underset{\mathbf{h}}{\text{minimize}} f(\boldsymbol{\beta} + \mathbf{h}) \tag{10.15}$$

$$\text{subject to} \quad \boldsymbol{\theta}_\alpha^T(\boldsymbol{\beta}_\alpha + \mathbf{h}_\alpha) \leq t \quad \text{and} \quad \mathbf{h} = \mathbf{P}^T \begin{pmatrix} \mathbf{h}_\alpha \\ \mathbf{0} \end{pmatrix} \tag{10.16}$$

Let the solution to this optimization problem be $\bar{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{h}$, then check $\bar{\boldsymbol{\beta}}$ for sign feasibility. If $\text{sign}(\bar{\boldsymbol{\beta}}_\alpha) = \boldsymbol{\theta}_\alpha$ then $\bar{\boldsymbol{\beta}}$ is sign feasible, which implies that $\|\bar{\boldsymbol{\beta}}\|_1 = \bar{\boldsymbol{\beta}}_\alpha^T \boldsymbol{\theta}_\alpha$ and hence by (10.15) and (10.16), $\bar{\boldsymbol{\beta}}_\alpha$ satisfies the constraint. If $\bar{\boldsymbol{\beta}}$ is not sign feasible, then it is not guaranteed that $\|\bar{\boldsymbol{\beta}}_\alpha\|_1 \leq t$, so a modification is needed.

(3) Iterations

- IF $\bar{\boldsymbol{\beta}}$ is not sign feasible then
  1. Find the smallest γ, $0 < \gamma < 1$ for which $\exists$ some $k \in \alpha$ such that $0 = \beta_k + \gamma h_k$. This moves to the first new zero component in descent direction $\mathbf{h}$.
  2. Update α by deleting k, and set $\boldsymbol{\beta} = \boldsymbol{\beta} + \gamma \mathbf{h}$. Reset $\boldsymbol{\beta}_\alpha$ and $\boldsymbol{\theta}_\alpha$, which will both now be feasible in terms of the constraint and solve the main step for a new $\mathbf{h}$.
  3. Iterate until $\bar{\boldsymbol{\beta}}$ is sign feasible.
- ELSE $\bar{\boldsymbol{\beta}}$ is sign feasible
  1. Calculate

$$\bar{\mathbf{v}} = \frac{\mathbf{X}^T \bar{\mathbf{r}}}{\|\mathbf{X}^T \bar{\mathbf{r}}\|_\infty} = \mathbf{P}^T \begin{pmatrix} \bar{\mathbf{v}}_1 \\ \bar{\mathbf{v}}_2 \end{pmatrix}$$

$$\text{where} \quad \bar{\mathbf{r}} = \mathbf{r}(\bar{\boldsymbol{\beta}}) = \mathbf{y} - \mathbf{X}\bar{\boldsymbol{\beta}}$$

2. -IF $|(\bar{\mathbf{v}}_1)_i| = |\theta_i|$ for $i \in \alpha$ and $-1 \leq (\bar{\mathbf{v}}_2)_i \leq 1$ for $i \notin \alpha$ then $\bar{\boldsymbol{\beta}}$ is a solution.
STOP.

-ELSE
(a) Find s such that $(\bar{\mathbf{v}}_2)_s$ has maximal value.
(b) Update α by adding s to it. Update $\boldsymbol{\beta}_\alpha$ by appending 0 as the last element. Append $\text{sign}(\bar{\mathbf{v}}_2)_s$ to $\boldsymbol{\theta}_\alpha$.
(c) Solve the main step and iterate.

10.6 LASSO applied to gasoline example

For the LASSO method the optimal model was found by iterating through values of t chosen on an equally spaced grid on the logarithmic scale, see Figure 10.2.

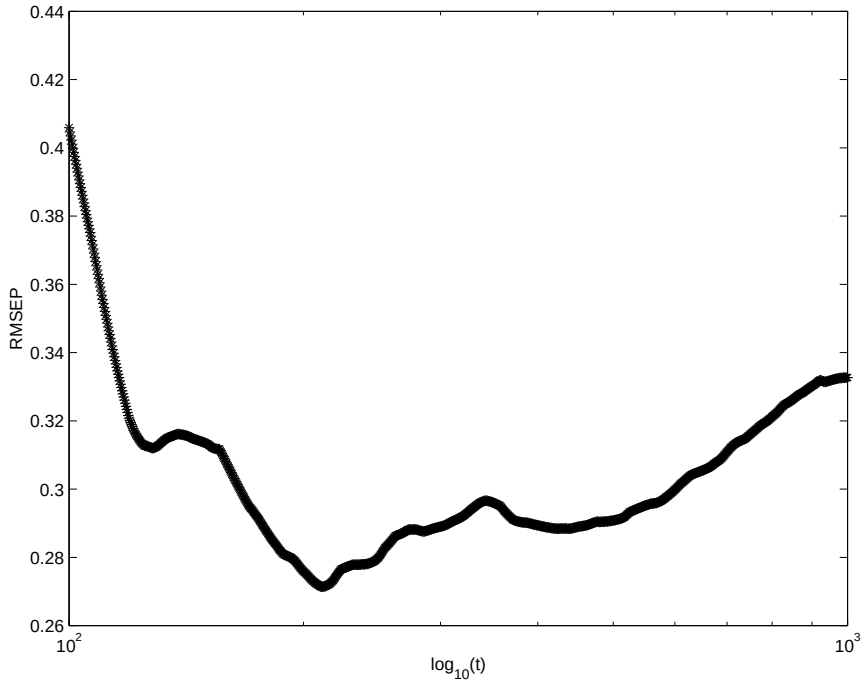

Figure 10.2: RMSEP as a function of $\log_{10}(t)$

Method	Regularization parameter	RMSEP
MLLS	$\sum(\lvert\beta_i\rvert)$=3783.4; $\lambda = 2.7e-15$	0.70
LASSO	$\sum(\lvert\beta_i\rvert)$=210.5; $\lambda = 3.7e-03$	0.27

Table 10.1: RMSEP-values for the MLLS and the LASSO method.

Like the previous method, the forward selection method, LASSO produces an estimate where most of the parameters are exactly equal to zero, see Figure 10.3. The forward selection method and LASSO select variables within approximately the same range of the spectrum, except for the very last part of the spectrum which is only included in the LASSO estimate. The forward selection method produces a slightly better RMSEP-value than LASSO, which might indicate that the last part of the spectrum contains no further information regarding the prediction of octane. That the Ridge method does well here could hint at the fact that information about the octane number probably is spread out over the whole spectrum, since Ridge regression implicitly assumes that the parameters are normally distributed.[5] However, variable selection methods tend to work best in situations characterized by true parameter vectors with components consisting of a very few (relatively) large (absolute) values[6].

[5]This interpretation has been used in [64]
[6]Frank and Friedman, [15] p. 110

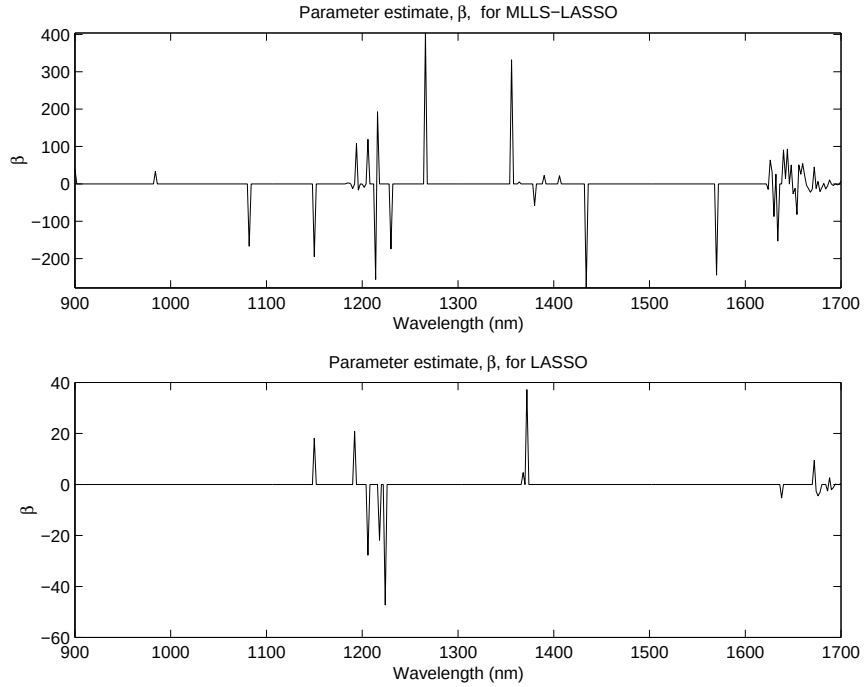

Figure 10.3: Parameter estimates for the MLLS solution for the LASSO (Top figure), and LASSO (Bottom figure)

Chapter 11
Adaptive Ridge Regression

11.1 Introduction

Adaptive Ridge Regression, (ARR), is a special form of Ridge regression, balancing the quadratic penalization on each parameter of the model. ARR is equivalent to LASSO in the sense that both procedures produce the same estimate.

11.2 Theory

Adaptive Ridge is a modification of the Ridge estimate, (5.1), which is defined by the quadratic constraint $\sum_{j=1}^{p} \beta_j^2 \leq t$. As mentioned in 5.2 it is the solution to

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2 \right] \tag{11.1}$$

where λ is the Lagrange multiplier varying with the bound t on the norm of the parameters. As mentioned in 5.3 the Bayes prior distribution for the Ridge estimate is a normal distribution with variance proportional to $1/\lambda$.

If all covariates are not equally relevant this is not appropriate. In [21] the following modification to (11.1) is proposed

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}}) = \arg\min_{(\boldsymbol{\beta},\boldsymbol{\lambda})} \left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \sum_{j=1}^{p} \lambda_j \beta_j^2\right] \tag{11.2}$$

Here, each coefficient has its own prior distribution. The priors are normal distributions with variances proportional to $1/\lambda_j$. To avoid simultaneous estimation of these p hyper-parameters by trial, the constraint

$$\frac{1}{p}\sum_{j=1}^{p}\frac{1}{\lambda_j} = \frac{1}{\lambda} \quad , \lambda_j > 0 \tag{11.3}$$

is applied on $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^T$, where λ is a predefined value. This constraint is a link between the p prior distributions. Their mean variance is proportional to $1/\lambda$. The adaptivity stems from the fact that values of λ_j are automatically induced from the data. Adaptivity refers here to penalization balance on each coefficient, $\hat{\beta}_j$, not to the tuning of λ.

11.3 The equivalence to LASSO

ARR and LASSO are equivalent, in the sense that they yield the same estimate. To see this the ARR estimate is defined by another parametrization. (11.2) and (11.3) in their present form may lead to divergent solutions for $\boldsymbol{\lambda}$, ($\lambda_j \to \infty$). Thus new variables are defined

$$\gamma_j = \sqrt{\frac{\lambda_j}{\lambda}}\beta_j \ , \quad \text{and} \quad c_j = \sqrt{\frac{\lambda}{\lambda_j}} \quad \text{for} \quad j = 1, \ldots, p \tag{11.4}$$

Then optimization problem (11.2) with constraint (11.3) can be written as

$$\begin{cases} (\hat{\mathbf{c}}, \hat{\boldsymbol{\gamma}}) = \arg\min_{(\mathbf{c},\boldsymbol{\gamma})} \left[(\mathbf{y} - \mathbf{X}\mathbf{c}\boldsymbol{\gamma})^T(\mathbf{y} - \mathbf{X}\mathbf{c}\boldsymbol{\gamma}) + \lambda \sum_{j=1}^{p} \gamma_j^2\right] \\ \text{subject to} \sum_{j=1}^{p} c_j^2 = p \quad , \quad c_j \geq 0 \end{cases} \tag{11.5}$$

with the following optimality conditions

$$\forall j, \begin{cases} \sum_{i=1}^{n} x_{ij}\big(\sum_{k=1}^{p} \hat{\beta}_k x_{ik} - y_i\big) + \frac{\lambda}{p}\text{sign}(\hat{\beta}_j)\sum_{k=1}^{p}|\hat{\beta}_k| = 0 \\ \text{or } \hat{\beta}_j = 0 \end{cases} \tag{11.6}$$

The optimality conditions are the normal equations of the problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \frac{\lambda}{p} (\sum_{k=1}^{p} |\beta_k|)^2 \right] \tag{11.7}$$

The estimate (11.7) is equivalent to the LASSO estimate, (10.2). The ARR estimate is thus the LASSO estimate. The only difference in their definition is that ARR uses the constraint $(\sum_{k=1}^{p} |\beta_k|)^2/p \leq t^2$ instead of $\sum_{k=1}^{p} |\beta_k| \leq t$.

Proof of equivalence:[1]

The corresponding Lagrangian, L, of (11.5) is

$$L(\mathbf{c}, \boldsymbol{\gamma}) = \left[(\mathbf{y} - \mathbf{X}\mathbf{c}\boldsymbol{\gamma})^T (\mathbf{y} - \mathbf{X}\mathbf{c}\boldsymbol{\gamma}) + \lambda \sum_{j=1}^{p} \gamma_j^2 \right] + \nu \left(\sum_{j=1}^{p} c_j^2 - p \right) - \xi^T \mathbf{c} \tag{11.8}$$

which for notational reasons will be written as

$$L(\mathbf{c}, \boldsymbol{\gamma}) = C_{emp}(\mathbf{c}, \boldsymbol{\gamma}) + \lambda \sum_{j=1}^{p} \gamma_j^2 + \nu \left(\sum_{j=1}^{p} c_j^2 - p \right) - \xi^T \mathbf{c} \tag{11.9}$$

where ν and ξ are the Lagrange multipliers corresponding respectively to the equality and the positivity constraints on $\{c_j\}$ from (11.5). The normal equations to (11.9) are thus

$$\begin{cases} \dfrac{\partial L}{\partial \gamma} = \dfrac{\partial C_{emp}(\mathbf{c}, \boldsymbol{\gamma})}{\partial \gamma} + 2\lambda\gamma \\ \dfrac{\partial L}{\partial \mathbf{c}} = \dfrac{\partial C_{emp}(\mathbf{c}, \boldsymbol{\gamma})}{\partial \mathbf{c}} + 2\nu\mathbf{c} - \xi \end{cases} \tag{11.10}$$

From the relation $\boldsymbol{\beta} = \mathrm{diag}(\mathbf{c})\boldsymbol{\gamma}$, a relation between the partial derivatives of C_{emp} with respect to $\mathbf{c}$ and $\boldsymbol{\gamma}$ is stated

$$\begin{cases} \dfrac{\partial C_{emp}}{\partial \boldsymbol{\gamma}} = \mathrm{diag}(\mathbf{c}) \dfrac{\partial C_{emp}}{\partial \beta} \\ \dfrac{\partial C_{emp}}{\partial \mathbf{c}} = \mathrm{diag}(\boldsymbol{\gamma}) \dfrac{\partial C_{emp}}{\partial \beta} \end{cases} \tag{11.11}$$

[1] The proof of equivalence is unpublished material, but is available at www.hds.utc.fr/~grandval.

From this system the following equation is formed

$$\text{diag}(\boldsymbol{\gamma})\frac{\partial C_{emp}(\mathbf{c},\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = \text{diag}(\mathbf{c})\frac{\partial C_{emp}(\mathbf{c},\boldsymbol{\gamma})}{\partial \mathbf{c}} \tag{11.12}$$

This equation is used to derive a relationship between $\hat{c}_j$ and $\hat{\gamma}_j$, independently of C_{emp} and the Lagrange multipliers:

$$\begin{cases} \text{diag}(\hat{\boldsymbol{\gamma}})\dfrac{\partial L}{\partial \boldsymbol{\gamma}} = \text{diag}(\hat{\boldsymbol{\gamma}})\dfrac{\partial C_{emp}(\hat{\mathbf{c}},\hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\gamma}} + 2\lambda\text{diag}(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\gamma}} \\ \text{diag}(\hat{\mathbf{c}})\dfrac{\partial L}{\partial \mathbf{c}} = \text{diag}(\hat{\mathbf{c}})\dfrac{\partial C_{emp}(\hat{\mathbf{c}},\hat{\boldsymbol{\gamma}})}{\partial \mathbf{c}} + 2\nu\text{diag}(\hat{\mathbf{c}})\hat{\mathbf{c}} - \text{diag}(\hat{\mathbf{c}})\boldsymbol{\xi} \end{cases} \tag{11.13}$$

A Lagrange multiplier is zero for inactive constraints, therefore $\text{diag}(\hat{\mathbf{c}})\xi = 0$. As (11.12) holds for $(\hat{\mathbf{c}},\hat{\boldsymbol{\gamma}})$, and optimality of $(\hat{\mathbf{c}},\hat{\boldsymbol{\gamma}})$ implies $\frac{\partial L}{\partial \boldsymbol{\gamma}} = \frac{\partial L}{\partial \mathbf{c}} = 0$, then, from (11.13) it follows that

$$\forall j\ \hat{c}_j^2 = \frac{\lambda}{\nu}\hat{\gamma}_j^2 \tag{11.14}$$

The equality constraint in (11.5) on $\{c_j\}$ implies:

$$\forall j\ \hat{c}_j = \frac{\sqrt{p}|\hat{\gamma}_j|}{\sqrt{\sum_{k=1}^{p}\hat{\gamma}_k^2}} \tag{11.15}$$

This equation is used to give the optimality conditions as a function of the original variables $\hat{\beta}_j$. As $|\hat{\beta}_j| = \hat{c}_j|\hat{\gamma}_j|$, it follows that

$$|\hat{\beta}_j| = \frac{\sqrt{p}\hat{\gamma}_j^2}{\sqrt{\sum_{k=1}^{p}\hat{\gamma}_k^2}} \Rightarrow \frac{|\hat{\beta}_j|}{\sum_{k=1}^{p}|\hat{\beta}_k|} = \frac{\hat{\gamma}_j^2}{\sum_{k=1}^{p}\hat{\gamma}_k^2} \Leftrightarrow \hat{c}_j^2 = \frac{d|\hat{\beta}_j|}{\sum_{k=1}^{p}|\hat{\beta}_k|} \tag{11.16}$$

This value of $\hat{c}_j$ is now plugged into the first equation of system (11.10) evaluated at $(\hat{\mathbf{c}},\hat{\boldsymbol{\gamma}})$, using the first equation of system (11.11):

$$\forall j\ \hat{c}_j\frac{\partial C_{emp}}{\partial \beta_j}(\hat{\beta}_j) + 2\lambda\hat{\gamma}_j = 0 \tag{11.17}$$

Therefore, either $\hat{c}_j = \hat{\gamma}_j = \hat{\beta}_j = 0$ or $\dfrac{\partial C_{emp}}{\partial \beta_j}(\hat{\beta}_j) + 2\lambda\dfrac{\hat{\gamma}_j}{\hat{c}_j} = 0$. From

(11.16), $\hat{\gamma}_j/\hat{c}_j$ can be written using $\boldsymbol{\beta}$ as follows:

$$\begin{aligned}\frac{\hat{\gamma}_j}{\hat{c}_j} &= \hat{\gamma}_j \hat{c}_j \frac{1}{\hat{c}_j^2} \\ &= \hat{\beta}_j \frac{\sum_{k=1}^p |\hat{\beta}_k|}{p|\hat{\beta}_j|} \\ &= \frac{1}{p}\text{sign}(\hat{\beta}_j) \sum_{k=1}^p |\hat{\beta}_k|. \end{aligned} \tag{11.18}$$

The optimality conditions are thus

$$\forall j, \quad \begin{cases} \dfrac{\partial C_{emp}}{\partial \beta_j}(\hat{\beta}_j) + 2\dfrac{\lambda}{p}\text{sign}(\hat{\beta}_j) \sum_{k=1}^p |\hat{\beta}_k| = 0 \\ \\ \text{or} \quad \beta_j = 0, \end{cases} \tag{11.19}$$

which are recognized as the normal equation of

$$C_{emp}(\boldsymbol{\beta}) + \frac{\lambda}{p}\left(\sum_{k=1}^p |\hat{\beta}_k|\right)^2 = 0 \tag{11.20}$$

for any empirical cost C_{emp}. □

11.4 Adaptive Ridge Regression applied to gasoline example

By iterating through values of λ on the logarithmic scale, the following results are obtained:

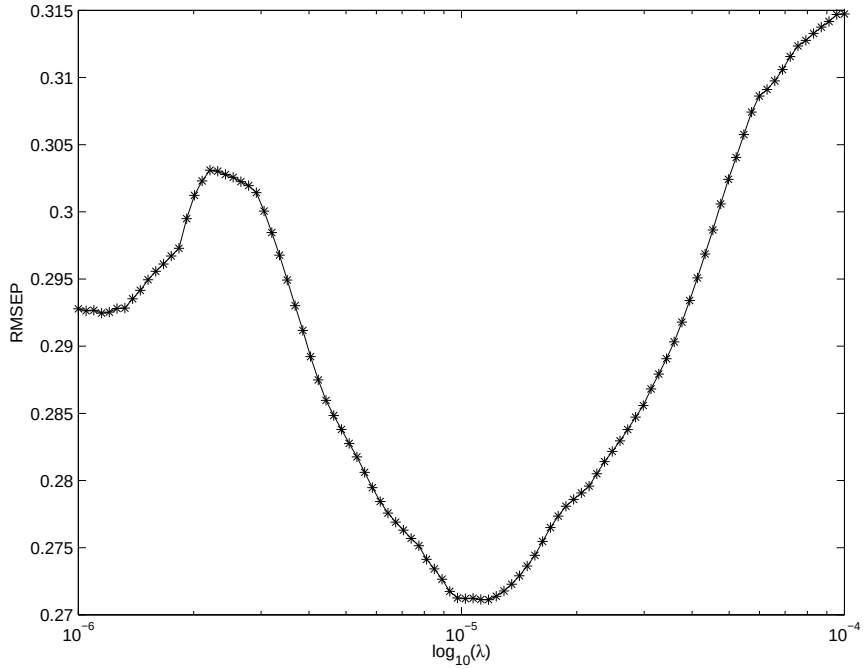

Figure 11.1: RMSEP as a function of $\log_{10}(\lambda)$

Method	Regularization parameter	RMSEP
Adaptive Ridge	$\sum(\lvert\beta_i\rvert)$=222.5; $\lambda = 1.2e-5$	0.27

Table 11.1: RMSEP-values for the Adaptive Ridge method.

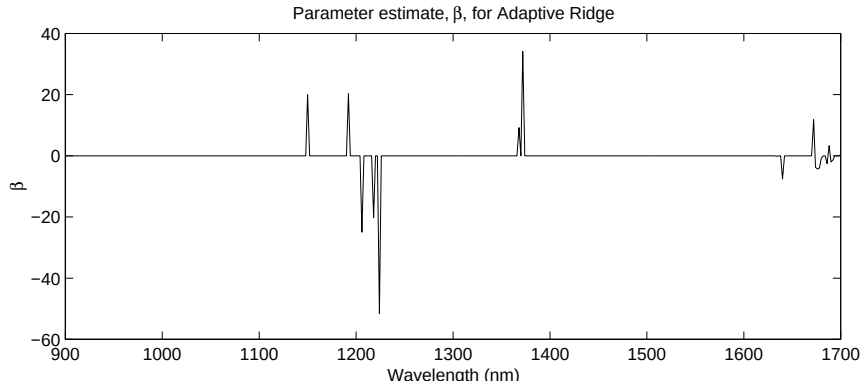

Figure 11.2: Parameter estimates for the Adaptive Ridge solution.

Due to the discretization of λ, the ARR result, $(\sum(|\beta_i|))$, is not exactly identical to the LASSO result.

The publicly available algorithms to produce LASSO solutions when dealing with a singular design matrix is at the moment, (at least to my knowledge), restricted to the following:

1. *lasso2* (http://www.stat.unipq.it/pub/stat/stalib/S/lasso2)
2. *arrfit* (http://www.hds.utc.fr/~grandval/arrfit.m)[2]
3. *lasso* (http://www.imm.dtu.dk/~hoe/files/lasso.m)[3]

They are all based on different theory but produce exactly the same solution. *lasso2* is considerably faster than *arrfit* and *lasso*, but has the disadvantage of a more complicated installation procedure compared to the other two algorithms that can be used directly as any other standard Matlab function.

[2]This algorithm provides the Adaptive Ridge solution.
[3]This algorithm has been used in [64].

Chapter 12
Basis-Function Regression

12.1 Introduction

As a continuous-wavelength alternative the linear combination of the spectral values can be replaced with an integral over the range of the wavelengths of an unknown coefficient-function multiplied by the spectral measurements. The unknown function can then be approximated by a linear combination of some basis functions (e.g. B-splines). The problem then becomes a linear regression problem where the number of regressors depend on the number of basis functions and not the number of wavelengths.

The approach was first suggested by Hastie and Mallows, [22], who focused on smoothing splines for estimation of the coefficient-function. Similarly Goutis, [20], used smoothing splines to estimate a coefficient-function in the case where the predictive information is related to the second derivative of the spectrum. Marx and Eilers, [41], project the spectral measurements onto a moderate number of equally spaced B-spline bases. This approach is very similar to the approach presented here. However, the difference being that (i) the underlying model is formulated using an integral over the wavelengths, and (ii) the number of basis functions is not restricted to be less than the number of observations. For near-continuous measurements (i) is largely a technicality which allows, in a simple way, to study what happens if the predictive ability is related to derivatives of the spectra

rather than the actual spectra.

12.2 Model

The spectra are measured at a number of wavelengths $\lambda_j;\ j = 1, \ldots, p$. The measurements of the characteristic quantity is called $y_i;\ i = 1, \ldots, n$ and the measured spectrum corresponding to y_i is called $a_i(\lambda_j);\ j = 1, \ldots, p$. The model, (3.1), is here presented in a slightly altered version.

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j a_i(\lambda_j) + e_i;\ i = 1, \ldots, n \tag{12.1}$$

where $e_i;\ i = 1, \ldots, n$ are the model errors which are assumed to be independently identical distributed (iid.) random variables, and $\beta_j;\ j = 0, \ldots, p$ are some coefficients which must be determined from data. Model (12.1) is a linear regression model. However, as measurement equipment get more advanced the spectra are measured at an increasing number p of wavelengths, so that each spectrum often can be considered known for every wavelength $\lambda \in [\underline{\lambda}, \overline{\lambda}]$. Therefore, the number of regressors p is often magnitudes larger than the number of observations n. Conceptually, it could be more convenient to use a model which explicitly regard the spectra as functions $a_i(\lambda);\ i = 1, \ldots, n$ of the bandwidth λ. As a generalization of (12.1) it is convenient to replace the summation with an integral over the interval of wavelengths, i.e. to use the model

$$y_i = \beta_0 + \int_{\underline{\lambda}}^{\overline{\lambda}} \beta(\lambda) a_i(\lambda) d\lambda + e_i, \tag{12.2}$$

where the coefficient β_0 and the *function* $\beta(\cdot)$ must be determined from data, c.f. Section 12.3. It is interesting to note that if it is suspected that some predictive ability is related to the first- and second-order derivatives of the spectra rather than the spectra itself (12.2) can still be used if the range of wavelengths over which the spectra is measured is wide enough.

To see this consider the model

$$y_i = \beta_0 + \int_{\underline{\lambda}}^{\overline{\lambda}} \left(\phi_0(\lambda) a_i(\lambda) + \phi_1(\lambda) \frac{da_i}{d\lambda}(\lambda) + \phi_2(\lambda) \frac{d^2 a_i}{d\lambda^2}(\lambda) \right) d\lambda + e_i, \tag{12.3}$$

which take into account both the actual spectra and its first- and second order derivatives. Assuming that the derivatives exists, simple calculations (partial integration) show that (12.3) can be written

$$\begin{aligned} y_i &= \beta_0 \\ &+ \left(\phi_1(\overline{\lambda}) - \frac{d\phi_2}{d\lambda}(\overline{\lambda})\right) a_i(\overline{\lambda}) - \left(\phi_1(\underline{\lambda}) - \frac{d\phi_2}{d\lambda}(\underline{\lambda})\right) a_i(\underline{\lambda}) \\ &+ \phi_2(\overline{\lambda})\frac{da_i}{d\lambda}(\overline{\lambda}) - \phi_2(\underline{\lambda})\frac{da_i}{d\lambda}(\underline{\lambda}) \\ &+ \int_{\underline{\lambda}}^{\overline{\lambda}} \left(\phi_0(\lambda) - \frac{d\phi_1}{d\lambda}(\lambda) + \frac{d^2\phi_2}{d\lambda^2}(\lambda)\right) a_i(\lambda)d\lambda + e_i. \end{aligned} \tag{12.4}$$

Given that the range of the wavelengths is so large that all important wavelengths are covered then $\phi_1(\overline{\lambda}) = \phi_1(\underline{\lambda}) = \phi_2(\overline{\lambda}) = \phi_2(\underline{\lambda}) = 0$. In this case the second line in (12.4) vanish and the term inside the parenthesis in the integral is a function of λ which can be handled by $\beta(\lambda)$ in (12.2). If not all important wavelengths are covered it is necessary to extent (12.2) with regression terms containing $a_i(\overline{\lambda})$, $a_i(\underline{\lambda})$, $\frac{da_i}{d\lambda}(\overline{\lambda})$, and $\frac{da_i}{d\lambda}(\underline{\lambda})$ in order to take first- and second-order derivatives of $a_i(\lambda)$ into account.

12.3 Approximations

To be able to determine the scalar β_0 and the function $\beta(\cdot)$ in (12.2) from data the function is approximated by a linear combination of a set of basis functions, such as B-spline basis functions, natural spline basis functions, or wavelet basis functions [4].

$$\beta(\lambda) = \mathbf{B}^T(\lambda)\boldsymbol{\theta}, \tag{12.5}$$

where $\mathbf{B}(\lambda) = [b_1(\lambda) \dots b_m(\lambda)]^T$ are the basis functions and $\boldsymbol{\theta} = [\theta_1 \dots \theta_m]^T$ are some coefficients to be determined from data. With (12.2) and (12.5) simple calculations show that

$$y_i = \beta_0 + \sum_{k=1}^{m} \theta_k x_{ki} + e_i, \tag{12.6}$$

where

$$x_{ki} = \int_{\underline{\lambda}}^{\overline{\lambda}} b_k(\lambda)a_i(\lambda)d\lambda;\ k = 1, \dots, m;\ i = 1, \dots, n, \tag{12.7}$$

does not depend on $\boldsymbol{\theta}$ and can be determined from the measurements of the spectra at the wavelengths $\lambda_j;\ j = 1, \ldots, p$ by use of the trapezoid rule of integration.

$$x_{ki} = \frac{1}{2}\sum_{j=1}^{p-1}(\lambda_{j+1} - \lambda_j)\left[b_k(\lambda_j)a_i(\lambda_j) + b_k(\lambda_{j+1})a_i(\lambda_{j+1})\right] \tag{12.8}$$

It is seen that, handled this way, the calibration problem is not dependent on p as long as the spectra is measured at fine enough intervals to allow the integrals in (12.7) to be evaluated with reasonable precision. Furthermore, although $p > n$, the number of basis functions can often be chosen so that $m < n$, whereby (12.6) becomes an ordinary regression problem. One may choose to use $n < m < p$, in this case PCR, PLS, Ridge regression, LASSO, and other shrinkage methods may be applied.

As noted in [41], the application of models like (12.6) regularize estimation as compared to models like (12.1). However, depending on the spectra, the regressors in (12.6) may still be near-collinear. Figure 12.1 shows a cubic B-spline basis with six equally spaced knots covering the interval 900 to 1700 nm, this results in $m = 8$. It is seen that the basis-functions are non-zero only for wavelengths around their maximum, this is the key feature by which (12.5) becomes a good approximation. However, if $a_i(\lambda)$ is constant across i for some wavelengths then the nature of the basis-functions may result in collinearity of the regressors (12.7). It is therefore suggested that instead of using model (12.6) directly the regressors are replaced by their principal components. If variable selection techniques are then applied to the principal components both problems where $m < n$ and $m \geq n$ can be handled.

The type of basis functions used influence the type of functions which can be approximated by (12.5). A B-spline basis of order n result in $\beta(\cdot)$ having continuous derivatives up to order n, i.e. a cubic B-spline basis is of order 2. This also holds for a natural spline basis, but here $\beta(\cdot)$ has the additional property that it is linear outside $[\underline{\lambda}, \overline{\lambda}]$. Opposed to this a wavelet basis can be used to approximate a function with sharp peaks.

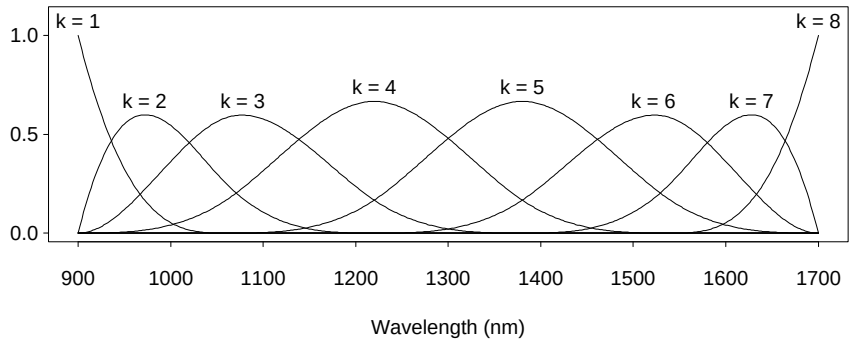

Figure 12.1: Cubic B-spline basis with six equally spaced knots (four internal) covering the interval 900 to 1700 nm.

12.4 Basis Function Regression applied to gasoline example

The results from Ridge, PCR, PLS, LASSO and FSR are summarized in Table 12.1. PLS, PCR and Ridge produce the best results. Figure 12.2 shows the parameter estimates plotted against their corresponding wavelengths. The estimates are obtained using the tuning-parameters listed in Table 12.1 together with the full data set.

Method	Regularization parameter	RMSEP
Ridge	k = 0.002	0.24
PCR	No. of components = 13	0.23
PLS	No. of components = 7	0.23
LASSO	$\sum(\lvert\beta\rvert)$=210.5	0.27
FSR	No. of variables = 17	0.25

Table 12.1: RMSEP-values for the regularization methods.

The model defined by (12.6) and (12.8), is now used with $b_1(\lambda), \ldots, b_k(\lambda)$ generated using a linear, quadratic and cubic B-spline basis with knots placed equidistantly over the range of wavelengths. If the number of basis-functions is restricted to be less than the number of observations, n, it results in a standard linear regression problem which can be solved using

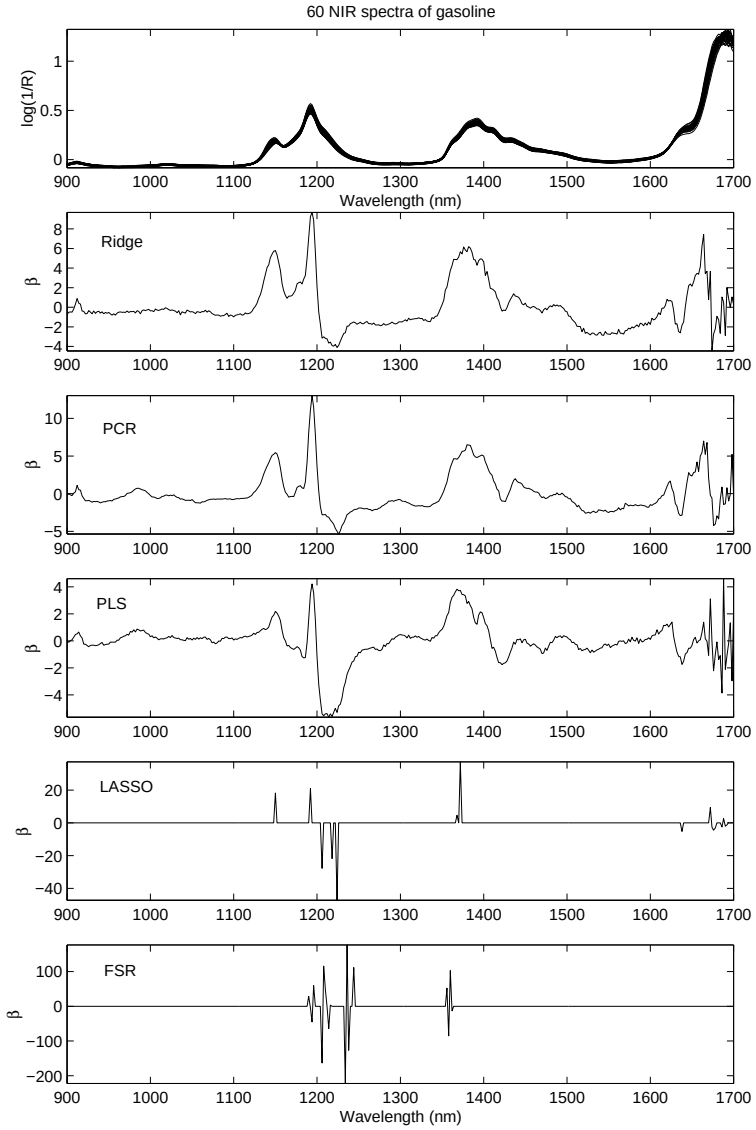

Figure 12.2: The 60 NIR spectra, together with the parameter estimates for Ridge, PCR, PLS, LASSO.and forward selection regression.

ordinary least squares. If the number of basis-functions is allowed to exceed the number of observations PCR, PLS, Ridge, LASSO or FSR can be applied.

The same setup as mentioned earlier is used to find the best model. The approach is straightforward, find the RMSEP-values for a fixed regularization parameter and varying number of basis-functions, now fix the regularization parameter to another value and find new RMSEP-values. This produces a matrix of RMSEP-values; find the smallest RMSEP-value and the corresponding value for the regularization parameter and the number of basis-functions. As an example LASSO has the optimum for $\sum(|\theta|)$=17.15 and 33 internal knots for the cubic B-spline basis; Figure 12.3 indicates the curvature of the RMSEP-surface around the optimum.

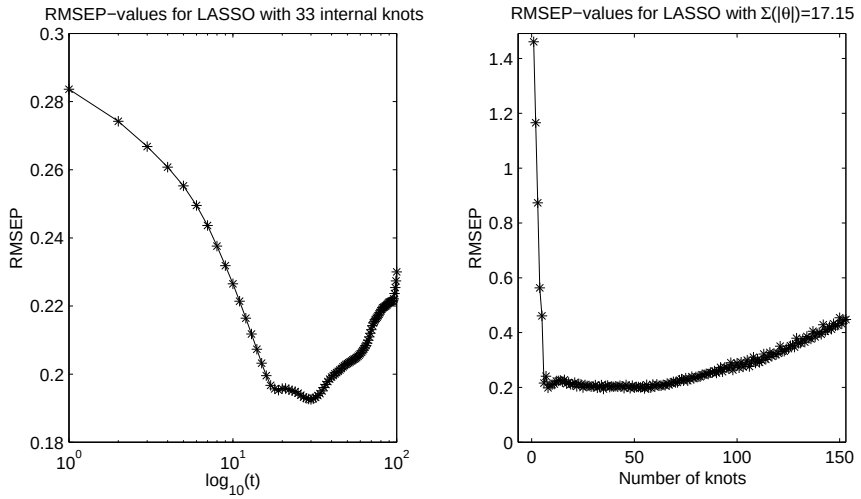

Figure 12.3: The RMSEP-values for LASSO for fixed number of internal knots (left) and fixed value for $\sum(|\theta|)$ (right) using a cubic B-spline basis.

The RMSEP-values are listed in Tables 12.2, 12.3 and 12.4. Contrary to the traditional methods it is seen that LASSO and FSR in combination with the spline bases perform best and that all the spline methods are superior to the traditional methods listed in Table 12.1. Comparing the best results until now with Tables 12.2, 12.3 and 12.4 results in a 17%

reduction in RMSEP-values when using spline basis functions. The simple OLS-solution results in a 13% reduction.

Method	Regularization parameter	Knots	RMSEP
Linear Spline-OLS		8	0.20
Linear Spline-Ridge	k = 0.0110	8	0.20
Linear Spline-PCR	No. of components = 7	8	0.20
Linear Spline-PLS	No. of components = 7	8	0.20
Linear Spline-LASSO	$\sum(\|\theta\|) = 12.65$	38	0.19
Linear Spline-FSR	No. of variables = 9	57	0.19

Table 12.2: RMSEP-values for some regularization methods combined with linear basis-functions.

Method	Regularization parameter	Knots	RMSEP
Quad. Spline-OLS		6	0.20
Quad. Spline-Ridge	k=0.0083	6	0.20
Quad. Spline-PCR	No. of components = 7	6	0.20
Quad. Spline-PLS	No. of components = 7	6	0.20
Quad. Spline-LASSO	$\sum(\|\theta\|) = 9.54$	48	0.19
Quad. Spline-FSR	No. of variables = 11	80	0.19

Table 12.3: RMSEP-values for some regularization methods combined with quadratic basis-functions.

Figure 12.4, 12.5 and 12.6 show the estimates of $\beta(\lambda) = \mathbf{B}^T(\lambda)\boldsymbol{\theta}$. The estimates are obtained using the tuning-parameters listed in Tables 12.2, 12.3 and 12.4 together with the full data set. For the OLS solutions curves indicating two times the pointwise standard error are also shown (obtained by disregarding that the number of internal knots are selected by use of cross-validation). For all but LASSO and FSR the estimates are quite similar. Comparing the standard error bands of the OLS-solution with the LASSO and the FSR-solution reveals that LASSO and FSR selects basis-functions corresponding to wavelengths for which the OLS-solution is significantly different from zero.

Method	Regularization parameter	Knots	RMSEP
Cubic Spline-OLS		4	0.21
Cubic Spline-Ridge	k=0.0008	4	0.20
Cubic Spline-PCR	No. of components = 7	4	0.20
Cubic Spline-PLS	No. of components = 7	4	0.20
Cubic Spline-LASSO	$\sum(\lvert\theta\rvert)$=17.15	33	0.19
Cubic Spline-FSR	No. of variables = 11	37	0.19

Table 12.4: RMSEP-values for some regularization methods combined with cubic basis-functions.

12.5 Comments

When the number of basis functions are low the knot placement may have large influence; it may move the valleys and peaks[1]. To avoid this the smoothing splines solution used by [20] and [22] may be applied. The P-spline approach by [41] provides a mix between these two approaches. All these approaches result in estimates of the parameter-function which have approximately the same degree of smoothness for all wavelengths for which the spectral measurements are performed. There is no reason to believe that this is desirable.

The B-spline-LASSO approach is one solution to the problem just outlined. Another solution would be to use wavelet basis functions together with LASSO. Since wavelets cover a large range of scales and positions, they may be more appropriate than B-splines. As yet another solution an adaptive knot-placement procedure could be applied together with standard linear regression. It is however not clear how to construct such a procedure.

For people using the traditional multivariate calibration techniques the main problem of applying the techniques presented here is the generation of the spline bases. In S-PLUS[2] and R[3] these can be generated with the built-in functions `bs` (B-splines) or `ns` (natural splines). In Matlab[4] one can use `bsplval.m` by Dr. Graeme A. Chandler, Mathematics Department, The University of Queensland, Australia. A ZIP-archive containing this function can be downloaded as www.maths.uq.edu.au/~gac/mn309/mfilez.zip

[1] See [23] pp. 251-254
[2] (www.splus.mathsoft.com)
[3] (www.r-project.org)
[4] (www.mathworks.com)

and in www.maths.uq.oz.au/~gac/mn309/bspl.html examples of how to apply it can be found.

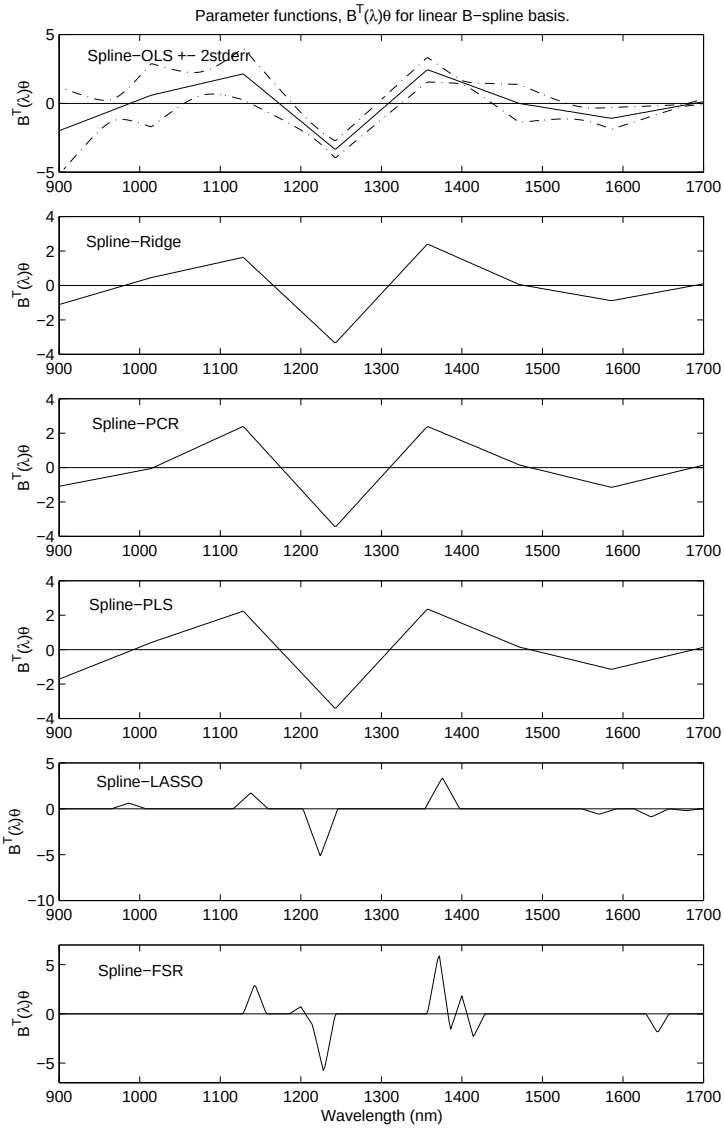

Figure 12.4: Estimated parameter-functions using OLS, Ridge, PCR, PLS, LASSO and forward selection regression together with linear B-spline bases.

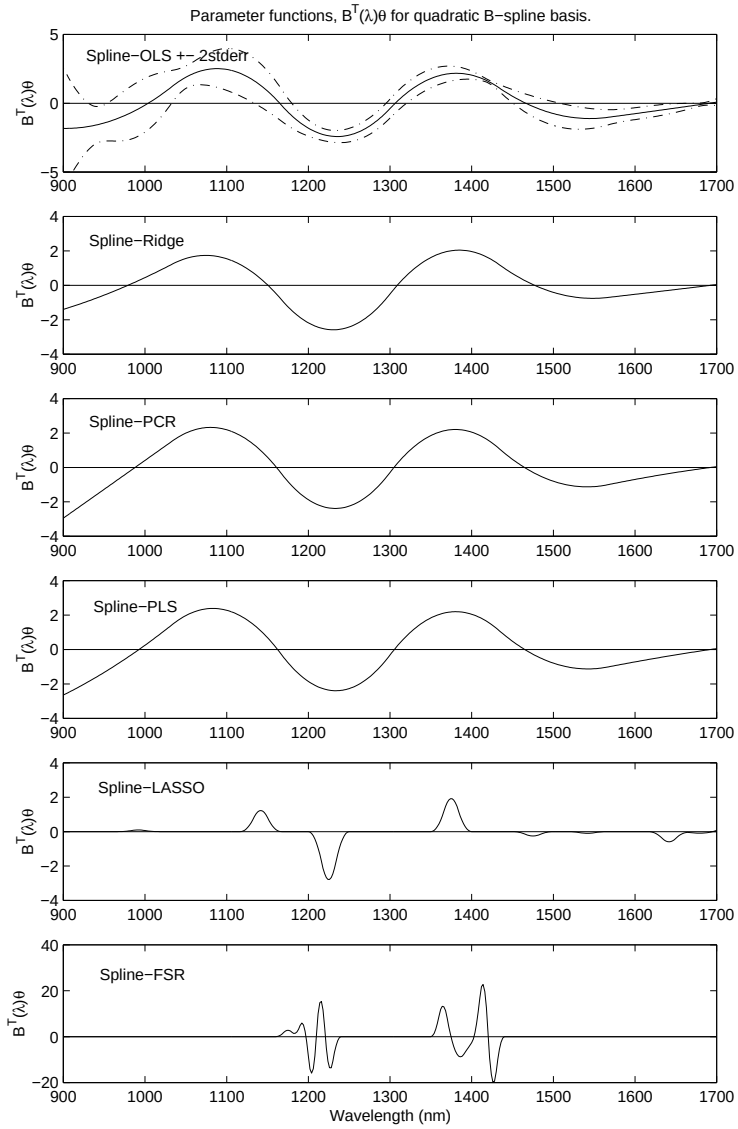

Figure 12.5: Estimated parameter-functions using OLS, Ridge, PCR, PLS, LASSO and forward selection regression together with quadratic B-spline bases.

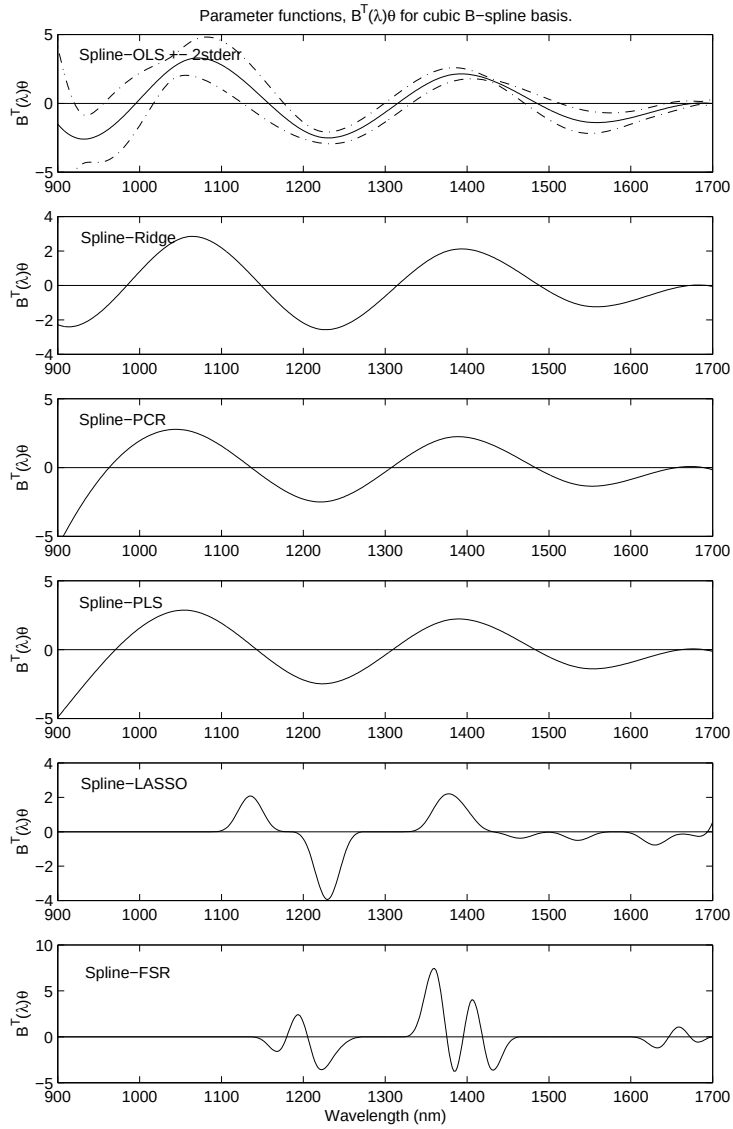

Figure 12.6: Estimated parameter-functions using OLS, Ridge, PCR, PLS, LASSO and forward selection regression together with cubic B-spline bases.

12.6 Range selection using the BFR estimates

Prior knowledge about which wavelengths are important for the prediction of a certain response is very valuable. Such knowledge would help to reduce the complexity of the problem, and it might even make it possible to do an exhaustive search for an optimal subset if the number of important wavelengths is small enough[5]. A method for selecting specific intervals of the spectrum using PLS has been suggested, see [48].

An attempt to identify an important wavelength range for prediction of octane is done by using the estimates resulting from using forward selection combined with the B-spline bases. The idea is that where the estimate is different from zero actually identifies the regions that are important for prediction of octane. In Figure 12.4 the estimate for the linear spline-FSR method is used to select a new and smaller range onto which the methods MLLS, Ridge, PCR, PLS, LASSO and FSR will be applied. The range obtained from the estimate is

$$[1128 : 1156,\ 1186 : 1242,\ 1358 : 1428,\ 1628 : 1658]$$

There are 96 of the original 401 wavelengths contained in this range, see Figure 12.7 for a graphical display of the range. The results for using this range are shown in Table 12.5. The parameter estimates can be seen in Figure 12.8. To check whether the results for the selected range are purely coincidential the methods have also been applied to the complementary set of wavelengths, see Table A.1 for RMSEP-values and Figure A.2 for the parameter estimates.

[5] With the computational power available today it is possible to perform exhaustive search for all model sizes, if the number of variables is less than 30.

Method	Regularization parameter	RMSEP	% improvement
MLLS	$\|\|\hat{\boldsymbol{\beta}}\|\|_2 = 827.8$	0.37	-9%
Ridge	$k = 4.32 \times 10^{-4}$	0.20	17%
PCR	No. of components = 7	0.19	17%
PLS	No. of components = 5	0.19	17%
LASSO	$\sum(\|\theta\|) = 236.45$	0.20	26%
FSR	No. of variables = 17	0.25	0%

Table 12.5: RMSEP-values for some regularization methods on the reduced range. The %-wise reduction in the RMSEP-value for each of the methods is also tabulated.

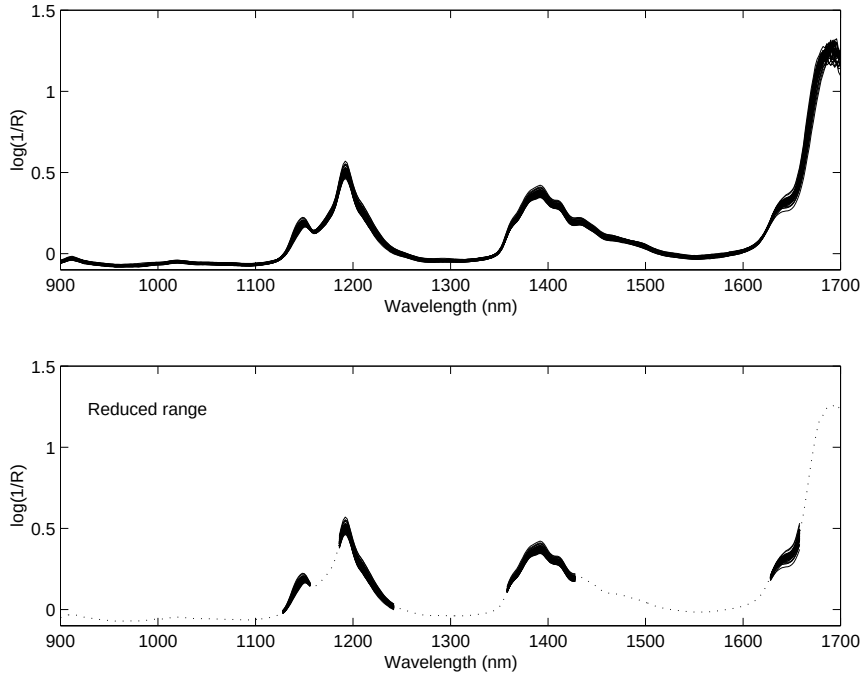

Figure 12.7: The full range spectra (top figure) and the selected range for prediction of octane (bottom figure).

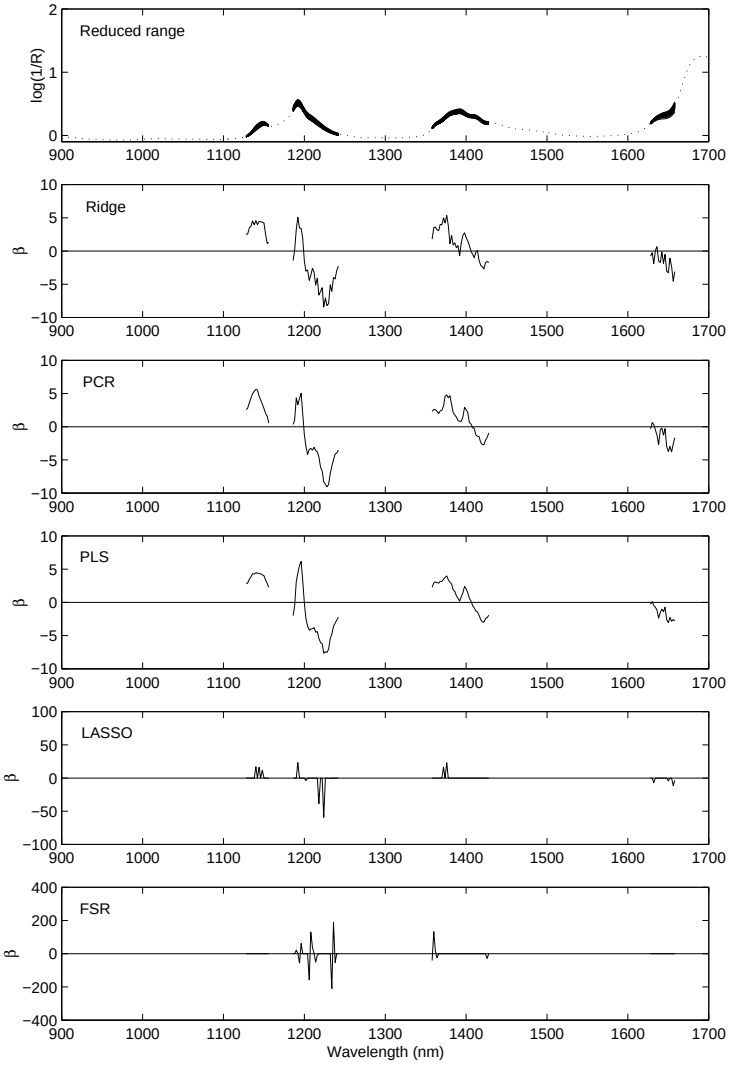

Figure 12.8: The 60 NIR spectra on the reduced range, together with the parameter estimates for Ridge, PCR, PLS, LASSO and forward selection regression.

The results clearly show that information about which part of the spectra that is important for the prediction of octane, indeed is contained in the regions found by using the linear spline-FSR estimate. Using the traditional methods on the reduced range results in large reductions of the RMSEP-values for all the methods except MLLS[6] and forward selection regression. The wavelengths selected by the forward selection method on the full spectrum are all but the last two included in the reduced range, so of course the new model does not deviate much from the full-spectrum model[7]. Removing uninformative parts of the spectrum also reveals how unstable the MLLS estimate is when used for prediction purposes. Due to the nature of the basis function regression it is only possible, with the present implementation, to apply the method to a connected range of wavelengths. Therefore it is not possible to show results for the spline methods on the selected range in this case.

That the basis function regression also can be used to select a range on which the traditional methods do better than in the full-spectrum case, makes it possible to develop faster instruments by employing a few critical regions of the entire spectrum. A reduction in the number of explanatory variables also helps in the interpretation of the models.

12.7 Summary for the gasoline example

The results for the gasoline example are summarized in Table 12.6. When the traditional methods are applied to the full spectrum they are quite successfull in shrinking the solution away from the MLLS solution in directions that captures the variation of octane. LASSO and FSR are also quite successfull in selecting single wavelengths which carry information regarding the variation of octane.

LASSO is not the best method when applied to the full spectrum, but in combination with the spline basis functions LASSO is always among the two best methods. The spline methods are in all cases better than the traditional methods based on the full spectrum. This indicates that a smoothing of the parameter estimates is desirable in this case. When using the linear spline-FSR estimate as a tool to select certain regions of

[6] The estimate for MLLS applied to the reduced spectrum can be seen in Figure A.1

[7] The wavelengths selected by the forward selection method in the two cases are shown in Table A.2

the spectrum all results except for forward selection, are improved substantially. The best methods on the reduced range result in a 17% reduction in the RMSEP-value compared to the best full-spectrum method.

Table 12.6: Results for the gasoline example.

Method	Regularization param.	Knots	RMSEP
MLLS	$\|\hat{\boldsymbol{\beta}}\|_2 = 217.7$		0.34
Ridge	$k = 0.002$		0.24
PCR	No. of comps.= 13		0.23
CPCR	No. of comps.= 7		0.26
PLS	No. of comps.= 7		0.23
CSR	Eigenv./factors=19/7		0.23
FSR	No. of var. = 17		0.25
LASSO	$\sum(\|\beta_i\|)$=210.5		0.27
Linear Spline-OLS		8	0.20
Linear Spline-Ridge	k = 0.0110	8	0.20
Linear Spline-PCR	No. of comps. = 7	8	0.20
Linear Spline-PLS	No. of comps. = 7	8	0.20
Linear Spline-LASSO	$\sum(\|\theta\|) = 12.65$	38	0.19
Linear Spline-FSR	No. of var. = 9	57	0.19
Quad. Spline-OLS		6	0.20
Quad. Spline-Ridge	k=0.0083	6	0.20
Quad. Spline-PCR	No. of comps. = 7	6	0.20
Quad. Spline-PLS	No. of comps. = 7	6	0.20
Quad. Spline-LASSO	$\sum(\|\theta\|) = 9.54$	48	0.19
Quad. Spline-FSR	No. of var. = 11	80	0.19
Cubic Spline-OLS		4	0.21
Cubic Spline-Ridge	k=0.0008	4	0.20
Cubic Spline-PCR	No. of comps. = 7	4	0.20
Cubic Spline-PLS	No. of comps. = 7	4	0.20
Cubic Spline-LASSO	$\sum(\|\theta\|)$=17.15	33	0.19
Cubic Spline-FSR	No. of var. = 11	37	0.19
Reduced MLLS	$\|\hat{\boldsymbol{\beta}}\|_2 = 827.8$		0.37
Reduced Ridge	$k = 4.32 \times 10^{-4}$		0.20
Reduced PCR	No. of comps. = 7		0.19
Reduced PLS	No. of comps. = 5		0.19
Reduced LASSO	$\sum(\|\theta\|) = 236.45$		0.20
Reduced FSR	No. of var. = 17		0.25

Chapter 13
Wheat example

13.1 Data

This set of data origins from a Near-InfraRed, (NIR), analysis of wheat. It contains 100 samples with specified protein and moisture content. Samples were measured using diffuse reflectance (R) as $\log(1/R)$ from 1100 nm to 2500 nm in 2 nm intervals ($n = 100$ and $p = 701$)[1]. The spectra are shown in Figure 13.1.

[1]The data set can be obtained from
ftp://ftp.clarkson.edu/pub/hopkepk/chemdata/kalivas

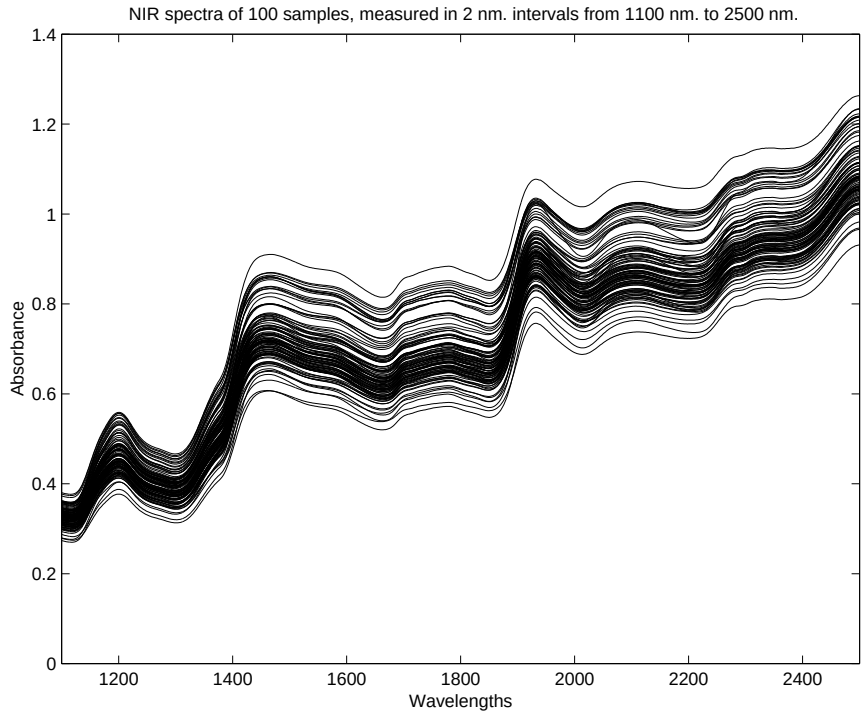

Figure 13.1: 100 NIR spectra for wheat, measured in 2nm. intervals from 1100nm. to 2500nm.

13.1.1 Pretreatment of data

In Figure 13.1 the spectra have clearly shifted, this is due to unequal particle sizes. By differencing the columns of $\mathbf{X}$ the constants and sudden shifts which are not important to the regression are removed, [41] p. 2. Figure 13.2 shows the first-order differenced spectra.

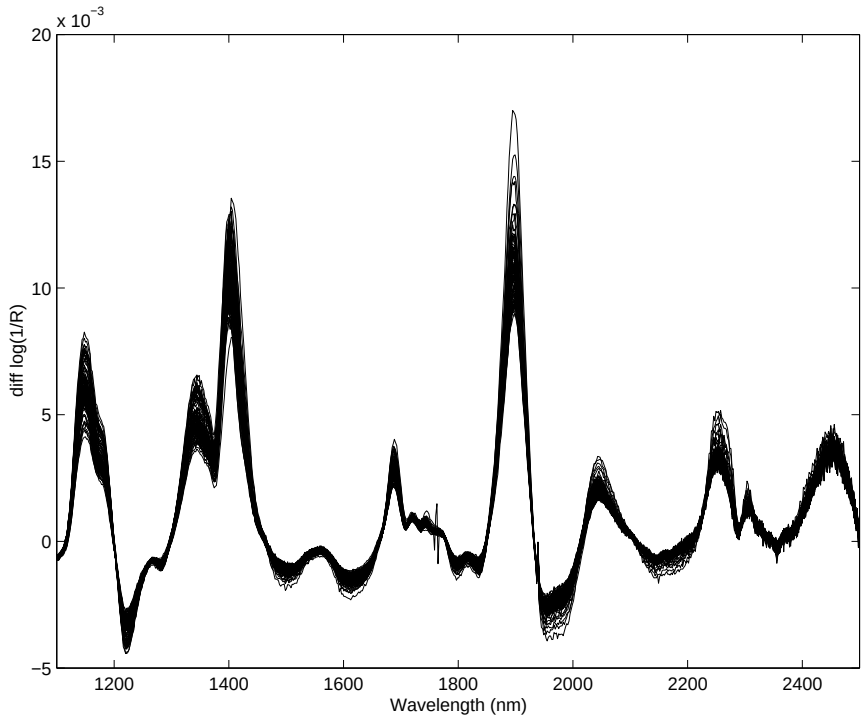

Figure 13.2: The first-order differenced spectra.

13.1.2 Setup for 5-fold cross-validation

The data is split into five different sets, the calibration part consists consecutively of 4 different parts, and the validation data is the part left out of the calibration data. When splitting the data it is important to construct the groups in a way that the response-variables span approximately the same levels. For the wheat example this is achieved by sorting the moisture and protein levels in ascending order and then numbering them successively from 1 to 5 in order to get five sets that cover approximately the same range. The relatively large number of observations, (100), makes it possible to reserve some of the observations for an external validation set. After the observations have been sorted in ascending order every third observation is extracted for the external validation set, see Figure 13.3 and 13.4.

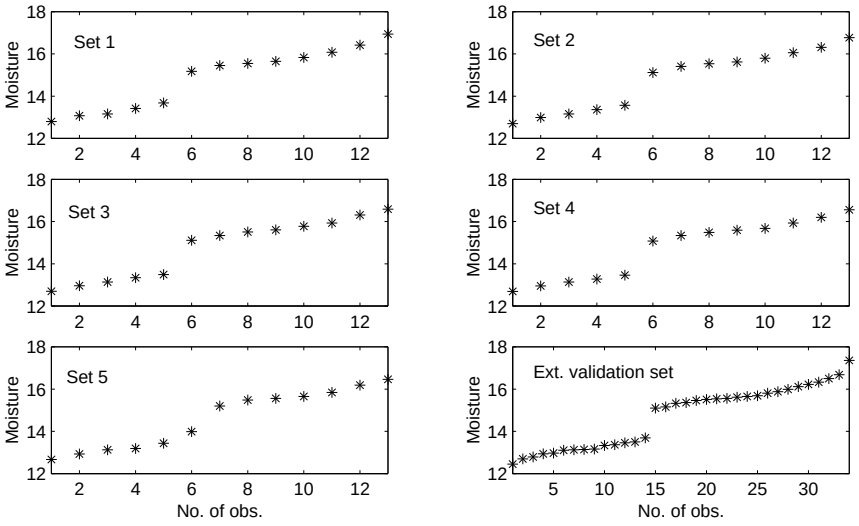

Figure 13.3: The 5-fold cross-validation splits, and the external validation set for moisture.

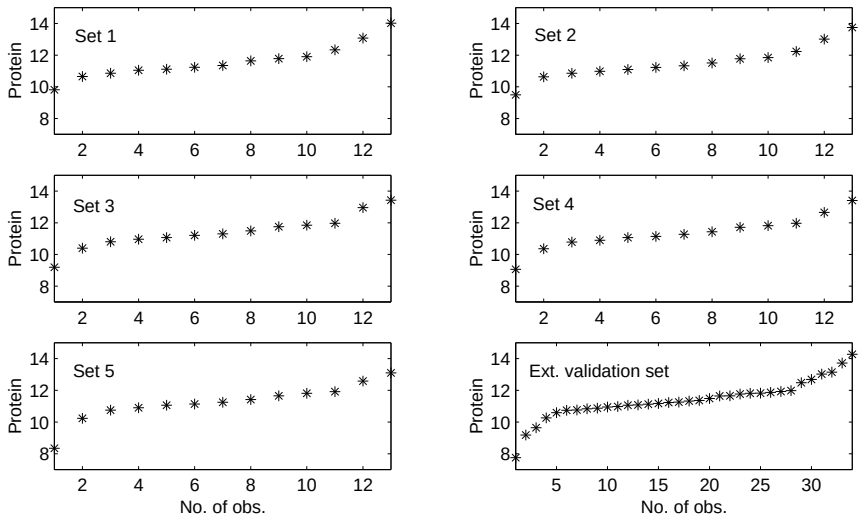

Figure 13.4: The 5-fold cross-validation splits, and the external validation set for protein.

13.2 Results

For all the methods the tuning parameter is determined using the cross-validated RMSEP-values. The estimate used to predict the external validation set is based on the full data set excluding the external validation set.

13.2.1 MLLS applied to wheat example

The parameter estimates for the MLLS solution for moisture and protein are shown in Figure 13.5.

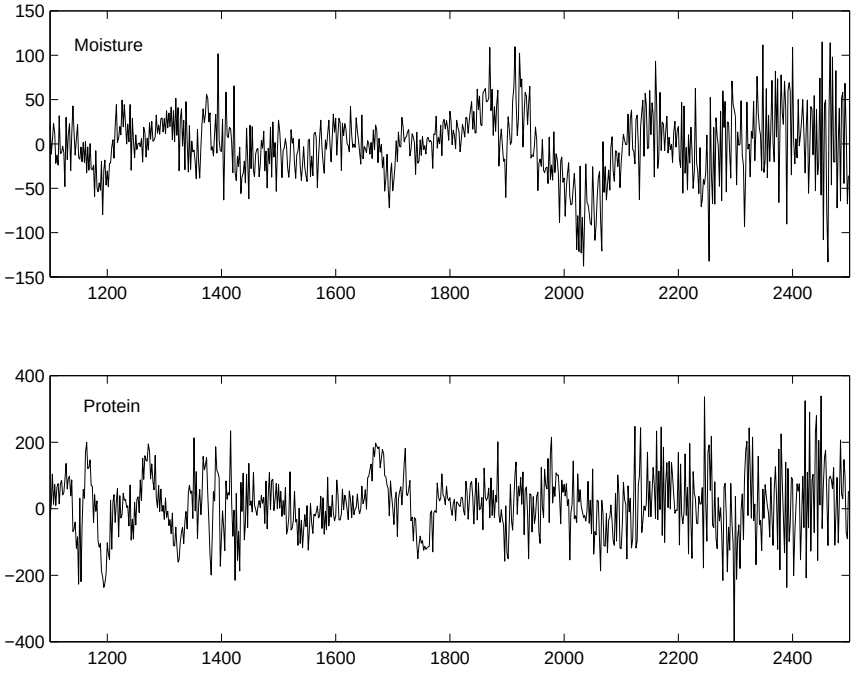

Figure 13.5: Parameter estimates, $\hat{\boldsymbol{\beta}}$, for MLLS.

The estimates in Figure 13.5 oscillates as in the gasoline example. The reason for this is explained in Section 5.4. The RMSEP-values are shown in Table 13.1 and 13.2 for moisture and protein respectively.

Method	$\|\hat{\boldsymbol{\beta}}\|_2$	RMSEP
MLLS	$\|\hat{\boldsymbol{\beta}}\|_2 = 1.01 \times 10^3$	0.27

Table 13.1: RMSEP-value for the MLLS solution for the moisture model.

Method	$\|\hat{\boldsymbol{\beta}}\|_2$	RMSEP
MLLS	$\|\hat{\boldsymbol{\beta}}\|_2 = 2.55 \times 10^3$	0.54

Table 13.2: RMSEP-value for the MLLS solution for the protein model.

13.2.2 Ridge applied to wheat example

For the Ridge method the optimal value was found by iterating through values of k chosen on an equally spaced grid on the logarithmic scale, see Figure 13.6.

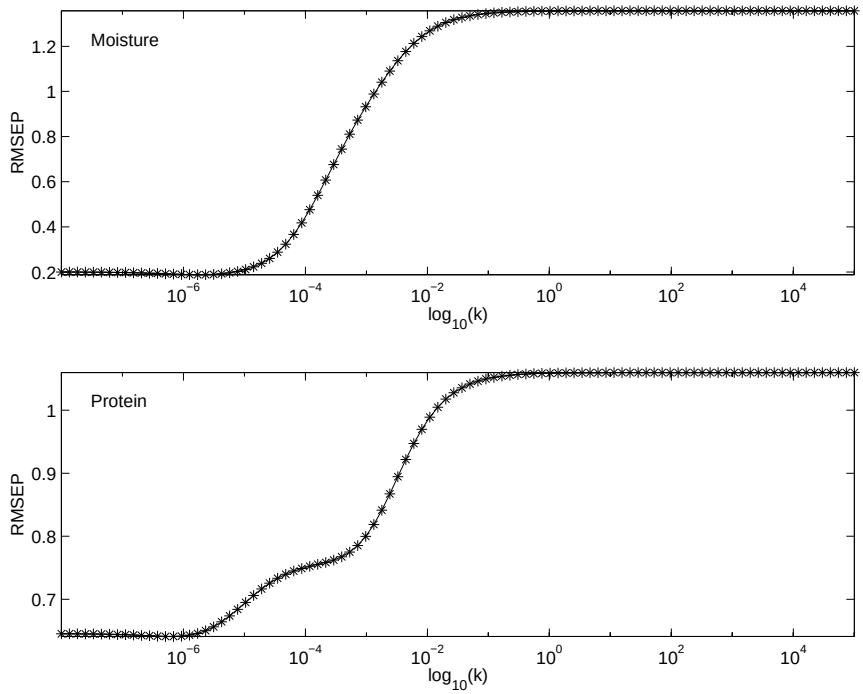

Figure 13.6: The RMSEP values for the Ridge method applied to the moisture and protein data as a function of $\log_{10}(k)$.

The RMSEP-values for the moisture and the protein data are shown in Table 13.3 and 13.4. The Ridge estimates for moisture and protein re-

Method	Regularization parameter	$\|\hat{\boldsymbol{\beta}}\|_2$	RMSEP
Ridge	$k = 1.79 \times 10^{-6}$	$\|\hat{\boldsymbol{\beta}}\|_2 = 7.42 \times 10^2$	0.27

Table 13.3: Ridge results for moisture.

Method	Regularization parameter	$\|\hat{\boldsymbol{\beta}}\|_2$	RMSEP
Ridge	$k = 6.28 \times 10^{-7}$	$\|\hat{\boldsymbol{\beta}}\|_2 = 2.08 \times 10^3$	0.54

Table 13.4: Ridge results for protein.

spectively is shown in Figure 13.7. Comparing the Ridge estimate to the MLLS estimate for moisture they look very much alike. The extra regularization implied by the Ridge method is reflected in the squared length of $\hat{\boldsymbol{\beta}}$. The regularization parameter, k, is chosen very small in both cases. Remember that, as $k \to 0$ the Ridge estimate is shown earlier to approximate the MLLS estimate. The RMSEP-values for MLLS and Ridge are the same for both moisture and protein.

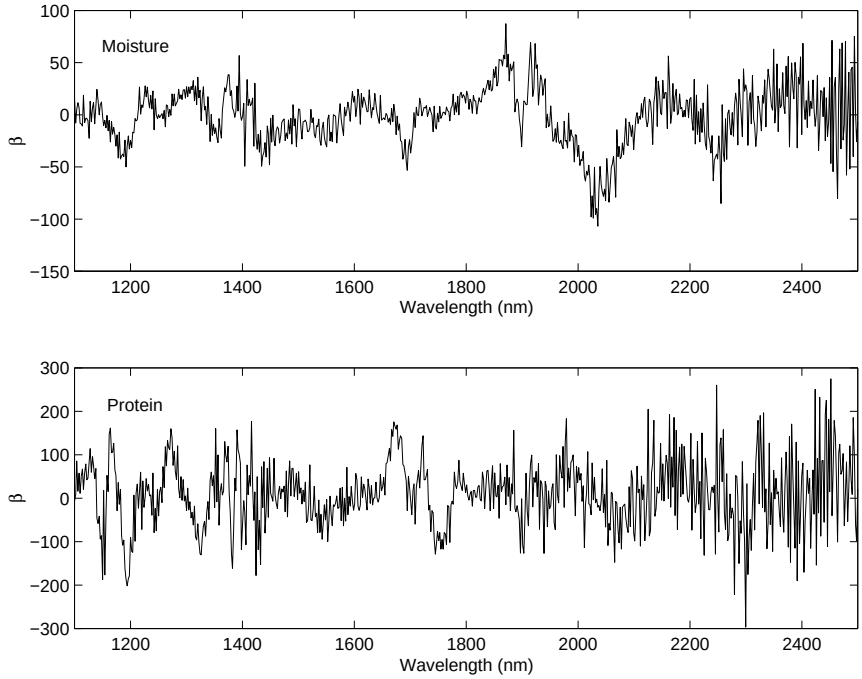

Figure 13.7: Parameter estimates, $\hat{\boldsymbol{\beta}}$, for Ridge.

13.2.3 PCR applied to wheat example

The PCR method with three different selection strategies was tested for a number of principal components. The result of this for both the moisture and the protein model is seen in Figure 13.8. The RMSEP-values for the

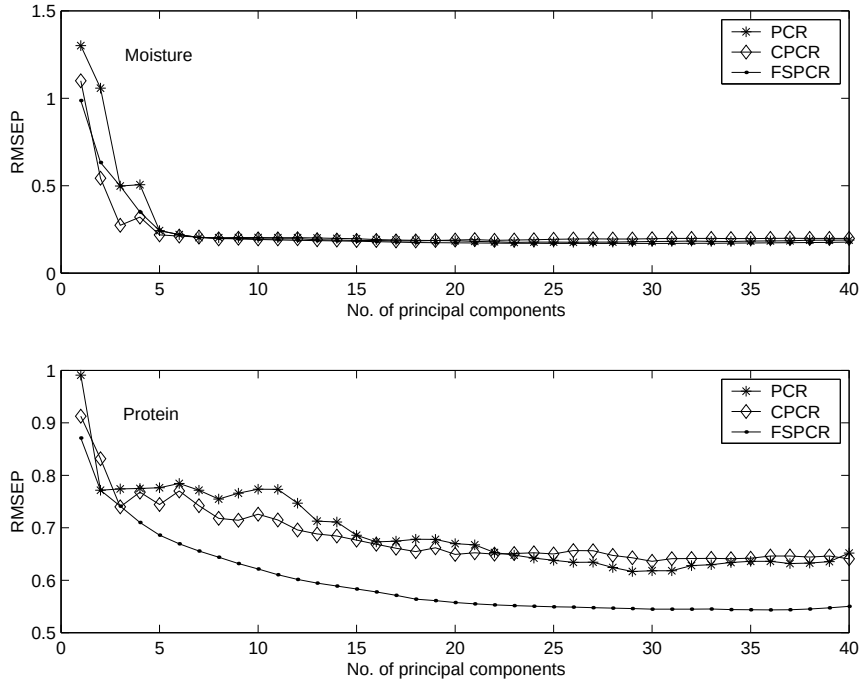

Figure 13.8: The RMSEP values for the moisture and protein model as a function of the number of principal components for the three different selection strategies.

prediction of moisture and protein, are shown in Table 13.5 and 13.6. The parameter estimates for the three selection strategies are shown in Figure B.1, B.2 and B.3.

Method	Regularization parameter	RMSEP
PCR	No. of components = 23	0.27
CPCR	No. of components = 17	0.28
FSPCR	No. of components = 27	0.26

Table 13.5: PCR results for moisture.

Method	Regularization parameter	RMSEP
PCR	No. of components = 29	0.57
CPCR	No. of components = 30	0.56
FSPCR	No. of components = 36	0.60

Table 13.6: PCR results for protein.

Table 13.7: Order of the PC's chosen by the CPCR and FSPCR method for moisture (only the first 15 are shown). The absolute correlation is also shown.

CPCR order for moisture	$\|r\|$	FSPCR order for moisture
3	0.6773	3
2	0.6450	2
5	0.3030	1
1	0.0985	5
7	0.0603	4
17	0.0489	6
16	0.0442	7
6	0.0400	20
21	0.0338	21
12	0.0333	14
52	0.0303	23
14	0.0298	22
22	0.0254	17
38	0.0251	12
4	0.0249	13

Table 13.8: Order of the PC's chosen by the CPCR and FSPCR method for protein (only the first 15 are shown). The absolute correlation is also shown.

CPCR order for protein	$\|r\|$	FSPCR order for protein
2	0.5973	2
1	0.4394	1
10	0.2081	12
12	0.2022	13
11	0.1942	15
13	0.1841	41
8	0.1830	22
19	0.1798	37
18	0.1666	24
7	0.1484	16
4	0.1341	29
5	0.1268	20
24	0.1190	23
46	0.0977	14
56	0.0973	28

Comments to the PCR methods

For the prediction of moisture the forward selection strategy is the best. FSPCR and PCR choose the same first 7 components. The CPCR method has 5 of the first 7 components equal to the other two methods. The components corresponding to some of the largest singular values are also the ones with the largest absolute correleation with moisture. Of the first 15 components chosen by CPCR and FSPCR 12 components are the same.

For protein the PCR and CPCR methods perform best, but the results for MLLS and Ridge are better. Only the first 2 components are the same for the three methods and of the first 15 chosen by CPCR and FSPCR only 5 components are the same. Regardless of the selection approach, prediction of protein requires more components than prediction of moisture and the selected components are much different. That the three methods choose so differently among the principal components could imply that information lies in all the principal components. The smoothness in the RMSEP-values with increasing number of components is an artifact of the selection criterion. In Figure 13.8 FSPCR for protein is consistently better than PCR and CPCR. If this is the result of overfitting it would explain the bad prediction result when the FSPCR estimate is used to predict the external validation set.

13.2.4 PLS applied to wheat example

The RMSEP-value for increasing number of PLS components is shown in Figure 13.9.

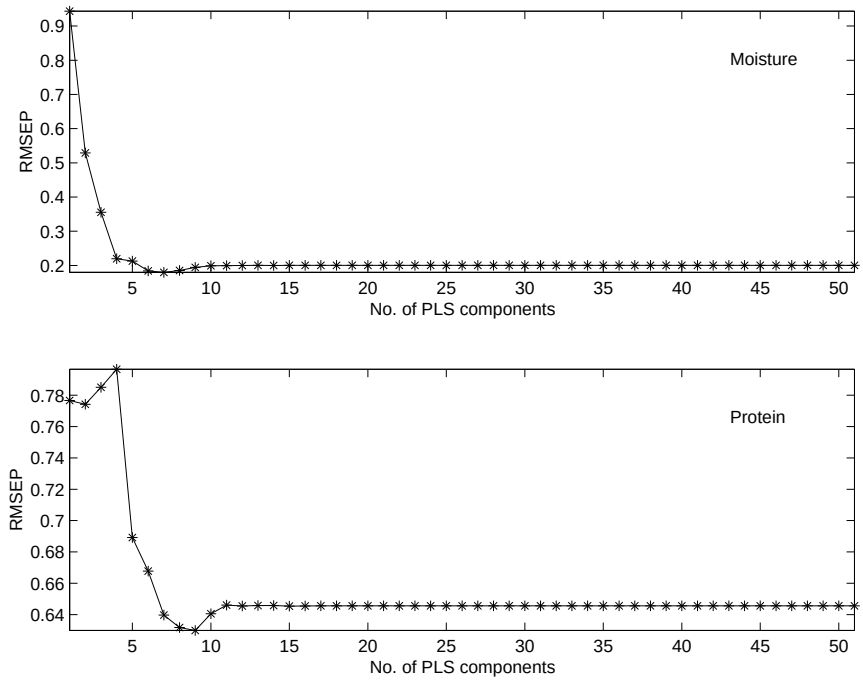

Figure 13.9: The RMSEP value as a function of the number of PLS components for the moisture model, (top figure), and protein model, (bottom figure).

Method	Regularization parameter	RMSEP
PLS	No. of components = 7	0.28

Table 13.9: PLS results for moisture.

Method	Regularization parameter	RMSEP
PLS	No. of components = 9	0.55

Table 13.10: PLS results for protein.

The tendency for PLS to use fewer components than PCR is very outspoken for both the moisture and the protein model. In both cases the PCR methods use from three to four times more components than PLS but as mentioned in Section 7.5 this is (in general) neither an advantage nor disadvantage. The resulting parameter estimates are shown in Figure 13.10, they look very much alike the ones produced by MLLS, Ridge and the PCR methods.

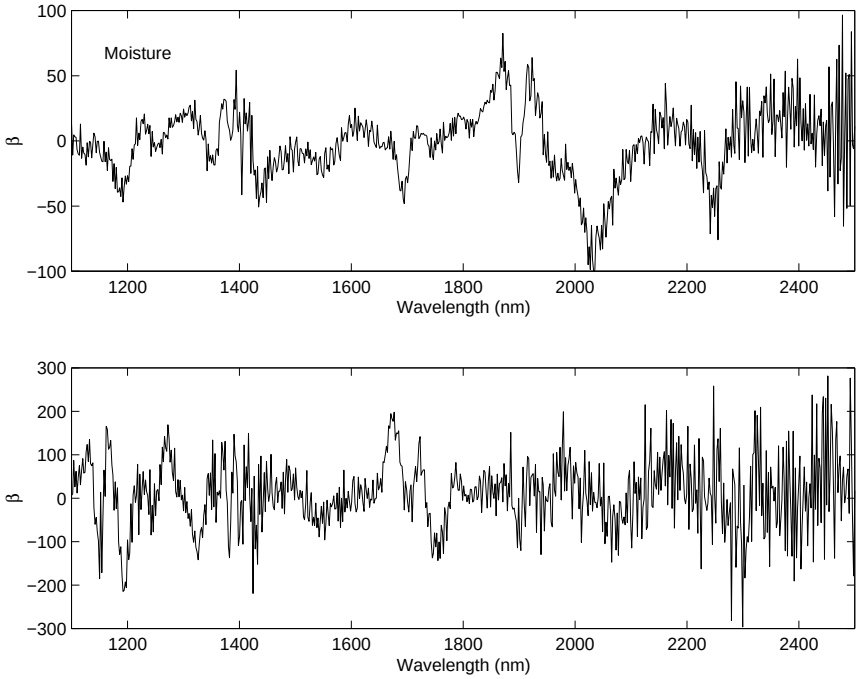

Figure 13.10: Parameter estimates, $\hat{\boldsymbol{\beta}}$, for PLS

13.2.5 CSR applied to wheat example

Method	Regularization parameter	RMSEP
CSR	No. of eigenvectors and factors = 35 and 34	0.28

Table 13.11: CSR results for moisture.

Method	Regularization parameter	RMSEP
CSR	No. of eigenvectors and factors = 50 and 34	0.54

Table 13.12: CSR results for protein.

13.2.6 FSR applied to wheat example

For a number of variables ranging from 1 to rank($\mathbf{X}$) the RMSEP-values for the different models are shown in Figure 13.11 As seen in Table 13.13 FSR finds that the optimal number of variables for predicting moisture is 7. From the estimate in Figure 13.12 one can see that all the variables are selected from within two small areas. Since the RMSEP-value for FSR is larger than for the other methods, more information regarding moisture is most likely contained in other regions as well. For protein FSR selects 12

Method	Regularization parameter	RMSEP
FSR	No. of variables = 7	0.34

Table 13.13: RMSEP-values for the forward selection method for the moisture model.

variables, which result in a RMSEP-value equal to the one obtained using MLLS, Ridge and the CSR method.

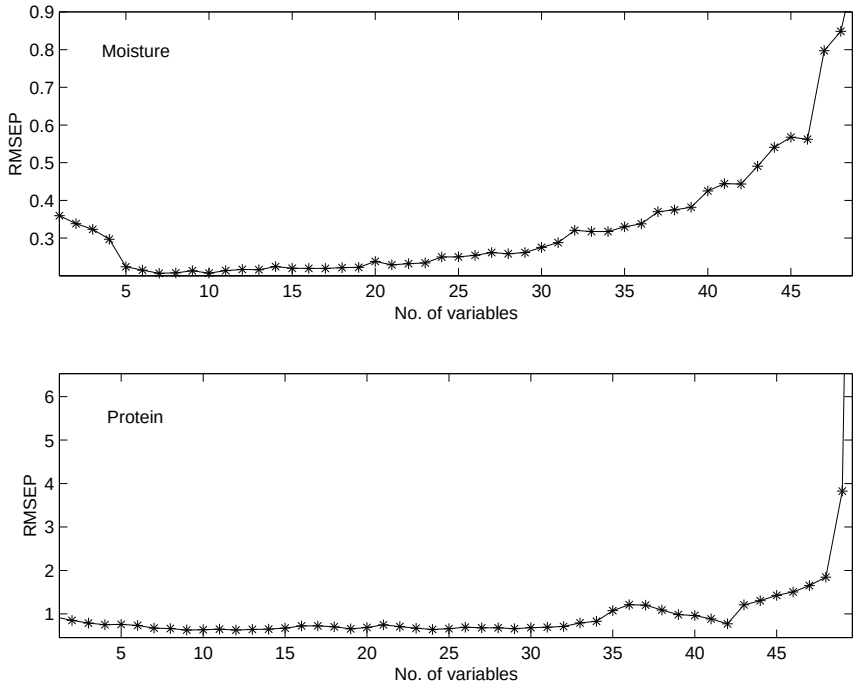

Figure 13.11: RMSEP as a function of the number of variables

Method	Regularization parameter	RMSEP
FSR	No. of variables = 12	0.54

Table 13.14: RMSEP-values for the forward selection method for the protein model.

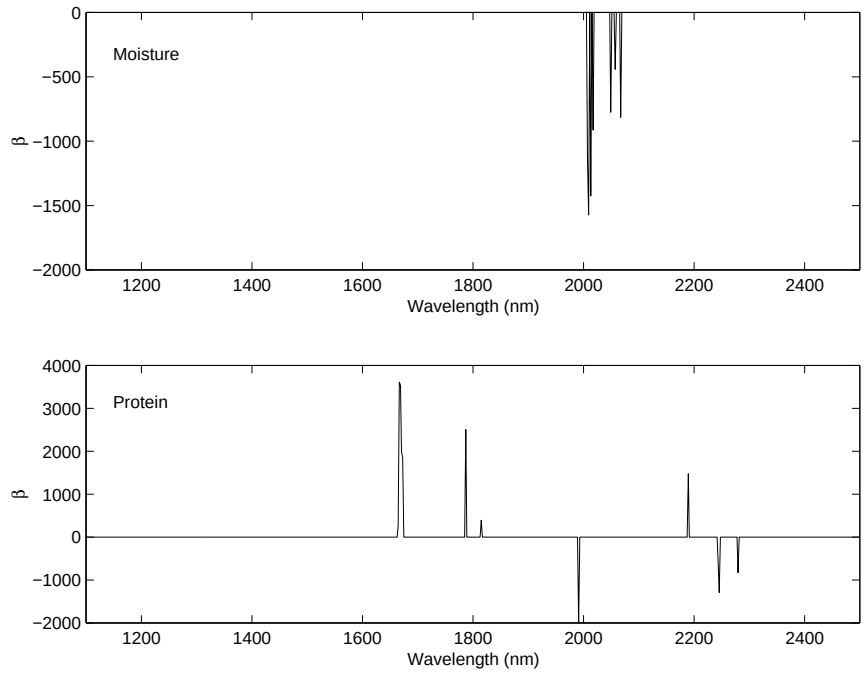

Figure 13.12: Parameter estimates, $\hat{\boldsymbol{\beta}}$, for the forward selection method.

13.2.7 LASSO applied to wheat example

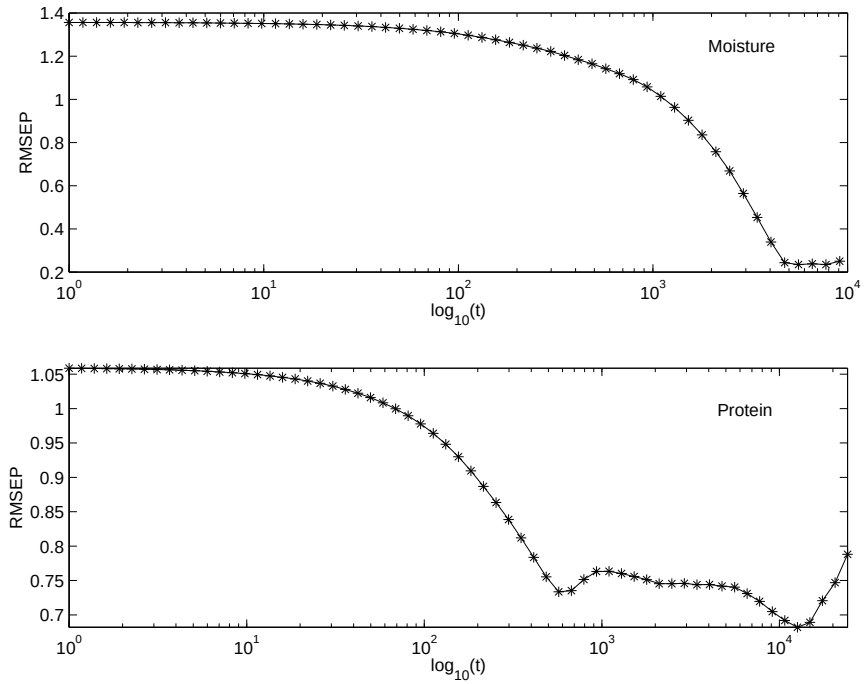

Figure 13.13: The RMSEP values for the LASSO method applied to the moisture and protein data as a function of $\log_{10}(t)$.

The LASSO method is better at predicting moisture than FSR but is inferior to the other methods. The LASSO estimate has the largest absolute values for wavelengths neighboring those selected by FSR. Besides that, LASSO selects variables in a few other areas of the spectrum. Since the Ridge method works better than both LASSO and FSR the important information regarding prediction of moisture is probably spread throughout the entire spectrum, see [64].

For prediction of protein the LASSO method selects variables from nearly all areas of the entire spectrum. LASSO and FSR provide competitive results for the prediction of protein compared to the other methods.

Method	Regularization parameter	RMSEP
LASSO	$\sum(\lvert\beta_i\rvert) = 8.09 \times 10^3$; $\lambda = 8.44 \times 10^{-5}$	0.30

Table 13.15: LASSO results for moisture.

Method	Regularization parameter	RMSEP
LASSO	$\sum(\lvert\beta_i\rvert) = 1.33 \times 10^4$; $\lambda = 4.20 \times 10^{-4}$	0.56

Table 13.16: LASSO results for protein.

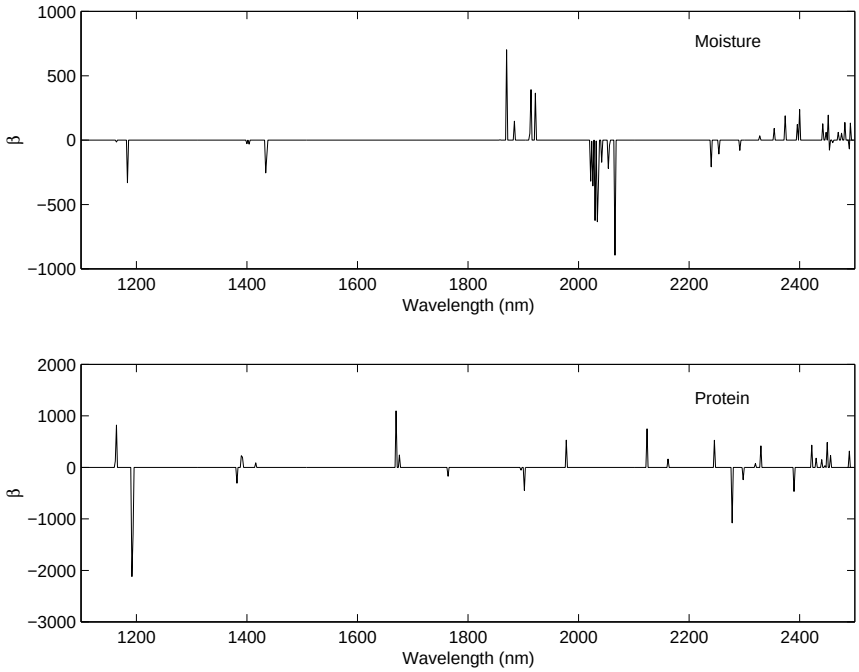

Figure 13.14: Parameter estimates, $\hat{\boldsymbol{\beta}}$, for LASSO

13.2.8 BFR applied to wheat example

As in the gasoline example the basis functions used are linear, quadratic and cubic B-spline basis-functions. The results for the prediction of moisture using OLS, Ridge, PCR, PLS LASSO and FSR in combination with linear, quadratic and cubic B-spline basis-functions are shown respectively in Table 13.17, 13.19 and 13.21. The results for the prediction of protein with the same methods are shown in Table 13.18, 13.20 and 13.22.

Method	Regularization parameter	Knots	RMSEP
Linear Spline-OLS		13	0.31
Linear Spline-Ridge	$k = 2.95 \times 10^{-5}$	40	0.30
Linear Spline-PCR	No. of components = 14	45	0.32
Linear Spline-PLS	No. of components = 8	31	0.33
Linear Spline-LASSO	$\sum(\lvert\theta\rvert)$= 546.2	80	0.30
Linear Spline-FSR	No. of variables = 12	58	0.31

Table 13.17: RMSEP-values for moisture for some regularization methods combined with the linear B-spline basis-function regression.

Method	Regularization parameter	Knots	RMSEP
Linear Spline-OLS		14	0.77
Linear Spline-Ridge	$k = 1.0 \times 10^{-10}$	192	0.47
Linear Spline-PCR	No. of components = 51	191	0.46
Linear Spline-PLS	No. of components = 27	191	0.44
Linear Spline-LASSO	$\sum(\lvert\theta\rvert)$=3.43×10^3	106	0.37
Linear Spline-FSR	No. of variables = 7	122	0.53

Table 13.18: RMSEP-values for protein for some regularization methods combined with the linear B-spline basis-function regression.

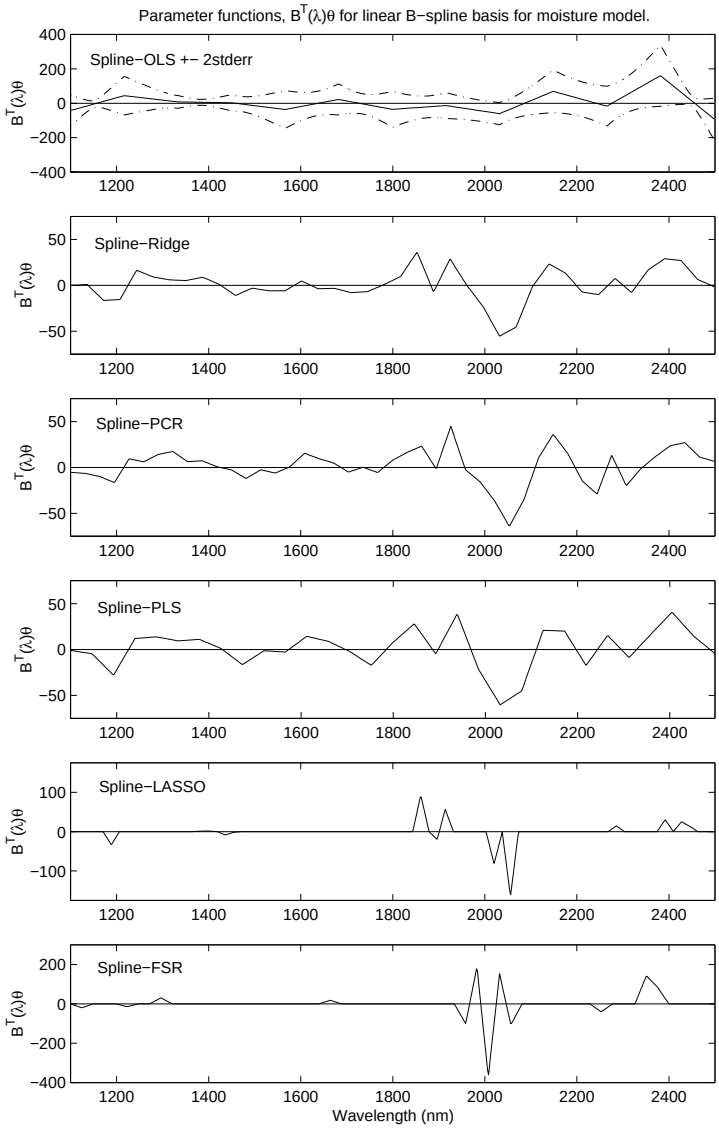

Figure 13.15: Estimated coefficient-functions using OLS, Ridge, PCR, PLS, LASSO, and FSR together with linear B-spline bases for the moisture model.

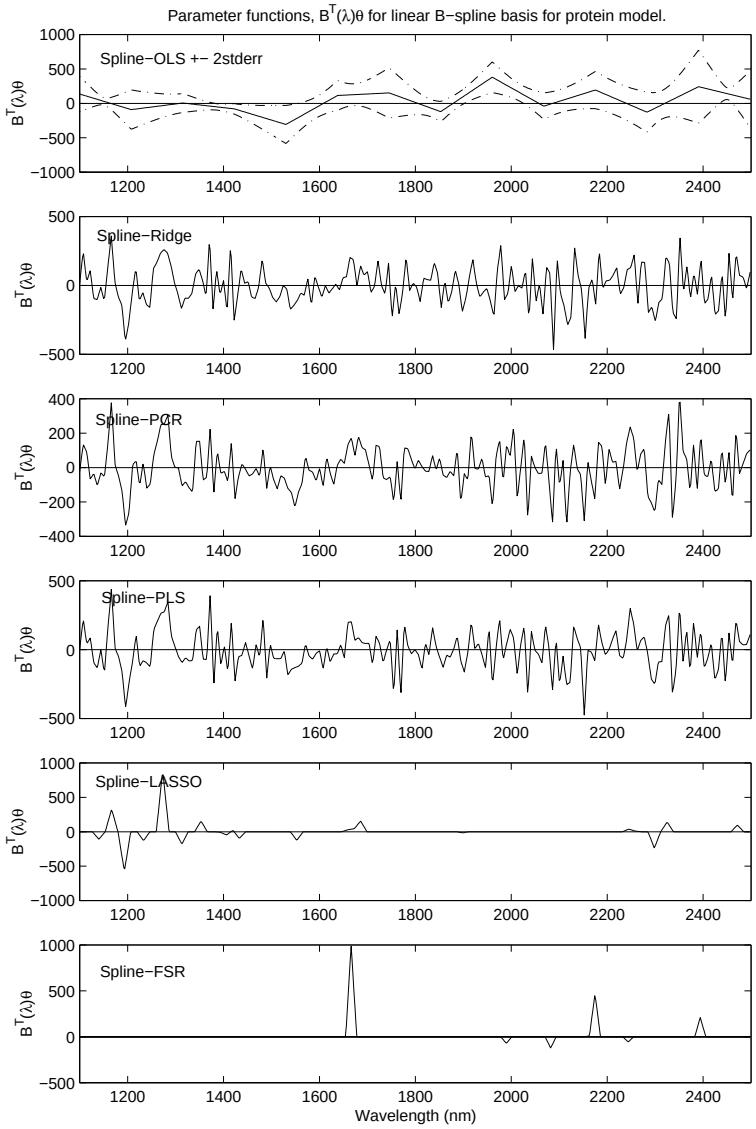

Figure 13.16: Estimated coefficient-functions using OLS, Ridge, PCR, PLS, LASSO, and FSR together with linear B-spline bases for the protein model.

Method	Regularization parameter	Knots	RMSEP
Quad. Spline-OLS		13	0.27
Quad. Spline-Ridge	$k = 2.2 \times 10^{-5}$	38	0.30
Quad. Spline-PCR	No. of components = 14	38	0.32
Quad. Spline-PLS	No. of components = 7	58	0.31
Quad. Spline-LASSO	$\sum(\lvert\theta\rvert)$=167.7	23	0.31
Quad. Spline-FSR	No. of variables = 3	17	0.35

Table 13.19: RMSEP-values for moisture for some regularization methods combined with the quadratic B-spline basis-function regression.

Method	Regularization parameter	Knots	RMSEP
Quad. Spline-OLS		12	0.60
Quad. Spline-Ridge	$k = 1.0 \times 10^{-11}$	219	0.47
Quad. Spline-PCR	No. of components = 50	284	0.51
Quad. Spline-PLS	No. of components = 26	219	0.47
Quad. Spline-LASSO	$\sum(\lvert\theta\rvert)$=4.5×10^3	143	0.34
Quad. Spline-FSR	No. of variables = 8	139	0.54

Table 13.20: RMSEP-values for protein for some regularization methods combined with the quadratic B-spline basis-function regression.

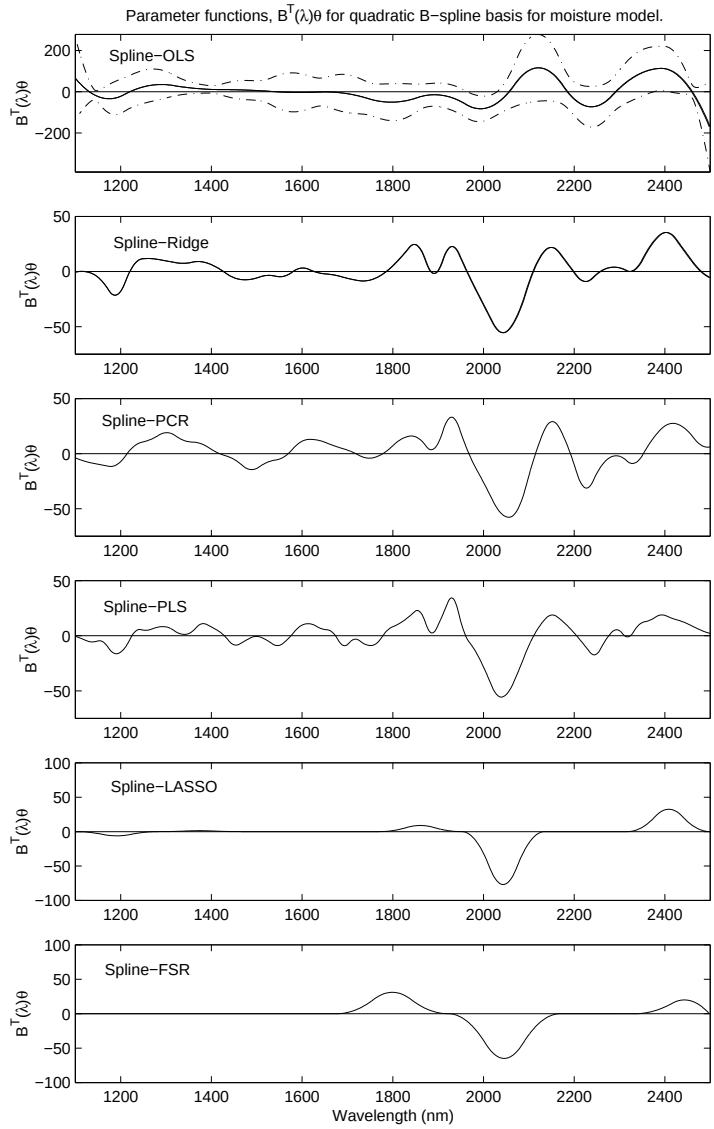

Figure 13.17: Estimated coefficient-functions using OLS, Ridge, PCR, PLS, LASSO, and FSR together with quadratic B-spline bases for the moisture model.

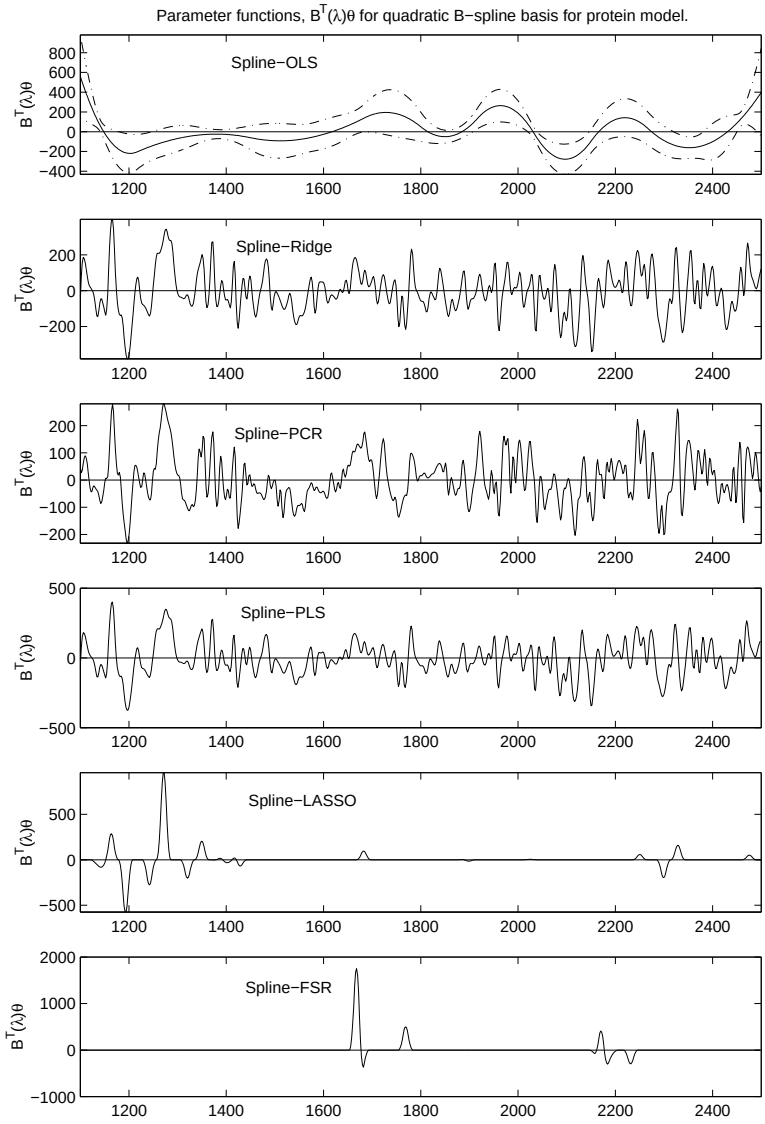

Figure 13.18: Estimated coefficient-functions using OLS, Ridge, PCR, PLS, LASSO, and FSR together with quadratic B-spline bases for the protein model.

Method	Regularization parameter	Knots	RMSEP
Cubic Spline-OLS		11	0.30
Cubic Spline-Ridge	$k = 2.02 \times 10^{-5}$	43	0.30
Cubic Spline-PCR	No. of components = 14	43	0.32
Cubic Spline-PLS	No. of components = 7	69	0.31
Cubic Spline-LASSO	$\sum(\lvert\theta\rvert)$=215.4	27	0.31
Cubic Spline-FSR	No. of variables = 5	37	0.29

Table 13.21: RMSEP-values for moisture for some regularization methods combined with the cubic B-spline basis-function regression.

Method	Regularization parameter	Knots	RMSEP
Cubic Spline-OLS		25	0.50
Cubic Spline-Ridge	$k = 1.0 \times 10^{-12}$	266	0.46
Cubic Spline-PCR	No. of components = 51	263	0.49
Cubic Spline-PLS	No. of components = 20	266	0.48
Cubic Spline-LASSO	$\sum(\lvert\theta\rvert)$=4.7×10^3	128	0.37
Cubic Spline-FSR	No. of variables = 8	153	0.56

Table 13.22: RMSEP-values for protein for some regularization methods combined with the cubic B-spline basis-function regression.

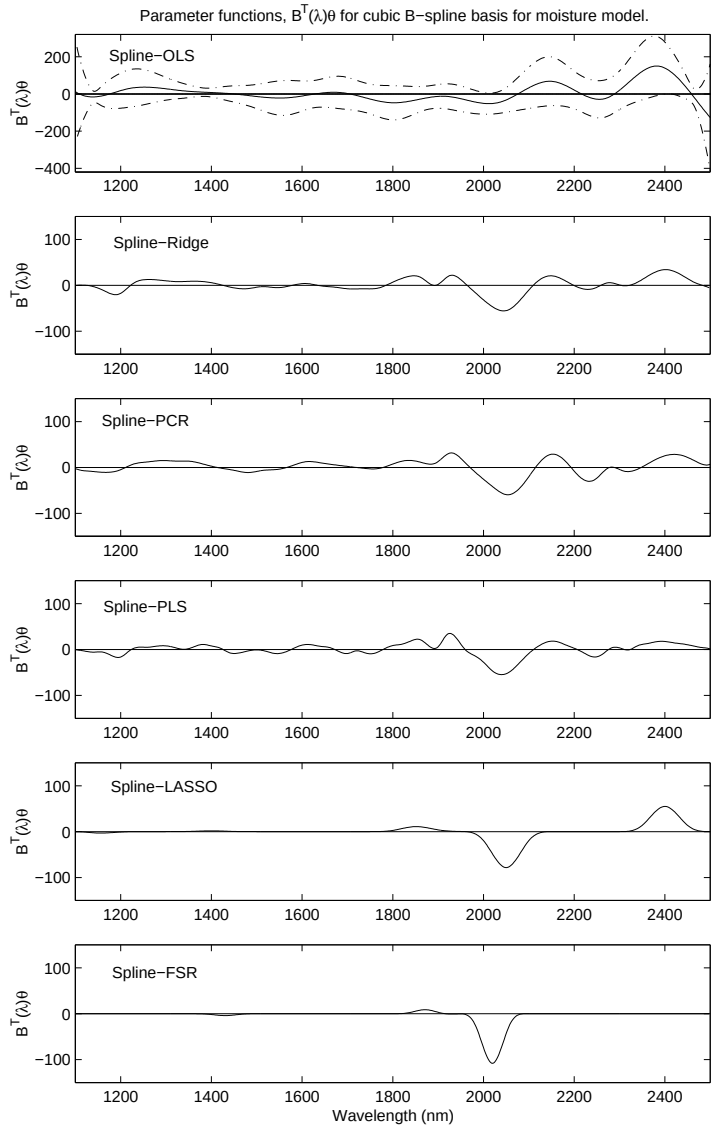

Figure 13.19: Estimated coefficient-functions using OLS, Ridge, PCR, PLS, LASSO, and FSR together with cubic B-spline bases for the moisture model.

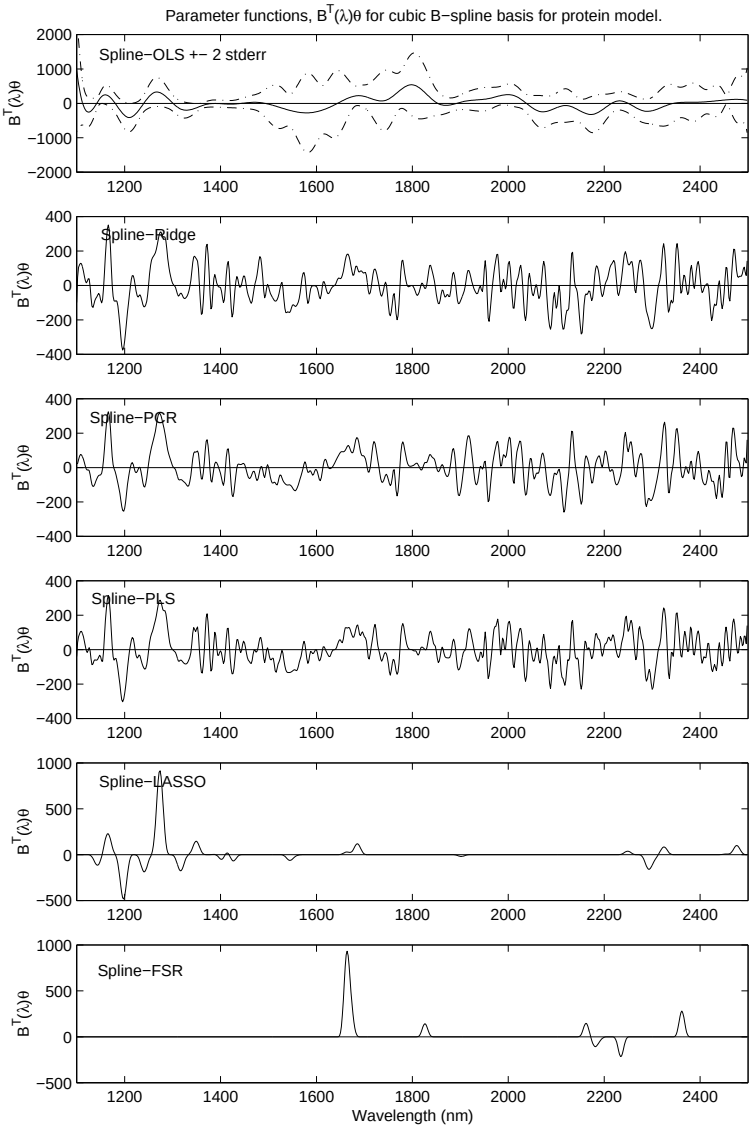

Figure 13.20: Estimated coefficient-functions using OLS, Ridge, PCR, PLS, LASSO, and FSR together with cubic B-spline bases for the protein model.

Summary of BFR results

The spline methods do in general not perform as well on the moisture data as the traditional methods. The B-spline basis functions are obviously not a suitable transformation. The simple OLS solution does however in a single case produce a competitive result. The predictive ability of the methods Ridge, PCR and PLS is in all cases diminished when used in combination with the B-spline bases. Forward selection is improved in combination with the linear and cubic B-splines. When LASSO is used in combination with the B-spline basis the prediction results are the same as for the normal LASSO.

For protein the picture is quite different, first of all the number of knots chosen by the methods is much higher than for moisture. The spline-OLS solution, which is constrained to select no more basis functions than the number of observations, clearly deviates from the other solutions. The spline methods perform very well on the protein data. LASSO is greatly improved and produce the best results. Ridge, PCR and PLS are also improved substantially in all cases. The spline-LASSO estimate now indicates that most of the variation of protein can be explained using the first part of the spectrum, whereas the spline-FSR estimates indicate the same regions as the FSR estimate. This is also reflected in the results for FSR. The best spline-method results in a 37% reduction of the RMSEP-value compared to the best otained previously.

13.2.9 Range selection using the BFR estimates

Moisture

The best of the spline results for the moisture case is the cubic Spline-FSR method. In Figure 13.19 the estimate for the cubic Spline-FSR method indicates that just one area of the spectrum is of interest, this contradicts the other estimates which indicate that important information is contained in the last part of the spectrum as well. Therefore the estimate from the cubic Spline-LASSO method is used to select a new and smaller range to which the methods MLLS, Ridge, PCR, PLS, LASSO and FSR will be applied. The range obtained from the estimate is

$$[1778 : 2128, \;\; 2298 : 2478]$$

There are 267 of the original 701 wavelengths contained in this range[2]. The results for using this range is shown in Table 13.23 and the parameter estimates can be seen in Figure 13.21. To check whether the results for the selected range are purely coincidential the methods have also been applied to the complementary set of wavelengths, see Table B.1 for the RMSEP-values. The RMSEP-values using the traditional methods on the

Method	Regularization parameter	RMSEP	% improvement
MLLS	$\|\|\hat{\boldsymbol{\beta}}\|\|_2 = 1.53 \times 10^3$	0.29	-7%
Ridge	$k = 1.42 \times 10^{-6}$	0.26	4%
PCR	No. of components = 12	0.26	4%
PLS	No. of components = 5	0.27	4%
LASSO	$\sum(\|\theta\|) = 8.11 \times 10^3$	0.30	0%
FSR	No. of variables = 7	0.34	0%

Table 13.23: RMSEP-values for moisture on the reduced range. The %-wise reduction in the RMSEP-value for each of the methods is also tabulated.

full spectrum were not improved in combination with the B-spline bases, but using the spline estimate to select a new set of wavelengths did improve the methods slightly. The slight improvement could be the result of a simpler model i.e. decreased variance of the parameter estimates, so whether or not the spline-LASSO estimate actually identifies an important set of wavelengths is not certain.

[2]See Figure B.4 for graphical display of the range

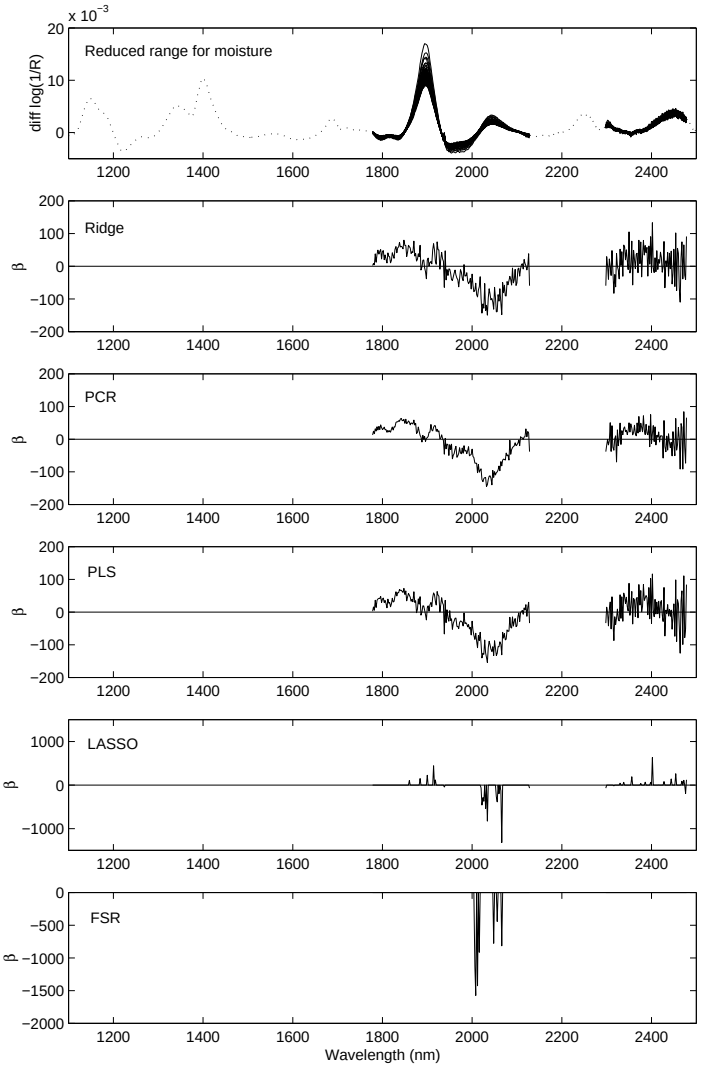

Figure 13.21: The 100 NIR spectra on the reduced range and the parameter estimates for Ridge, PCR, PLS, LASSO and forward selection regression for the moisture model.

Protein

The best of the spline results for the protein case come from the Spline-LASSO method. The best result is obtained in combination with the quadratic B-spline basis. In Figure 13.18 the estimate from the quadratic Spline-LASSO method is used to select a new and smaller range to which the methods MLLS, Ridge, PCR, PLS, LASSO and FSR will be applied. It turns out that the small peaks in the estimate carry only little or none information for the prediction of protein. This observation makes it difficult to use an automized procedure for selecting the range. By selecting the areas where the estimate has the largest peaks the following range is obtained:

$$[1122 : 1206, \;\; 1228 : 1284]$$

There are 72 of the original 701 wavelengths contained in this range[3]. The results for using this range is shown in Table 13.24 and the parameter estimates can be seen in Figure 13.22. To check whether the results for the selected range are purely coincidential the methods have also been applied to the complementary set of wavelengths, see Table B.2 for the RMSEP-values. For Ridge, PCR and PLS the best results are obtained when applied

Method	Regularization parameter	RMSEP	% improvement
MLLS	$\|\|\hat{\boldsymbol{\beta}}\|\|_2 = 2.83 \times 10^4$	0.68	-26%
Ridge	$k = 3.0 \times 10^{-8}$	0.32	41%
PCR	No. of components = 16	0.30	47%
PLS	No. of components = 6	0.29	47%
LASSO	$\sum(\|\theta\|) = 5.46 \times 10^4$	0.38	32%
FSR	No. of variables = 34	0.38	30%

Table 13.24: RMSEP-values for protein on the reduced range. The %-wise reduction in the RMSEP-value for each of the methods is also tabulated.

to the reduced range, they are also the overall best results for prediction of protein. FSR is also greatly improved when forced to select variables from the first part of the spectrum. The MLLS solution has for both moisture and protein the same predictive power in the full spectrum situation. The above result clearly shows that when uninformative parts of the spectrum are removed, the shrinkage induced by the other regularization methods

[3]See Figure B.4 for graphical display of the range

result in estimates with much better predictive ability than the MLLS solution[4].

[4]See Figure B.5 and B.6 for the MLLS solutions for respectively moisture and protein when applied to the reduced range.

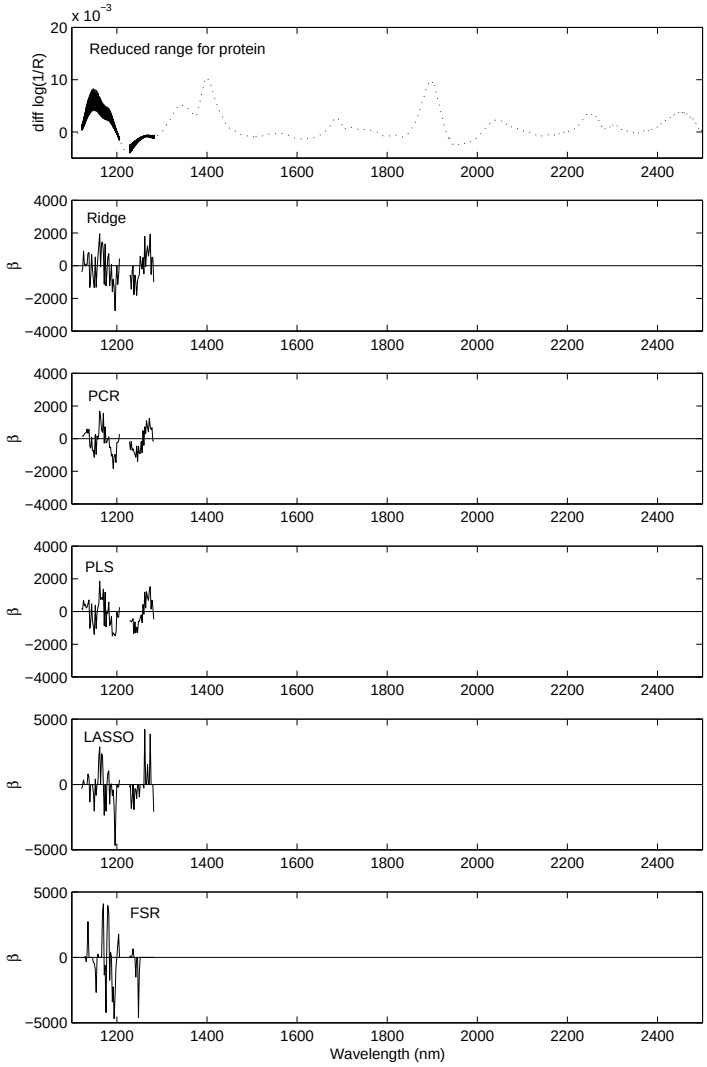

Figure 13.22: The 100 NIR spectra on the reduced range and the parameter estimates for Ridge, PCR, PLS, LASSO and forward selection regression for the protein model.

13.3 Summary for the wheat example

The results for the wheat example are summarized in Table 13.25 and 13.26. When used for predicting moisture the methods generally performed worse or just as good when used in combination with the B-spline bases. Since Ridge regression is among the best methods, the information regarding moisture is probably spread throughout the entire spectrum. Shrinking the solutions away from the MLLS solution by using e.g. Ridge, PCR or PLS does not have any predictive advantage. Even though the selection of a smaller range does result in slight improvements, it is very likely just to be the result of a simpler model, which compensates for the loss of information by being more robust. This is also indicated in Table B.1 which shows that the methods do almost as well on the complementary set of wavelengths.

For protein none of the methods do better than MLLS, this is also indicated by the small regularization for most of the methods. The spline methods do in all cases improve the traditional methods based on the full spectrum. This indicates that a smoothing of the parameter estimates indeed is desirable in this case. When using the Spline-LASSO estimate as a tool to select certain regions of the spectrum all results, except for LASSO, are further improved. The best method on the reduced range result in a 47% reduction in the RMSEP-value compared to the best full-spectrum method.

Method	Regularization param.	Knots	RMSEP
MLLS	$\|\hat{\boldsymbol{\beta}}\|_2 = 1.01 \times 10^3$		0.27
Ridge	$k = 1.79 \times 10^{-6}$		0.27
PCR	No. of comps. = 23		0.27
CPCR	No. of comps. = 17		0.28
FSPCR	No. of comps. = 27		0.26
PLS	No. of comps. = 7		0.28
CSR	Eigv/factors = 35/34		0.28
FSR	No. of var. = 7		0.34
LASSO	$\sum(\|\beta_i\|) = 8.09 \times 10^3$		0.30
Linear Spline-OLS		13	0.31
Linear Spline-Ridge	$k = 2.95 \times 10^{-5}$	40	0.30
Linear Spline-PCR	No. of comps. = 14	45	0.32
Linear Spline-PLS	No. of comps. = 8	31	0.33
Linear Spline-LASSO	$\sum(\|\theta\|)$= 546.2	80	0.30
Linear Spline-FSR	No. of var. = 12	58	0.31
Quad. Spline-OLS		13	0.27
Quad. Spline-Ridge	$k = 2.2 \times 10^{-5}$	38	0.30
Quad. Spline-PCR	No. of comps. = 14	38	0.32
Quad. Spline-PLS	No. of comps. = 7	58	0.31
Quad. Spline-LASSO	$\sum(\|\theta\|) = 167.7$	23	0.31
Quad. Spline-FSR	No. of var. = 3	17	0.35
Cubic Spline-OLS		11	0.30
Cubic Spline-Ridge	$k = 2.02 \times 10^{-5}$	43	0.30
Cubic Spline-PCR	No. of comps. = 14	43	0.32
Cubic Spline-PLS	No. of comps. = 7	69	0.31
Cubic Spline-LASSO	$\sum(\|\theta\|)$=215.4	27	0.31
Cubic Spline-FSR	No. of var. = 5	37	0.29
Reduced MLLS	$\|\hat{\boldsymbol{\beta}}\|_2 = 1.53 \times 10^3$		0.29
Reduced Ridge	$k = 1.42 \times 10^{-6}$		0.26
Reduced PCR	No. of comps. = 12		0.26
Reduced PLS	No. of comps. = 5		0.27
Reduced LASSO	$\sum(\|\theta\|) = 8.11 \times 10^3$		0.30
Reduced FSR	No. of var. = 7		0.34

Table 13.25: Results for the wheat example for moisture.

Method	Regularization param.	Knots	RMSEP
MLLS	$\|\hat{\boldsymbol{\beta}}\|_2 = 2.55 \times 10^3$		0.54
Ridge	$k = 6.28 \times 10^{-7}$		0.54
PCR	No. of comps. = 29		0.57
CPCR	No. of comps. = 30		0.56
FSPCR	No. of comps. = 36		0.60
PLS	No. of comps. = 9		0.55
CSR	Eigv./factors = 50/34		0.54
FSR	No. of var. = 12		0.54
LASSO	$\sum(\|\beta_i\|) = 1.33 \times 10^4$		0.56
Linear Spline-OLS		14	0.77
Linear Spline-Ridge	$k = 1.0 \times 10^{-12}$	192	0.47
Linear Spline-PCR	No. of comps. = 51	191	0.46
Linear Spline-PLS	No. of comps. = 27	191	0.44
Linear Spline-LASSO	$\sum(\|\theta\|)$=3.43×10^3	106	0.37
Linear Spline-FSR	No. of var. = 7	122	0.53
Quad. Spline-OLS		12	0.60
Quad. Spline-Ridge	$k = 1.0 \times 10^{-12}$	219	0.47
Quad. Spline-PCR	No. of comps. = 50	284	0.51
Quad. Spline-PLS	No. of comps. = 26	219	0.47
Quad. Spline-LASSO	$\sum(\|\theta\|)$=4.5×10^3	143	0.34
Quad. Spline-FSR	No. of var. = 8	139	0.54
Cubic Spline-OLS		25	0.50
Cubic Spline-Ridge	$k = 1.0 \times 10^{-12}$	266	0.46
Cubic Spline-PCR	No. of comps. = 51	263	0.49
Cubic Spline-PLS	No. of comps. = 20	266	0.48
Cubic Spline-LASSO	$\sum(\|\theta\|)$=4.7×10^3	128	0.37
Cubic Spline-FSR	No. of var. = 8	153	0.56
Reduced MLLS	$\|\hat{\boldsymbol{\beta}}\|_2 = 2.83 \times 10^4$		0.68
Reduced Ridge	$k = 3.0 \times 10^{-8}$		0.32
Reduced PCR	No. of comps. = 16		0.30
Reduced PLS	No. of comps. = 6		0.29
Reduced LASSO	$\sum(\|\theta\|) = 5.46 \times 10^4$		0.38
Reduced FSR	No. of var. = 34		0.38

Table 13.26: Results for the wheat example for protein.

Chapter 14
Summary

Although the examples presented here do not cover all possible calibration situations, they give a reasonable idea of how the considered calibration techniques perform.

The solutions produced by Ridge, PCR, PLS and CSR when applied directly to the gasoline example are all very similar and they all give similar RMSEP-values. The CPCR solution looks more like the MLLS solution, this is due to the inclusion of a principal component corresponding to one of the smaller eigenvalues. Nevertheless the exclusion of other principal components results in a better prediction compared to MLLS. A characteristic property of these methods is that none of them will produce an estimate where any of the parameters are equal to zero.

The LASSO method has been shown to produce parameter estimates where some of the values are zero while others are quite large (compared to e.g. PLS or Ridge estimates). Ridge regression limits the squared length of the estimate, (L_2 norm), whereas LASSO limits the absolute length of the estimate, (L_1 norm). LASSO is conceptually placed between Ridge regression and subset selection. The subset selection procedure used here is forward selection. FSR finds a solution by selecting a set of explanatory variables of a specified size, that minimizes the ordinary least squares criteria. When LASSO and FSR is applied to the gasoline example they select variables within approximately the same range of the spectrum, except for the very

last part of the spectrum which is only included in the LASSO estimate. LASSO has the largest RMSEP-value and FSR is almost as good as Ridge, PCR and PLS.

In the wheat example NIR spectra are used to predict the amount of moisture and protein in wheat. When applied to predict moisture MLLS, Ridge, PCR, CPCR, FSPCR, PLS and CSR have more or less the same predictive ability. The solutions for LASSO and FSR are quite different which is reflected in the RMSEP-values. LASSO does almost as good as the other methods whereas the 7 variables chosen by FSR from the cross-validation result in the largest RMSEP-value.

For prediction of protein LASSO and FSR choose quite different subsets of variables but both methods predict just as well as the other methods. One should note that for both moisture and protein the MLLS solution gives just as good predictions as any of the other methods applied to the full spectrum. Shrinking the solution away from the MLLS solution using Ridge, PCR or PLS does not have any predictive advantage.

The basis function regression using B-spline basis functions was applied to the gasoline data. The spline results are similar whether linear, quadratic or cubic B-spline bases are used. The number of internal knots chosen is relatively small. This results in very smooth parameter functions. For Spline-OLS, -Ridge, -PCR and -PLS the estimates look very similar. The prediction results are also similar and in all cases better than what was obtained using the previous methods. For Spline-LASSO and -FSR the number of internal knots chosen is much larger than for the other methods. This results in estimates where small parts of the function is different from zero. Both Spline-LASSO and -FSR have sligthly smaller RMSEP-values than the other spline methods.

When using the parameter function from Spline-FSR to select ranges of the spectrum and thereafter applying Ridge, PCR, PLS and LASSO to the reduced set of explanatory variables smaller RMSEP-values are obtained compared to the full spectrum case. For FSR the same result is obtained because the wavelengths selected from the full spectrum are almost all contained in the reduced set of wavelengths. The RMSEP-values obtained using the spline methods are the same as the ones obtained using the reduced range.

When predicting moisture using the basis function regression no reduction in the RMSEP-value is obtained. It is possible that another basis function

would be more appropriate (e.g. wavelet basis). When the Spline-LASSO estimate is used to select a reduced set of wavelengths only a small reduction in the RMSEP-value is achieved and this is most likely just the effect of a simpler model which compensates for the loss of information by being more robust.

Smoothing the estimate is an advantage when predicting protein. In all cases reduction in the RMSEP-value is achieved. The best result is for Spline-LASSO which results in a 37% reduction of the best RMSEP-value. Relatively many basis functions are needed to predict well. This shows the importance of combining a regularization method with the basis function regression and thereby being able to choose a number of basis functions which exceed the number of samples. The Spline-OLS result clearly reflects the need for many basis functions in this case. The selection of a smaller set of wavelengths is very successfull here which is reflected in the 47% reduction in the best RMSEP-value from the full spectrum case.

14.1 A calibration strategy

A strategy to select a suitable calibration technique for a given calibration problem will be suggested here. It should be mentioned that the proposed strategy cannot be considered general, as it is based on a limited, though relatively representative, set of data.

The strategy is based on the grouping among the methods revealed by the results above.

1. • Ridge
   • PLS
   • PCR
2. • LASSO
   • FSR (or any other variable selection method)
3. • Spline-LASSO
   • Spline-FSR (or any other variable selection method)

The proposed strategy is simply to use one method from each group presented above. The reason for using methods from group 1 and 3 is a consequence of the results obtained by the gasoline and wheat example. The reason for also using a method from group 2 is to cover the case where only a few explanatory variables carry all the information of variation in

the response variable.[1] If a good result is obtained using one of the spline methods it is implied that a smaller range should be extracted and tested on one of the methods from both group 1 and 2. Thereby a simpler and more robust model can be obtained.

14.2 Further enhancements

- As mentioned in Section 12.5 the B-spline basis approach is just one solution to the basis function regression. Another solution would be to use wavelet basis functions together with methods like e.g. LASSO. Since wavelets cover a large range of scales and positions, they may be more appropriate than B-splines. Polynomials or cosinusoids are also possible alternatives.

- As yet another improvement an adaptive knot-placement procedure could be developed, thereby making non-equidistant knot placement possible.

- Finally, make the basis function regression work on the reduced range if it is not continous.

[1] An example where only 3 out of 926 explanatory variables yield the best prediction is found in [64]

Chapter 15
Conclusion

The aim has been to present some of the traditional and newest methods for multivariate calibration and compare the quality of prediction obtained with these methods. Furthermore, a new method that replaces the linear combination of the spectral values with an integral over the range of the wavelengths of an unknown coefficient-function multiplied by the spectral measurements, has been introduced. The idea is not new, the approach was first suggested by Hastie and Mallows, [22]. Marx and Eilers, [41], project the spectral measurements onto a moderate number of equally spaced B-spline bases. This approach is very similar to the approach presented here. However, the main difference here is that the number of basis functions is not restricted to be less than the number of observations.

The NIR data sets analyzed here, gasoline and wheat, has been published as intended reference data sets, [30]. Although this analysis is far from being exhaustive, the data sets do represent typical calibration problems that can be encountered in practice.

The conclusions made are only valid for the calibration situation studied, i.e., when new samples will be situated within the calibration domain. The results for Ridge, PCR and PLS are considered as benchmark results for comparison with other calibration techniques. They are the most widely used methods for calibration. Numerous studies have shown that these methods produce similar estimates and prediction results when applied to problems involving data with high collinearity in which the variance of the

estimate tends to dominate the bias, see e.g. [7], [15], [56] or [63]. The results from the examples presented here confirm this conclusion. The results for the different selection strategies for PCR indicate that the top-down selection strategy is the most stable, that is also concluded by Kalivas [36]. Cyclic subspace regression, CSR, is most importantly a simple algorithm that provides not only solutions for PCR, PLS and MLLS but also a finite number of other related methods. In the examples presented here the CSR solution always obtains results that are just as good as for Ridge, PCR and PLS. The newest developments of LASSO by Osborne [49], has been presented here. The new theory leads to a very fast algorithm that makes it possible to perform the calibration on a standard PC. Applying LASSO directly to the examples here do not result in better predictions than by using Ridge, PCR and PLS. In [64] it has been shown that LASSO works better than Ridge and PLS when only a moderate number of wavelengths are needed to predict the response variable. FSR also works best if all the variation of the response variable can be described by just a few explanatory variables. The new method presented here, basis function regression, leads in two of the cases to better results than the benchmark methods. Here B-spline basis functions have been used. Generally it works best in combination with either LASSO or FSR. It has been shown that the estimates resulting from combining LASSO or FSR with the B-spline bases can be used to identify smaller parts of the entire spectrum that contains explanatory variables which are important for predicting the response variable. Large improvements of the benchmark results can sometimes be gained by applying e.g. Ridge to this reduced set of explanatory variables.

Based on the results a calibration strategy has been proposed with the basis function regression as an important new tool.

All the methods presented here have been implemented in Matlab and will be made publicly available on http://www.imm.dtu.dk/~hoe.

Finally I hope this thesis can act as a help to future students of this area by being used as an introduction to the traditional multivariate methods of calibration and as a stepstone to future research into new methods or improvements of the existing.

Appendix A
Gasoline

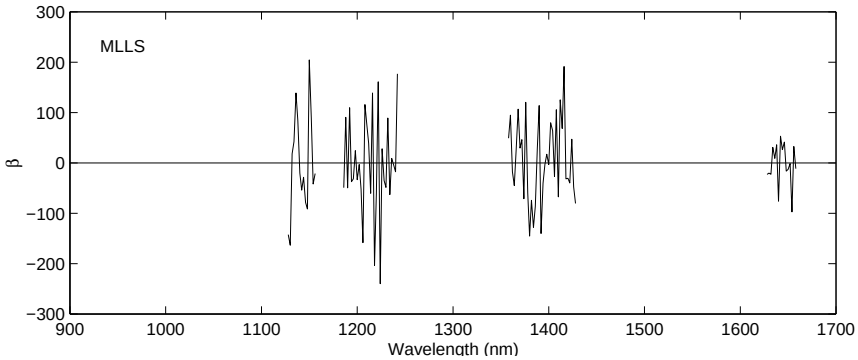

Figure A.1: Parameter estimates, $\hat{\boldsymbol{\beta}}$, for MLLS applied on the reduced range.

Full spectrum	Reduced Spectrum
1208	1208
1196	1196
1214	1214
1190	1190
1216	1216

1192	1192
1210	1210
1194	1194
1206	1206
1362	1362
1234	1234
1360	1360
1236	1236
1358	1358
1238	1238
1356	1364
1244	1426

Table A.2: Wavelengths selected by the forward selection method.

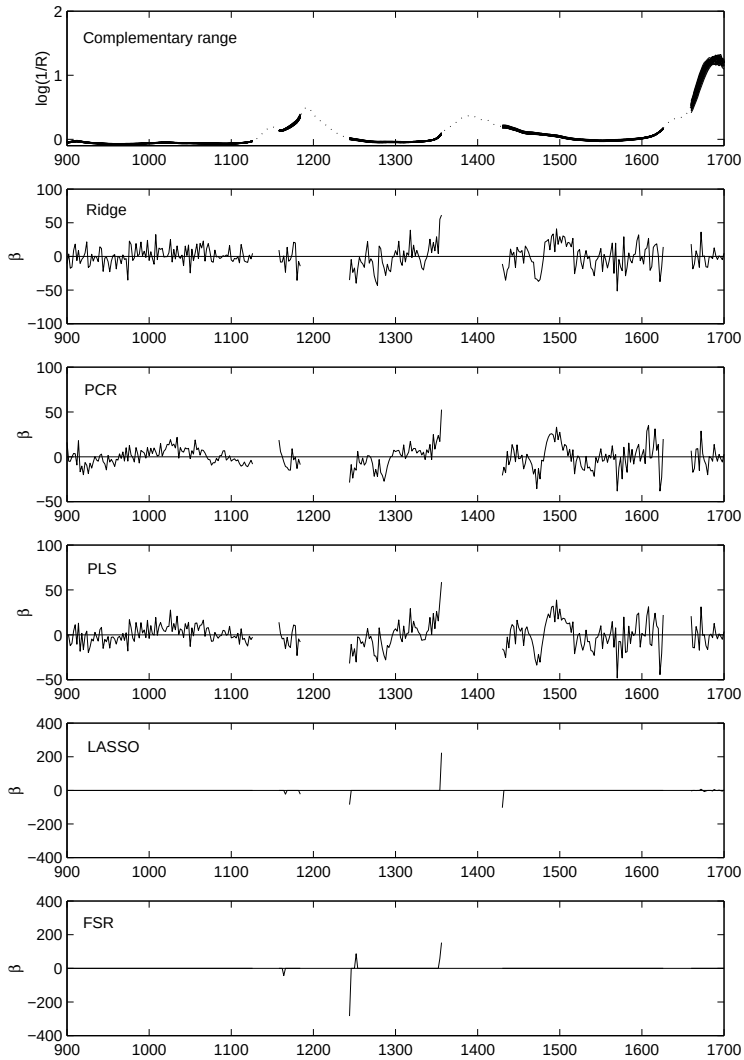

Figure A.2: The 60 NIR spectra for the complementary set of wavelengths, together with the parameter estimates for Ridge, PCR, PLS, LASSO and forward selection regression.

Method	Regularization parameter	RMSEP	% improvement
Ridge	$k = 1.84 \times 10^{-6}$	0.27	-13%
PCR	No. of components = 43	0.26	-13%
PLS	No. of components = 23	0.27	-17%
LASSO	$\sum(\lvert\theta\rvert) = 497.70$	0.30	-11%
FSR	No. of variables = 5	0.31	-24%

Table A.1: RMSEP-values for some regularization methods on the complementary set of wavelengths.

Appendix B
Wheat

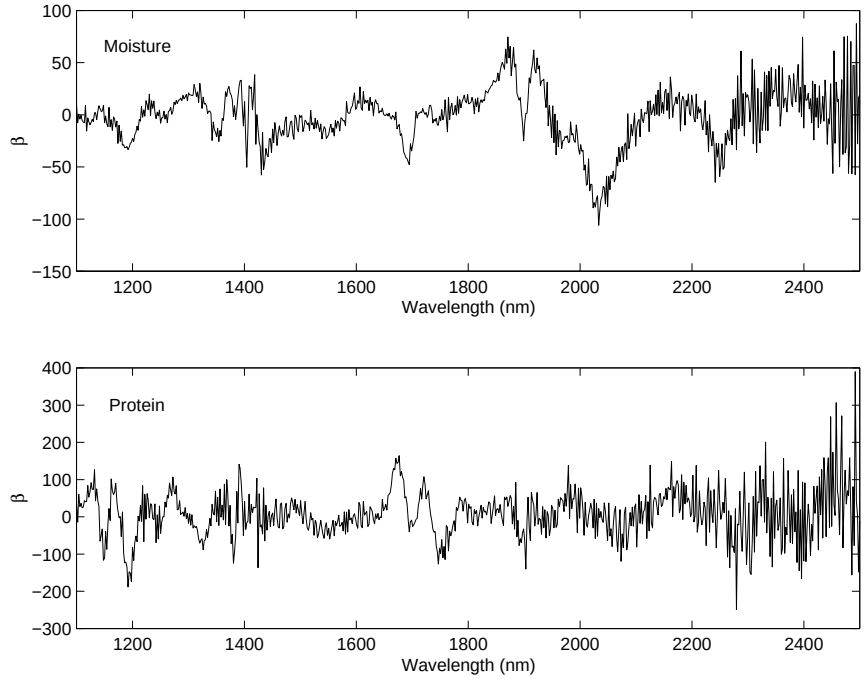

Figure B.1: Parameter estimates, $\hat{\boldsymbol{\beta}}$, for PCR.

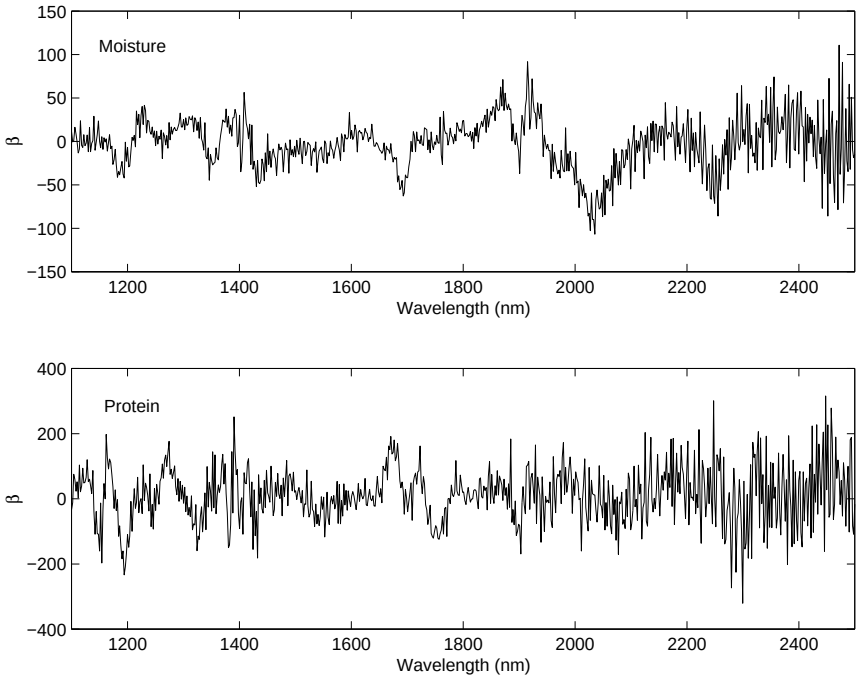

Figure B.2: Parameter estimates, $\hat{\boldsymbol{\beta}}$, for CPCR.

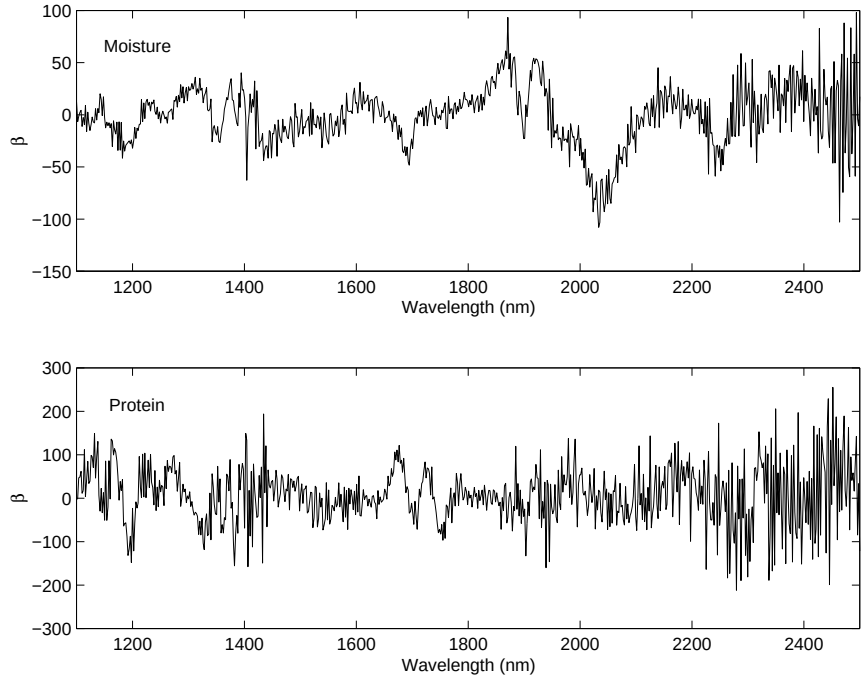

Figure B.3: Parameter estimates, $\hat{\boldsymbol{\beta}}$, for FSPCR

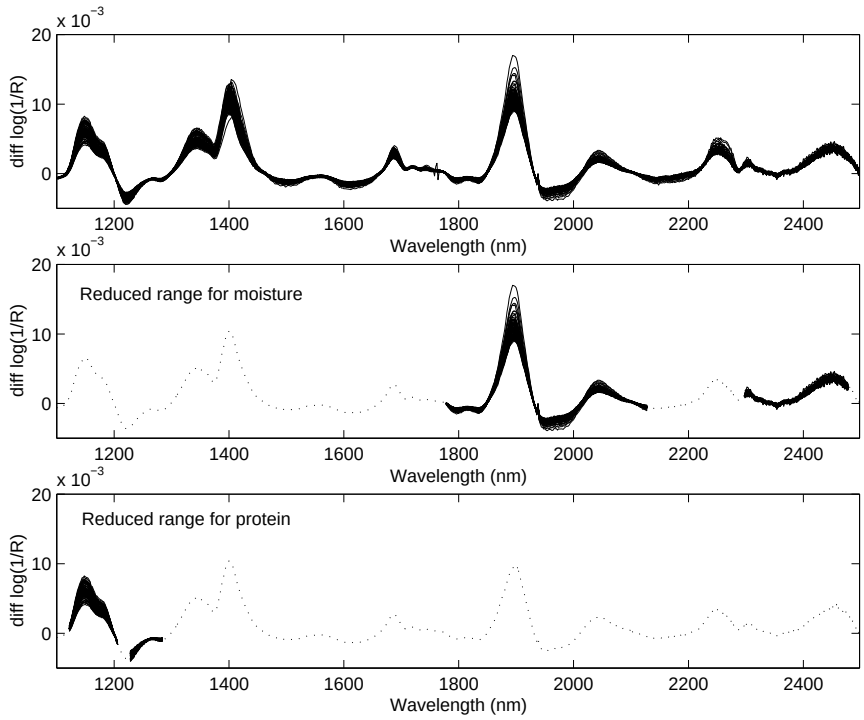

Figure B.4: The first-order differenced spectra (top figure), the selected range for moisture (middle figure) and the selected range for protein (bottom figure).

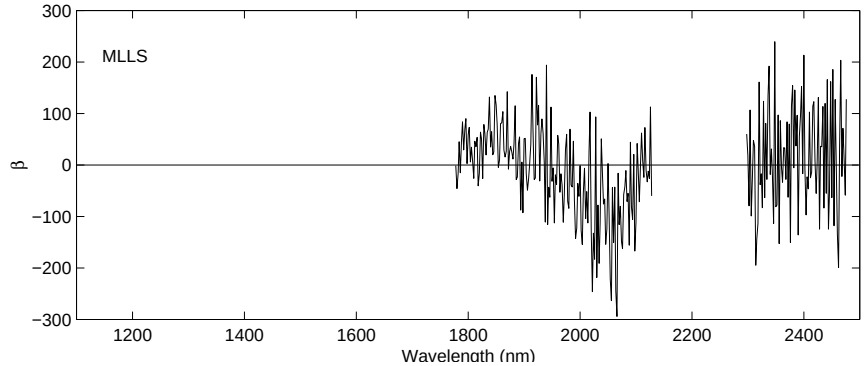

Figure B.5: Parameter estimates, $\hat{\boldsymbol{\beta}}$, for MLLS applied on the reduced range for moisture.

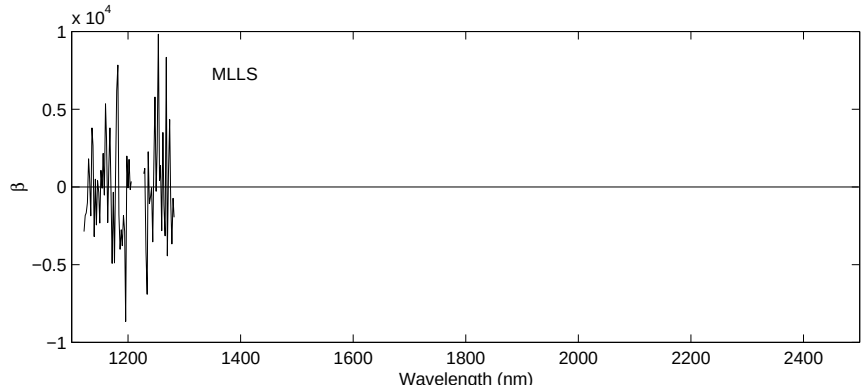

Figure B.6: Parameter estimates, $\hat{\boldsymbol{\beta}}$, for MLLS applied on the reduced range for protein.

Method	Regularization parameter	RMSEP	% improvement
Ridge	$k = 8.69 \times 10^{-7}$	0.28	-4%
PCR	No. of components = 29	0.28	-4%
PLS	No. of components = 7	0.28	0%
LASSO	$\sum(\lvert\theta\rvert) = 9.46 \times 10^3$	0.30	0%
FSR	No. of variables = 13	0.32	6%

Table B.1: RMSEP-values for moisture on the complementary set of wavelengths.

Method	Regularization parameter	RMSEP	% improvement
Ridge	$k = 6.87 \times 10^{-7}$	0.62	-15%
PCR	No. of components = 30	0.64	-12%
PLS	No. of components = 8	0.61	-11%
LASSO	$\sum(\lvert\theta\rvert) = 1.49 \times 10^4$	0.68	-21%
FSR	No. of variables = 9	0.57	-6%

Table B.2: RMSEP-values for protein on the complementary set of wavelengths.

Appendix C
Some Matlab functions

C.1 Ridge Regression

```
function [b] = ridge(X,y,k)

% function [b] = ridge(X,y,k)
% Input:  X is a (n x p) matrix with p explanatory variables.
%         y is a (n x 1) vector with the response variables.
%         k is the Ridge parameter.
% Output: b is the parameter estimate for Ridge Regression.

[n,p] = size(X);

[n1,collhs] = size(y);

if n~=n1,
    error('The number of rows in Y must...
           equal the number of rows in X.');
end

b = inv(X'*X + k*eye(p))*X'*y;
```

C.2 Principal Components Regression

```
function [b] = pcr(X,y,comp)

% function [b] = pcr(X,y,comp)
% Input:  X is a (n x p) matrix with p explanatory variables
%         y is a (n x 1) vector with the response variables
%         comp is the number of PC's to be used.
% Output: b is the parameter estimate for the
%         principal component regression.

[n,p] = size(X);

[n1,collhs] = size(y);

if n~=n1,
    error('The number of rows in Y must...
          equal the number of rows in X.');
end

[U,S,V] = svd(X);
lambda = diag(diag(S));

b=V(:,1:comp)*inv(lambda(1:comp,1:comp))*U(:,1:comp)'*y;


function [b,index,Y] = cpcr(X,y,comp)

% function [b] = cpcr(X,y,comp)
% Input:  X is a (n x p) matrix with p explanatory variables
%         y is a (n x 1) vector with the response variables
%         comp is the number of PC's to be used.
% Output: b is the parameter estimate for the principal
%         component regression using a correlation strategi
%         on the PC's and the y's.
%         index contains the number corresponding to the
%         size of the eigenvalues.
%         Y contains the correlation values.
```

```
[n,p] = size(X);

[n1,collhs] = size(y);

if n~=n1,
    error('The number of rows in Y must...
           equal the number of rows in X.');
end

index=[];
[U,S,V] = svd(X);

lambda = diag(diag(S));
nmy=norm(y);
r=rank(X);

for i=1:r
   corr(i) = abs((U(1:n1,i)'*y)/(norm(U(1:n1,i))*nmy));
end
[Y,I] = sort(corr');
index=I(end-comp+1:end);

b=V(:,I(end-comp+1:end))*inv(lambda(I(end-comp+1:end),...
  I(end-comp+1:end)))*U(:,I(end-comp+1:end))'*y;
```

C.3 Partial Least Squares Regression

```
function [b] = pls(X,Y,comp)

% Input:  X is a (n x p) matrix with p explanatory variables
%         y is a (n x 1) vector with the response variables
%         comp is the number of components to be used.
% Output: b is the parameter estimate for the
%         partial least squares regression.

[n,p] = size(X);
[n1,p1] = size(Y);
```

```
if n~=n1,
    error('The number of rows in Y must...
          equal the number of rows in X.');
end

x = X;
y = Y;
I = eye(p);
w = zeros(p,comp);
u = zeros(n,comp);

for i = 1:comp
   w(:,i) = x'*y;
   w(:,i) = w(:,i)/norm(w(:,i));
   r(:,i) = x * w(:,i);
   x      = x -  r(:,i) * w(:,i)';
   y      = y - r*(r\y);
end

cw = r\Y;
b  = w * cw;
```

C.4 Cyclic Subspace Regression

```
function [b] = csr(X,y)

% function [b] = csr(X,y)
% Input:  X is a (n x p) matrix with p explanatory variables
%         y is a (n x 1) vector with the response variables
% Output: b is the parameter estimate for the
%         cyclic subspace regression.

% Step 1: Perform SVD on X
[U,S,V] = svd(X);
[n,m] = size(X);
Xs = X;
ys = y;
```

```
% Step 2
k = rank(X);

for l=1:k
   % Perform CSR algorithm
   X = Xs;
   P = U(:,1:l)*U(:,1:l)';
   y = P*ys;
   W = zeros(length(y),1);
   Z = zeros(length(X'*y),1);
   for i = 1:l
      a = X'*y;
      z = a/norm(a);
      Z = [Z z];
      b = X*z;
      w = b/norm(b);
      W = [W w];
      X = (eye(size(U))-w*w')*X;
      y = (eye(size(U))-w*w')*y;
   end
   W = W(:,2:l+1);
   Z = Z(:,2:l+1);
   for j=1:l
      XP = W(:,1:j)*W(:,1:j)'*Xs*Z(:,1:j)*Z(:,1:j)';
      b = pinv(XP)*ys;
end
```

C.5 Forward Selection Regression

```
function [b,SS] = fsr(X,y,k)

% FUNCTION [b] = fsr(X,y)
% Input:  X is a (n x p) matrix with p explanatory variables
%         y is a (n x 1) vector with the response variables
%         k is the number of variables to choose.
% Output: b is the parameter estimate for the
%         Forward selection regression.
```

```
%         SS is a vector containing the chosen variables.

[n,p] = size(X);

[n1,collhs] = size(y);

if n~=n1,
    error('The number of rows in Y must equal ...
           the number of rows in X.');
end

nmy = norm(y);
set = 1:p;
pcorr = zeros(length(set),1);
for i = set
corr(i)  = norm(X(:,i)'*y,1)^2/(norm(X(:,i))^2*nmy);
end

[Y1,I1] = max(abs(corr));
SS(1) = I1;
Xbest = X(:,I1);
betabest = inv(Xbest'*Xbest)*Xbest'*y;

set = setdiff(set,I1);

for j = 2:k

   for i = 1:length(set)
       taeller(i) = norm( (X(:,set(i)) - ...
                    ( inv(X(:,set(i))'*X(:,set(i)) ) *...
                    X(:,set(i))' * Xbest * Xbest))' * ...
                    ( y-Xbest*betabest ),1 )^2;

       naevner(i) = norm(X(:,set(i)) - ( inv(X(:,set(i))'*...
                         X(:,set(i)) ) * X(:,set(i))' *...
                         Xbest * Xbest ))^2;

       pcorr(i,1) = (taeller(i)/naevner(i))/...
                     norm(y-Xbest*betabest)^2;
```

```
   end

   [Y2,I2] = max(abs(pcorr));
   Xbest = X(:,set(I2));
   betabest = inv(Xbest'*Xbest)*Xbest'*y;
   SS(j) = set(I2);
   set = setdiff(set,set(I2));
   pcorr = zeros(length(set),1);
end
b=inv(X(:,SS)'*X(:,SS))*X(:,SS)'*y;
```

C.6 Adaptive Ridge Regression

```
function [beta,msr] = arrfit(X,y,lambda,precision)
%ARRFIT Adaptive Ridge Regression linear fit to data
%       ARRFIT(X,y,lambda) finds the coefficients BETA, of
%       the linear fit to the data, X(i,:)*BETA ~= y(i),
%       minimizing the following expression:
%
%       sum((X*BETA-y).^2) + lambda * sum(abs(BETA))^2
%
%       [BETA,MSR] = ARRFIT(X,y,lambda,precision) returns the
%       coefficients BETA and the mean squares residuals MSR.
%
%       X is the vector or matrix of input data, y is the
%       vector of output data.
%       lambda (default=1) is a scalar or vector of
%       penalization coefficients. If lambda is a vector,
%       each column of BETA and MSR corresponds to the
%       respective value of lambda. precision (default=1e-2)
%       is an optional parameter of the procedure. It is a
%       measure of the absolute and relative precisions
%       required for BETA.

%       22/06/98 Y. Grandvalet
```

```
if nargin < 4;
   precision = 1e-2;
   if nargin < 3;
      lambda = 1;
      if nargin < 2;
        error('ARRFIT requires at...
               least two input arguments.');
      end;
   end;
end;
precision = precision.^2;

% Check that matrix (X) and vector (y)
% have compatible dimensions

[n,d]   = size(X);
[ny,dy] = size(y);
if ny~=n,
    error('The number of rows in y must equal...
           the number of rows in X.');
end
if dy ~= 1,
    error('y must be a vector, not a matrix');
end

% Check that (lambda) has correct dimensions

[nl,dl] = size(lambda);
if dl ~= 1 & nl ~= 1,
    error('lambda must be a scalar or vector.');
end
[nl] = max([nl,dl]);

% Check that (precision) has correct dimensions

if length(precision) ~= 1,
    error('precision must be a scalar.');
end
```

```
% Initializations

beta = zeros(d,nl);

XX = (X'*X);
Xy = (X'*y);

for i=1:nl
    if  lambda(i)==Inf;
        beta(:,i) = zeros(d,1);
    else;
        Lambda = lambda(i)*ones(d,1);
        U = chol(XX + diag(Lambda));
        betanew = U\(U'\Xy);
        stop    = 0;
        while (~stop);
             betaold      = betanew;
             normbetaold  = abs(betaold)./mean(abs(betaold));
             ind          = find( normbetaold > precision );
             Lambda(ind)  = (d*lambda(i))./normbetaold(ind);
             betanew      = zeros(d,1);
             U            = chol(XX(ind,ind) +...
                            diag(Lambda(ind)));
             betanew(ind) = U\(U'\Xy(ind));
             stop         = max( abs(betaold-betanew)./...
                            (1+abs(betanew)) ) < precision;
        end
        beta(:,i) = betanew;
    end;
end;

if nargout > 1
   msr = sum( (X*beta - y(:,ones(nl,1))).^2 )/n;
end;
```

Bibliography

[1] Akaike, H. (1974). *A New Look at Statistical Model Identification*, IEEE Transactions on Automatic Control, 19, 716-723.

[2] Almøy, T. (1994). *A simulation study on comparison of prediction methods when only a few components are relevant*, Computational Statistics & Data Analysis 21, 87-107.

[3] Björkström, A. Sundberg, R. (1996). *A generalized view on continuum regression.*, Research report no. 189, university of Stockholm .

[4] Boor, C. de (1978). *A Practical Guide to Splines*, Springer Verlag.

[5] Breiman, L. (1996). *Heuristics of instability and stabilization in model selection*, The Annals of Statistics, Vol 24, No. 6, pp. 2350-2383.

[6] Breiman, L. & Spector, P. (1992). *Submodel Selection and Evaluation in Regression. The X-Random Case*, International Statistics Review, Vol 60, No. 3, pp. 291-319.

[7] Brown, P.H. (1993). *Measurement, Regression, and Calibration*, Oxford Science Publications.

[8] Butler, N.A. Denham, M.C. (2000). *The peculiar shrinkage properties of partial least squares regression*, J.R. Statist. Soc. B. 62, part3, pp.585-593.

[9] Clark, D.I. & Osborne, M.R. (1988). *On Linear Restricted and Interval Least-Squares Problems*, IMA Journal of Numerical Analysis 8, 23-36.

[10] Conradsen, K. (1984). *En introduktion til statistik 2*, Forelæsningsnote, IMSOR, DTU.

[11] Draper, N.R. & Smith, H. (1981) *Applied Regression Analysis*, 2nd edn. Wiley, New York.

[12] Efron, B. (1983) *Estimating the Error Rate of a Prediction Rule: Improvement on Cross-validation*, Journal of the American Statistical Association. Vol. 78 pp. 316-331.

[13] Feiveson, Alan.H. (1994) *Finding the best regression subset by reduction in nonfull-rank cases*, SIAM J. Matrix Anal. Appl. Vol. 15, No. 1, pp. 194-204.

[14] Fletcher, R. (1993). *Practical methods of optimization*, John Wiley & Sons.

[15] Frank, I.E. Friedman, J.H. (1993). *A statistical view of some chemometrics regression tools*, Technometrics, 35, pp.109-135.

[16] Furnival, G.M. Wilson, R.W. (1974). *Regressions by Leaps and Bounds*, Technometrics, Vol. 16, pp.499-511.

[17] Gill, P.E. Murray, W. & Wright, M.H. (1981). *Practical Optimization*, London:Academic Press.

[18] Gill, P.E. Murray, W. & Wright, M.H. (1991). *Numerical Linear Algebra and Optimization*,Vol.1, Addison-Wesley Publishing Company.

[19] Golub, G.H. & Van Loan, C. (1989). *Matrix computations*, The Johns Hopkins University Press.

[20] Goutis, C. (1998). *Second-derivative functional regression with applications to near infra-red spectroscopy*, Journal of Royal Statistical Society B. Vol. 60, part 1, pp. 103-114.

[21] Grandvalet, Y (1998). *Least Absolute Shrinkage is Equivalent to Quadratic Penalization*, In Niklasson, L. Bodén, M. & Ziemske, T. editors, ICANN´98, volume 1 of Perspectives in Neural Computing, pages 201-206. Springer.

[22] Hastie, T.J. Mallows, C. (1993). *Comment on "A Statistical View of Some Chemometrics Regression Tools"*, Technometrics, vol. 35, no. 2. pp. 140-143.

[23] Hastie, T.J. Tibshirani, R.J. (1990). *Generalized additive models*, Chapman & Hall.

[24] Helland, I. S. (1988). *On the structure of partial Least Squares Regression*, Simulation and Computation, Vol. 17, No. 2, 581-607.

[25] Helland, I. S. Almøy, T. (1994). *Comparison of Prediction Methods When Only a Few Components Are Relevant.*, Journal of the American Statistical Association, Vol. 89, No. 426, Theory and Methods.

[26] Helland, I. S. Naes, T. (1993). *Relevant Components in Regression*, Scandinavian Journal of Statistics, Vol. 20, pp. 239-250.

[27] Helland, I. S. Naes, T. Isaksson, T. (1995). *Related versions of the multiplicative scatter correction method for preprocessing spectroscopic*

data, Chemometrics and Intelligent Laboratory Systems, 29, pp. 233-241.

[28] Hoerl, A. E. & Kennard, R. W. (1970). *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, 12, 55-67.

[29] Joliffe, I. T. (1982). *A note on the use of Principal Components in Regression*, Applied Statistics 31, No. 3, 300-303.

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

[31] Kalivas, J.H. (1999). *Cyclic Subspace Regression with analysis of the hat matrix*, Chemometrics and Intelligent Laboratory Systems 45, 215-224.

[32] Kalivas, J.H. Bakken, G.A. Houghton, T.P. (1999). *Cyclic Subspace Regression with analysis of wavelength-selection criteria*, Chemometrics and Intelligent Laboratory Systems 45, 225-239.

[33] Kalivas, J.H. Brenchley J.M. Lang, P.M. & Nieves, R.G. (1998). *Cyclic Subspace Regression*, Journal of Multivariate Analysis 65, 58-70.

[34] Kalivas, J.H. Brenchley J.M. Lang, P.M. & Nieves, R.G. (1998). *Stabilization of Cyclic Subspace Regression*, Chemometrics and Intelligent Laboratory Systems 41, 127-134.

[35] Kalivas, J.H. Lang, P.M. (1997). *Response to "Comments on Interrelationships between sensitivity and selectivity measures for spectroscopic analysis" by K.Faber et al.*, Chemometrics and Intelligent Laboratory Systems 38, 95-100.

[36] Kalivas, J.H. Xie, Y.L. (1997). *Evaluation of principal component selection methods to form a global prediction model by principal component regression*, Analytica Chimica Acta 348, 19-27.

[37] Madsen, H.(1995). *Tidsrækkeanalyse*, IMM DTU.

[38] Madsen, K. Nielsen, H.B. Tingleff, O. (1999). *Optimization with constraints*, J. No. H40. IMM DTU.

[39] Mallows, C. L. (1973). *Some comments on* C_p, Technometrics, 15, pp.661-675.

[40] Manne, R. (1987). *Analysis of two Partial-Least-Squares algorithms for multivariate calibration*, Chemometrics and Intelligent Laboratory Systems, pp.187-197.

[41] Marx, B. D. & Eilers, P. H. C. (1999). *Generalized Linear Regression on Sampled Signals and Curves: A P-Spline Approach*, Technometrics, Vol. 41, No. 1, pp.1-13.

[42] Messick, N.J. Kalivas, J.H. & Lang, P.M. (1997). *Selecting factors for Partial Least Squares*, Microchemical Journal, no.55, pp.200-207.

[43] Miller, A. J. (1984). *Selection of subsets of regression variables (with discussion)*, Journal of the Royal Statistical Society A, 147(3):389-425.

[44] Miller, A. J. (1990). *Subset Selection in Regression*, Chapman and Hall, London.

[45] Nielsen, H.B. (1999). *Algorithms for linear optimization. An introduction*, J. No. ALO. IMM DTU.

[46] Nielsen, H.B. (1998). *Cubic splines*, J. No. H45. IMM DTU.

[47] Nielsen, H.Aa. Rasmussen, M. Madsen, H. (2001). *Calibration with near-continous spectral measurements*, 23rd Symposium of applied statistics, University of Copenhagen, Copenhagen, January 2001.

[48] Nørgaard, L. Saudland, A. Wagner, J. Nielsen, J.P. Munck, L. & Engelsen, S.B. (2000). *Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy*, Applied Spectroscopy, Vol. 54, No. 3.

[49] Osborne, M.R. Presnell, B. & Turlach, B.A. (2000). *On the Lasso and its Dual*, Journal of Computational and Graphical Statistics. Vol.9 no.2 pp. 319-337.

[50] Osborne, M.R. Presnell, B. Turlach, B.A. (2000). *A new approach to variable selection in least squares problems*, IMA Journal of Numerical Analysis 20, 389-403.

[51] Otto, M. (1997). *Statistical comparison of calibration methods in multicomponent analysis*, J. Anal. Chem. 359: 123-125.

[52] Roecker, E.B. (1991). *Prediction Error and its Estimation for Subset-Selected Models*, Technometrics, Vol. 33, No. 4, pp. 459-468.

[53] Sadegh, P. Nielsen, H. Aa. & Madsen, H. *A Semi-parametric Approach for Decomposition of Absorption Spectra in the Presence of Unknown Components*, IMM DTU.

[54] Shao, J.(1993). *Linear Model Selection by Cross-validation*, Journal of the American Statistical Association. Vol. 88, No. 422, Theory and Methods pp. 486-494.

[55] Stone, M.(1974). *Cross-validation Choice and Assessment of Statistical Predictions*,Journal of the Royal Statistical Society. Ser. B, 36, pp. 111-147, Theory and Methods.

[56] Sundberg, R.(1999). *Multivariate calibration - direct and indirect regression methodology*,Scandinavian Journal of Statistics. Vol. 26, pp. 161-207.

[57] Tibshirani, R.(1996). *Regression shrinkage and selection via the lasso*, J.R. Statist. Soc. B, 58, 267-288.

[58] Van Huffel, S. & Vandewalle, J. (1991). *The Total Least Squares Prob-*

lem, SIAM

[59] Van Loan, F. (1997). *Introduction to scientific computing*, Prentice-Hall, Inc.

[60] Wenjiang, J.Fu (1998). *Penelized Regressions: The Bridge Versus the Lasso*, Journal of Computational and Graphical Statistics. Vol.7 no.3 pp. 397-416.

[61] Wold, S. Ruhe, A. Wold, H. & Dunn, W.J. (1984). *The collinearity problem in linear regression. The Partial Least Squares (PLS) approach to generalized inverses*, SIAM J. SCI. STAT. COMPUT. Vol.5 no.3 pp. 735-744.

[62] Wu, W. (2000). *Fast regression methods in a Lanczos (or PLS-1) basis. Theory and applications*, Chemometrics and Intelligent Laboratory Systems. Vol.51 pp. 145-161.

[63] Öjelund, H. (1997). *Multivariable calibration of environmental sensors.* IMM-EKS-1997-40.

[64] Öjelund, H. Madsen, H. & Thyregod, P. (2000). *Calibration with Absolute Shrinkage*, to appear in Journal of Chemometrics.

[65] Öjelund, H. Brown, P.J. Madsen, H. & Thyregod, P. (2000). *Prediction Based on Mean Subset*, To be submitted.