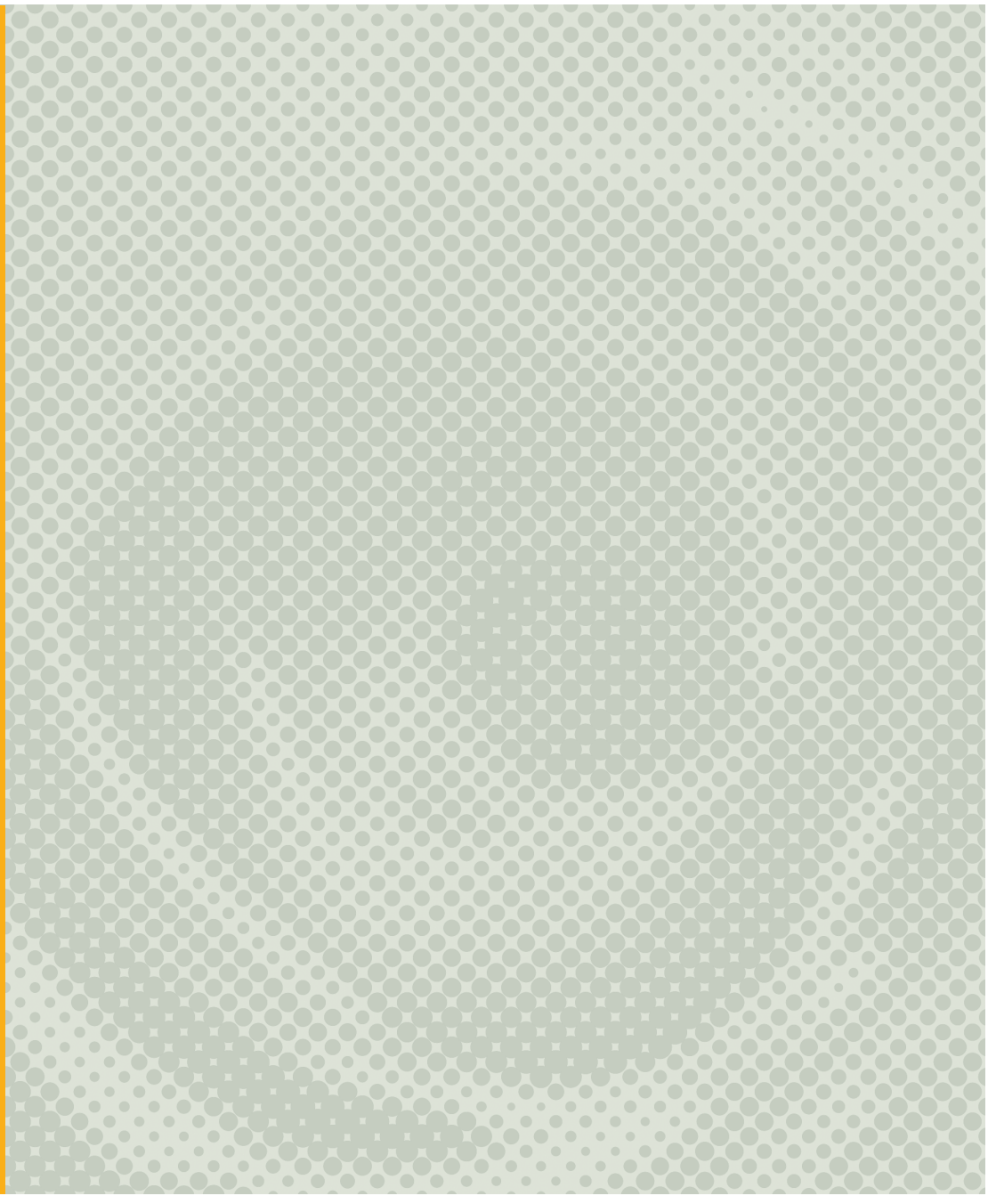
Development and evaluation of prediction equations for NIR instrument, measuring fat in Atlantic Salmon (Salmo salar L.) fillets, using Multivariate Methods.

Ólafur Hjörtur Kristjánsson

Norwegian University of Life Sciences
Department of Chemistry, Biotechnology and food science
Master Thesis 60 credits 2012



UNIVERSITETET FOR MILJO- OG BIOVITENSKAP

Development and evaluation of prediction equations for NIR instrument, measuring fat in Salmon fillets, using Multivariate Methods.

by

Ólafur Hjörtur Kristjánsson

A thesis submitted in partial fulfillment for the degree in Master of Science in Bioinformatics and Applied Statistics

in the
Bioinformatics and Applied Statistics
Institutt for kjemi, bioteknologi og matvitenskap.

May 2012

Declaration of Authorship

I, Ólafur Hjörtur Kristjánsson, declare that this paper titled, "Development and evaluation of prediction equations for NIR instrument, measuring fat in Salmon fillets, using Multivariate Methods" and the work presented in it is my own. I confirm that:

- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

UNIVERSITETET FOR MILJO- OG BIOVITENSKAP

Abstract

Bioinformatics and Applied Statistics
Institutt for kjemi, bioteknologi og matvitenskap.

Master of statistics

by Ólafur Hjörtur Kristjánsson

Knowledge of fat in salmon is extremely important to salmon breeder and the whole salmon industry. By monitoring fat in salmon fillet, huge amount of money will be saved. Several methods are available to determine fat in salmon fillets. Stofnfiskur Iceland decided to buy the NIR instrument Qmonitor which was installed in there slaughter line. When applying existing prediction model to results obtained by Qmonitor the prediction of fat was wrong. Aim of this thesis is to develop a new valid prediction model which will be applied to results obtained by the NIR instrument Qmonitor when measuring fish from all families in the nucleus of Stofnfiskur for breeding purposes. This thesis will provide background of NIR, breeding and problems of modeling fat in salmon fillet. Main goal is to discuss methods needed to explore the data, develop prediction model and validate the prediction model obtained. Use of recently developed CPLS will then be introduced in order to reduce the prediction error of existing methodology when creating prediction model. All methods will be compared and there qualities and drawback discussed. Three datasets are presented in the thesis were two of them where made for this thesis and one comes from paper defining methods used when modeling QMonitor data.

In the paper where the method of picking out five 14 mm plugs from the fillet to capture the variation of fat in the fillet a RMSEP value reported was 1.96. By using Canonical Partial Least Squares with the additional response a location of the plug, the RMSEP of the same dataset was 1.75. On the dataset made for this thesis to develope prediction model for the QMonitor in Iceland CPLS had the best performance obtaining RMSEP value of 1.8. Additional values which improved the prediction model where additional information about the plugs such as thickness of the plug, moisture in the plug and weight of the plug.

Acknowledgements

After a interesting study of prediction models and writing of this thesis I have developed a statistic goggles. Statistic googles are something which is not possible to buy but with help of great people is possible to develop and will come to place in the future !

First I like to thank my main supervisor Trygve Almøy who has guided me through the world of statistics and tolerated me in the intensive working periods when I was in Norway during the writings. Kristian Hovde Liland is very clever statistician which helped be to solve lot of programming problems and introduced me to CPLS for which I like to thank him.

I like to thank the people at Nofima for the help to create and provide datasets for this thesis and all their assistances in measuring and datahandling the fish. I want to thank Martin Høy and Jens Petter Wold for introducing me into the world of near infrared spectroscopy and let me have their Matlab scripts to gather datasets.

Without all the support from Stofnfiskur this thesis would not have been possible. I want to thank the people in Stofnfiskur for the help to measure the fish and gather the data. Especially I would like to thank the CEO Jónas Jónasson for the support and believed in me to implement NIR technology in Stofnfiskur.

Finally I would like to thank Theodór Kristjánsson for correcting the breeding part of the thesis and Petur Sæmundsen for English corrections.

Ås, Mai 2012

Ólafur Hjörtur Kristjánsson

Contents

List of Figures

List of Tables

Abbreviations

ABS	**A**bsorbance Value
BLUP	**B**est **L**inear **U**nbiased **P**rediction
BLUE	**B**est **L**inear **U**nbiased **E**stimation
CCD	**C**harge **C**coupled **D**evice
CPLS	**C**anonial **P**artial **L**east **S**quare
CPLSR	**C**anonial **P**artial **L**east **S**quare **R**egression
LF-NMR	**L**ow **F**ield **N**uclear **M**agnetic **R**esonance
LS	**L**east **S**quare
NIR	**N**ear **I**nfrared
NMR	**N**ucelar **M**agnetic **R**esonance
SIP	**S**election **I**ndex **P**rocedure
OD	**O**ptical **D**ensitie
PC	**P**rinciple **C**omponent
PCA	**P**rincipal **C**omponent **A**nalysis
PCR	**P**rincipal **C**omponent **R**egression
PLS	**P**artial **L**east **S**quares
PLSR	**P**artial **L**east **S**quares **R**egression
RMSEP	**R**oot **M**ean **S**quared **E**rror of **P**rediction
SNV	**S**tandard **N**ormal **V**ariate

Symbols

$\mathbf{A}$	The product $\mathbf{X}'\mathbf{X}$
$E()$	Expected value
$\mathbf{E}$	Eigenvectors of $\mathbf{X}'\mathbf{X}$ (Loadings)
$\mathbf{F}$	Residual matrix
$\mathbf{I}$	Identy matrix.
I	Raw spectra measure
I_0	Calibration spectra value
i	Elements number
(i)	Without element i
K	Number of batches in K-fold cross validation.
k	Method dependent number.
n, q, p, q	Nr of rows columns in matrix/vector
$\mathbf{P}$,$\mathbf{p}_a$	Bilinear factor loadings for $\mathbf{X}$-variables.
$\mathbf{Q}$,$\mathbf{q}_a$	Bilinear factor loadings for $\mathbf{y}$-variables.
R^2_{pred}	Coefficient of determination for prediction
$\mathbf{T}$,$\mathbf{t}_a$	Bilinear factor scores for objects in $\mathbf{X}$-space
$\mathbf{V}$,$\mathbf{v}_a$	Compression weights for $\mathbf{X}$-variables
$\mathbf{W}$,$\mathbf{w}_a$	Pls factor loading weights for $\mathbf{X}$-variables
$\mathbf{X}$	Explanatory matrix (Matrix given with uppercase bold letter)
$\mathbf{x}$	Column from the matrix $\mathbf{X}$
$\mathbf{x}_0$	Explanatory vector obtained from new sample
$\bar{\mathbf{x}}$	Mean value of columns of $\mathbf{X}$
x	Element from the explanatory matrix $\mathbf{X}$
$'$	Transposed sign of matrix/vector ($\mathbf{X}'$)
$^{-1}$	Inverse of matrix/vector ($\mathbf{X}^{-1}$)

$\mathbf{y}$	Vector of responses (Vectors given with lowercase bold letter)
$\mathbf{y}_{(i),k}$	Prediction of y_i using model developed without obs. i
y	Observation from vector $\mathbf{y}$
$\bar{\mathbf{y}}$	Mean value of vector $\mathbf{y}$
$\hat{\mathbf{y}}$	Predicted value of $\mathbf{y}$
θ	Paramter (Parameters given with greek letter)
$\boldsymbol{\theta}$	Vector of parameters
$\hat{\theta}$	Estimated paramter
$\hat{\boldsymbol{\theta}}$	Estimated vector of parameters
Ξ	Condition number
ψ	Conditional number
$\boldsymbol{\varepsilon}$	Error of model
λ	Eigenvalue
Σ	Population covariance matrix
$\sum$	Summation
τ	Eigenvalue

Chapter 1

Introduction

Statistics play a big role in the worlds development. It is a powerful tool to gather information from observations made on resources and determine future observations. One aspect of statistics is the use in agriculture, specially in animal breeding.

The objective of this thesis is to develop and validate a prediction model for a Near Infrared Spectroscopic (NIR) instrument used to predict fat and color of salmon fillets in the breeding company Stofnfiskur in Iceland. The results from the NIR machine will be used for animal breeding. Discussion of prediction and validation methods when modeling data from NIR machines will be carried out.

1.1 Stofnfiskur

Stofnfiskur is a breeding company located in Iceland. The company emphasis on selective breeding on Atlantic salmon *(Salmo salar)*, Arctic charr *(Salvelinus alpinus)* and Atlantic cod *(Gadus morhua)*. Stofnfiskur was established in 1991 using brood stock of Atlantic Salmon imported from Norway to Iceland around 1980. Breeding of Arctic charr was added to the production in the late 90's in collaboration with Holar Agriculture University. Atlantic cod breeding program was established in the year 2003 [1]. In Stofnfiskur, implementation of animal breeding is used to improve the brood stock in the nucleus in order to sell improved eyed salmon, arctic carr and cod eggs and fryes. Main emphasis of Stofnfiskur production is on salmon eggs. Stofnfiskur is one of few breeding companies in the world who is capable of selling salmon eggs all year round

from disease free environment produced in land based breeding stations using water from geothermal and freshwater boreholes. The breeding stations are at six locations in Iceland, distributed in the southwest of Iceland.

FIGURE 1.1: Stofnfiskur brood stock farm in Vogar Iceland

Stofnfiskur has exported salmon eggs mainly to Canada, Chile, Faroe Islands, Denmark, Ireland and Norway since 1996. Stofnfiskur main markets are showed in figure 1.2 [2].

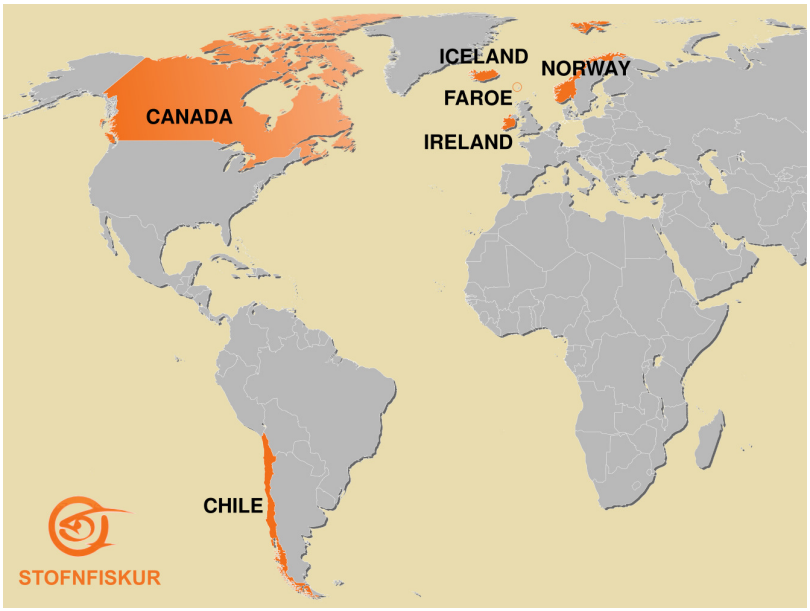


FIGURE 1.2: Main markets were Stofnfiskur sells their production

This thesis is supported by Stofnfiskur and results will be used for further studies carried out in Stofnfiskur using the technology and methods described in this thesis.

1.2 Animal breeding

The farmer has the goal to find the best animals in the herd which will be mated to create next generation. He could select his animals based on knowledge about the herd or visual properties of the animals. To maximize his profit he should hire an animal breeder. Theory of animal breeding provide the breeder with a tool where statistical methods can be applied to find the best individuals among the heard. By implementing statistics in animal breeding the grading of the animals, referred as breeding value will be estimated closer to the animals true performance when all effects have been accounted for which can affect the breeding value, referred as true breeding value. Correlation between the predicted breeding value and the true breeding value is referred as accuracy [3].

Before statistics were applied in breeding of animals, accuracy and genetic gain, which is a measure of genetical improvement between generations, was lower than today. The reason for lower genetic gain was mainly because the individuals were selected based on their phenotype value. Usually without regarding any other informations, such as relation among the herd and additional effects which have an impact on the observed phenotype value. An example of a trait in animal breeding is the weight of an animal. Selecting the heaviest animal without help of statistics for next generation, would recommend selecting biggest individual. This individual could be the only big individual among its family and could have gained their size due to additional effects in the environment which result in higher growth. Offspring from such animals are likely to be small, because they will inherit genes which provide less growth in average environment qualities [4].

Several statistical methods are practiced in breeding when next generation of breeding candidates are selected. First application of statistics in breeding was done by Fairfield Smith 1936 in plant breeding by using selection index procedure (SIP) [4]. Lanoy Nelson Hazel developed the SIP method for selecting animals to next generation in 1945. In SIP method individuals receive score for each measured trait, which weight is put on

according to genetic gain and economic importance, which defines vector of parameters in linear regression [4].

In 1950 Charles Roy Henderson developed the best linear unbiased estimates (BLUE) for fixed effects and best linear unbiased prediction (BLUP) for random effects in mixed model. Mixed model equation was applied into animal breeding when enough computer power was available around 1970's. Relationship matrix was developed which describes kinship of the herd and the covariance of the breeding values. Mixed model solved couple of problems which could effect the measure obtained (history of breeding values taken from this book) [4].

For example, group of fish measured at different time has usually different mean value of the trait measured within each measure date, then other days during the measuring season, the time affects the measure obtained from each fish. Fixed effects in the mixed model contains information about additional effects, which are not relevant when selection for measured trait is done. Results from the BLUP equation in animal breeding is refferd as breeding value. Breeding value calculations using BLUP increased the accuracy of the breeding value of each trait measured which was a huge step forward in selecting animals when producing a population for next generation [4].

The main goal of breeding is to move the mean value of the characteristic measured in a population to desired direction based on economic importance [4].

1.3 Breeding of Norwegian Atlantic Salmon

Breeding of Atlantic salmon started in Norway 1971 when Akvaforsk started their genetic research with Atlantic Salmon. First they gathered wild salmon from 40 Norwegian river strains all around Norway and one river in Sweden to form a base population. Research station was build in Sunndalsøra 1971 were the main activity of breeding in Norway still takes place [5].

First problem in salmon breeding was to find which fish manage to survive and grow in domestication. When stable population was establish selection started by calculating individuals selection index. When enough computer power was available to calculate BLUP the combination of breeding values on each trait, family breeding values and

selection index are combined based on genetic gain, accuracy and economic importance [4].

Akvafork did comparison of salmon selected for 5 generations and wild salmon from the Namsen river to measure how much has been accomplished by breeding. The improvement of selected salmon against wild is shown in table 1.1 [3].

Trait	Improvement in selected over wild (%)
Growth rate	+113
Feed consumption	+40
Protein retention	+9
Energy retention	+14

TABLE 1.1: Genetic gain in Atlantic salmon over five generations

1.4 Traits in salmon breeding

Traits which selection of individuals for next generation is based on in salmon breeding programs today are mainly, body size at harvest, disease resistance, early sexual maturity and quality traits [6].

Selection by body weight is done by measuring the breeding candidates when the average size of the population is around 3 kg, which is a usual slaughter size.

Selection against sexual maturity has the goal to delay the maturation. If maturation appears before slaughter the fillets are less saleable. When fish matures and prepares spawning, he stops eating, then growth stops and the fillets get lean because the fish is saving energy (fat) to produce eggs. In addition the color of the fillet reduces because the salmon eggs retains their color from the fillets.

Selection against diseases is also practiced. Diseases selected against is Furunculosis, Infectious Salmon Anemia (ISA) and Infectious Pancreatic Necrosis (IPN). Selecting individuals for disease resistance is by done exposing $10 - 20$ individuals, which are full sips of the breeding candidates and measure performance of families. By the family means individuals for next generation are selected [4].

Quality traits in a breeding plan are color of the fillet and fat in the fillet. Color is measured as % of astaxanthin in total chemicals of the fillet. Average value of astaxanthin is 7% today. The aim of breeding is to select individual with dark colored flesh. Fish

having dark flesh will give possibilities to reduce amount of added color to the feed. Astaxanthin is the most expensive part of feed [6].

Fat in fillet is measured as % fat of all chemicals in the fillet. Average fat in 4 kg fish is 17% [6]. Depending on markets, selection is used to reduce or increase the fat in fillet. To reduce production cost it is favorable to reduce fat in the fillet. Leaner fish needs less feed to retain the energy needed to live and uses more of the feed to grow [6]. Increased demand for salmon fillets in sushi demands fish with high fat to get the raw fillet soft to bite. Therefore selection of fish to increase fat is also done. In addition, a selection against deformities is under constant inspection.

All those traits can be inherited from parent to offspring. All at different level but inheritance has been found in most traits, which is expected when quantitative geneticists are selecting breeding candidates for next generation [3].

1.5 Methods to measure fat and pigment in salmon fillets

In this paper the focus is on application of multivariate statistic methods to measure the trait fat in fillet. Fat in fillet has been measured through the years with different methodologies. Six instruments exist today to measure fat in fish. Three of them measure the exterior on live fish and three of them the fillet of slaughtered fish.

First method which measures on live fish is the Near Infrared Spectroscopy (NIR) probe called QPoint [7] produced by the company QVision, which sends light 1 cm into the skin. In the middle of the machine is a NIR detector which collect light in the infrared range of the electromagnetic spectrum. Based on the values from the NIR detector fat is predicted in the whole fillet. Only a prototype has been made by QVision. Results published have shown inaccurate results [8].

Second machine that measures fat on live fish is Torry Fat meter [9]. Torry fat meter is small hand-helt equipment which is easy to operate. Torry Fat meter is based on the Nuclear Magnetic Resonance (NMR) technique. Torry fat meter obtains quite different results when measurements are done on same fish at same spot repeatedly. It has been the experience in Stofnfiskur.

Third machine is a low field Nuclear Magnetic Resonance (LF-NMR) which measures conductive of electric waves sent into the sample, coming from a magnetic and predicts fat based on the conduciveness. It has been showing quite good results which need to be confirmed [10].

Best result obtained for breeding purposes would to measure fat and color on living fish where the values would be obtained directly on the individuals among which the breeding candidates are selected. The problem of measuring fat on live fish is the impact of the skin color which changes the light reflectance when methods using light reflectance are used. When magnetic methods are used the problem is a thick fat layer under the skin which is not of interest when fat is measured in the fillet. To obtain more reliable breeding estimates of fat and pigment, a measure is done among full sibs of the breeding candidates after slaughtering and filleting. Based on value of sibs, family means are obtained which selection is based on when selecting breeding candidates.

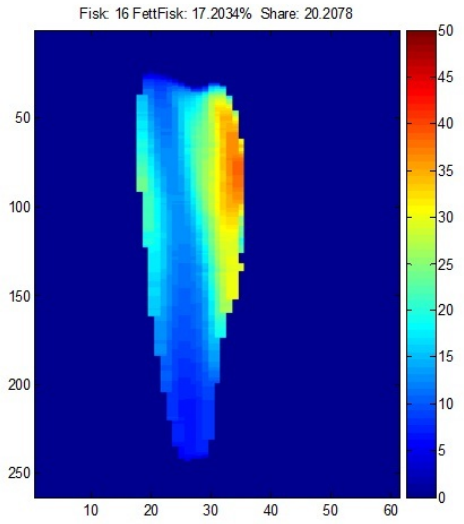


FIGURE 1.3: Predicted fat in the whole fillet [7]

First out of three instruments used where the fish needs to be slaughtered is Computed Tomograph (CT) which is very expensive instrument. It is impossible to use CT in the field, but it gives good results [11]. Second instrument used to measure fat in fillet is PhotoFish developed by Nofima Marin. PhotoFish is a box with a good camera in the top which collects values of Red, Green and Blue (RGB). PhotoFish predicts both fat and color in the fillet based on these RGB values [8]. Third machine measuring fillet

is QMonitor, developed by Nofima Mat and TiTech. QMonitor uses Near Infrared Spectroscopy (NIR) to measure fat in the fillet and another detector in the visible electromagnetic spectrum to measure pigment in the fillet. The NIR scanner transmits light to the sample and back scatter of the illuminated light is measured by the two detectors in the instrument. Based on the result image fat and color are predicted in the whole fillet as shown in figure 1.3 [7].

Results from QMonitor has been published showing quite good estimates of fat and color in salmon fillets [12]. Only the last two instruments are capable of measuring color in the fillet which is one of the main demand of buyers when they buy salmon fillet in a shop [13]. Therefore color has to be obtained when selecting breeding candidates.

1.6 Comparison of equipment to measure fat and pigment

In fall 2010 a comparison on available instruments methods to measure fat and pigment was conducted at Nofima Ås to determine which is the best available instrument to predict fat and color in Salmon. A total of 45 fish from Stofnfiskur were filleted and measured in Nofima Ås. Fat and color values were obtained from PhotoFish and QMonitor based on existing prediction equations for both machines. After measuring the fillets using these two instruments the skin was removed of each fillet and they minced. Out of the minced 30 gr. samples were collected and sent to Nofima Sundalsøra to obtain fat and color values by chemical method [8]. Values obtained at the lab in Sunndal are considerate as the true fat and color value for the whole fillet. These true values were compared to the values predicted from the machines to determine their quality. Values predicted by QMonitor were closer to the chemical values. Similar results have been published [8]. Based on this trial Stofnfiskur decided to invest in QMonitor from Qvision which was installed in Stofnfiskur.

Quality of NIR machines is how cheap it is to obtain explanatory variables. Compared to the lab method where the sample has to be destroyed and expensive dangerous solvents are used to determine the fat in the fillet. PhotoFish require all observed values to be sent to Nofima Marin in order to get fat values from the fillets. Usually the only cost of NIR machine is the startup cost which includes buying the machine and obtain chemical values from a few samples to build a prediction model. NIR machines are capable of

measure a lot of material without preparation, dangerous solvent, labor and additional cost [14, 15].

1.7 Problems

The QMonitor in Iceland which Stofnfiskur bought did not have the same light strength and is not calibrated the same way as the QMonitor standing in Nofima. The lamps in the QMonitor in Iceland are 15 years younger and the mirrors which reflects the light to the sample and in the light detector are new. In the fall of 2011, 2500 fish were measured at Stofnfiskur using the new Qmonitor in Iceland. Values predicted using prediction model developed for the QMonitor standing in Norway, are far to high and not likely to be true based on literature [4]. It is very important for Stofnfiskur's breeding work to know the results from the new QMonitor.

Several methods have been used to create prediction model for Qmonitor spectral image, which contains approximately 16.250 pixels (259 pixels $\times$ 65 pixels), depending on fillet size. Each of the 15 channels in the light detector inside the instrument capture one image each on different light strength. The first method when the instrument was first developed 2005 was to obtain one average spectra image over each channel resulting in 15 values as the explanatory variables. Average chemical fat values obtained in lab by chemical extraction of fat for the whole fillet was used as a response. The weakness of that method is information about fat variation in the fillet is terminated.

Variation of fat in fillet is high. In the belly area fat is high and in the backbone fat is low. Solution to this weakness is described in paper from 2009 [12] where the idea is to create prediction model which predicts fat for every pixel (where each pixel contains one value on each channel of the machine) of the spectral image obtained from the QMonitor and average all the predicted fat values of each pixel is obtained and reported as the fat value of the whole fillet. The model is developed by picking out $5 - 6$ plugs which are, 15 mm in diameter from the fillet and measure the fat content using LF-NMR as a response. Corresponding pixels are selected from the spectral image and prediction model for each pixel is developed using multivariate calibration. Development was done on this method to determine salt content of salmon in paper published 2009 [11] by measuring water in the plugs instead of fat which is highly negatively correlated to salt.

1.8 Datasets

Three dataset are used in this thesis to understand the method published [12] and develop prediction model for the QMonitor in Iceland.

The first dataset contains of the data collected by the authors of the paper where the method used to create model to predict fat in salmon fillet was described [11]. It contains fat measures of five plugs from both fillets of 15 fishes weighting $2-5$ kg. Spectra values are from images captured by the QMonitor standing in Nofima in Norway. The spectral image values were preprocessed by Standard Normal Variate (SNV) preprocessing. This dataset is used to learn the methods of predicting fat in each pixel of the fillet.

Dataset 2 was recorded in Nofima Norway on fish from Stofnfiskur to see how the existing model performed using fish from Stofnfiskur. More weight and size variation was of the fish measured than in the paper [12]. In total of 43 fish weighing from $1-8$ kg were measured. Six plugs were collected from the right fillet of each fish. Because lack of time, fat values using LF-NMR were obtained for all plugs in three fish and two randomly selected plugs from the remaining fish. Data was collected using QMonitor standing in Nofima. Spectra values were collected on same location as the plug were taken from the fish and SNV preprocessing applied on each observation.

Creators of the paper which describes the modeling method [12] assisted when collecting the data. When the fish was measured, 13 additional measures under guidance of experienced scientist were conducted which is possible to use in recently developed prediction method which have not been tried before on spectral image from QMonitor.

Using the knowledge from the sampling of Dataset 2, Dataset 3 was created based on measurements from the new QMonitor standing in Iceland and plugs collected in Iceland. In total of five plugs from left fillet of 24 fish weighing $1-6$ kg were measured using LF-NMR in Nofima Norway. To improve prediction model 8 additional responses were collected.

1.9 Objective

The main objective of this thesis is using multivariate statistics to translate the results from QMonitor to fat values which is possible to use with confidence for future breeding work in Stofnfiskur. In addition main prediction and validation methods will be discussed. Method described in the paper [12] will be discussed and were used to create usable dataset, to develop a prediction model for the QMonitor in Iceland.

Descriptive statistics will be carried out on all datasets to explore their limits. Principle Components Analysis (PCA) will be tried on the preprosessed spectra values obtained from QMonitor. Multivariate prediction methods such as Principle Component Regression (PCR) and Partial Least Squares Regression (PLSR) will be explored and recently developed method using Canonical correlation denoted Canonical Partial Least Squares (CPLS) will be tried. Validation on the results obtained and the ability to predict will be calculated using root mean square error of prediction using leave one out cross validation and K-fold cross validation. Different test and calibration sets will be defined based on the data to explore the prediction ability of the data.

Main purpose of this thesis will be to develop prediction model for a NIR instrument QMonitor located at Stofnfiskur in Iceland by multivariate method which has not been used before on QMonitor data to improve published methods.

1.10 Software used

In this thesis calculations were carried out in R 2.14 [16], Matalb 7,5 student version and written using LaTeX.

Chapter 2

Multivariate Statistics

The main objective of this thesis is to address most of the topics needed to model NIR data for prediction.

The importance of multivariate statistics has been increasing with increased computer power and development of machinery which can retrieve more advanced results. These complicated measurements can be translated with help of multivariate statistics to simple values.

Example of a problem which NIR technology solves, could be a factory that needs to know the amount of chemicals in their products to monitor the production and report key figures of the product to buyer. Measuring amount of chemicals at lab on every product is very expensive and demands trained personnel which use dangerous solvents. Determining amount of chemicals in a lab is also destructive to the product and it can not be sold after measuring. This problem can be solved by using NIR technology. NIR instrument reports large datasets which without help of multivariate statistics would be hard to translate into preferred values.

Notations are given in the symbols table in the beginning of the thesis.

2.1 Statistical model

To describe outcome from a sample, a statistical model is used. Response of n samples are stored in $n \times 1$ response vector $\mathbf{y}$. Explanatory variables are stored in an $n \times p$

matrix $\mathbf{X}$, where p is number of columns in $\mathbf{X}$, where each column contains measure obtained from samples $1 \cdots n$. The relationship of $\mathbf{y}$ and $\mathbf{X}$ is described with an unknown parameter vector θ. The regression model is stated in equation 2.1 [17].

$$\mathbf{y} = g(\mathbf{X}, \theta) + \boldsymbol{\varepsilon} \tag{2.1}$$

The term $\boldsymbol{\varepsilon}$ in equation 2.1 is referred as error. Usually the expectation of the error is assumed to by 0, denoted $E(\boldsymbol{\varepsilon}) = 0$. Variance of the error describes the covariance structure between samples in the response vector $\mathbf{y}$. If there is no covariance between the samples and they all have equal variance the variance of the error is denoted $var(\boldsymbol{\varepsilon}) = \Sigma = \sigma^2\mathbf{I}$. Where $\mathbf{I}$ is $n \times n$ identity matrix.

Usually for prediction modeling the relation of the $n \times 1$ vector $\mathbf{y}$ and $n \times p$ matrix $\mathbf{X}$ is assumed linear. The parameter vector θ will be referred as $p \times 1$ vector $\boldsymbol{\beta}$. The linear relation of $\mathbf{y}$ and $\mathbf{X}$ defined in equation 2.2.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.2}$$

In a linear model the error is defined as the difference between $\mathbf{X}\boldsymbol{\beta}$ and $\mathbf{y}$, denoted $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$. Expectation of $\mathbf{y}$ is defined as $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ giving the definition $\boldsymbol{\varepsilon} = \mathbf{y} - E(\mathbf{y})$ of the error.

NIR instrument can not report exact value of $\mathbf{y}$. What can be reported is likely result of $\mathbf{y}$ if the true value of $\mathbf{y}$ would be accessible. This likely result reported of $\mathbf{y}$ is referred as a prediction of $\mathbf{y}$, denoted $\hat{\mathbf{y}}$. The result is not predicted out of nowhere, it is predicted after firmly estimating $\boldsymbol{\beta}$ in equation 2.2 by $\hat{\boldsymbol{\beta}}$. To estimate $\hat{\boldsymbol{\beta}}$ true values of $\mathbf{y}$ are used along with its explanatory values stored in $\mathbf{X}$ obtained from the NIR machine.

The main difference between prediction and estimation is, in estimation, values of the response vector and the explanatory matrix are needed in order to estimate the model parameters. In prediction only explanatory variables and the estimated parameters are needed, but to evaluate the prediction, response values are also needed. After estimating

$\hat{\boldsymbol{\beta}}$ using the explanatory variables and the response, the estimated parameters obtained are used without modification for prediction. When new explanatory variables of a sample is introduced, new prediction for the sample are obtained using the estimated parameters. Prediction model is given by equation 2.3.

$$\hat{y} = \bar{y} + \hat{\boldsymbol{\beta}}' (\mathbf{x}_0 - \bar{\mathbf{x}}) \tag{2.3}$$

Where $\mathbf{x}_0$ is $p \times 1$ explanatory vector obtained from new sample. $\bar{\mathbf{x}}$ is vector of column means of the explanatory matrix from the calibration dataset, $\bar{\mathbf{y}}$ is mean of response in the calibration dataset and $\hat{\boldsymbol{\beta}}$ obtained from the calibration dataset.

2.2 Criteria for model validation

There exist many criteria to validate prediction model. The two main criteria used are root mean square of prediction (RMSEP) and coefficient of determination for prediction denoted R^2_{pred}.

2.2.1 Root Mean Squared Error of Prediction (RMSEP)

When creating a prediction model, the aim is to create a robust model which is capable of meeting explanatory variables in the future and predict response close to what will be observed if the true future response of the future explanatory variables will be obtained. One measure to quantify the quality of prediction model when it meets the future explanatory variables is by calculating the distance between future predicted response $\hat{\mathbf{y}}$ and future true responses $\mathbf{y}$. It is impossible to know the future distance but it can be estimated. The average future distance is defined as θ^2 and is given by equation 2.4.

$$\theta^2 = E(\mathbf{y} - \hat{\mathbf{y}})^2 \tag{2.4}$$

Main interest is in the distance between $\mathbf{y}$ and $\hat{\mathbf{y}}$ in the future. It does not matter if the distance between $\mathbf{y}$ and $\hat{\mathbf{y}}$ is negative or positive. Effect of negative and positive difference is removed by squaring the distance. When squaring the distance, outliers have huge influence on the distance measure obtained by θ^2 which is not of main interest. By taking square root of the squared estimated difference between $\mathbf{y}$ and $\hat{\mathbf{y}}$ outliers do not have as much influence on our quantification of prediction quality of the prediction model. Definition of θ is given by equation 2.5

$$\theta = \sqrt{E(\mathbf{y} - \hat{\mathbf{y}})^2} \tag{2.5}$$

Observations are then in original scalar by taking the square root. To know θ would be nice and would solve the evaluation problem, but it is not possible to obtain because there is no access to future explanatory variables to determine $\hat{\mathbf{y}}$. If it would be possible to know future explanatory variables, a knowledge of future values of $\mathbf{y}$ would also be needed. Then it is impossible to calculate θ directly. Example, if there exist fat values of salmon fillets in the future there would be no need for NIR technique to determine it. Only data available is the current data. Therefore quantification of the prediction quality of the model can only be quantified on the current data. The prediction model is expected to meet explanatory variables similar to those who are already in the dataset, in the future. Therefore an estimation of $\hat{\theta}$ is done by eliminate part of the dataset. Create prediction model based on remaining data. Address new prediction model to the eliminated explanatory variables, predict $\hat{\mathbf{y}}$ based on those explanatory variables and calculate the average distance between the predicted $\hat{\mathbf{y}}$ and the eliminated $\mathbf{y}$. Estimation of $\hat{\theta}$ is defined in equation 2.6.

$$RMSEP_k = \hat{\theta}_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{(i),k})^2} \tag{2.6}$$

In equation 2.6, n is number of samples, y_i is observation i from the response vector $\mathbf{y}$ and $\hat{y}_{(i),k}$ is the prediction of the eliminated response i using model estimated without information from explanatory and response variables in row i of our dataset using

method k. Batch i containing more then one observation and corresponding explanatory variables can also be removed instead of only one observation from $\mathbf{y}$ and $\mathbf{X}$ if it is appropriate. When $RMSEP_k$ is plotted, it is usually plotted against some complexity factor k of the method used to create the prediction model. Estimation of $\hat{\theta}_k$ is referred as $RMSEP_k$ which stands for root mean square error of prediction using special method k (will be defined later). This is the most common measure for validation of prediction model. $RMSEP_k$ estimates the average distance between $\mathbf{y}$ and $\hat{\mathbf{y}}$ based on our current data. $RMSEP_k$ sums all the square distances between y_i and $\hat{y}_{(i),k}$, divides by number of samples to know the average distance between y and $\hat{y}_{(i),k}$.

2.2.2 $R^2_{\mathbf{pred(k)}}$

Another measure of quality of prediction commonly used is $R^2_{\text{pred(k)}}$ reffered as coefficent of determination which gives an idea how much of the variation can be expected to be explained in the new data by the model. Which is given by equation 2.6 [18].

$$R^2_{\text{pred(k)}} = \left(1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_{(i),k}\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}\right) \tag{2.7}$$

Where k is method dependent factor and is defined along with method used. If $R^2_{\text{pred(k)}}$ is low we should rather report $\bar{\mathbf{y}}$ as our prediction then using our prediction model to predict $\hat{\mathbf{y}}$ because it is performing equally good as $\bar{\mathbf{y}}$. If $R^2_{\text{pred(k)}}$ is close to 1 it indicates that our prediction model is performing better than reporting $\bar{\mathbf{y}}$ as our prediction. Note, that there is one to one correspondence between $RMSEP_k$ and $R^2_{\text{pred(k)}}$. Therefore it is not considerate necessary to report both. In this thesis results obtained from $RMSEP_k$ will be reported and used as quality criteria for prediction models. If comparing prediction quality of datasets containing different variation among the response variables, $R^2_{\text{pred(k)}}$ gives comparable prediction quality among the datasets because it eliminates the variation differences.

2.3 Validation of prediction quality

To measure the quality of a prediction model, a validation is done. In order to predict new response, the prediction model will be subjected to new set of explanatory variables.

Hopefully the new explanatory variables will be similar to the explanatory variables used to create the prediction model. It is hard to measure how well the new model will predict the new sample, because there does not exist any true response to compare the predicted response and estimate the difference. Solution to this problem is to eliminate responses and corresponding explanatory variables. Predict new response based on the eliminated explanatory variables, and compare the predicted response to the eliminated response and estimate their deviation. Then quality of the prediction has been quantified to some extent.

2.3.1 Leave one out Cross validation

Leave one out cross validation aims to quantify the quality of the prediction model by predicting the response i when the sample i has been removed from the dataset. This is done for all samples.

Algorithm of leave one out cross validation is as following [19]:

1. Response and explanatory variables of sample i are eliminated from the dataset. Creating $(n-1)\times 1$ explanatory vector $\mathbf{y}_{(i)}$ which is the original response vector $\mathbf{y}$ without sample i and $(n-1)\times p$ response matrix $\mathbf{X}_{(i)}$ without row i of the original explanatory matrix.

2. Estimate $\hat{\boldsymbol{\beta}}_{(i),k}$ by using $\mathbf{y}_{(i)}$ and $\mathbf{X}_{(i)}$ which are the original dataset without sample i, using method k.

3. Predict response of sample i, denoted $\hat{y}_i$ using the model in equation 2.3 and introduce the equation to row i of the original explanatory matrix $\mathbf{X}$.

4. Repeat the algorithm for $i = 1 \cdots n$ until every predicted value of $n \times 1$ vector $\hat{\mathbf{y}}$ has been obtained.

When the algorithm has obtained every value of $\hat{\mathbf{y}}$ the distance between $\mathbf{y}$ and $\hat{\mathbf{y}}$ is calculated in order to quantify the quality of the model.

2.3.2 K-fold cross validation

Leave one out cross validation can underestimate the prediction error when the prediction model will be exposed to more than one row of explanatory variables at once. Leave one out cross validation does underestimate the prediction error because the eliminated sample belongs to batch where other samples are already included in the dataset. Then some information is already in the dataset about the eliminated sample which will not happen in the real world if the model would get exposed to new observation. Solution to the problem of underestimation in leave one out cross validation is to divide the n observations into K batches were $K \leq n$ (K used for number of batches because of tradition. K here is not he same as the method dependent parameter k). By eliminating batches of the responses and explanatory variables in each step of the estimation algorithm defined in chapter 2.4.1 instead of one line at a time (if $K < n$). Then estimate the prediction ability of the prediction model for those batches. If the dataset does not constrain the K to be a constant it gives better idea of model quality to let K vary and see how our prediction quality measures develops.

Example of a K-fold cross validation is when measuring a fish fillet with NIR machine. Each fillet contains five plugs. When new fish is measured the prediction model will be exposed to five measurements at time. The five plugs on the same fish are regarded as one batch which is more realistic because the NIR machine will get exposed the whole fish, not only one plug out of the whole fillet. The other four plugs also contain information about the fifth plug. The n observations will be divided to K batches, defined by number of fish where each fish consist of j observations.

2.3.3 Calibration and test sets

Dividing the dataset to a calibration and a test set is well known method in order to quantify quality of a prediction model. Best predictor is estimated from the calibration set using leave one out or k-fold cross validation. By development of $RMSEP_{\text{cal},k}$ in the calibration dataset in relation to increasing level of complexity level k the optimal prediction model is chosen. The parameter $\hat{\boldsymbol{\beta}}_{\text{cal},k}$ is estimated for each complexity level k. Usually $\hat{\boldsymbol{\beta}}_{\text{cal},k}$ is chosen by the lowest $RMSEP_{\text{cal},k}$ in combination to complexity

parameter k. Which complexity level k should exactly be choosen can be decided by different methodologies which are available. It is common to chosen the level of complexity where value of $RMSEP_{\text{cal},k}$ starts to flatten out when it is plotted against complexity. There also exist some penalty methods where the goal is to keep the complexity as low as possible in relation to $RMSEP_{\text{cal},k}$. Relation of complexity and estimation error is illustrated in figure 2.2 and should be kept in mind when level of complexity is decided. When $\hat{\boldsymbol{\beta}}_{\text{cal},k}$ is found we expose it to the explanatory variables $\mathbf{X}_{\text{test}}$ of the test set and predict $\hat{\mathbf{y}}_{\text{test}}$ by equation 2.8.

$$\hat{\mathbf{y}}_{\text{test}} = \mathbf{X}_{\text{test}}\hat{\boldsymbol{\beta}}_{\text{cal}} \tag{2.8}$$

Quality of the prediction model made by the calibration set is quantified by estimating $RMSEP_{\text{test}}$ which measures the deviation of $\hat{\mathbf{y}}_{\text{test}}$ and the response values of the test set $\mathbf{y}_{\text{test}}$ by equation 2.6. $RMSEP_{\text{cal}}$ is compared to the $RMSEP_{\text{test}}$ found when the optimal $\hat{\boldsymbol{\beta}}_{\text{cal(k)}}$ was used for explanatory variables of the test set. If value of $RMSEP_{\text{cal}}$ and $RMSEP_{\text{test}}$ are far from each other, the prediction model estimated by the calibration set is not robust. The model is not capable to get exposed to new data and predict values close to true response values or highly influent outliers are among the test or calibration set. If those outliers are due to error when obtaining the values, they should be eliminated. If no outliers are detected and the deviations between $RMSEP_{\text{cal}}$ and $RMSEP_{\text{test}}$ is large the model is not capable of getting exposed to new explanatory variables and predict them sufficiently.

2.4 Least square estimation

The classical method to estimate $\hat{\boldsymbol{\beta}}$ in order to predict $\hat{\mathbf{y}}$ is least square estimation. Estimation of $\hat{\boldsymbol{\beta}}$ is defined in equation 2.9 [17].

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{2.9}$$

In a NIR instrument are many channels, each measuring at its own light strength. Difference in light strength between adjacent channels is usually rather low. Resulting in

correlation between adjacent channels to be rather high. For one response can be many different channels explaining each response. Explanatory variables obtained as a measure of light reflectance are on small interval and are then close to each other. When noise and other errors have been removed and the measurements have been standardized by prepossessing the data, the shape of each wave look similar. Each channel of the instrument measure more information about certain chemicals than other channels resulting in low spread of observations. If similar samples are measured, the waves after prepossessing should more or less have similar appearance. [20]

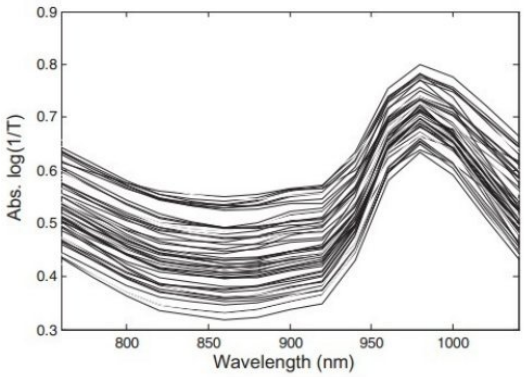


FIGURE 2.1: Average NIR transflectance spectra (log(1/T)) from the 70 dried salted coalfish [21].[22]

Result from a NIR instrument is a explanatory matrix, which has many highly correlated columns of data. When such explanatory matrix is used in multiple linear regression it is regarded as multicollinearity. Consequence of trying to invert nearly linearly dependent explanatory matrix containing NIR data in order to obtain least square estimation of $\hat{\boldsymbol{\beta}}$ in equation 2.8 becomes unstable or impossible. Another problem with least square estimation of $\hat{\boldsymbol{\beta}}$ is that sometimes are more columns in the explanatory matrix than observations. Example is when NIR machine has thousands of channels each measuring at different light strength. Calibration dataset will essentially contain fewer observations, then $n < p$. When $n < p$ it is impossible to invert $\mathbf{X}'\mathbf{X}$ which makes least square estimation impossible to use in order to estimate $\hat{\boldsymbol{\beta}}$.

2.5 Reduction in dimensions

In order to estimate $\hat{\boldsymbol{\beta}}$ it is recommended to reduce number of explanatory variables, without losing much information. Then the problem when $n < p$ is solved. Then it will be possible to have better estimate of $\hat{\boldsymbol{\beta}}$. A general problem when each response is explained by many explanatory variables, is the complexity of regression model. The complexity is rather high if all components/variables of the regression model are included. It is possible to reduce complexity in ordinary regression models by an elimination of components/variables of the model by several methods, but then information is removed. By no reduction on the regression model, the regression model will be complex, which will most likely include a lot of estimation errors as can be seen in figure 2.2 [20].

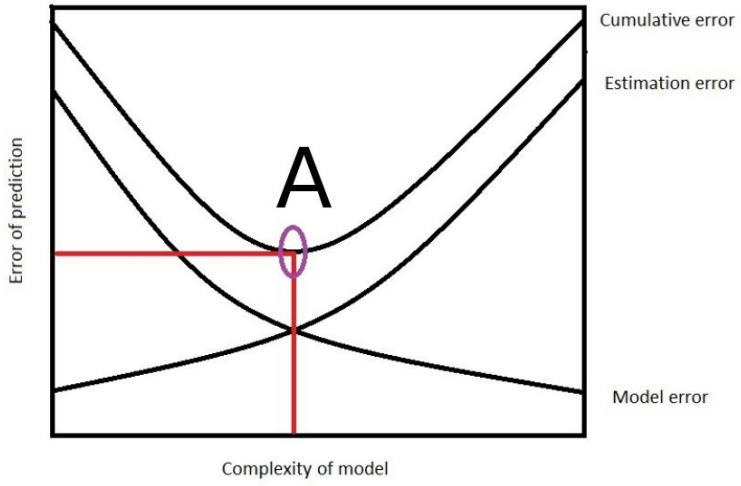


FIGURE 2.2: Cumulative model error

Figure 2.2 shows that too complex model introduce problems when to many parameters have to be estimated. Estimating to many parameters has many sources of estimation errors. Reducing complexity to much, reduced number of parameters estimated, which introduce increased model error. Then the prediction model will predict poorly. The aim is to find optimal complexity were the prediction error is in the lowest point. In this point optimal number of parameters are estimated and the model is complex enough to give reasonable prediction.

Problem of finding a optimal complexity level of our model without eliminating too much informations can be solved by Principle Component Regression (PCR), Partial Least Square Regression (PLSR) and finally a new development of PLSR a Canonical

Partial Least Squares Regression (CPLSR). They all tend to create a few components which capture most of the variation in the explanatory matrix $\mathbf{X}$ or the covariation between the explanatory matrix $\mathbf{X}$ and response matrix $\mathbf{y}$. Still the prediction model will be complex enough without introducing to much model error. These methods can help to come close to the point A in figure 2.2.

2.5.1 Eigenvectors, eigenvalues and coliniearity

If we have the $k \times k$ matrix $\mathbf{A}$ and $k \times k$ identity matrix $\mathbf{I}$. The solutions $\lambda_1, \lambda_2, \cdots, \lambda_k$ of the polynomial equation $|\mathbf{A} - \lambda I| = 0$ are defined as eigenvalues. Let's assume that there exists nonzero vector $\mathbf{e}$ such that

$$\mathbf{Ae} = \lambda \mathbf{e} \tag{2.10}$$

then we define $\mathbf{e}$ as eigenvector of the matrix $\mathbf{A}$ for corresponding eigenvalue λ.

We define $\mathbf{A} = \mathbf{X}'\mathbf{X}$ where each value of $\mathbf{X}$ is centered by subtracting its columns means.

Multicollinearity of the $\mathbf{X}$ can be measured by eigenvalues of $\mathbf{A}$. $\mathbf{A}$ is a symmetrical and positive definite, hence all the eigenvalues are positive. If some of the eigenvalues are small, then there is dependences among some columns of $\mathbf{X}$. One measure which has been used to verify the dependence is condition number of $\mathbf{A}$. Which is defined as

$$\Xi = \frac{\lambda_{\max}}{\lambda_{\min}} \tag{2.11}$$

where Ξ is measure spread in the eigenvalue spectrum of $\mathbf{A}$. If Ξ is less than 100 there is no problem with multicollinearity. Number between 100 and 1000 indicates some multicollinearity, and above 1000 indicates high level of multicollinearity [18]. Value of Ξ is strongly dependent on value of p. If $p > n$ then $\lambda_{\min} = 0$ which is impossible to divide by.

2.5.2 Principle Components Analysis (PCA)

In order to reduce the size of the explanatory matrix $\mathbf{X}$ we can use Principle Component Analysis (PCA). If our $n \times p$ explanatory matrix is a set of $n \times 1$ vectors

$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_p]$, then a smaller set of $n \times 1$ vectors $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2 \cdots \mathbf{t}_g]$ where $g < p$ can replace our original $\mathbf{X}$ with out of loosing much information. The first principle $\mathbf{t}_1$ component spans the direction of highest variation in $\mathbf{X}$. The next component $\mathbf{t}_2$ is orthogonal on the first one spanning the most of remaining variation of $\mathbf{X}$. This goes on until all variation has been explained by the components. Then $g = p$. Usually most of the variation is explained only be few components, then $g < p$. An important property of $\mathbf{T}$ are that the columns are orthogonal to each other. This gives a more stable estimate of the model parameters because of lower variance without introducing bias. Also problems of multicollinearity has then been reduced or removed. Number of error sourced will also reduce in modeling the data as shown in figure 2.2. Relation of our explanatory variables $\mathbf{X}$ and set of principle components $\mathbf{T}$ is described by equation 2.12 [20].

$$\mathbf{X} = \mathbf{T}\mathbf{E}' + \mathbf{F} \tag{2.12}$$

Where $\mathbf{E}$ is a $p \times g$ matrix referred as loadings. Loadings are defined as the eigenvectors of the covariance matrix $\mathbf{X}'\mathbf{X}$ were $\mathbf{X}$ has been centered. Principle component is then defined in equation 2.13.

$$\mathbf{T} = \mathbf{X}\mathbf{E} \tag{2.13}$$

The loadings reveal the correlation among components and their contribution to the columns of $\mathbf{X}$. Value of loadings is between 1 and -1 because they are scaled to unit length. The $n \times p$ matrix $\mathbf{F}$ in equation 2.12 is the residual matrix. If $g = p$ then $\mathbf{F} = 0$. In that situation all covariation of $\mathbf{X}$ is explained by $\mathbf{T}\mathbf{E}'$. When $g < q$ then $\mathbf{F}$ exist and contains residual which is not explained by $\mathbf{T}\mathbf{E}'$. Figure 2.3 show main ideas of Principle Components where two components are used to replace $n \times 3$ explanatory matrix $\mathbf{X}$ [23].

Result from the principle components are the scores, which are stored in the $n \times g$ matrix $\mathbf{T}$ and will replace $\mathbf{X}$. Scores are distance from values of $\mathbf{X}$ projected on the component to the center of the component.

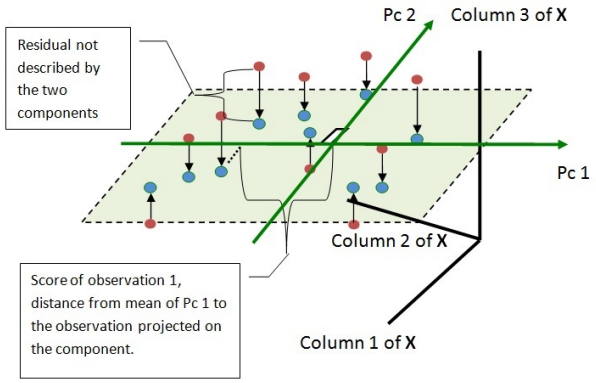


FIGURE 2.3: Main parts of PC

Decision on how many components to retain to replace $\mathbf{X}$ is based on how many components are needed to capture most of the variation of $\mathbf{X}$. There exists few methods to evaluate number of components. One is to look at proportion between eigenvalues of $\mathbf{X}'\mathbf{X}$ which is the covariance matrix of $\mathbf{X}$. The ratio is defined in equation 2.14

$$\psi_g = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_g}{\lambda_1 + \lambda_2 + \lambda_3 + \cdots \lambda_p} \tag{2.14}$$

The ratio ψ_g is referred as the conditional number and is the ratio between explained and unexplained variation of $\mathbf{X}$ by g number of components. When value of ψ is close to one and value of ψ does not increase by increasing g the number of components is found. In the R-Code in appendix A the method of calculating all p eigenvectors of the covariance matrix of $\mathbf{X}$ to estimate $\hat{\boldsymbol{\beta}}$ and number of components to reatain is decided depending on prediction quality evaulated by RMSEP or R^2_{pred}.

2.5.3 Principle Component Regression (PCR)

This thesis aims to predict new values based on complex explanatory matrix $\mathbf{X}$. With help of regression of $\mathbf{y}$ on the principle components retained using methods of chapter 2.5.2 an estimation of $\hat{\boldsymbol{\beta}}$ is done in order to predict $\hat{\mathbf{y}}$. Regression using Principle Components is referred as Principle Component Regression (PCR). Estimation of $\hat{\boldsymbol{\beta}}$ is done by equation 2.15 [20]

$$\hat{\boldsymbol{\beta}} = \mathbf{Eq} \tag{2.15}$$

where $\mathbf{E}$ is the loading of $\mathbf{X}$ found by methods in chapter 2.5.2 and $\mathbf{q}$ are the loadings of $\mathbf{y}$ is found by least square regression of $\mathbf{y}$ on the scores. The least square solution of the $\mathbf{y}$ loadings is given with equation 2.16.

$$\mathbf{q} = (\mathbf{E}'\mathbf{X}'\mathbf{X}\mathbf{E})^{-1}\mathbf{E}'\mathbf{X}'\mathbf{y} \tag{2.16}$$

It is known that $\mathbf{T} = \mathbf{XE} = diag(\lambda_i)$. Using that the estimation of $\hat{\boldsymbol{\beta}}$ given in equation 2.17.

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \mathbf{Eq} \\ &= \mathbf{E}(\mathbf{E}'\mathbf{X}'\mathbf{X}\mathbf{E})^{-1}\mathbf{E}'\mathbf{X}\mathbf{y} \\ &= \mathbf{E}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{y} \\ &= \mathbf{E}(diag(\lambda_i))^{-1}\mathbf{T}'\mathbf{y} \end{aligned} \tag{2.17}$$

Prediction $\hat{\mathbf{y}}$ is done by using equation 2.2. Usually more components are retained then needed when using PCR because PCR only focus on the variation of $\mathbf{X}$ when predicting $\hat{y}$. RMSEP is estimated by equation 2.6 and based on the prediction ability decision on number of components to retain is decided.

2.6 Partial Least Squares (PLS)

Using PCR finds small set $\mathbf{T}$ of vectors replacing $\mathbf{X}$ by finding maximum covariance only of $\mathbf{X}$ and don't regard the information in $\mathbf{y}$ which is the goal to predict. There is no guarantee that major variation in $\mathbf{X}$ is connected to the variation in $\mathbf{y}$. Retaining components $\mathbf{T}$ using information both from $\mathbf{X}$ and $\mathbf{y}$ is done in partial least squares which similar to principle component but uses the maximum covariance of $\mathbf{X}$ and $\mathbf{y}$ and maximizes their shared variance. There exist many methods to calculate partial

least squares which calculate the scores of $\mathbf{X}$ and loadings of $\mathbf{X}$ and $\mathbf{y}$ from centered $\mathbf{X}$ and centered $\mathbf{y}$ and its deflated versions. Method used in this thesis will be based on algorithm given in the book Multivariate Calibration [20]. In the appendix is the R code for the algorithm.

2.6.1 The Partial Least Square algorithm

This algorithm is done over components k, where $k = 0$ means no method is used, only the data is centered. The $k = 1 \cdots p$ is number of components calculated and the number of deflation rounds on $\mathbf{X}$ and $\mathbf{y}$. The algorithm is as following [20].

1. In order to deflate $\mathbf{X}$ and $\mathbf{y}$ we find the maximum covariance of $\mathbf{X}_{k-1}$ and $\mathbf{y}_{k-1}$ and normalize it to unit length. Which is given in formula 2.18.

$$\mathbf{w}_k = \frac{\mathbf{X}_{k-1}\mathbf{y}_{k-1}}{\sqrt{\mathbf{w}_k\mathbf{w}_k'}} \tag{2.18}$$

   Result of equation 2.18 is $p \times 1$ vector which is referred as loading weights and contains weight of each variable on the covariance of $\mathbf{X}$ and $\mathbf{y}$. Normalisation of $\mathbf{w}_k$ is by dividing $\sqrt{\mathbf{w}_k\mathbf{w}_k'}$ and it sets $\mathbf{w}_k$ to unit length. All covariation of $\mathbf{y}_{k-1}$ and $\mathbf{X}_{k-}$ has then been moved to the scores.

2. Next step is to calculate the scores of $\mathbf{X}_{k-1}$ which is the projection of $\mathbf{X}_{k-1}$ onto $\mathbf{w}_k$ which compress $\mathbf{X}_{k-1}$ to fewer components. The scores are given with the formula 2.19

$$\mathbf{t}_k = \mathbf{X}_{k-1}\mathbf{w}_k \tag{2.19}$$

   The score vector in 2.19 is of size $n \times 1$.

3. Calculation of the $\mathbf{X}$ loadings is used to deflate $\mathbf{X}$. $\mathbf{X}$-loadings are calculated by project $\mathbf{X}_k$ onto $\mathbf{t}_k$. The formula is given in 2.20

$$\mathbf{p}_k = \frac{\mathbf{X}_{k-1}'\mathbf{t}_k}{\mathbf{t}_k'\mathbf{t}_k} \tag{2.20}$$

Size of the **X**-loading vector is $p \times 1$ and contains information of the covariance between $\mathbf{X}_{k-1}$ and $\mathbf{y}_{k-1}$ due to each variable. Dividing by $\mathbf{t}_k'\mathbf{t}_k$ normalizes $\mathbf{p}_a$.

4. To deflate $\mathbf{y}_{k-1}$ the $\mathbf{y}_{k-1}$-loadings are calculated. Here similarly $\mathbf{y}_{k-1}$ is projected onto $\mathbf{t}_k$ and normalize it by same method as was used for $\mathbf{p}_k$. Formula obtained for $\mathbf{y}_{k-1}$ loadings is 2.21

$$\mathbf{q}_a = \mathbf{y}_0'\mathbf{t}_a \frac{1}{\mathbf{t}_a'\mathbf{t}_a} \tag{2.21}$$

Now everything is in place to deflate $\mathbf{X}_{k-1}$ and $\mathbf{y}_{k-1}$.

5. Deflating $\mathbf{X}_{k-1}$ by removing values in the direction of the highest covariance of $\mathbf{y}_{k-1}$ and $\mathbf{X}_{k-1}$ in $\mathbf{X}_{k-1}$ space as follow

$$\mathbf{X}_k = \mathbf{X}_{k-1} - \mathbf{t}_k\mathbf{p}_k' \tag{2.22}$$

Now the $n \times p$ $\mathbf{X}_k$ matrix is prepared for next round of the algorithm if there exist any more information from y left.

6. Deflation of $\mathbf{y}$ is by removing values in the $\mathbf{y}_{k-1}$ space in direction of the highest covariance between $\mathbf{X}_{k-1}$ and $\mathbf{y}_{k-1}$. That is done by equation 2.23

$$\mathbf{y}_k = \mathbf{y}_{k-1} - \mathbf{t}_k\mathbf{q}_k' \tag{2.23}$$

For next round of the algorithm a deflated $\mathbf{y}$ is created which can be deflated again if there exist any variation to deflate on.

2.6.2 Partial Least Square Regression (PLSR)

Predicting new observation using components obtained by the algorithm in chapter 2.6.1 is by defining regression model which regresses $\mathbf{y}$ onto the components obtained. Estimation of $\hat{\boldsymbol{\beta}}$ is done in order to predict the new value by equation 2.2. Difference between estimating $\hat{\boldsymbol{\beta}}$ using PLSR compared to PCR is that we have a access to more exact loadings for $\mathbf{y}$ and we also have obtained the loading weights. Equation to estimate $\hat{\boldsymbol{\beta}}$ using results from PLS algorithm is described by equation 2.24 [20].

$$\hat{\boldsymbol{\beta}} = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}\mathbf{q} \tag{2.24}$$

where $\mathbf{W}$ is $p \times k$ matrix of loading weights for each component retained. The $p \times k$ matrix $\mathbf{P}$ containing loadings of $\mathbf{X}$ for each number of component $1 \cdots k$ and $1 \times k$ vector $\mathbf{q}$ containing the loadings of $\mathbf{y}$ for each component $1 \cdots k$. Solution of $\hat{\boldsymbol{\beta}}$ in equation 2.24 is used to predict $\mathbf{y}$ by equation 2.2.

2.7 Choosing number of Components

For both PLSR and PCR, number of components to retain is evaluated by $RMSEP_k$ in equation 2.6. Values of $RMSEP_k$ is plotted against k number of components retained. The point where value of $RMSEP_k$ is low compared to other values of $RMSEP_k$ in the graph and the value of $RMSEP_k$ is not lowered significantly by retaining one more component, number of components to retain is found. Nature of the prediction error is that it decreases faster using PLS because it has to take into account covariation between $\mathbf{y}$ and $\mathbf{X}$ which is of interest, and therefore prediction ability of principle components obtained by PLS is expected to perform better.

2.8 Using additional information

In this thesis the emphasis is mainly on using prediction models containing one response. When there exist additional responses within the dataset used for predicting, they should be used to obtain better prediction. Additional responses can be included with several developed PLS methods, such as multiresponse PLS (PLS2). Prediction model developed by using PLS2 will be capable of predicting a response which are not of interest. Running the PLS2 algorithm is similar to the PLS algorithm only the components are defined by the maximum variance between $\mathbf{X}$ and $\mathbf{Y}$.

Another method is Least Square Partial Least Squares (LS-PLS) where the additional responses are included as a column of the explanatory matrix. LS-PLS will still not serve the purpose of being able to apply the developed prediction model to the explanatory

matrix from a instrument in order to predict singe response, because this additional explanatory variables have to be measured when an object is measured by the NIR instrument. Then qualities of NIR machine will be dismissed.

2.8.1 Canonical Partial Least Squares (CPLS)

One solution of using additional responses to improve the prediction model, when there is only access to them when the model is developed, is by Canonical Partial Least Squares (CPLS). CPLS uses the additional data through Canonical correlation to adjust the prediction model.

The main idea behind Canonical correlation is to find two vectors a and b which maximizes the correlation between the matrices $\mathbf{Y}$ and $\mathbf{X}$ [24].

Before defining CPLS, the matrix $\mathbf{Y}$ and $\mathbf{Z}$ will be defined.

The matrix $\mathbf{Y}$ is a $n \times (q+1)$ matrix containing the original $n \times 1$ response vector $\mathbf{y}$ and the q additional $n \times 1$ vectors $\mathbf{y}$ of responses. The matrix $\mathbf{Z}$ is $n \times (p+1)$ matrix containing the product of $n \times p$ explanatory matrix $\mathbf{X}$ and $n \times (q+1)$ loading weight matrix $\mathbf{W}$. Definition of $\mathbf{Z}$ is given in equation 2.25

$$\mathbf{Z} = \mathbf{XW} \tag{2.25}$$

When applying CPLS in this thesis our response used is one dimensional vector. The additional responses are varying from being $n \times 1$ vector of additional response up to $n \times q$ number of additional responses.

In the Canonical Partial Least Square Regression the properties of Canonical correlation are applied to adjust the loading weights in the PLSR algorithm. The element a and the $n \times (q+1)$ vector $\mathbf{b}$ where q is number of additional responses are defined such they maximize the correlation between $\mathbf{Y}a$ and $\mathbf{Zb}$. This $\mathbf{b}$ is used to define the loading weights in the PLSR algorithm [25].

2.8.2 Modification on the PLS algorithm for CPLS

The main modification on the partial least square algorithm defined in equations 2.18 − 2.23 is regarding the calculations of covariance between $\mathbf{X}_{k-1}$ and $\mathbf{y}_{k-1}$ which gives the loading weights which was given with equation 2.18. Modification on the PLS algorithm in order to use CPLS is as following [26]

1. First modification is to define $p \times (q+1)$ loading weight matrix $\mathbf{W}_0$ where loading weight for every pair of $\mathbf{X}_{k-1}$ and $\mathbf{y} \in \mathbf{Y}_{k-1}$ is calculated. The number $q+1$ defines number of responses in $\mathbf{Y}_{k-1}$ where 1 is number of ordinary response and q is number of additional responses. p is number of columns in $\mathbf{X}_{k-1}$.

2. Defining $n \times (q+1)$ projection matrix $\mathbf{Z}_{k-1}$ were $\mathbf{X}_{k-1}$ is projected in smaller space $\mathbf{W}_0$. Then $\mathbf{X}_{k-1}$ has been reduced without losing any characteristics.

3. Concept of Canonical Correlation is then used by finding value a and $n \times (q+1)$ vector $\mathbf{b}$ which maximizes the correlation of $\mathbf{Y}_{k-1}a$ and $\mathbf{Z}_{k-1}b$.

$$max\left(corr\left(\mathbf{Y}_{k-1}\mathbf{a}, \mathbf{Z}_{k-1}\mathbf{b}\right)\right) \tag{2.26}$$

4. Solution vector $\mathbf{b}$ of the maximizing problem in equation 2.26 is then used to calculate new loading weights.

$$\mathbf{w}_k = \mathbf{W}_0\mathbf{b} \tag{2.27}$$

5. The PLS algorithm cointinues as described in equations 2.19 to 2.23 until next vector of loading weights need to be calculated.

Chapter 3

Measurements

Response values used in this thesis were obtained from 15 mm plugs removed from a salmon fillet. Fat value of those plugs were obtained using Low Field Proton Nuclear Magnetic Resonance (LF-NMR)(Shown in image 3.1). Explanatory values were obtained using the Near Infrared (NIR) instrument QMonitor. Aim of this chapter is to explain how those machines works and how the values they report are obtained and preprocessed.

3.1 Low Field proton Nuclear Magnetic Resonance (LF-NMR)

Response values reported in this thesis are fat values. Fat value is measured as % of fat out of all chemicals in the sample. Fat was measured on cylindrical plugs taken from the fillets without skin using core sampler on five or six predefined locations (locations defined in figure 4.1 and 4.8). Size of each plug was 15 mm in diameter. Length of each plug was according to thickness of the fillet where the plug was taken. Locations defined are supposed to span the variation of fat in the fillet. The plugs were measured using Low Field Proton Nuclear Magnetic Resonance (LF-NMR) with a Maran Ultra Resonance 0.5 Tesla equipped with a gradient probe (Oxford Instruments, Abingdon, UK). The calibration of the instrument is done by fish oil which is 100% fat. No preparation is needed to measure the fat in the plugs other than storing them in teflon box made for the LF-NMR scanner. If the sample did not fit the box it was cut longitudinally until it fit. Each sample was warmed in a thermostat for ten minutes at $40°C$ to get the

fat into a liquid form. Temperature of the magnetic is set to $40°C$. Weight of every plug is measured to retain fat and moisture value from the instrument. Method used for measuring is called "The one shot method" which was developed by Anved Teknolog As(Harstad Norway)[27]. Measurement time on each plug is three minutes. Error in measuring fat using LF-NMR is approximately $\pm 0.2\%$ fat which is good compared to hazardous chemistry methods on both fish and meat [28]. Quality of this method in addition to its accuracy is that no hazardous solvents and skills are needed to measure the fat and moisture using LF-NMR.

FIGURE 3.1: The LF-NMR machine in Nofima

Fundamental theory of NMR is based up on protons which have a spin, charge and a magnetic dipole. When they are subjected to a magnetic field they start to follow the magnetic field and reorient. The reorientation of the protons causes them to give out radio waves which can be detected and are measured [28]. By using multivariate statistical models the measures are translated to fat and moisture values.

3.2 Near Infrared Spectroscopy (NIR)

Near Infrared spectroscopy is a quite new technology even though Isaac Newton found out in 1665 that all colors in the electromagnetic spectrum are compressed in white light. He managed to separate them with a prism. There was not until in 1800, when William Herchel discovered that temperature increases towards red light in the electromagnetic spectrum. Herchel realization revealed that light contains more information

than only a color [29]. After the discovery of William Herchel there was no activity in the spectroscopic field until 1950. In 1950 Karl Norris, developed the technology to analyze agricultural goods [30]. Near Infrared radiation which has an harmonic motion has the length wavelength $760 - 2500\ nm$ in the electromagnetic spectrum. Figure 3.2 [31] shows the electromagnetic spectrum.

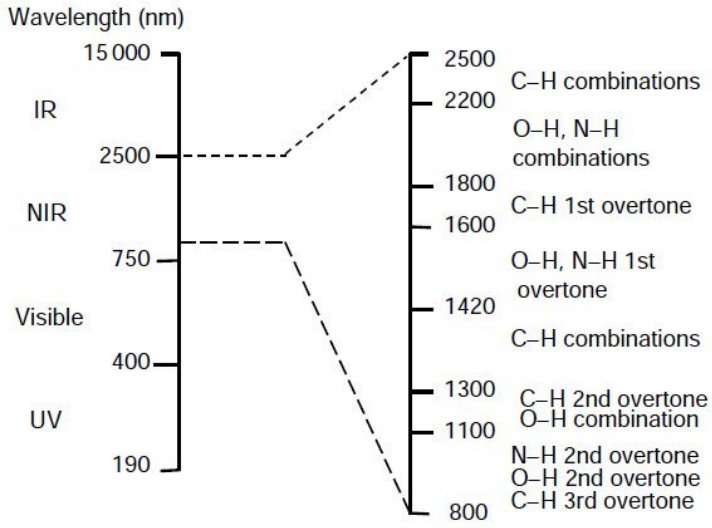


FIGURE 3.2: Electromagnetic spectrum [31]

Before defining wavelength, a vibration frequency has to be defined. Definition of vibration frequency is number of times the wave pattern in the electromagnetic spectrum is repeated in 1 second. Wavelength is defined by equation 3.1 [31].

$$\text{Wavelength} = \frac{\text{Velocity of light}}{\text{Frequency}} \tag{3.1}$$

Atoms of chemicals are held together by chemical bounds which vibrate. Frequency of the vibration is a function of the strength k between atoms and mass of the atoms m_1 and m_2. Each chemical has its inter-atomic distance. When the vibration frequency matches the inter-atomic distance there will be a transfer of energy which is plotted against wavelength. This plot is referred to as a spectrum. Relationship of potential energy and inter-atomic distance of those two atoms are parabolic [31]. When matching vibration frequency of a molecule it is referred to as overtone of chemical containing the molecules. The overtones appear on a precise location in the electromagnetic spectrum as

seen in figure 3.2[31]. When main interest of a measure in a sample is fat and water, the useful information in the electromagnetic spectrum is at wavelength range $760-1040\ nm$ which is centered around the second overtone of water. When fat in sample is of interest it can be measured through water because fat and water are negatively correlated. Main information about fat in NIR spectrum is at $920-930\ nm$ [27].

Near Infrared spectroscopy (NIR) is a commonly known technique when information is needed about amount of ingredients in food, animal feed, wool, textiles, chemicals and petroleum [29]. Main cost of measuring by using NIR technology is to buy the machine and the development of prediction model to use along with the measures obtained. Main maintenance cost is to replace the lamps when the are burnt out. After calibration and creation of prediction model the NIR machine is capable of measuring a huge amount of material [15]. Main advantage of NIR technology is that no special chemicals or dangerous solvents are needed when operating the machine. Usually no preparation is needed on the measured sample. NIR scanners have saved the industry a huge amount of money by being quick to measure, reducing lab work and by not requiring experienced worker to operate it. Main NIR methods for optics sampling are shown in figure 3.3.

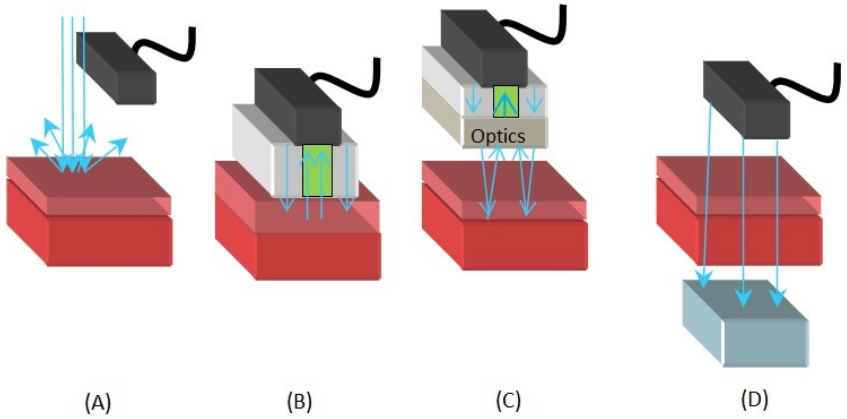


FIGURE 3.3: Main NIR methods [7, 33]

When installing a NIR machine to a production line the ideal setup is to use remote interactance ((A) 3.3). When using remote interactance it becomes possible to place the machine above the sample to measure. In remote reflectance is no information about the interior of the sample which is a drawback when determination of quantity of fat within salmon fillet. Solution could be interactance measure ((B) 3.3). It is hard to

implement interactance measure onto a conveyor belt which is a drawback. Interactance also only retrieve information from inside of the sample. Using transmission ((D)3.3) to measure NIR is difficult when thickness of sample vary. Then strength of the NIR light have to vary also which can be complicated when sample has unequal thickness. Measuring using remote contact free interactance ((C)3.3) is the optimum setup of NIR machine when measuring fat and color of salmon fillets. Having the aim to develop such a machine the product QMonitor was created.

3.2.1 QMonitor

In the year 1996 the company TiTech was established. TiTech is a company that makes NIR scanners to sort waste in order to recycle. In the year 2003 TiTech, Sintef and Matforsk started a project named QVision. Aim of the project was to make a scanner that could scan food products and report useful values for the industry. In the year 2005 the NIR scanner QMonitor was ready as a product for the food industry and the company QVision was established as a subsidiary company of TiTech [7]. One of the food products QMonitor was developed to scan where salmon fillets. The spectral images from QMonitor have been deciphered in order to report fat, [12], water [27] and salt content [11] in salmon fillets. In addition pigment is measured with a visible light detector in the region $460 - 760nm$ which also inside the QMonitor. QMonitor is also used for meat, crabs, dried salted coalfish [21] and french fries. In 2012 the name of QVision was change to Odenberg when the TiTech company was bought by Odenberg.[15].

When QMonitor managed to scan salmon fillets and predict fat based on developed prediction model a problem appeared. The method used to develop the prediction model did not take into account the high variation of fat in the fillet. The whole spectra image was averaged for modeling and then information about the fat variation was determinated. A suggestion of solution to the problem was published 2009 and was described in a paper[12]. The solution aims to capture the variation of fat by measuring fat on predifined locations of every fillet and retain the spectral image values from same location when developing a prediction model. The prediction model developed is then applied to every pixel of the spectral image and mean fat value for the whole fillet is reported.

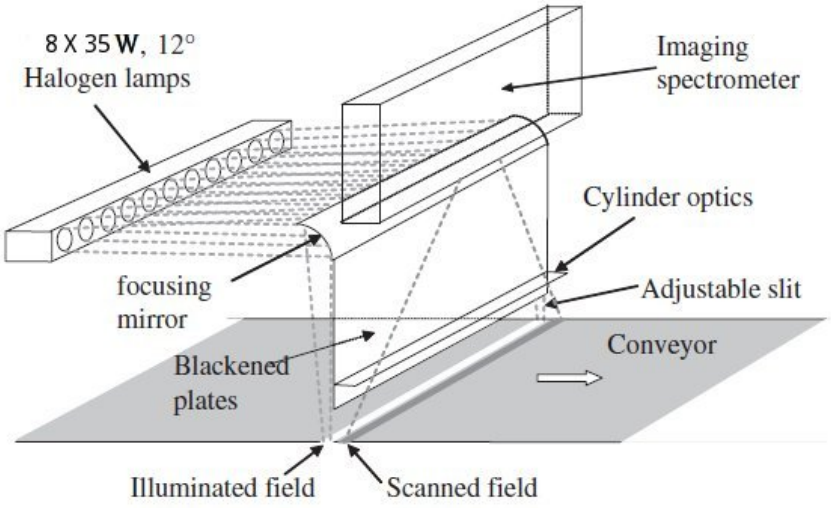


FIGURE 3.4: Setup of QMonitor [22]

QMonitor is an online imaging scanner shown in figure 3.4 which measures the backscatter of light from the sample. The scanner is placed 11.4 cm above a conveyor belt which the salmon fillet is laid on with the skin facing down. Light source in the machine consists of eight $35W$, 12° Reflecto industry halogen lamps. Each having light illuminating in same direction as the conveyor belt, horizontally within the Scanner. The light illuminates onto a mirror which change the direction of the light vertically on to the conveyor belt. The light goes $15 - 20$ mm into the salmon fillet. A cylindrical lens collects the light and is separated from the light source by 2 cm metal plate which also shield the detector from unwanted reflected light from the fish surface [11].

A Charged Couple Device (CCD) detector (Imaging spectrometer) translates the backscatter from the lens to spectral values. Speed of the conveyor belt can be up to $1,5\ ms^{-1}$ but when the data was collected in this thesis the speed was approximately $0,2\ ms^{-1}$. Size of each spectral image depends on the speed of the conveyor belt. With the speed of $0,2\ ms^{-1}$ the size of the image is about 60 pixels in perpendicular direction of the conveyor belt movement were each pixel is 1.2 cm and $300 - 400$ pixels in the direction of the conveyor belt movement with pixel size of 0.5 cm [11].

The CCD consist of 15 channels who all capture spectral image with 20 nm betweenfrom 760 nm to 1040 nm resulting in 15 spectral images shown in figure 3.5 which the multivariate statistics are used to translate.

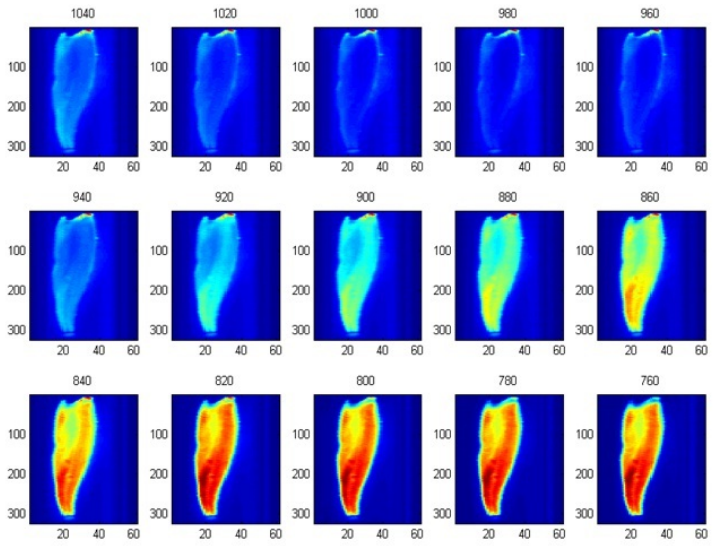


FIGURE 3.5: Spectra values obtained from salmon fillets [32]

3.3 Preprocessing of NIR data

Measuring chemical using NIR technology need preprocessing before applying multivariate statistics. The preprocessing removes unwanted information and standardizes the measures. Main preprocessing methods used on QMonitor data are calculation of absorbance values (ABS) and Standard Normal Variate value (SNV).

3.3.1 Raw value of spectra

Measure of reflectance of light is defined as raw value in spectroscopic data. Raw value is a ratio between the measure obtained and a known reference value. Reflectance is stated in equation 3.2

$$R = \frac{I}{I_0} \tag{3.2}$$

where I is the measure obtained by NIR machine. I_0 is a known reference value. During calibration of NIR machine a piece with known value is placed under the NIR detectors and reported values from the machines are adjusted to the values they are supposed to

report when this kind of standardized piece is measured. Calibration value is contained by the variables I_0. Reflectance value is non-linear which can introduce problems.

3.3.2 Absorbance (ABS) value

Using nonlinear observations can introduce problems which are not of inters. Solution is to make the reflectance values linear by equation 3.3.

$$ABS = -log_{10}\left(\frac{I}{I_0}\right) \tag{3.3}$$

Equation 3.3 is a linear function of the measure obtained which multivariate statistics can be applied to.

3.3.3 Standard Normal Variate (SNV) of spectra

The Standard Normal Variate (SNV) preprocessing applies to individual spectra and can therefore be done on the dataset before cross validation. SNV removes slope variation of the spectral waves, which contains non useful information [34]. In this thesis SNV correction is used to remove offset due to light scattering [27]. Calculation of SNV for every sample is stated in equation 3.4.

$$SNV(x_{ik}) = \frac{x_{ik} - \bar{x}_i}{\sqrt{\frac{\sum_{k=1}^{n}(y_{ik}-\bar{y}_i)^2}{n-1}}} \tag{3.4}$$

Each value of the explanatory matrix is standardized using SNV. Equation 3.4 retrieves the SNV of sample number i containing n wavelengths. Wavelength number k is within sample i [34]. Comparison of absorbance values and SNV values are shown in figure 3.6.

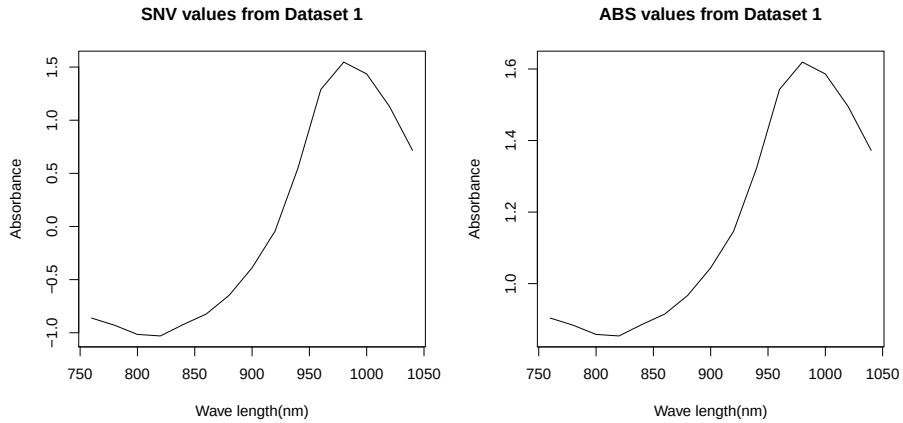


FIGURE 3.6: Raw and SNV preprocessed data from QMonitor

After applying SNV as in figure 3.6 the mean of the waves is 0 and values of the waves is centered at 0.

3.4 How NIR data is collected from QMonitor spectral images

During the master studies followed by this master thesis, different programs and data format have been used to measure salmon fillets using the NIR machine QMonitor. The first year .udp files were collected using program written in $C++$ by QVision and Nofima Mat. The udp. files are rather big compared to other data files. They contain a lot of data which is not used when fat and color values are obtained from QMonitor. To calculate the fat and pigment values a program which converts the .udp files to .bin files was made by Nofima Mat. Using Matlab the obtained regression coefficients are multiplied by observed values coming from the .bin image. Second program used published by QVision is QAnalaysis, version 2.1 which is used to calibrate the QMonitor and adjust the QMonitor according to .XML file which is loaded into the program. QAnalysis is capable of analyze individual value in the images and use existing prediction model to predict fat and color of the salmon fillet. The program was also capable of converting .udp files to .bin files. Version 2.2 was published were useless buttons and bugs in the first program were fixed. Finally the current version 2.3 was published which is capable of record data. All Matlab scripts which have been made for the QMonitor

came also in a folder included in version 2.3. The scripts included in versions 2.3 in combination to special made scripts were used to obtain the explanatory variables used in this thesis. The QAnalysis tools are built using Matlab compiler.

When measuring each salmon fillet, each measure is stored in .udp or .bin file. The file is loaded into a Matlab program made by Nofima Mat and QVision which allows the user to pick out values from selected locations in the spectral image. The raw, abs and SNV values are calculated and retrieved for those locations, shown in figure 3.7.

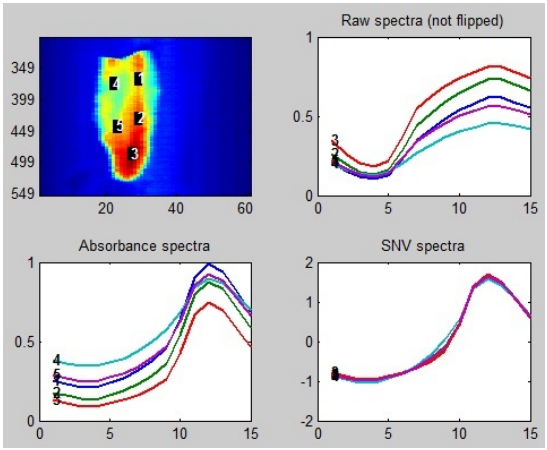


FIGURE 3.7: Program to pick out information from Spectral Image

Using this program in figure 3.7 and images captured by digital camera the explanatory variables in datasets 2 and 3 were created. Where result is $n \times 15$ matrix of SNV values from the selected points on the spectral image.

Chapter 4

Material

4.1 The datasets

Three datasets are used in this thesis. The first dataset used is originated from the paper describing the most recent sampling and calibration strategy when developing model for QMonitor measuring fat in fillet [12]. It is used to learn the methods when predicting fat in salmon fillet and understand how the NIR data samples and preprocessed, prior to model development. The NIR values from the spectral image in dataset 1 were obtained using the prototype of the NIR instrument QMonitor standing in Nofima, Ås Norway.

The second dataset was part of an experiment where Stofnfiskur and Nofima conducted a trial to find the best available method to measure fat and pigment in salmon fillets. Stofnfiskur did send fish to Nofima Norway where the fish was measured using the old QMonitor. Based on the results obtained, Stofnfiskur decided to buy QMonitor.

The third dataset was collected in Iceland at Stofnfiskur specially for this thesis in order to develop prediction model for their New QMonitor. NIR spectra values were obtained by the QMoinitor which will be referred as the New QMonitor. The fish in dataset three was a part of a larger experiment where additional measurements were conducted. The additional measures are stored in the dataset.

Each QMonitor measure consist of 15 images shown in figure 3.5 measuring light back scattering in 15 different light regions from 760 nm to 1040 nm. The point where the plug was collected is picked out on all 15 spectral images and used as an explanatory

variable in the regression. Preprocessing was to convert the raw values to abs values which were converted to SNV values.

The idea behind the locations of the plugs removed from the fillet is to span the fat variation in the fillet. Each plug is 15 mm in diameter and the thickness of each plug is according to the thickness of the fillet where the plug was collected. The plugs in all datasets which are used as a response were measured using $LF-NMR$ machine located in Nofima. In order to retain fat and moisture, weight of each plug had to be measured.

Number of observations are summarized in table 4.1.

Nr of	**Dataset 1**	**Dataset 2**	**Dataset 3**
Fish	15	43	24
Round weight	$2-5$ kg	$1-8$ kg	$1-6$ kg
Fillets per fish	2	1	1
Plugs per fillet	5	$2-6$	5
Observations	150	98	120
Broken observations	5	10	13
Total # observations	145	88	107
Fat range	$3.84-26.22$	$0.23-32.84$	$0.23-32.8$
Average	13.85	$14,25$	14.24
StDev	6.12	$6,6$	6.6

TABLE 4.1: Overview of the datasets

In table 4.1, the broken observations are values which are not likely to be true. Example of broken value are negative fat value in a plug or spectra values far from its neighbors. Broken observations were removed in analysis which is a common rule and is done commonly in NIR publications [12].

In PCR and PLSR the main interest is on eigenvalues and eigenvectors of $\mathbf{X}'\mathbf{X}$ to get an idea of how many components are needed and to see if it is possible to reduce the complexity of the prediction model by using subset of component to replace $\mathbf{X}$. When prediction of response $\hat{\mathbf{y}}$ is of interest the covariance and correlation between $\mathbf{X}$, $\mathbf{y}$ need to be illustrated. Also is it faverable to look at the covariance between the principle components and $\mathbf{y}$.

The dataset are made by different groups of people and companies, recorded at different time and on different locations in the world.

4.2 Dataset 1

Martin Høy and Jens Petter Wold are authors and co-authors of many papers concerning NIR technology [8, 12, 21, 22, 27]. They are the main source of information concerning Near Infrared Spectroscopy in this thesis. They have many years of experience and a Phd degree which comes into play when sampling data using the equipment needed to create the prediction model published. Martin Høy and Jens Petter Wold among others were involved when the machines were developed and the computer programs made to obtain the measures from the machine.

As seen in table 4.1 dataset 1 consists of 15 fish weighing from 2 to 5 kg round body weight. Both fillets from each fish were used, in total 30 fillets. The fish were machine filleted with skin still intact by a Norwegian salmon company (Bremnes Seashore AS) [12].

Five plugs were collected from each fillet on the locations in figure 4.1 [12]. They were measured by LF-NMR at Nofima by Frank Lundby, which is one of the most experienced person at Nofima Mat using LF-NMR scanners. Franks has conducted calibration studies in order to develop the best method on how to calibrate LF-NMR scanner on a standard way across labs [28].

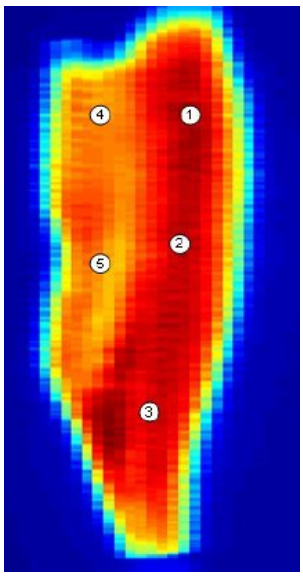


FIGURE 4.1: Plug locations [12]

From the 30 fillets, 5 plugs, in total 150 plugs were collected. Out of 150 plugs collected from the 15 fish only 5 of them were broken.

4.2.1 Statistics of Dataset 1.

To understand the variance structure of $\mathbf{X}$ the eigenvalues are obtained from the covariance matrix of $\mathbf{X}$ which is the product $\mathbf{XX'}$ and their main properties evaluated.

i	**Eigenval.**	**Variance ratio**	**Total**	**Cov(Pc,y)**	**Corr(X,y)**	**Cov(X,y)**
1	3.92	0.77	0.77	137.50	-0.18	-0.05
2	1.07	0.21	0.98	-8.57	-0.52	-0.11
3	0.06	0.01	0.99	-0.91	-0.81	-0.16
4	0.02	0.00	1.00	0.89	-0.92	-0.16
5	0.02	0.00	1.00	0.86	-0.92	-0.17
6	0.00	0.00	1.00	0.17	-0.75	-0.18
7	0.00	0.00	1.00	0.13	-0.09	-0.02
8	0.00	0.00	1.00	0.04	0.86	0.29
9	0.00	0.00	1.00	-0.04	0.95	0.64
10	0.00	0.00	1.00	0.04	0.92	0.39
11	0.00	0.00	1.00	-0.04	-0.62	-0.11
12	0.00	0.00	1.00	-0.01	-0.89	-0.29
13	0.00	0.00	1.00	0.00	-0.92	-0.20
14	0.00	0.00	1.00	0.02	0.12	0.01
15	0.00	0.00	1.00	0.00	0.40	0.11

TABLE 4.2: Covariance and correlation structure of dataset 1

In table 4.2 size of eigenvalue i in relation to other eigenvalues of $\mathbf{XX'}$ shows how much variation principle component i contains and is referred to as variance ratio. To see how much variation may be explained by the first i components. The ratio between the first i eigenvalues divided by sum of all eigenvalues is calculated. In table 4.2 it is referred to as Total and shows how much total variation may be explained by the first i components. When the predicted response is of main interest the covariance between the principle component i and $\mathbf{y}$ is calculated. It is defined as the inner product of eigenvector i of the covariance matrix $\mathbf{XX'}$ and $\mathbf{Xy}$ given in equation 4.1

$$Cov(Pc_i, \mathbf{y}) = e_i'\mathbf{X'y} \tag{4.1}$$

Where e_i in equation 4.1 is eigenvector i of the covariance matrix $\mathbf{XX'}$. When $\mathbf{X}$ is going to be used to predict $\mathbf{y}$ it is also good to look at the correlation and covariation between $\mathbf{y}$ and the columns of $\mathbf{X}$ which is also shown in table 4.2.

To illustrate how much variation may be explained by the first i components the Scree plot is made. The Scree plot is shown in figure 4.2.

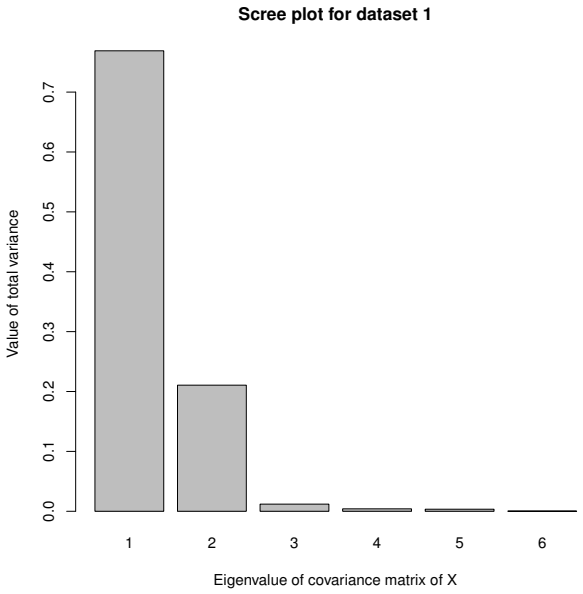


FIGURE 4.2: Scree plot of the eigenvalues of $\mathbf{XX}'$ in dataset 1

When estimating $\hat{\beta}$ in prediction a the choise of method depends on the correlation among the variables in $\mathbf{X}$. The correlation among the X variables is shown in table 4.3.

	s760	s780	s800	s820	s840	s860	s880	s900	s920	s940	s960	s980	s1000	s1020	s1040
s760	1.00	0.92	0.63	0.36	-0.09	-0.44	-0.91	-0.59	-0.18	0.01	0.80	0.44	0.14	-0.85	-0.88
s780	0.92	1.00	0.88	0.69	0.29	-0.09	-0.76	-0.86	-0.53	-0.36	0.93	0.74	0.49	-0.75	-0.91
s800	0.63	0.88	1.00	0.94	0.68	0.34	-0.43	-0.99	-0.84	-0.73	0.90	0.94	0.81	-0.52	-0.77
s820	0.36	0.69	0.94	1.00	0.88	0.61	-0.12	-0.95	-0.95	-0.89	0.78	0.97	0.94	-0.33	-0.60
s840	-0.09	0.29	0.68	0.88	1.00	0.91	0.36	-0.74	-0.96	-0.98	0.46	0.84	0.95	0.06	-0.21
s860	-0.44	-0.09	0.34	0.61	0.91	1.00	0.70	-0.42	-0.79	-0.88	0.09	0.57	0.80	0.39	0.16
s880	-0.91	-0.76	-0.43	-0.12	0.36	0.70	1.00	0.35	-0.13	-0.31	-0.61	-0.16	0.16	0.77	0.74
s900	-0.59	-0.86	-0.99	-0.95	-0.74	-0.42	0.35	1.00	0.88	0.77	-0.89	-0.95	-0.85	0.47	0.73
s920	-0.18	-0.53	-0.84	-0.95	-0.96	-0.79	-0.13	0.88	1.00	0.97	-0.63	-0.93	-0.98	0.12	0.41
s940	0.01	-0.36	-0.73	-0.89	-0.98	-0.88	-0.31	0.77	0.97	1.00	-0.47	-0.87	-0.99	-0.04	0.25
s960	0.80	0.93	0.90	0.78	0.46	0.09	-0.61	-0.89	-0.63	-0.47	1.00	0.84	0.59	-0.81	-0.94
s980	0.44	0.74	0.94	0.97	0.84	0.57	-0.16	-0.95	-0.93	-0.87	0.84	1.00	0.93	-0.45	-0.70
s1000	0.14	0.49	0.81	0.94	0.95	0.80	0.16	-0.85	-0.98	-0.99	0.59	0.93	1.00	-0.10	-0.39
s1020	-0.85	-0.75	-0.52	-0.33	0.06	0.39	0.77	0.47	0.12	-0.04	-0.81	-0.45	-0.10	1.00	0.93
s1040	-0.88	-0.91	-0.77	-0.60	-0.21	0.16	0.74	0.73	0.41	0.25	-0.94	-0.70	-0.39	0.93	1.00

TABLE 4.3: Correlation between the $\mathbf{X}$ variables in dataset 1

The nature of the fat in a fillet is that there is more fat in the belly and less fat in the backbone. Locations where the plugs were collected is displayed in figure 4.1. Distribution of fat within each location is shown in figure 4.3.

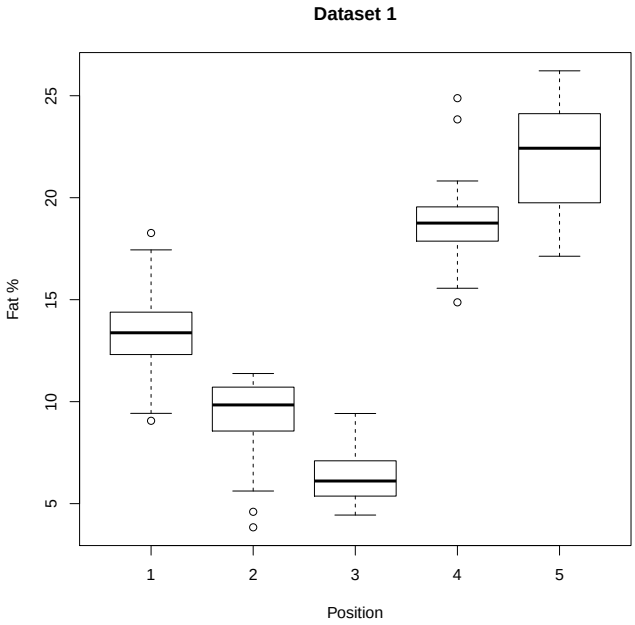


FIGURE 4.3: Fat in plug versus location of plug in dataset 1

In figure 4.3 the plugs from the belly flaps are plug D and E, referred to as 4 and 5 in figure 4.1 which contains the highest fat in the fillet and the plugs from the backbone are plugs A, B and C referred to as 1,2 and 3 in figure 4.1. In order to explore the explanatory matrix further, scores and loadings are retained and plotted in figure 4.4

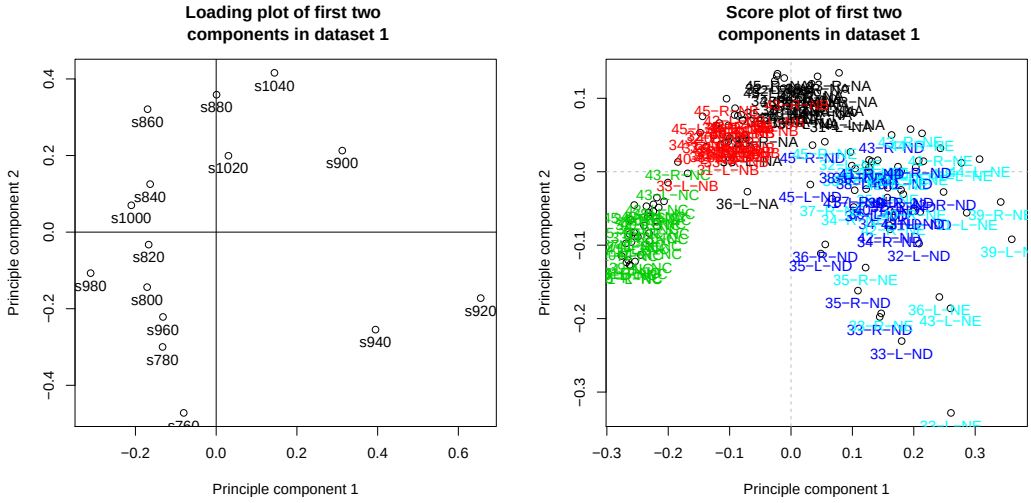


FIGURE 4.4: Score and loadings in dataset 1.

The loading plot in figure 4.4 show how much impact the first two components retain information from $\mathbf{X}$. Main information about fat is located in the columns of $\mathbf{X}$ where measures from the spectral images around 930 nm are located. The scores show if any grouping is among the data in $\mathbf{X}$. In figure 4.4 grouping among the locations of the plug appear. The observations are colored, according to where they were taken from the fillet in figure 4.1.

4.3 Dataset 2

In fall 2010 a comparison of fat and pigment measurement methods available in salmon breeding was done in cooperation with Nofima and Stofnfiskur which the author of this thesis was in charge of. Dataset 2 was recorded as a part of this comparison. A quality test was also done on fish from Stofnfiskur which is stored as additional responses. The fish was selected from 3 year old fish in Iceland. It was selected within a weight interval that spans likely weights of fish. The fat and pigment measurement instrument will get exposed in Stofnfiskur's nucleus. The fish was from seven different weight classes, $1-2$ kg, $2-3$ kg up to $7-8$ kg. Between 2 to 8 fish were in each weight class. Stofnfiskur did send in total 45 gutted fish to Nofima in Ås as seen in table 4.1. Length, ungutted and gutted weight were measured in Iceland.

After measuring the fish in Iceland it was put into styrofoam box on ice and sent to Norway by air on Monday in the first week of November 2010. On Thursday morning the fish arrived in Nofia Marin Ås. On the first day the fish was filleted and quality measures obtained. Example of quality measures are the gaping of the fillet, texture in the meat of the fillet and and value of Ph in the fillet after filleting. Then the fillets were measured using QMonitor standing in Nofima Mat. After measuring the fillets,

FIGURE 4.5: Measuring using old QMonitor at Nofima Ås.

the plugs were removed from the fillets. Thickness of each plug was measured by measuring average thickness of the fillet where the plug was collected. Figure 4.6 shows these measures.

FIGURE 4.6: Bjarne and Målfrid measuring plug thickness

The skin was removed from the fillets and they were minced and samples sent to Sundalsøra to determine the chemical values of fat and pigment in the fillets in order to see the prediction performance of the machines.

FIGURE 4.7: Mincing the fillet in order to take 30 gr sample.

In quality measures and filleting, two fillets were dismissed. Total number of plugs measured with LF-NMR was therefore 98. Fat is measured as % of total chemicals of the plug. Main statistics of the dataset are listed in table 4.4.

	Max	Min	Mean	Std. deviation	#
Ungutted weight of whole fish	7.46	1.44	4.59	1.75	43
Gutted weight of whole fish	6.8	1.32	4.19	1.6	43
Length of whole fish	83	49	69.28	9.22	43
Plug thickness	36	10	19.81	8.0	98
Plug weight	3.68	1.14	2.34	0.60	98
Plug Moisture %	78.48	-0.1	65.7	8.27	98
True Fat from whole fillet	21.59	6.84	16.63	3.00	98
Astataxthin from whole fillet	11.12	4.41	7.1	1.45	98

TABLE 4.4: Main statistics of additional responses in dataset 2

In table 4.4, value of fat and pigment in the fillets is measured as % of total chemicals, and amount of Astataxthin in *mg* out of total chemicals in 30 gr sample taken from whole minced fillet without skin.

The author of this thesis did measure the fat in the plugs with LF-NMR machine. The LF-NMR machine was calibrated using standardized fat samples for meat. Reason for using meat for calibration was because a student in Nofima Mat was asked to start up the LF-NMR scanner and he was used to use the meat standard samples for calibration. Calibration of LF-NMR scanner should be done using fish oil. The machine is supposed to report 100% fat in the calibration.

FIGURE 4.8: Cylindrical plugs removed from the fillet.

Only one fillet out of two were selected because of the high correlation between fillets. Decision was made to gather six 15 mm cylindrical plugs from each fillet to span the variation of fat in the fillet. The location are show in figure 4.9.

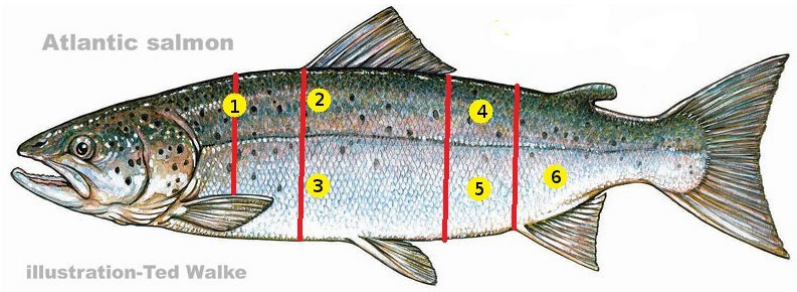


FIGURE 4.9: Position of plugs in dataset 2

Every plug from first three fish were measured. Thereafter only two plugs were measured from each fish at random in total 98 plugs out of 270 plugs, because of how time consuming it is to measure each plug. After measuring the plugs the author selected the corresponding pixels from the spectral image on the locations where the plugs were supposed to be from using a Matlab function created by QVision that retrieved SNV values for each location of the plug.

4.3.1 Statistics of Dataset 2

When PCR and PLSR is utilized relation of **y** and **X** is of most interest in addition to variation structure of **X**. Using same methods as in chapter 4.2.1 the results in table 4.5.

i	**Eigenval.**	**Variance ratio**	**Total**	**Cov(Pc,y)**	**Corr(X,y)**	**Cov(X,y)**
1	3.51	0.94	0.94	89.70	-0.48	-0.09
2	0.18	0.05	0.99	3.15	-0.83	-0.17
3	0.02	0.01	1.00	0.02	-0.90	-0.23
4	0.01	0.00	1.00	-0.17	-0.91	-0.21
5	0.00	0.00	1.00	0.42	-0.92	-0.15
6	0.00	0.00	1.00	0.48	-0.85	-0.09
7	0.00	0.00	1.00	-0.03	0.80	0.11
8	0.00	0.00	1.00	-0.03	0.90	0.39
9	0.00	0.00	1.00	-0.02	0.92	0.61
10	0.00	0.00	1.00	0.01	0.92	0.38
11	0.00	0.00	1.00	0.02	-0.78	-0.14
12	0.00	0.00	1.00	-0.03	-0.88	-0.33
13	0.00	0.00	1.00	0.00	-0.90	-0.24
14	0.00	0.00	1.00	0.01	0.00	0.00
15	0.00	0.00	1.00	0.00	0.69	0.17

TABLE 4.5: Covariance and correlation structure of Dataset 2

To understand to which degree the variation is explained the first eigenvalues are displayed using scree plot in figure 4.10.

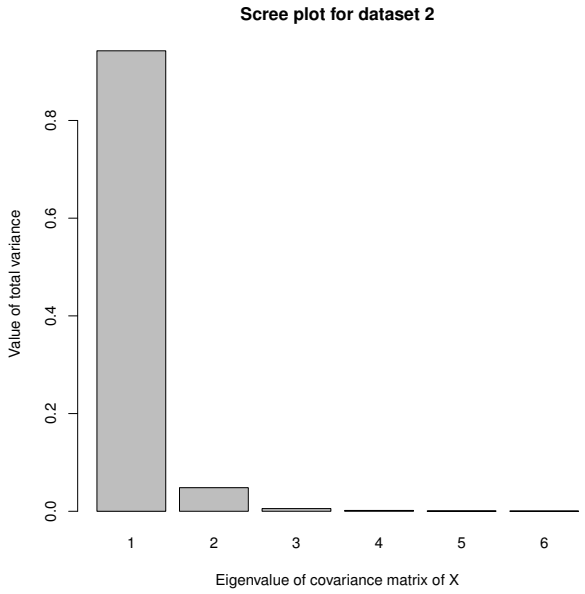


FIGURE 4.10: Scree plot of the eigenvalues of $\mathbf{X}\mathbf{X}'$ in dataset 2

When choosing method to estimate $\hat{\beta}$ the information about the correlation among the columns of $\mathbf{X}$ is needed. The correlation is shown in table 4.6.

	s760	s780	s800	s820	s840	s860	s880	s900	s920	s940	s960	s980	s1000	s1020	s1040
s760	1.00	0.84	0.62	0.53	0.41	0.13	-0.81	-0.62	-0.58	-0.55	0.78	0.69	0.55	-0.66	-0.89
s780	0.84	1.00	0.94	0.90	0.84	0.63	-0.99	-0.95	-0.92	-0.91	0.95	0.97	0.91	-0.35	-0.94
s800	0.62	0.94	1.00	0.99	0.96	0.83	-0.95	-1.00	-0.99	-0.99	0.92	0.99	0.98	-0.13	-0.84
s820	0.53	0.90	0.99	1.00	0.99	0.88	-0.91	-0.99	-0.99	-0.99	0.88	0.97	0.98	-0.06	-0.79
s840	0.41	0.84	0.96	0.99	1.00	0.94	-0.84	-0.96	-0.98	-0.98	0.83	0.93	0.97	0.06	-0.70
s860	0.13	0.63	0.83	0.88	0.94	1.00	-0.63	-0.84	-0.88	-0.88	0.61	0.77	0.88	0.34	-0.44
s880	-0.81	-0.99	-0.95	-0.91	-0.84	-0.63	1.00	0.95	0.92	0.90	-0.97	-0.97	-0.91	0.36	0.94
s900	-0.62	-0.95	-1.00	-0.99	-0.96	-0.84	0.95	1.00	1.00	0.99	-0.92	-0.99	-0.99	0.10	0.83
s920	-0.58	-0.92	-0.99	-0.99	-0.98	-0.88	0.92	1.00	1.00	0.99	-0.89	-0.98	-0.99	0.03	0.79
s940	-0.55	-0.91	-0.99	-0.99	-0.98	-0.88	0.90	0.99	0.99	1.00	-0.85	-0.97	-1.00	-0.00	0.77
s960	0.78	0.95	0.92	0.88	0.83	0.61	-0.97	-0.92	-0.89	-0.85	1.00	0.95	0.86	-0.46	-0.96
s980	0.69	0.97	0.99	0.97	0.93	0.77	-0.97	-0.99	-0.98	-0.97	0.95	1.00	0.97	-0.23	-0.90
s1000	0.55	0.91	0.98	0.98	0.97	0.88	-0.91	-0.99	-0.99	-1.00	0.86	0.97	1.00	0.01	-0.77
s1020	-0.66	-0.35	-0.13	-0.06	0.06	0.34	0.36	0.10	0.03	-0.00	-0.46	-0.23	0.01	1.00	0.61
s1040	-0.89	-0.94	-0.84	-0.79	-0.70	-0.44	0.94	0.83	0.79	0.77	-0.96	-0.90	-0.77	0.61	1.00

TABLE 4.6: Correlation of $\mathbf{X}$ in dataset 2

Plugs 1,2 and 4 where collected from the backbone which is leaner than the belly flaps were plug 3 and 5 were collected. Tail of a salmon also have low fat where plug 6 was collected. Distribution of fat within location is displayed in figure 4.11.

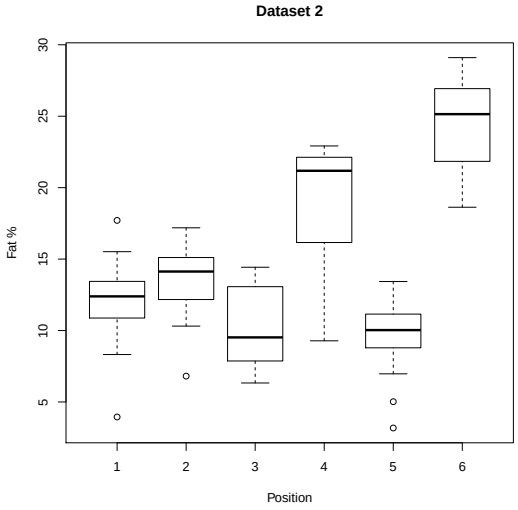


FIGURE 4.11: Fat in plug versus plug location in, dataset 2

To see the impact on the first two components on the columns of **X** the loading plot is retained. Main information about fat is located in the columns of **X** where measures from the spectral images around 930 nm are located. The Loadings are retained as descried in chapter 4.2.1. Loadings are plotted for dataset 2 in figure 4.12.

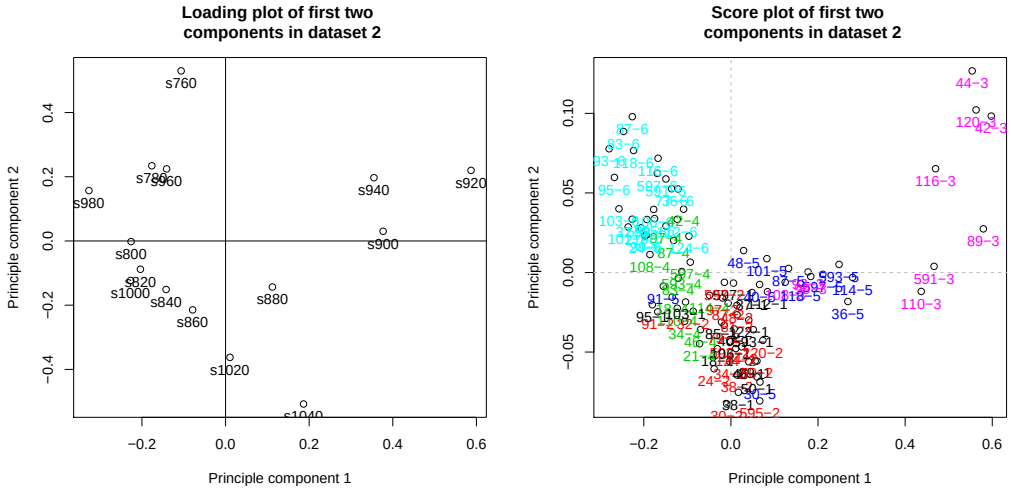


FIGURE 4.12: Score and loadings in dataset 2.

The scores are colored according to where the plugs come from the fillet the measures are retained from in figure 4.12.

4.4 Dataset 3

After the instrument comparison at Nofima Ås, Norway, Stofnfiskur decided to buy a QMonitor to measure fat and pigment in the nucleus. The new instrument was installed at Stofnfiskur in Iceland. The predicted values retained using existing prediction model on the data obtained from the New QMonitor were wrong according to literature.

Dataset 3 was created in Iceland using the new QMonitor. A number of 2300 fish were measured in fall of 2011 where predicted fat values are needed. Selection of fish to create the next generation will be done in summer 2012 which is the aim to use the results from this thesis to select. Out of these 2300 fish a sample of 24 fish were selected. After learning the procedure of sampling in Norway the author did all the sampling in Iceland. After removing the plugs the skin was removed and the fillet was minced and 30 gr. samples collected which were sent to Sunndalsøra to determine the fat and pigment in the fillets. Sunndalsøra reported an error when mincing the fillets. The error was the fillets were not minced enough. In order to create new prediction model, 24 fish were collected from weight classes $1-2$ kg, $2-3$ kg, $3-4$ kg, $4-5$ kg and $5-6$ kg. The weight interval $1-6$ kg which is similar to the weight of the fish the machine will get exposed to in the future when fat and pigment will be measured. When filleting the fish additional measures were retained.

FIGURE 4.13: Collecting Salmon, measuring round body weight and length

Additional measures were conducted on the fish which are shown in table 4.8 which were used in order to improve the prediction ability of the model. Main summaries of the additional measures are shown in table 4.7.

	Max	Min	Mean	Standard dev	#
Fillet weight, kg	1.95	0.22	1.52	0.30	48
Ungutted weight, kg	6.444	0.8	4.32	1.32	24
Length of whole fish, cm	83	0	69.0	15.7	120
Sex of fish, 1=male, 2=female	2	1	1.37	0.46	24
Intestine weight, gr	460	48	251.25	92.1	24
Plug weight	4.69	1.27	3.42	0.79	120
Plug moisture	79.3	-0.07	69.2	8.0	120

TABLE 4.7: Main statistics of additional responses in datset 3

Sex of the fish is recorded with 1 for male fish which has not matured, and 2 for female fish which has not matured. Most of the fish in this dataset were males, as can be seen in table 4.7. Intestine weight is weight of intestine without liver and heart.

FIGURE 4.14: Measuring weight of fillet

From each fish one fillet out of two were selected randomly. Ungutted weight and length of each fish was measured. Intestine weight without liver and heart was measured. In addition fillet weight was measured.

One fillet of each fish was used and five plugs were collected according to figure 4.1, further informations are in table 4.1 about the sampling. Visual image was captured of each fish in order to improve collection of the data on the spectral images.

Each fillet was measured using the QMonitor standing in Iceland seen in figure 4.15.

FIGURE 4.15: Measuring the fat using QMonitor in Stofnfiskur, Iceland

4.4.1 Statistics of Dataset 3

In order to explore the variation of **X** and the relation of **X** and **y** and the components made of **X** same methods are used as in chapter 4.2.1 to obtain the results in table 4.8.

XX_i'	**Eigenval.**	**Variance ratio**	**Total**	**Cov(Pc,y)**	**Corr(X,y)**	**Cov(X,y)**
1	2.29	0.92	0.92	87.97	-0.69	-0.11
2	0.13	0.05	0.97	-0.12	-0.91	-0.17
3	0.05	0.02	0.99	0.96	-0.93	-0.18
4	0.02	0.01	1.00	0.01	-0.92	-0.15
5	0.01	0.00	1.00	0.54	-0.93	-0.13
6	0.00	0.00	1.00	-0.04	-0.87	-0.09
7	0.00	0.00	1.00	-0.05	0.82	0.08
8	0.00	0.00	1.00	0.05	0.93	0.30
9	0.00	0.00	1.00	-0.04	0.94	0.53
10	0.00	0.00	1.00	-0.06	0.91	0.28
11	0.00	0.00	1.00	-0.01	-0.76	-0.13
12	0.00	0.00	1.00	-0.01	-0.94	-0.24
13	0.00	0.00	1.00	0.01	-0.89	-0.14
14	0.00	0.00	1.00	0.00	0.30	0.03
15	0.00	0.00	1.00	0.00	0.68	0.12

TABLE 4.8: Covariance and correlation structure of Dataseat 3

In order to see the proportion of the variance the first eigenvalues explain scree plot of the eigenvalues is displayed in figure 4.16.

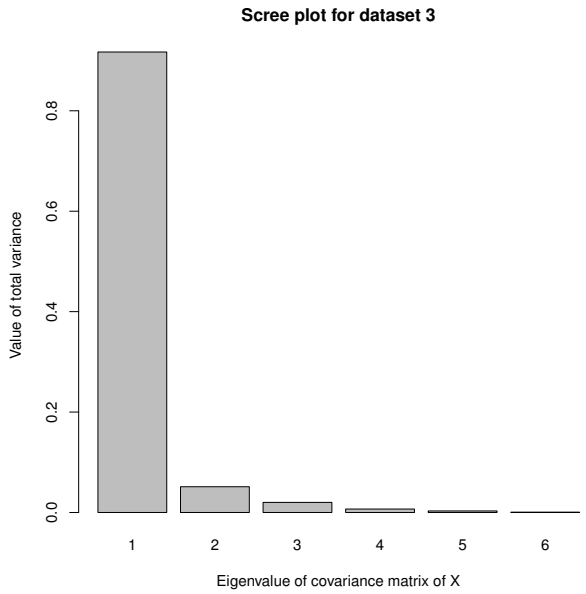


FIGURE 4.16: Scree plot of **X** in dataset 3

To select a method to estimate $\hat{\beta}$ the correlation among the columns of **X** is needed which is shown in table 4.9.

	s760	s780	s800	s820	s840	s860	s880	s900	s920	s940	s960	s980	s1000	s1020	s1040
s760	1.00	0.88	0.73	0.65	0.59	0.41	-0.90	-0.77	-0.73	-0.68	0.69	0.76	0.69	-0.38	-0.72
s780	0.88	1.00	0.96	0.92	0.88	0.75	-0.96	-0.97	-0.96	-0.93	0.76	0.95	0.93	-0.28	-0.73
s800	0.73	0.96	1.00	0.99	0.96	0.86	-0.92	-0.99	-0.98	-0.96	0.76	0.97	0.96	-0.25	-0.72
s820	0.65	0.92	0.99	1.00	0.98	0.88	-0.87	-0.97	-0.97	-0.93	0.78	0.96	0.93	-0.29	-0.72
s840	0.59	0.88	0.96	0.98	1.00	0.96	-0.79	-0.96	-0.97	-0.92	0.78	0.96	0.90	-0.29	-0.66
s860	0.41	0.75	0.86	0.88	0.96	1.00	-0.60	-0.86	-0.91	-0.85	0.70	0.87	0.82	-0.23	-0.52
s880	-0.90	-0.96	-0.92	-0.87	-0.79	-0.60	1.00	0.92	0.87	0.85	-0.74	-0.89	-0.86	0.31	0.77
s900	-0.77	-0.97	-0.99	-0.97	-0.96	-0.86	0.92	1.00	0.99	0.94	-0.81	-0.98	-0.93	0.30	0.72
s920	-0.73	-0.96	-0.98	-0.97	-0.97	-0.91	0.87	0.99	1.00	0.96	-0.77	-0.98	-0.95	0.25	0.67
s940	-0.68	-0.93	-0.96	-0.93	-0.92	-0.85	0.85	0.94	0.96	1.00	-0.60	-0.92	-0.99	0.03	0.53
s960	0.69	0.76	0.76	0.78	0.78	0.70	-0.74	-0.81	-0.77	-0.60	1.00	0.85	0.58	-0.79	-0.91
s980	0.76	0.95	0.97	0.96	0.96	0.87	-0.89	-0.98	-0.98	-0.92	0.85	1.00	0.91	-0.42	-0.80
s1000	0.69	0.93	0.96	0.93	0.90	0.82	-0.86	-0.93	-0.95	-0.99	0.58	0.91	1.00	-0.02	-0.56
s1020	-0.38	-0.28	-0.25	-0.29	-0.29	-0.23	0.31	0.30	0.25	0.03	-0.79	-0.42	-0.02	1.00	0.80
s1040	-0.72	-0.73	-0.72	-0.72	-0.66	-0.52	0.77	0.72	0.67	0.53	-0.91	-0.80	-0.56	0.80	1.00

TABLE 4.9: Correlation of **X** in dataset 3

To see which columns of **X** the components are retaining, information from the loading plot is retained. Loading is plotted in figure 4.17.

In figure 4.17 the scores are also plotted and grouping among plugs within locations is found.

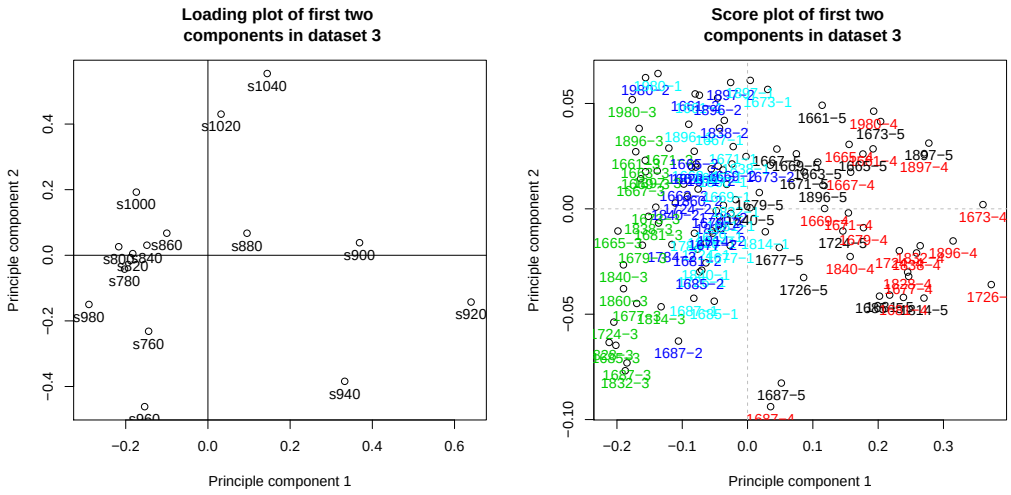


FIGURE 4.17: Score and loadings in dataset 3.

Distribution of fat within each plug location is also important and is shown in figure 4.18.

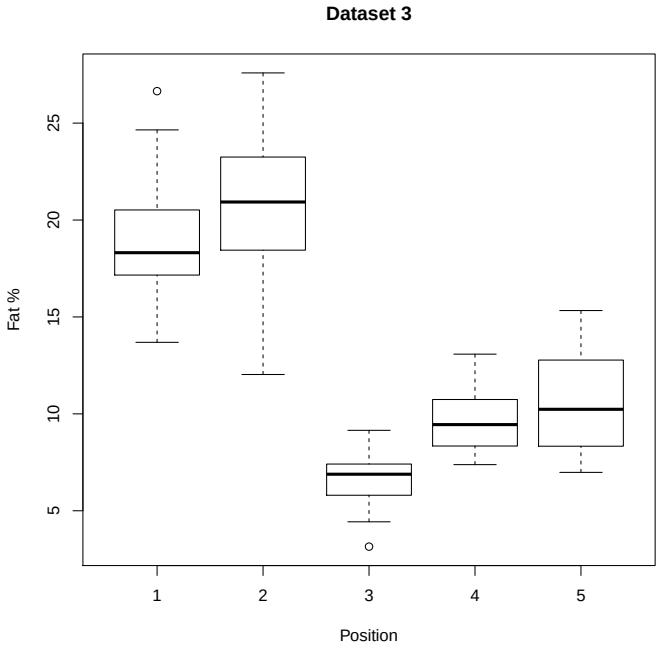


FIGURE 4.18: Fat in plug versus location of plug, Dataset 3

Where high fat is seen in the plugs from the belly and low fat in the plugs from the backbone.

4.5 General comments about the dataset prior to prediction

In all the dataset the first eigenvalue of $\mathbf{X}'\mathbf{X}$ is much larger than the other eigenvalues which means that the first component will explain much of the variation in $\mathbf{X}$. Correlation between $\mathbf{y}$ and the first component is also rather high for all the datasets which is a good quality when the aim in prediction is to predict $\mathbf{y}$ using principle components.

High correlation among the columns of $\mathbf{X}$ supports that using least square to estimate $\hat{\beta}$ will give a unstable estimates, which supports usage of other methods to estimate $\hat{\beta}$.

4.5.1 Box plot of fat within location of plug

Box plot of fat within location reveals higher fat values in the belly area and lower fat in the backbone and the tail. In dataset 1 and 3 where 5 plugs where collected from each fillet, a clear difference is between fat within locations. In dataset 2 are the plugs collected at 6 locations. Then the difference between fat within locations does not become as clear. The box plot reveals the fat variation within fillet which was terminated when the whole spectral image was averaged in earlier methods. Taking as many plugs as possible should give more information about the variation of fat in the fillet. How many are needed to capture most of it is unknown. In addition, knowing the optimal location to pick the optimal number of plugs to create a prediction model with as low prediction error as possible is a study that should be carried out and is publishable material if the solution exist. No significant difference were seen between using five or six plugs from each fillet.

4.5.2 Score and loadings

Information about fat is located in $920 - 930\ nm$ in the electromagnetic spectrum. The first component in all datasets had rather high positive loading value on the columns where this information about fat is located. Grouping of plug locations is also seen in the score plot of all datasets. The grouping is most clear in dataset 1 because the fillets have the smallest weight interval and two fillets are measured from each fish which are highly correlated. Also a highly educated scientist did the measures with fewer mistakes.

Reason for less grouping in figure 4.17 could be that the weight of the fish was more equally distributed among the weight classes in the $1-6$ kg fish than among the weight classes of the $1-8$ kg fish in dataset 2. Distribution of weight of a fillet among the weightiness in Dataset 1 is unknown. Judging by the score the weight of the fillets in dataset 1 have probably been equally distributed.

Correlation among the first component and y was high for the first component in all datasets. First component explained from $77-94\%$ of the total variation of $\mathbf{X}$ which is a good quality when the aim is to keep the complexity level as low as possible without losing information. Correlation and covariance between $\mathbf{X}$ and $\mathbf{y}$ was highest in the region in the spectrum where the information about fat was located. All this results support a prediction model using two or fewer components, because they are explaining most of the variation in $\mathbf{X}$. The scores of dataset 2 are most condense with one plug location further from the others. These few plugs from this other location have probably been too few in contrast to the other because of a bad random function in Excel that was supposed to select two plugs randomly from each fillet, but did not do a better work as shown. Grouping is detectable in the score plot of dataset 3.

Correlation of among the columns of $\mathbf{X}$ in all datasets is rather high. Due to this high level multicollinearity a $\mathbf{X}^{-1}$ does not exist. Therefore it is not possible to obtain as good estimator by least square estimation. Other estimation methods of $\hat{\beta}$ should be considerate.

Chapter 5

Results

The dataset were created using Matlab. Matlab scripts are not in the Appendix because they were borrowed from Martin Høy[33]. Calculations were carried out using the statistical program R [16]. Functions were made to calculate Principle component regression. The function used leave one out cross validation and K-fold cross validation. The cross validation could use batches containing all measures on each fish, or batches where all plugs from each location were stored. Based on PLS algorithm, in this thesis a PLS program was written which uses the same cross validation arguments as the PCR function.

Function was written to try test and calibration sets. Arguments which can be passed into this function is how; big part of the observations, how many fishes, or how many plug locations should be stored in the calibration set or stored in the test set. Ratio between test and calibration set is passed into the function. The calibration set is cross validated by the same methodology as used in the PCR function.

Function was made that tried all combinations of additional variables trough CPLS and found which combination gave the lowest prediction error.

5.1 PCR

Using the principle component regression function a RMSEP value was calculated for the first five components for all datasets shown in figure 5.1.

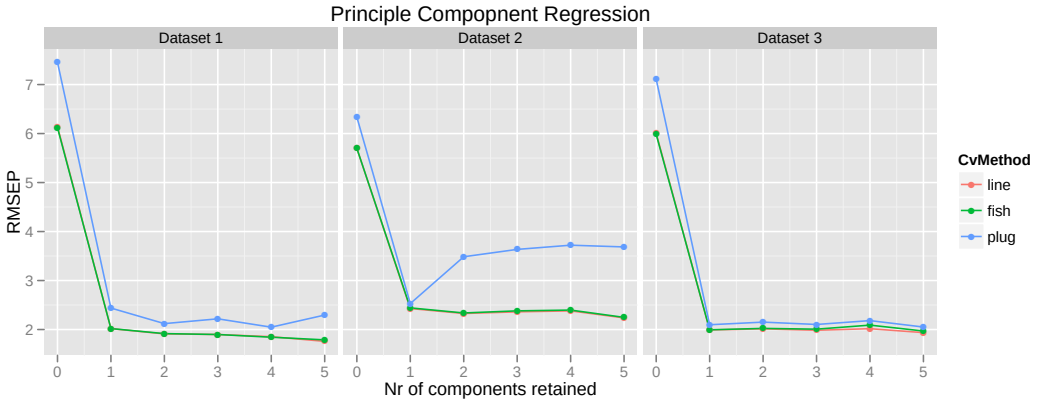


FIGURE 5.1: PCR on all dataset using Cv with different segmentation.

In figure 5.1 CvMethod indicates the method used for cross validation. The colors in 5.1 are, "line" means leave one out cross validation, "fish" means K-fold cross validation where all plugs from each fish are stored in seperate segments, and "plug" stands for K-fold cross validation where each location of plugs is put into one segment. This notation is used through the result chapter. Because of the zero component model in figure 5.1 predicts worse than other models zero component model is excluded in further calculations. By excluding the zero component model the figure 5.2 is obtained.

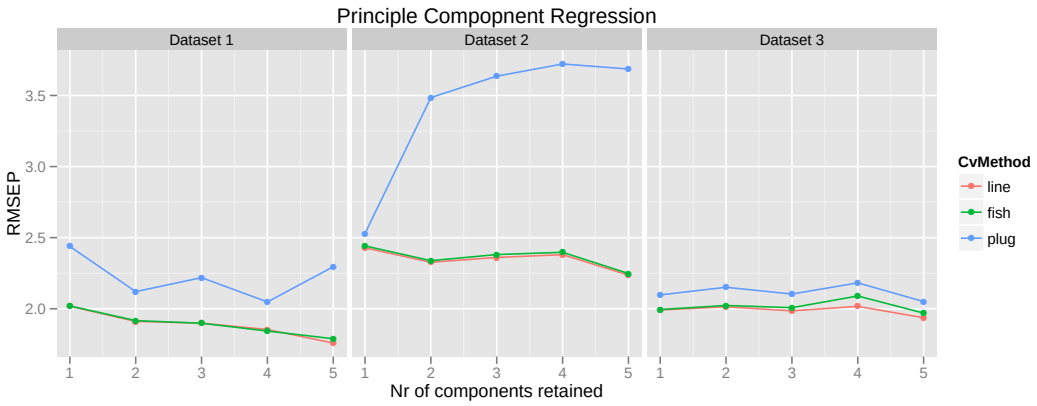


FIGURE 5.2: PCR on all dataset using Cv with different segmentation.

In figure 5.2 K-fold cross validation using plug segments performs poorly. The other cross validations methods performs better and similar to each other.

5.2 PLSR

Using the PLSR function in appendix A giving the following results which was obtained on the three datasets using the first five components with same cross validation as in chapter 5.1.

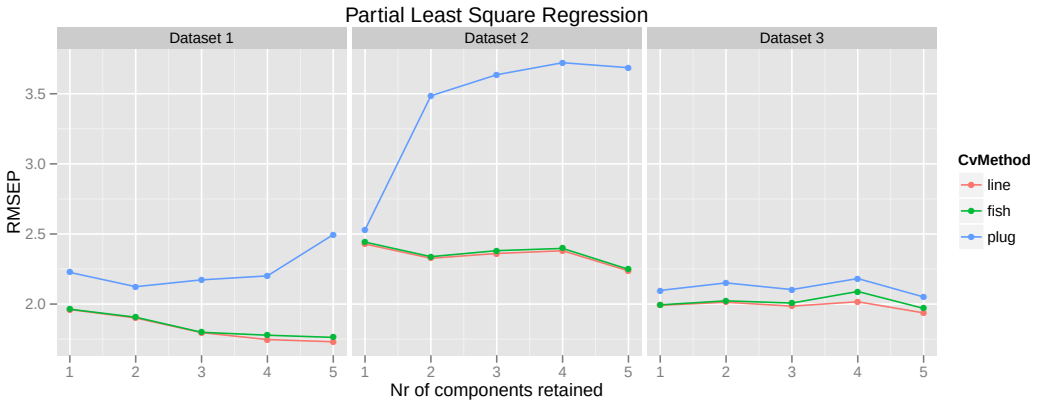


FIGURE 5.3: PLSR on all dataset using Cv with different segmentation.

Cross validating with plug in each segment seems a bad idea in figure 5.3. Cross validation segmented by plug is therefore excluded in further analysis. Comparison of PCR against PLSR using the leave one out cross validation, referred as a line in the plots, and K-fold cross validation storing each fish in one segment is done in figure 5.4.

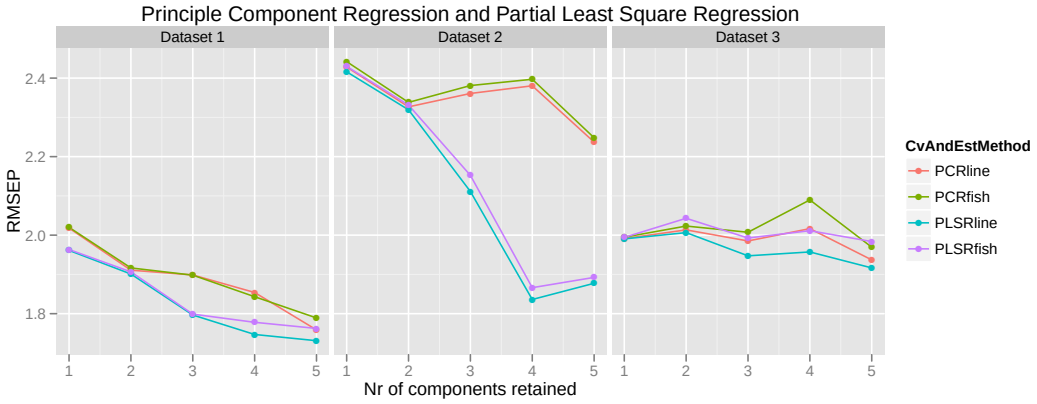


FIGURE 5.4: PLSR vs PCR on all dataset using Cv with different segmentation.

In figure 5.6 the PLSR has a lower RMSEP values. Something strange is happening in dataset 2. The RMSEP value in all datasets reduces when more components are

retained.Looking closer at PLSR using the leave one out cross validation the figure 5.5 was obtained.

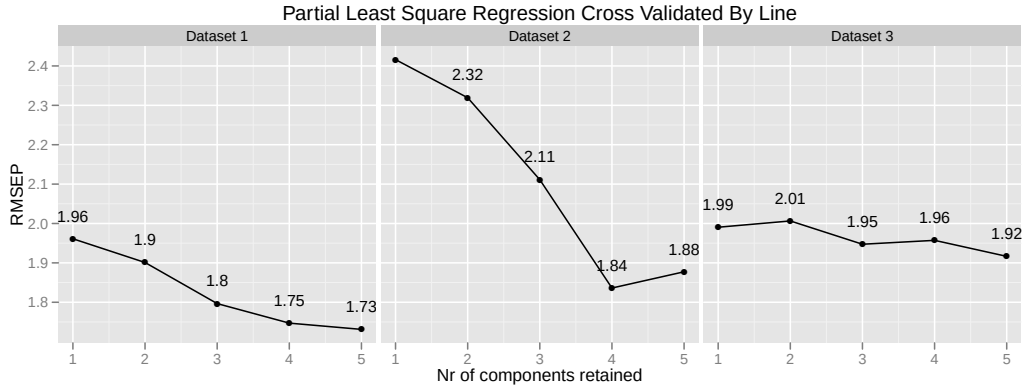


FIGURE 5.5: PLSR on all dataset leave on out Cv.

In order to investigate the prediction performance further plot of $\mathbf{y}$ against the new predicted $\hat{\mathbf{y}}$ using the leave one out cross validation gives the figure 5.6. Linear model was fitted by $\hat{\mathbf{y}}$ and $\mathbf{y}$ in order to estimate the slope and the intercept of the model to understand how the model will predict lean and fat plugs. A $1:1$ line was plotted to compare to the line defined by the model. To get a comparison across datasets the R^2_{pred} value was also obtained.

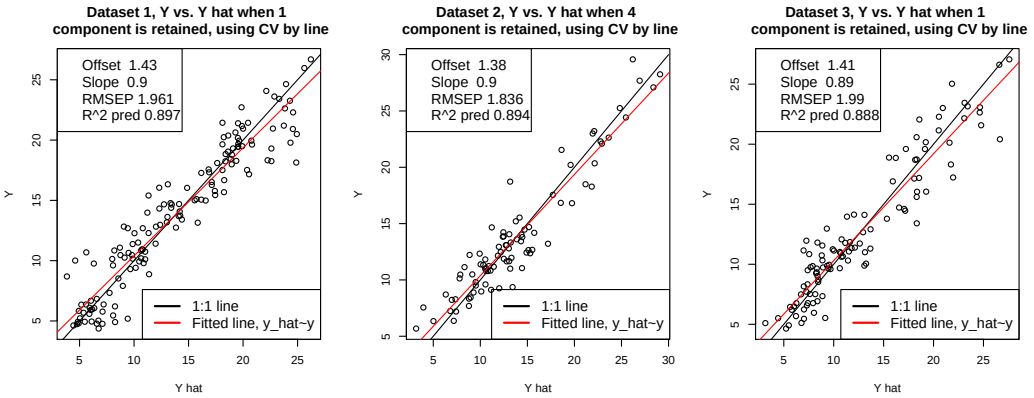


FIGURE 5.6: $\mathbf{y}$ vs $\hat{\mathbf{y}}$ with lowest RMSEP, using leave on out Cv.

Even though the RMSEP values for the datasets using the leave one out cross validation are different the plot of $\mathbf{y}$ against $\hat{\mathbf{y}}$ are similar with almost the same R^2 values. The RMSEP values are all similar, but four components are needed in dataset 2 which should be kept in mind when figure 5.6 is evaluated.

RMSEP value using PLSR were calculated using K-fold cross validation where segmentation of the data is defined by a fish. The result is shown in figure 5.7.

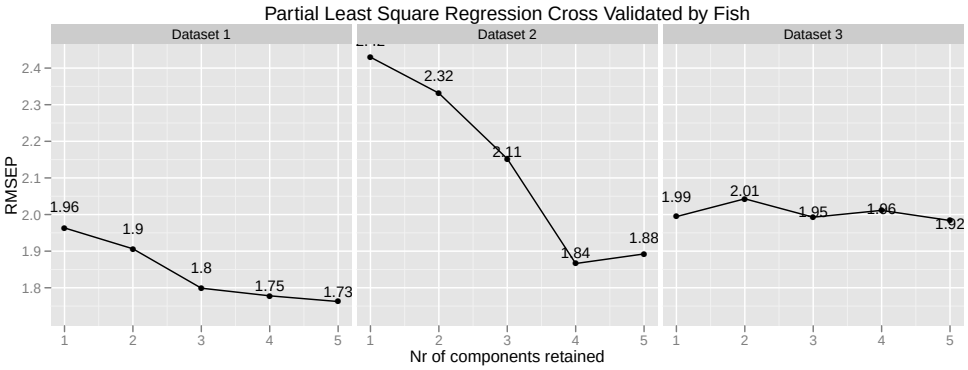


FIGURE 5.7: PLSR on all datasets using fish as segment for Cv.

In order to know more about the prediction performance a plot of $\mathbf{y}$ versus $\hat{\mathbf{y}}$ was made where linear model were fitted using $\hat{\mathbf{y}}$ and $\mathbf{y}$. Also the R^2_{pred} was calculated as in figure 5.6. Then figure 5.8 was obtained.

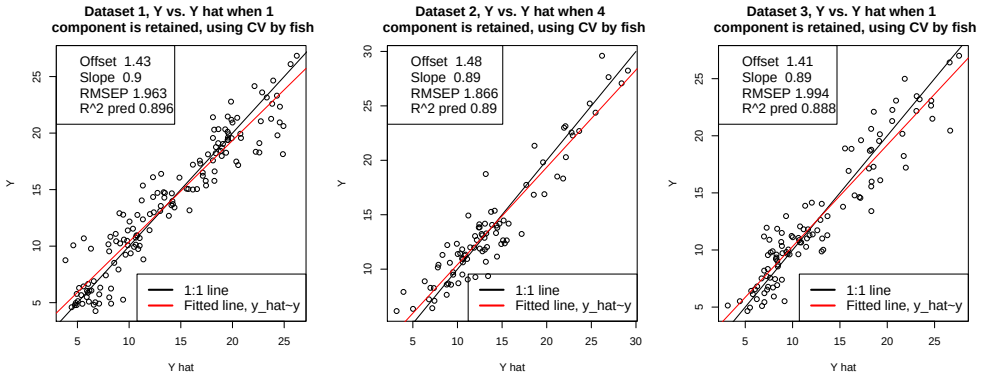


FIGURE 5.8: $\mathbf{y}$ vs $\hat{\mathbf{y}}$ with lowest RMSEP using fish as segmentation Cv.

In figure 5.8 the models seem to perform similar as seen on the $RMSEP$ and R^2_{pred} values. Dataset 2 needed 4 components to optain this result.

In most cases the PLSR performed better, therefore PCR is excluded from further calculations. Cross validation by K-fold cross validation where the segments are defined by fish and the leave one out cross validation seems to perform similarly. In this thesis it is considered most reasonable to use K-fold cross validation in cross validation. Then the leave one out cross validation will be excluded from further calculations for PLSR.

5.3 Calibration and Test set

By using the test function which is stored in appendix A the data was divided up to calibration and test set where $\hat{\beta}$ corresponding to the lowest value of RMSEP was chosen and exposed to the test set. Values of $\hat{\mathbf{y}}$ were obtained using the $\hat{\beta}$ from the calibration set and the data from the test set. Then prediction error was estimated by equation 2.6.

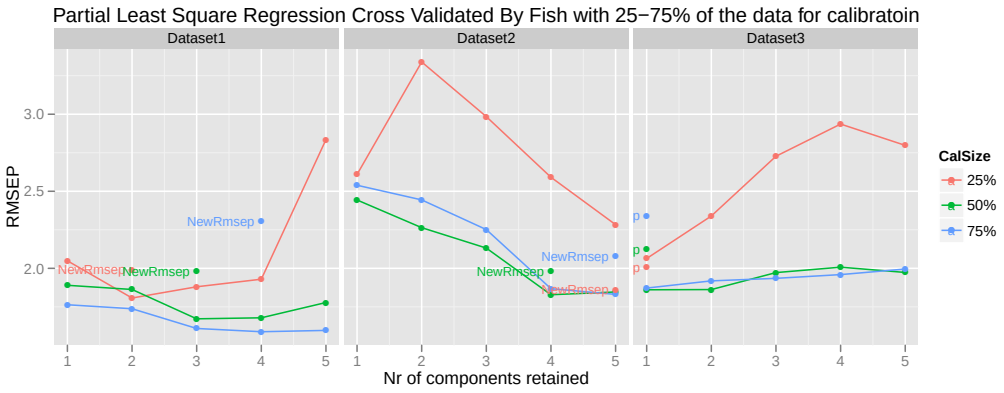


FIGURE 5.9: Calibration set containing first $25-75\%$ of the data, using fish for segmentation of the Cv.

In graph 5.9 the first $25-75\%$ of the dataset is used as calibration set. RMSEP values are calculated for all number of components. The test set is the remaining $75-25\%$ of the dataset.

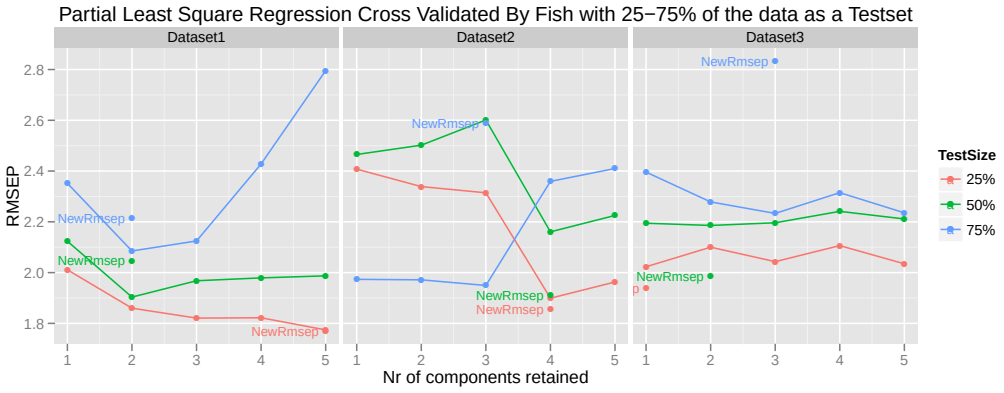


FIGURE 5.10: Test set containing first $25-75\%$ of the data, using segmentation by fish for the Cv.

Then the opposite was tried in figure 5.10. First $25-75\%$ of the datasets are used as the test set and the remaining $75-25\%$ are used as the calibration set.

When K-fold cross validation is done where the batches are defined by fish it can happen that one fish gets divided between the test and the calibration set. In following graphs the calibration set was defined as the first $25-75\%$ of the fish in the dataset and the calibration set as the remaining $75-25\%$ fish in the dataset.

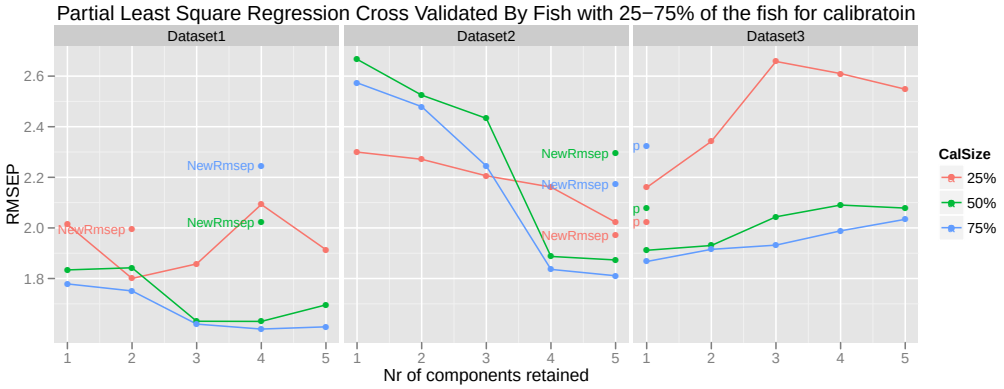


FIGURE 5.11: Calibration set containing first $25-75\%$ of the fish, using segmentation by fish for the Cv.

In figure 5.11 the devision of data was by fish. The opposite devision is calculated in graph 5.12.

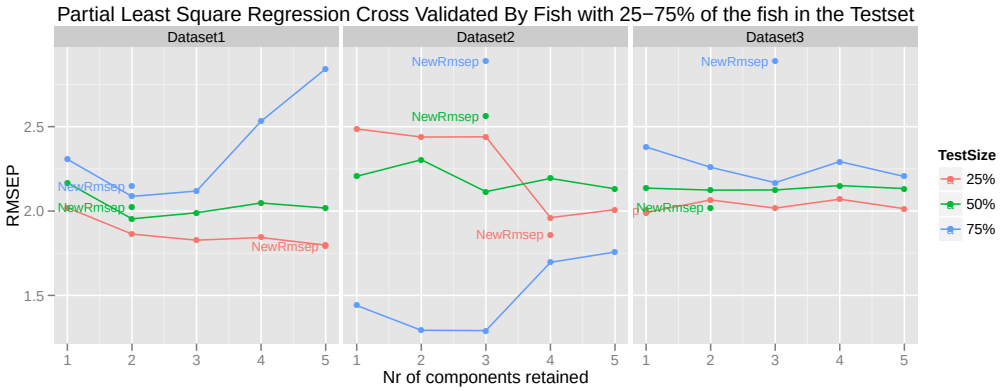


FIGURE 5.12: Test set containing first $25-75\%$ of the fish, using segmentation by fish for the Cv.

First $25-75\%$ of the fish was defined as the test set and remaining $75-25\%$ was defined as the calibration set where the $\hat{\beta}$ was retained from and introduced to the first $75-25\%$ fishes in the dataset and the RMSEP value evaluated based on the $\hat{\mathbf{y}}$ and $\mathbf{y}$ from the test set.

5.4 CPLSR

Additional responses were tried to improve the prediction performance. The leave one out cross validation was introduced again since it has larger influence using CPLSR.

5.4.1 Dataset 1

In dataset 1 was only one addition response, the location of the plug. By supplying the CPLS with the plug location the following RMSEP values were obtained in contrast to the RMSEP values obtained by PLSR.

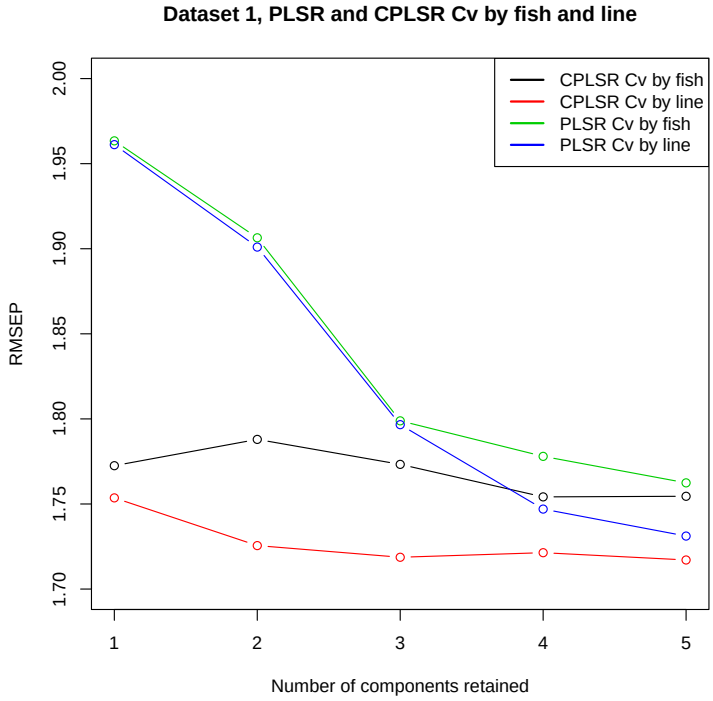


FIGURE 5.13: PLSR vs CPLSR using plug location as additional response on dataset 1.

In figure 5.13 the RMSEP is lower for one component using CPLS.

5.4.2 Dataset 2

In dataset 2 13 additional variables exist. A program was made in R which is located in the appendix which tried all possible combinations of additional responses and calculated all RMSEP values using the first five components. The combination containing the lowest RMSEP values were retained.

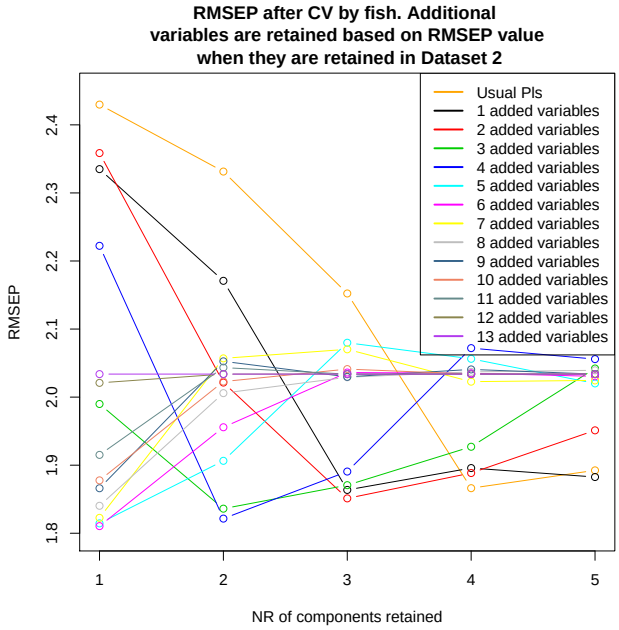


FIGURE 5.14: CPLSR, All RMSEP dataset 2, using segmentation by fish for the Cv

In figure 5.14 all the RMSEP values are plotted for the additional responses with the lowest RMSEP value when K-fold cross validation using fish as the segments was used.

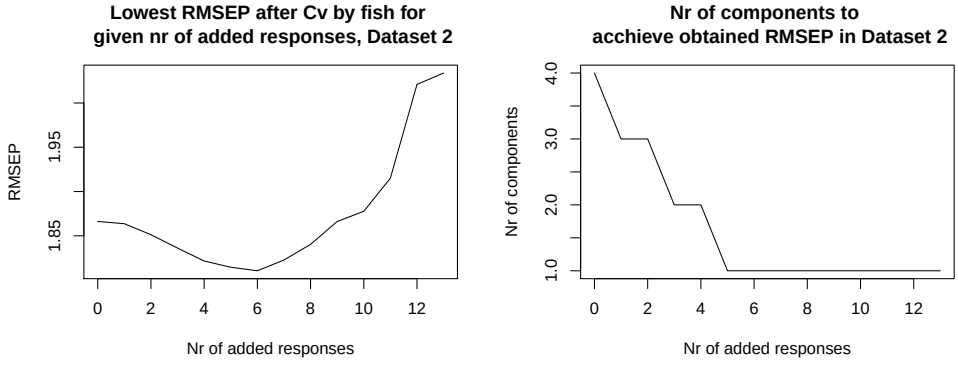


FIGURE 5.15: Lowest RMSEP using segmentation by fish for the Cv, Dataset 2

In plot 5.15 the left side is development of the lowest RMSEP value when variables are added and on the right side is the plot of how many components were needed to obtain the RMSEP value displayed on the left side of the plot.

PlugThick							
PlugW	PlugMoust						
PlugThick	PlugMoust	Temp					
GutW	PlugMoust	Ph	Temp				
UngW	PlugThick	TrueF	Cant	Ph			
Length	PlugMoust	TrueF	Asta	Ph	Temp		
UngW	GutW	PlugMoust	TrueF	Asta	Ph	Temp	
UngW	GutW	Length	PlugMoust	Idoxan	Asta	Cant	Ph

TABLE 5.1: Combination of additional responses containing the lowest RMSEP value using segment defined by fish for Cv, dataset 2

The first 8 combinations of additional responses which had the lowest RMSEP value is shown in table 5.1. PlugThick is the thickness of the plug. PlugMoust is the moisture in the plug optained by the LF-NMR scanner. Temp is the temperature of the fillet after filleting. PlugW is weight of the plug.

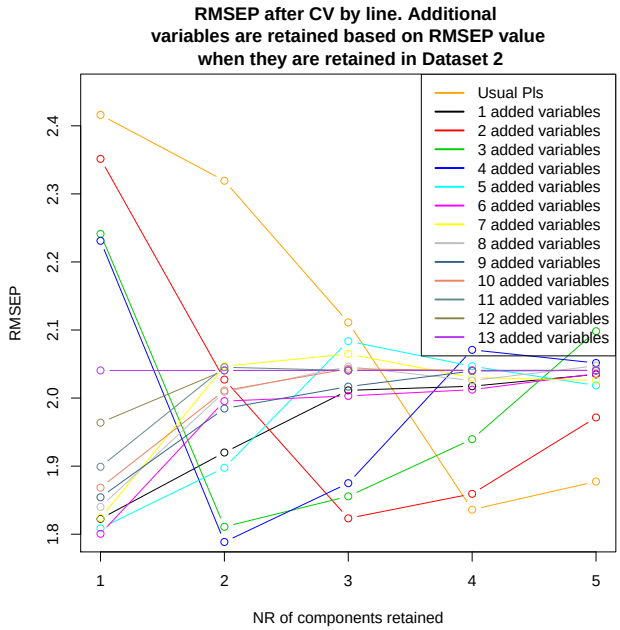


FIGURE 5.16: CPLSR, All RMSEP dataset 2 using leave on out Cv

All combinations of additional data was tried where the leave one out cross validation was tried. Best combinations were retained which contained the lowest RMSEP among all

combinations for given number of added variables. In plot 5.16 development of RMSEP for those optimal combinations is plotted in contrast of PLSR.

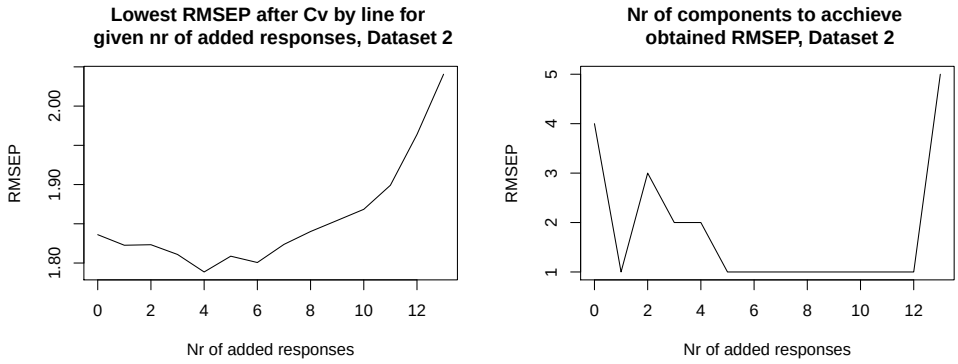


FIGURE 5.17: Lowest RMSEP using leave on out Cv Dataset 2

In plot 5.17 development of the lowest RMSEP value was plotted against number of added variables. On the right side number of principle components needed to obtain these lowest RMSEP values when the leave one out cross validation was used on dataset 2.

Posit							
PlugMoust	Idoxan						
PlugMoust	Ph	Temp					
UngW	PlugMoust	Ph	Temp				
UngW	PlugThick	TrueF	Cant	Ph			
UngW	Length	PlugW	Asta	Ph	Temp		
UngW	GutW	PlugW	TrueF	Asta	Ph	Temp	
UngW	GutW	Length	PlugThick	PlugW	Asta	Cant	Ph

TABLE 5.2: Combinations containing the lowest RMSEP value using leave one out CV, dataset 2

The first 8 combinations containing the lowest RMSEP values using CPLSR when using the leave one out cross validation are shown in table 5.2 where Posit is the position of the plug and Idoxan is a chemical in the fillet. Ph is the ph value in the fillet when it was filleted.

5.4.3 Dataset 3

Best combinations were calculated by trying all combinations of the 8 additional responses and choosing the optimal combination by the lowest RMSEP, when the K-fold cross validations was used and the segments were defined by fish. Result is displayed in figure 5.18

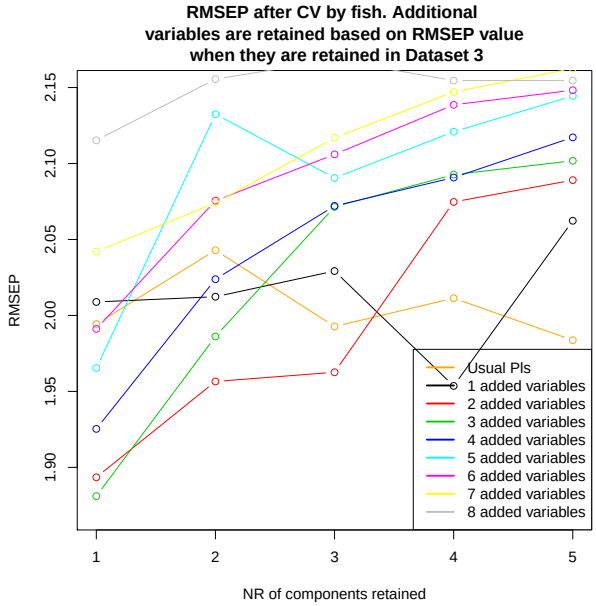


FIGURE 5.18: CPLSR, All RMSEP using segmentation by fish for the Cv, dataset 3

In figure 5.19 the development of the lowest RMSEP value is plotted against no. of additional responses retained. On the right side is the number of components needed to retain this RMSEP value.

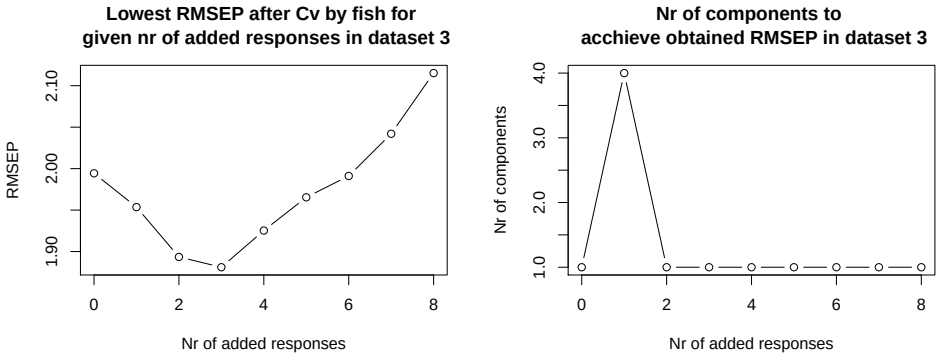


FIGURE 5.19: Lowest RMSEP using segmentation by fish for the Cv, Dataset 3

The additional responses which contained the lowest RMSEP values are shown in table 5.3 when K-fold cross validation was used.

Length							
FillW	IntW						
FillW	Sex	IntW					
UngW	Sex	IntW	PlugW				
FillW	UngW	Sex	PlugW	PlugMo			
FillW	UngW	Length	Sex	IntW	PlugW		
FillW	UngW	Length	Sex	IntW	PlugW	PlugMo	
FillW	UngW	Length	Sex	IntW	PlugW	PlugMo	Posit

TABLE 5.3: The sets of additional responses containing the lowest RMSEP value when using segmentation by fish for Cv, Dataset 3

In table 5.3 Length is the length of the fish un gutted, FillW is the weight of the fillet, IntW is the weight of the intestine of the fish. Sex is the sex of the fish where 1 is male and 2 is female.

Best combinations were calculated by trying all combinations of the 8 additional responses and choosing the optimal combination by the lowest RMSEP among the combinations when the leave one out cross validations was used.

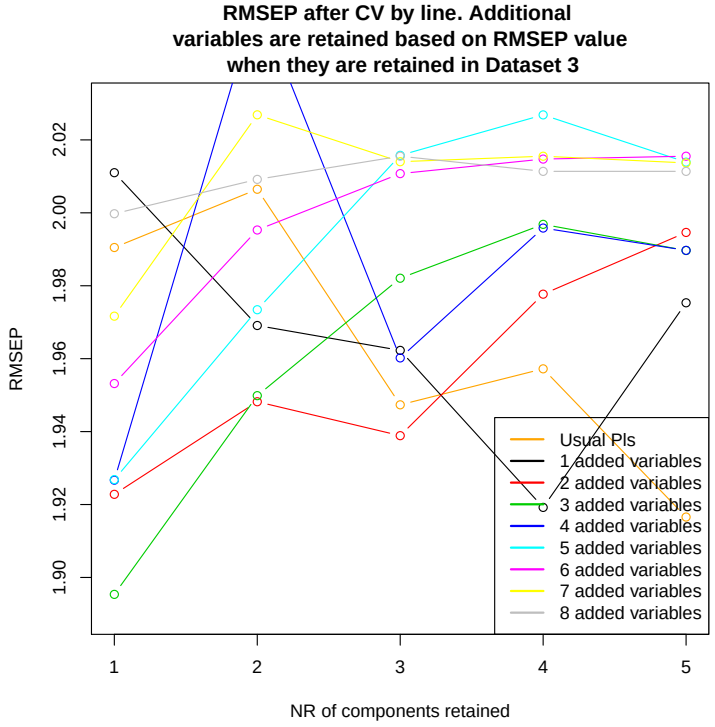


FIGURE 5.20: CPLSR, All RMSEP using leave on out Cv, dataset 3

In figure 5.21 the development of the lowest RMSEP value is plotted against number of additional responses retained. On the right side is number of components needed to retain this RMSEP value.

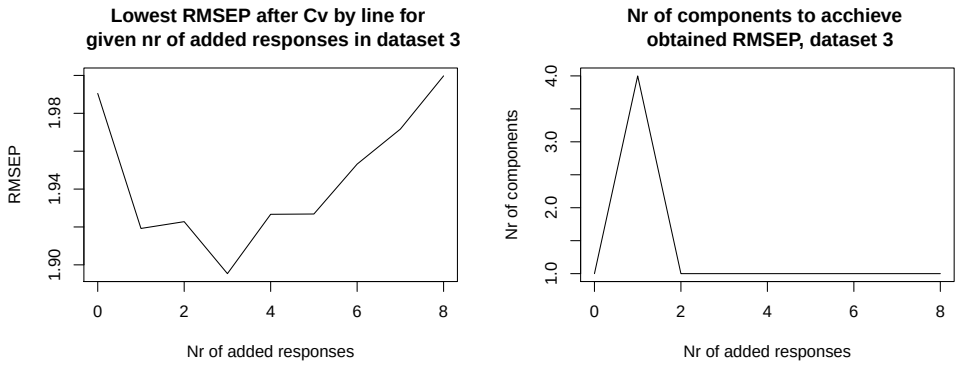


FIGURE 5.21: Lowest RMSEP using leave one out Cv, Dataset 3

The pairs who had the lowest RMSEP values are shown in table 5.4 when leave the one out cross validation using CPLSR was used in figure 5.21.

PlugMo							
FillW	IntW						
FillW	UngW	IntW					
Length	Sex	IntW	PlugW				
FillW	UngW	Length	PlugMo	Posit			
UngW	Length	Sex	IntW	PlugMo	Posit		
FillW	UngW	Length	Sex	PlugW	PlugMo	Posit	
FillW	UngW	Length	Sex	IntW	PlugW	PlugMo	Posit

TABLE 5.4: The combination of additional responses containing the lowest RMSEP value using leave one out CV, dataset 3

In table 5.4 PlugMo is moisture in the plug, FillW is the weight of the fillet, IntW is the weight of the intestines.

5.5 Developed Models

The aim is to develop a model for the QMonitor in Iceland which dataset 3 is made by. In figure 5.22 optimal combination of additional variables when the leave one out cross validation and K-fold cross validation are used.

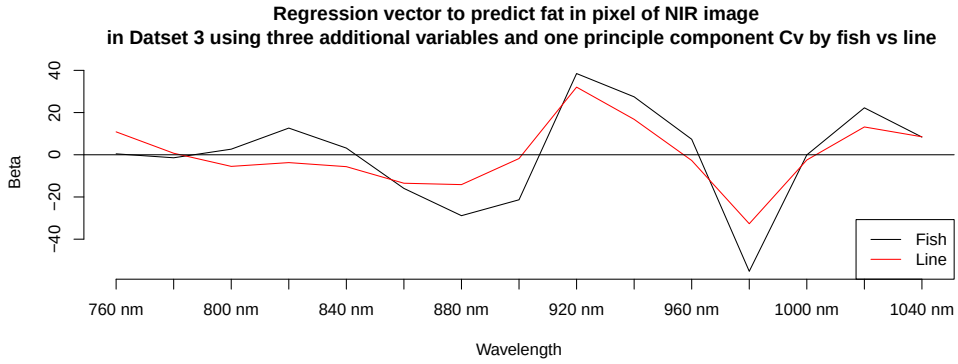


FIGURE 5.22: $\hat{\beta}$ in dataset 3 with the lowest RMSEP using CPlSR

In figure 5.22 information about fat located in $920 - 930\ nm$ is weighted up and information about water is weighted down because of its high negative correlation.

The old beta obtained by PLSR using one component in dataset 1 compared to the new model obtained using CPLSR supplied with three additional responses and retaining one component is displayed in figure 5.23.

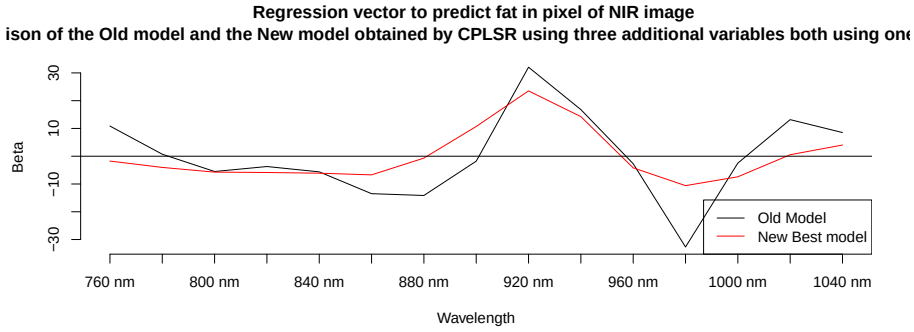


FIGURE 5.23: $\hat{\beta}$ in dataset 3 with the lowest RMSEP using CPlSR

Plot of y from dataset 3 against the $\hat{\mathbf{y}}$, predicted using the old PLSR model is shown in figure 5.24

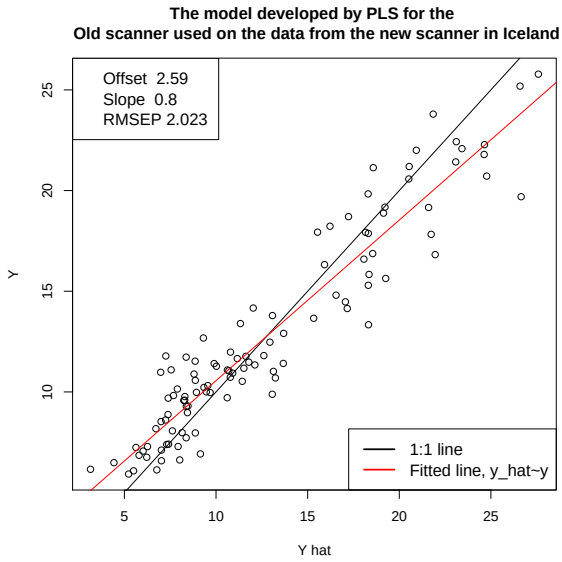


FIGURE 5.24: $\hat{\beta}$ in dataset 3 with the lowest RMSEP using CPlSR

Plot of $\mathbf{y}$ against $\hat{\mathbf{y}}$ predicted using the new prediction model developed by CPLSR using three additional responses and retaining one component is shown in figure 5.25.

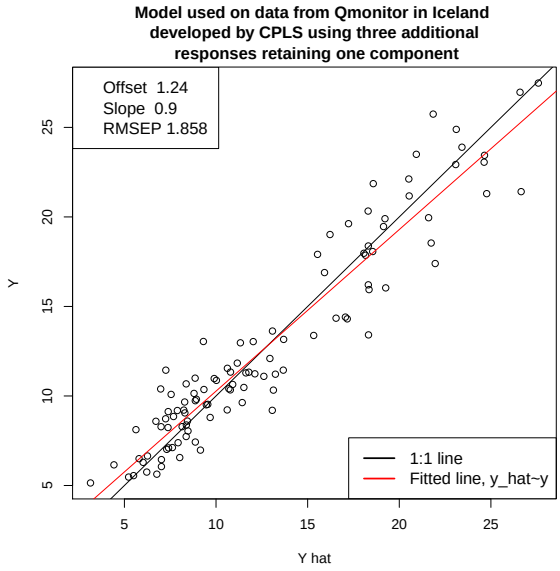


FIGURE 5.25: $\hat{\beta}$ in dataset 3 with the lowest RMSEP using CPlSR

The model used in figure 5.25 will be used in Stofnfiskur breeding work in the future, where the CPLS were supplied with the additional responses weight of fillets, ungutted weight and intestines weight.

Chapter 6

Discussion

6.1 Discussion.

In this thesis three datasets were of interest. Several summary statistics were carried out to fully understand the behavior of the dataset and the variation structure. The variation structure was explored among the explanatory matrix and the covariation between the components and the explanatory matrix before modeling were carried out. Several prediction methods were tried on the data and their quality evaluated using several validation methods that are capable of measuring prediction performances. Hopefully, close to an optimal model was developed and compared to existing model which will be used for predicting fat in salmon fillet of the salmon in the breeding work of Stofnfiskur in Iceland.

6.1.1 General comments about the data

In dataset 1 the weight interval of the fish was rather small. Choosing fish from such a small interval will not be as in the real world where the fat measuring instrument is designed to be used in the future. Weight of a fish is normally distributed. In a fish farm the whole tank of fish is slaughtered when the mean weight is around 4 kg. Among that group will be fish smaller and larger than $2 - 5$ kg. Therefore the RMSEP value tends to be underestimated when it is reported as 1.96. The model developed in this paper [11] was applied on the measures obtained on the fish in dataset 2. The weight of the fish in dataset 2 is ranging from $1 - 8$ kg which is more realistic. The original

idea when collecting dataset 2 was to challenge the instruments available because most of the published papers obtaining fat values from fish on a small weight interval which is not realistic. The RMSEP between the predicted fat values from dataset 2 using the model from the paper [11] and the chemical values obtained on the fish in dataset 2 was 2.5. Calculations were not carried out within this thesis, but were done when deciding which instrument Stofnfiskur should invest.

In dataset 1 both fillets of 15 fishes were used. Correlation of fat in fillets within the same fish is considerably high when breeding work is done. In previous fat measurement methods only fat value from one fillet of each fish is reported. Measure can almost be introduced as a repeated record for that fish. Then 75 plugs should only be needed to report the prediction quality. This correlation between fillets should be calculated to estimate leak of information between fillets.

Because of correlation between plugs and the small weight interval of the fish the score plot in the principle component analysis has a nice look regarding grouping of plug location. It has more grouping compared to the other datasets.

6.1.2 Broken observation

In this thesis outlier detection has not been discussed. There exist several methods to detect outliers. No outlier detection calculations were done. The data reported as broken was removed because of unrealistic values. The LF-NMR scanner sometimes reported zero fat or negative fat. When measuring with LF-NMR scanner the sample is not destroyed after measuring so measuring the sample many times should give similar results. When minus or zero value was obtained the calibration sample, a fish oil was measured. Result from the fish oil was always 100% fat. Then the broken sample was measured again and still a negative or zero value was obtained.

After collecting the spectra values from the spectral images a model was fitted and new fat values were predicted using cross validation for all datasets. After prediction a plot of $\mathbf{y}$ versus $\hat{\mathbf{y}}$ was obtained. Values which were extremely far from the straight line were removed. Reason for they were far from a straight line is the wrong fat values or the spectra values has a high probability of being wrong. Either the spectral image was broken or the spectral values were obtained on wrong place on the image.

6.1.3 Cross validation methods

Three cross validation methods were tried on all datasets. A classical leave one out cross validation, a K-fold cross validation were the batches were defined by fish and the third cross validation method was to use leave one plug out. Then all plugs from same location of each fish were put into each batch.

K-fold cross validation where the batches were defined by plug did not perform well. The $\hat{\mathbf{y}}$ were unstable. It can be seen from the box plot of fat values in all the datasets the average fat in each location vary. When a plug with high fat is removed in cross validation, leaner plugs are used to predict it. Same applies for lean plugs. After seeing how badly a K-fold cross validation performed using batches defined by plug the K-fold cross validation were considerated irrelevant and were not used in further analysis.

The leave one out cross validation was then used. The problem of using the leave one out cross validation is information of other plugs originated from the same fish as the removed plug. The information leaks from the remaining plugs and prediction of the removed plug will be too good compared to what will be observed in the real world. RMSEP will underestimate the prediction error.

The third cross validation method used was K-fold cross validation where each batch contains all plugs from each fish. Number of batches were the same as number of fish in the dataset and number of samples within a batch were the same as number of plug from each fillet. When predicting values of the removed batch in the cross validation, no information were in the dataset about the removed fillet which is more like the real world than the leave one out cross validation when the prediction model will be exposed to salmon fillet in the future.

In the paper [12] where the methodology of how to create a prediction model is described, the cross validation methods used is the leave one out because the K-fold cross validation where each batch contained fish did perform equally when using PLSR. In this thesis K-fold validation where batches contained plugs from each fish performed similarly as the leave one out cross validation for the PCR and PLSR. Using CPLSR, a significant difference between RMSEP values obtained were observed. The leave one out cross validation for CPLSR were lower in most cases and did recommend to use other additional responses to obtain lowest prediction error than using K-fold cross validation

where batches were defined by plugs of each fish. To decide to use the leave one out cross validation in stead of K-fold cross validation is a dangerous conclusion and would underestimate the prediction error.

Conclusion of this paper is that using K-fold cross validation where each segment contains all plugs from one fish should be used when prediction model is developed and evaluated.

6.1.4 Calibration and test sets

To estimate the quality of the prediction model the dataset was divided into calibration and test set. That is a good method to see how our prediction model will perform in the future when it meets new explanatory variables where the model was created without information about this future observation.

After revealing the high correlation among the first components and $\mathbf{y}$, and the scree plot of all the datasets only first five components were considerated when looking at the calibration and test set.

Dividing the data irrespective of fish, had lower RMSEP values than dividing the data by fish. Dividing by fish is more realistic because that situation is more like the real world. The model will get exposed to whole fillets, not only some plugs out of the fillet. Information is then leaking between the test and calibration set. Therefore dividing the data irrespective of fish underestimates the prediction error.

Using only 25% of the fish or the dataset irrespective of fish obtains unstable $\hat{\boldsymbol{\beta}}_{\text{cal}}$ resulting in unstable $\hat{\mathbf{y}}$ values. The RMSEP values change much in most cases when components were retained in all the datasets. Quality of having small calibration set, the test set gets big which results in good estimate of the model performance. Estimating prediction performance of very unstable $\hat{\boldsymbol{\beta}}_{\text{cal}}$ retained from a small calibration set reveals the upper bound of the prediction quality.

Having the first 50% of the data or the fish results in more stable estimate of $\hat{\boldsymbol{\beta}}_{\text{cal}}$ than only using 25% of the data or the fish for calibration. The test set gets smaller resulting in poorer estimate of the prediction quality of the model when it will get exposed to new explanatory matrix in the future. Using 50% of the data gave more stable estimate of $\hat{\boldsymbol{\beta}}_{\text{cal}}$ in all the datasets in this thesis.

When using 75% of the data the most stable $\hat{\boldsymbol{\beta}}_{\text{cal}}$ were obtained. Most of the data is then used to develop the model and very little data is left to estimate the model quality.

Using the first $25-75\%$ of the data or fillets as a calibration set gives more stable $\hat{\boldsymbol{\beta}}_{\text{cal}}$ than using the first $25-75\%$ of the data or fish as a test set. To understand why, a larger dataset should be recorded. It could also be evaluated by letting the division between prediction and test set vary more in order to make conclusion of why the robustness of $\hat{\boldsymbol{\beta}}_{\text{cal}}$ is so dependent which $25-75\%$ of the dataset are used for calibration.

When only 25% of the data was used, the RMSEP was unstable and poor because 25% of the data are too few observations. Then each observation has a huge effect on the RMSEP value. When 50% of the data was used for calibration similar for of the RMSEP curve appeared as when CV was used indicating good quality. In dataset 1 when 75% of the data was used for calibration the new RMSEP value obtained when the $\hat{\beta}_{\text{cal}}$ was introduced to the test set, the RMSEP value obtained were rather far from where the $\hat{\beta}_{\text{cal}}$ was selected. It indicates that last 25% of the data are quite different from the first 75% of the data.

To use the idea of creating test and calibration set the aim is to create same situation as when measuring fillets in the future. Therefore test and calibration sets were tried where $25-75\%$ first fish in the dataset was used for calibration of the model.

Highest RMSEP value of all datasets was close to 2.6 compared to 3.5 when the division on the data was by fish. In dataset 1 a 25% of fish are only four fishes which reveals why the model is unstable when it meets calibration set containing other eleven fishes.

When first $25-75\%$ of the fish were used for calibration higher values were obtained, indicating that the last part of the dataset contains fish that is different from the first part of the fish in dataset 1.

In all situations quality of the prediction model varies a lot when devision of the data to test and calibration is done indicating that the prediction error obtained by usual cross validation is underestimating the prediction error.

6.1.5 Prediction methods

Three methods where tried to estimate $\hat{\beta}$. The first method was to use Principle Component Regression. Then Components are defined by the highest variation within the explanatory matrix. Variation of $\mathbf{y}$ is not taken into account which can be bad when the correlation to the response is of main interest. In figure 5.1 a zero component model is shown to ensure that using PCR to estimate $\hat{\beta}$ instead of reporting the average will perform better. The zero component model obtained higher prediction error than other models. In figure 5.2 a development of RMSEP is revealed when the zero component model is excluded. In dataset 1 the RMSEP reduces when components are added to the model. In dataset 2 there is a minor change in value of RMSEP when components are retained. Change of RMSEP when components are added in dataset 3 is almost none. Possible reason for flat RMSEP curve in dataset 2 and 3 is how much of variation is captured by the first component. By adding components to the model not much more is captured of the variation of the explanatory matrix.

To include the covariation between $\mathbf{X}$ and $\mathbf{y}$, a PLSR is used to estimate $\hat{\beta}$. Then components are defined by the highest covariation between $\mathbf{y}$ and $\mathbf{X}$. A PLSR is shown in figure 5.3 for all datasets. Appearance of the RMSEP curve looks similar as for PCR just with lower values because of better estimation of $\hat{\beta}$. The difference can be seen clearer between PLSR and PCR figure 5.4 where cross K-fold validation with segments defined by plug location is excluded because of its poor ability to estimate the prediction error. PLSR performs better showing lower RMSEP value for all number of components. Difference between PLSR and PCR is not high in dataset 1 at component 2, in dataset 2 when component 1 and component 1 and 2 are retained. In dataset 3 there is small difference between PLSR and PCR when 1 component is retained.

In order to know more about the prediction ability of the models, a plot of $\mathbf{y}$ against $\hat{\mathbf{y}}$ was made for every datasets least square. Line was fitted to the plotted values. If this fitted line has the same slope and offset as a $1:1$ line, the model would predict the observations perfectly without bias. All the models have a slope of the fitted line 0.9 meaning that if predicted value is high it will be lower than the true y and if the predicted value is low it will probably be higher than the true y. If it would be possible to observe fat value as 0 from the spectral image the model would predict fat value of 1.4%. If the slope would be higher than 1 and the intercept higher than 0 a high

predicted value would be higher than true y and low predicted value would be lower than true y. Even though the RMSEP is different between the datasets, the value of R^2_{pred} is the same for all of them. In R^2_{pred} the effect of variation in RMSEP value has been terminated by dividing the standard deviation of the difference between $\hat{\mathbf{y}}$ and $\mathbf{y}$. Then assessment has been done on how well the model will predict, not exactly the value of the average distance between $\mathbf{y}$ and $\hat{\mathbf{y}}$ values.

6.1.6 Including additional responses

To improve the estimation of $\hat{\beta}$ the Canonical Partial Least Squares were applied using additional data. The CPLS seems to reach the lower bound of the RMSEP obtained by PLSR using fewer components. Then complexity is decreased which is favorable.A model developed by PLSR seems to need more components than the CPLS model which creates more sources of error. The additional responses supplied tends to explain the variables which are contained in the residual in the PLSR model, resulting in that not as many components are needed to span the variation between $\mathbf{y}$ and $\mathbf{X}$.

In this thesis a function in R was made that tried all combinations of additional responses in order to find if there was something connecting the additional responses which obtained the lowest RMSEP values. What connected the additional responses was that they were all giving more information about the plug. Main responses obtaining the lowest RMSEP were moisture in the plug, weight of the plug and location of the plug.

In order to try methods using additional responses it would be better to have more measures included in dataset 1 to see if similar additional responses are having effect on the prediction obtained. A study of finding the optimal additional response could be carried out. In this thesis number of 13 additional responses are stored in dataset 2 and 8 additional responses for dataset 3. Result from this thesis gives idea of which additional responses obtained. In this thesis additional responses which gave more information about the plug are removed.

In dataset 1 the only additional data was location of the plug. By using the location of the plug the RMSEP value obtained using one component was the same as when PLSR uses four components. In figure 5.13 it can be seen that the cross validation method has more effect in CPLSR that in PLSR. The cross validation by line in CPLSR

is underestimating the prediction error. By using CPLSR the model published in the paper [11] has been improved. The RMSEP value reported was 1.96 which can be lowered using CPLSR to 1.76.

In dataset 2 existed more additional variables. Best combination of additional variables was found by a function which selected the combination having the lowest prediction error. Using 6 additional variables gave the lowest prediction error, only using one component model. Traditional PLSR reached similar values using 4 component model.

The CPLSR method is more sensitive for cross validation method than PLSR and PCR. Reason for that could be that the additional responses introduce higher level of leaking between batches of data. The leave one out cross validation gave lower prediction error in all cases. The function that found the additional response which resulted in the lowest prediction error, gave different additional responses depending on cross validation method. The use of K-fold cross validation is considered as the optimal cross validation method in this thesis. Then leaking between batches is removed, which seems to has a higher magnitude when the additional responses are present.

For Dataset 3 existed 8 additional responses for the fish. When K-fold cross validation where segments were defined by fillet, the lowest RMSEP value were obtained in one component model using three additional responses. Traditional PLSR did not show as good result as the CPLSR method for the first five components.

The additional responses which improved the prediction model were information about the location of the plug, moisture in the plug and thickness of the plug.

If there had been time, a calculation of CPLSR should also had been done on the test and calibration set to investigate the behavior of $\hat{\beta}$ when using CPLSR.

6.2 Main Results

Conclusion from this thesis is that when modeling pixel values in a NIR image by collecting plugs from the fillet, measure its fat value and retaining corresponding spectral values is to use CPLSR when developing the model. CPLSR demands fewer components than PLSR and PCR. K-fold cross validation should be used to estimate the prediction

quality in stead of the leave one out cross validation. The data used should be divided up to test and calibration set to know how much variation in $\hat{\beta}$ may be expected.

The model developed for the NIR scanner in Iceland, developed by using CPLSR, puts weights on the regions where information of fat is contained in the explanatory matrix. The used additional responses were weight of the fillets, ungutted weight and intestines weight.

The RMSEP value when using the $\hat{\beta}$ developed by PLSR on dataset 1 on the explanatory variables of dataset 3 obtained was 2.023. By using the model obtained by CPLS using dataset 3 the RMSEP value was 1.858. When the prediction of the old model was plotted against the true value of fat in dataset 3 a offset of a fitted line with intercept 2.59 and slope 0.8 meaning that low predictions are to high and high predicted values are to low in the real world. The slope was 0.9 and the offset was 1.24 of using the new model on the data obtained from the QMonitor in Iceland.

It was found, when developing CPLSR model as many additional variables should be collected as possible. There was not much consistence between datasets when looking at the cross validation methods. For a QMonitor with same configurations as the one in Iceland the additional responses weight of fillets, ungutted weight and intestines weight should be used to improve the prediction model.

6.3 Further studies

Variation of fat in fillet is rather high. Plugs for modeling are picked out of the fillet by a system that the person who makes the model thinks capture the variation of fat the best. In this paper two systems are tried. To pick six plugs or pick five plugs on different locations but over all on similar places. It is not knowng if more or less plugs are needed to improve or obtain same prediction quality. There is also no knowledge of where to take the plugs. One solution would be to obtain a chemical value for the whole fillet. Then divide the fillet once, and by variable selection find which part of the fillet is more representative. When that is found, the fillet is divided again and by variable selection selects the most representative region on the fillet. This is repeated until the fillet has been divided such that each grid contains only one pixel. In the beginning of this thesis, this was the topic to solve, with the aim to publish the solution. The problem which

needes to be solved is how to divide the fillet into graph. The division has to be the same for every size and shape of the fillet. The goal is to find where to pick the plugs and optimum number of plugs needed to create a good prediction model.

There exist a lot of programs around, that have been made for the QMonitor machine. Main purpose of these programs is to capture data and analyses. There is not much programming needed to combine those programs to similar programs I'm using in this thesis.

Understanding why Canonical Partial Least squares tends to use fewer components to retain same prediction quality as the PLSR reports using more components. The datasets presented in this thesis would be useful to try, because they have different background data included. To find out which background data would be ideal to record when making a model for QMonitor and why this data is better than other data.

To somehow estimate the effect of quality of the measures would be nice to quantify. There are many sources of error when doing the measures on the samples, and when the data is collected from the spectral images. The fourth dataset was included in this thesis to get a better knowledge about which observations were broken by nature, and which were broken by wrong collection of the spectral values. The dataset was excluded because if kept in too much data would than be included in this thesis. The data was also predicting very poorly. One reason was because the program used to collect the spectral image is quite new and containins some bugs.

To fix up the programs made in this thesis would be nice. Because of lack of time they could not been made more publishable. They should be able to cope whith any kind of segmentation and be usable in more analysis than only in analysis of salmon fillets.

Appendix A

R-code

```
# Set the working directory to where the Excel Sheets are located.

setwd("C:/Users/olafur/Documents/OliMaster")

# The packages used to get acess to:
# R.oo          The function trim which is done on the excel data
# RODBC   Used to import Excel sheets containing data
# pls           Gives acess to the CPLSR function.

# Packages used to obtain nice plots are utils, ggplot2 and geor

# Before using the packages they need to be installed by the command
# install.packages('packages name')

library(R.oo)
library(RODBC)
library(pls)
library(utils)
library(ggplot2)
library(geoR)

# The scriptes used to read in the data from the datasets. The data was
# storred in Excel sheets containing the spectra value and all responses.
# In all the reading scripts the pat in setwd() have to be set to the
# folder where the files are located. After reading in the data the
# location of the plugs is set as factors Then reduced datasets
# are created for all dataset where broken observations are removed.

#------------------------- Plug 1 -------------------------------------

setwd("C:/Users/olafur/Documents/oliMaster/Plug1")
channel <- odbcConnectExcel("plug1.xls")
```

```
plug1 <- sqlQuery(channel, "select * from [plug1$]")
close(channel)
rm(channel)
row.names(plug1)<-plug1$name
plug1<-data.frame(NIR=I(as.matrix(plug1[,25:39])),Fat=plug1$Fat,PlugPos=factor(
    plug1$Position))
plug1$Posit<- I(model.matrix(~y-1, data.frame(y=factor(plug1$PlugPos))))
plug1$PlugPos<-NULL

###### Removing broken data from dataset 1.
rem<-c("34-L-NE","45-L-NE","39-L-ND","44-L-ND","34-L-NC")
which.rem<-which(!is.na(match(row.names(plug1),rem)))
new.plug1<-plug1[-which.rem,]

#------------------------------ Plug 2 ------------------------------------

setwd("C:/Users/olafur/Documents/oliMaster/Plug2")
channel2 <- odbcConnectExcel("plug2.xls")
plug2 <- sqlQuery(channel2, "select * from [plug2Best$]")
close(channel2)
rm(channel2)
row.names(plug2)<-plug2$name
plug2<-data.frame(NIR=I(as.matrix(plug2[,5:19])), Fat=plug2$fat,PlugPos=factor(
    plug2$plug),UngW=plug2$ungw,GutW=plug2$guttW,Length=plug2$length,PlugThick=
    plug2$plugTh,PlugW=plug2$plugW,PlugMoust=plug2$plugMo,TrueF=plug2$TrueF,
    Idoxan=plug2$Idox,Asta=plug2$Asta,Cant=plug2$Cant,Ph=plug2$Ph,Temp=plug2$Temp
    )
plug2$Posit<- I(model.matrix(~y-1, data.frame(y=factor(plug2$PlugPos))))
plug2$PlugPos<-NULL

###### Removing broken data from dataset 2
rem<-c
    ("85-2","79-3","77-3","597-3","87-3","593-3","122-5","81-5","593-6","599-3")
which.rem<-which(!is.na(match(row.names(plug2),rem)))
new.plug2<-plug2[-which.rem,]

#------------------------------ Plug 3 ------------------------------------

setwd("C:/Users/olafur/Documents/oliMaster/Plug3")
channel3 <- odbcConnectExcel("plug3.xls")
plug3 <- sqlQuery(channel3, "select * from [Plug3AllCorr$]")
close(channel3)
rm(channel3)
row.names(plug3)<-plug3$name
plug3<-data.frame(NIR=I(as.matrix(plug3[,7:21])),Fat=plug3$Fat,PlugPos=factor(
    plug3$Plug),FillW=plug3$FillW,UngW=plug3$UngW,Length=plug3$Length,Sex=
    plug3$Sex,IntW=plug3$Intest,PlugW=plug3$PlugW,PlugMo=plug3$PlugMo)
plug3$Posit<- I(model.matrix(~y-1, data.frame(y=factor(plug3$PlugPos))))
```

```
plug3$PlugPos<-NULL
setwd("C:/Users/olafur/Documents/OliMaster")

###### Removing broken data from dataset 3
rem<-c
    ("1726-3","1838-5","1832-5","1980-5","1784-5","1784-4","1784-3","1828-5","1663-4","1814-4","

which.rem<-which(!is.na(match(row.names(plug3),rem)))
new.plug3<-plug3[-which.rem,]
#------ All dataset ready for analysis.

rm(rem)
rm(which.rem)


#---------- First analysis ----------------------------------------
# Function which takes Y and X data.
# Calculates the eigenvalues of X'X and its qualities and
# the covariance between Pc and Y. Then the covariance and
# correlation between comlumns of X and y
#------------------------------------------------------------------

# Example: FirstAnalyses(new.plug1$NIR,new.plug1$Fat)

FirstAnalyses<-function(X,Y){
    res<-list()
    ratio<-matrix(0,nrow=dim(X)[2],ncol=6,dimnames<-list(c(),c("  Eigenvalue ","
    Ratio of variance "," Total ratio "," Covariance of Pc and Y","Correlation
   of X and Y","Covariance between X and Y")),byrow=T)
    # First do scaling
    CentX<-scale(X,scale=F)
    CentY<-scale(Y,scale=F)
    # Correlation of Y and X.
    s<-t(CentX)%*%CentY
    # Correlation of principle component and Y
    E<-eigen(t(CentX)%*%CentX)
    for(i in 1:dim(X)[2]){
        ratio[i,1]<-E$value[i]
        ratio[i,2]<-E$value[i]/sum(E$value)
        ratio[i,3]<-sum(E$value[1:i])/sum(E$value)
        ratio[i,4]<-t(E$vector[,i])%*%s
        }
    ratio[,5]<-cor(CentX,CentY)
    ratio[,6]<-cov(CentX,CentY)
    res<-ratio

    return(res)
}
```

```
#---------- Principal components regression ----------
# Function which takes Y and X data and retrieves Scores,
# loadings, b_hat and y_hat. Also it retrives the values
#  used in the scree plot. ncom is number of component prefered.
#-----------------------------------------------------------

# Example Pc(plug2$Fat,plug2$NIR,ncom=5)

Pc<-function(Y,X,ncom){
      res<-list()
      y<-as.matrix(Y)
      x<-X
      scores<-matrix(nrow=dim(y)[1],ncol=ncom)
      y_hat<-matrix(nrow=dim(y)[1],ncol=ncom)
      b_hat<-matrix(nrow=dim(x)[2],ncol=ncom)
      inter<-c()
      rmsep<-c()
      scree<-c()
            for( b in 1:ncom){
                  P<-eigen(t(scale(x,scale=F))%*%scale(x,scale=F))$vectors[,1:b,drop
    =FALSE]
                  scree[b]<-eigen(t(scale(x,scale=F))%*%scale(x,scale=F))$values[b]/
    sum(eigen(t(scale(x,scale=F))%*%scale(x,scale=F))$values)
                  scores<-scale(x,scale=F)%*%P[,1:b]
                  b_hat[,b]<-P%*%solve(t(P)%*%t(scale(x,scale=F))%*%scale(x,scale=F)
    %*%P)%*%t(P)%*%t(scale(x,scale=F))%*%(y-mean(y))
                  l<-1
                  for(k in 1:length(y)){
                        y_hat[k,b]<-mean(y)+t(b_hat[,b])%*%(x[l,]-colMeans(x))
                        l <-l+1
                        }
                  }
            res$loding<-P
            res$scores<-scores
            res$scree<-scree
            res$inter<-mean(y)
            res$beta<-b_hat
            res$y_hat<-y_hat
      return(res)
}

#----------Partial Least Square Regression ----------
# Function which takes Y and X data and calculate the
# parts of PLS using the which is described in the
# chapter "Partial Least Square Regresion algoritm"
# ncom is number of component prefered.
# Then method is what is going to be
```

```
#-------------------------------------------------------------

# Example  Pls(plug2$Fat, plug2$NIR, ncom=5)

Pls<-function(Y,X,ncom){
        res<-list()
        y<-as.matrix(Y)
        x<-X
        y_hat<-matrix(nrow=dim(y)[1], ncol=ncom)
        b_hat<-matrix(nrow=dim(x)[2], ncol=ncom)
        inter<-c()
        rmsep<-c()
        y_i<-y
        x_i<-x
        out_y<-y
        out_y2<-y-mean(y)
        out_x<-x
        out_x2<-scale(x, scale=F)
        P<-W<-matrix(nrow=dim(out_x)[2], ncol=ncom)
        T<-matrix(nrow=dim(out_x2)[1], ncol=ncom)
        Qa<-c()
        for( b in 1:ncom){
            w_a<-t(out_x2)%*%out_y2
            cc<-c(1/sqrt(t(out_y2)%*%out_x2%*%t(out_x2)%*%out_y2))
            w_a<-cc*w_a
            W[,b]<-w_a
            t_a<-out_x2%*%w_a
            T[,b]<-t_a
            p_a<-t(out_x2)%*%t_a%*%(1/(t(t_a)%*%t_a))
            P[,b]<-p_a
            q_a<-t(out_y2)%*%t_a%*%(1/(t(t_a)%*%t_a))
            Qa[b]<-q_a
            out_x2<-out_x2-t_a%*%t(p_a)
            out_y2<-out_y2-t_a%*%q_a
            b_hat[,b]<-W[,1:b,drop=FALSE]%*%solve(t(P[,1:b,drop=FALSE])%*%W
    [,1:b,drop=FALSE])%*%Qa[1:b]
            l<-1
            for(k in 1:length(y)){
                y_hat[k,b]<-mean(out_y)+t(b_hat[,b])%*%(x_i[l,]-colMeans(
    out_x))
                inter[k]<-mean(out_y)
                l<-l+1
                }
            }
        res$y_hat<-y_hat
        res$ymean<-mean(y)
        res$xmean<-colMeans(x)
        res$b_hat<-b_hat
```

```
        res$LoadWei<-W
        res$Scores<-T
        res$xLoad<-P
        res$yLoad<-Qa
    return(res)
}

#-- Cross validation using Principle componet Regression ----------------
# Function which takes Y and X data and calculates rmsep
# between y and hat(y) where hat(y) was calculated using
# leave one out CV leave one segment out. Parameters
# used are:
# method :  Takes the values line for leave on out CV,
#             fish using segmetns defined by fish and
#             plug using segments defined by plug.
#-------------------------------------------------------------------------

PcCv<-function(Y,X,ncom,method="line"){
        res<-list()
        y<-as.matrix(Y)
        x<-X
        y_hat<-matrix(nrow=dim(y)[1],ncol=ncom)
        inter<-c()
        rmsep<-c()
        scores<-matrix(nrow=dim(y)[1],ncol=ncom)
        b_hat<-matrix(nrow=dim(x)[2],ncol=ncom)
        scree<-c()
        if(method=="line"){
            nn<-dim(y)[1]
            numbers<-1:nn
        }
        if(method=="fish"){
            numbers<-match(gsub(".*","",rownames(X)),unique(gsub(".*","",
    rownames(X))))
            fish<-unique(trim(gsub(".*","",rownames(X))))
            nn<-length(fish)
        }
        if(method=="plug"){
            numbers<-match(trim(gsub(".*-","",rownames(X))),unique(trim(gsub
    (".*-","",rownames(X)))))
            fish<-unique(trim(gsub(".*-","",rownames(X))))
            nn<-length(fish)
        }
        for(i in 1:nn){
            line<-numbers == i
            y_i<-y[line,]
            x_i<-x[line,,drop=FALSE]
            out_y<-y[!line,]
```

```
                out_y2<-y[!line,]-mean(y[!line,])
                out_x<-x[!line,]
                out_x2<-scale(x[!line,],scale=F)
                lines<-which(line)
                      for( b in 1:ncom){
                            P<-eigen(t(out_x2)%*%out_x2)$vectors[,1:b,drop=FALSE]
                            scree[b]<-eigen(t(out_x2)%*%out_x2)$values[b]/sum(eigen(
    t(out_x2)%*%out_x2)$values)
                            scores<-scale(x,scale=F)%*%P[,1:b]
                            b_hat[,b]<-P%*%solve(t(P)%*%t(out_x2)%*%out_x2%*%P)%*%t(
    P)%*%t(out_x2)%*%out_y2
                            l<-1
                            for(k in lines){
                            y_hat[k,b]<-mean(out_y)+t(b_hat[,b])%*%(x_i[l,]-colMeans
    (out_x))
                            inter[k]<-mean(out_y)
                            l <-l+1
                            }
                }
          }
          res$y_hat<-y_hat
          res$b_hat<-b_hat
          res$loding<-P
          res$varRatio<-scree
          res$scores<-scores
          res$inter<-sqrt(mean((y-inter)^2))
          res$rmsep <-sqrt(apply((matrix(y,dim(X)[1],ncom)-y_hat)^2,2,mean))
          res$Rsqrt<-1-(apply((matrix(y,dim(X)[1],ncom)-y_hat)^2,2,sum)/apply((
    matrix(y,dim(X)[1],ncom)-mean(y))^2,2,sum))

     return(res)
}

#--Cross validation using Partial Least Square Regression ----
# Function which takes Y and X data and calculates rmsep
# between y and hat(y) where hat(y) was calculated
# using leave one out CV leave one segment out. Parameters
# used are:
# method :      Takes the values line for leave on out
#               CV, fish using segmetns defined by
#               fish and plug using segments defined by plug.
#-------------------------------------------------------------

# Example: PlsCv(new.plug2$Fat,new.plug2$NIR,ncom=5,method="fish")
PlsCv<-function(Y,X,ncom,method="no"){
      res<-list()
      y<-as.matrix(Y)
      x<-X
```

```
if(method=="no"){
        for(i in 1:ncom){
                Pc<-eigen(t(scale(x,scale=F))%*%scale(x,scale=F))$values
                res$X<-c(res$X,sum(Pc[1:i])/sum(Pc))
                }
        }else{
        y_hat<-matrix(nrow=dim(y)[1],ncol=ncom)
        b_hat<-matrix(nrow=dim(x)[2],ncol=ncom)
        inter<-c()
        rmsep<-c()
        if(method=="line"){
                nn<-dim(y)[1]
                numbers<-1:nn
        }
        if(method=="fish"){
                numbers<-match(gsub("-.*","",rownames(X)),unique(gsub("-.*","",
rownames(X))))
                fish<-unique(trim(gsub("-.*","",rownames(X))))
                nn<-length(fish)
        }
        if(method=="plug"){
                numbers<-match(trim(gsub(".*-","",rownames(X))),unique(trim(gsub
(".*-","",rownames(X)))))
                fish<-unique(trim(gsub(".*-","",rownames(X))))
                nn<-length(fish)
        }
        for(i in 1:nn){
                line<-numbers == i
                y_i<-y[line,]
                x_i<-x[line,,drop=FALSE]
                out_y<-y[!line,]
                out_y2<-y[!line,]-mean(y[!line,])
                out_x<-x[!line,]
                out_x2<-scale(x[!line,],scale=F)
                P<-W<-matrix(nrow=dim(out_x)[2],ncol=ncom)
                T<-matrix(nrow=dim(out_x2)[1],ncol=ncom)
                Qa<-c()
                lines<-which(line)
                for( b in 1:ncom){
                        w_a<-t(out_x2)%*%out_y2
                        cc<-c(1/sqrt(t(out_y2)%*%out_x2%*%t(out_x2)%*%out_y2))
                        w_a<-cc*w_a
                        W[,b]<-w_a
                        t_a<-out_x2%*%w_a
                        T[,b]<-t_a
                        p_a<-t(out_x2)%*%t_a%*%(1/(t(t_a)%*%t_a))
                        P[,b]<-p_a
                        q_a<-t(out_y2)%*%t_a%*%(1/(t(t_a)%*%t_a))
```

```
                    Qa[b]<-q_a
                    out_x2<-out_x2-t_a%*%t(p_a)
                    out_y2<-out_y2-t_a%*%q_a
                    b_hat[,b]<-W[,1:b,drop=FALSE]%*%solve(t(P[,1:b,drop=FALSE])
    %*%W[,1:b,drop=FALSE])%*%Qa[1:b]
                    l<-1
                    for(k in lines){
                        y_hat[k,b]<-mean(out_y)+t(b_hat[,b])%*%(x_i[l,]-colMeans
    (out_x))
                        inter[k]<-mean(out_y)
                        l<-l+1
                        }
                    }
                }
          res$y_hat<-y_hat
          res$ymean<-mean(y)
          res$xmean<-colMeans(x)
          res$b_hat<-b_hat
          res$LoadWei<-W
          res$Scores<-T
          res$xLoad<-P
          res$yLoad<-Qa
          res$inter<-sqrt(mean((y-inter)^2))
          res$rmsep <-sqrt(apply((matrix(y,dim(X)[1],ncom)-y_hat)^2,2,mean))
          res$Rsq<-1-(apply((matrix(y,dim(X)[1],ncom)-y_hat)^2,2,sum)/apply((
    matrix(y,dim(X)[1],ncom)-mean(y))^2,2,sum))
          }
      return(res)
}
#------------ Test Set Function ---------------------------------------
# Function which divide the X and Y data into two datasets,
# devleopes Beta_cal and selects the one with the lowest
# value and introduce it to X_test and calculate
# y_hat and retain the RMSEP. Parameters are:
# Size              Values between 0 and 1 defining
#                   the ration between cal and testset
# val               Takes pls and pc. Selects validation
#                   method on the calibration set.
# val.cv.meth  Takes line, fish og plug. Is the
#                   CV method of the calibration set.
# test.meth         Takes fish and part. Divide data
#                   by fish or irrespective of content.
# Test              Takes cal and test. Defines what the
#                   first part of data is used as.
# remo              Takes F and T. If value is T values
#                   of y vs y_hat is plotted from the
#                   calibration set and selected values retained.
#-------------------------------------------------------------------
```

```
# Test(new.plug1$Fat,new.plug1$NIR,ncom=5,size=0.5,val="no",val.cv.meth="line",
    test.meth="part",Test="test",remo="F")

Test<-function(Y,X,ncom="no",size=0.5,val="pls",val.cv.meth="line",test.meth="
    fish",Test="cal",remo="F"){
     res<-list()
     y_hat<-c()

     if(test.meth=="part"){
          cal.lines<-1:round(length(Y)*size)
          elimin<-1:dim(X)[1]==cal.lines
     }
     if(test.meth=="fish"){
          numbers<-match(gsub(".*","",rownames(X)),unique(gsub(".*","",rownames
    (X))))
          fish<-1:length(unique((trim(gsub(".*","",rownames(X))))))
          cal.lines<-round(length(fish)*size)
          elimin<-numbers<=cal.lines
          nn<-length(fish)
     }
     if(test.meth=="plug"){
          numbers<-match(trim(gsub(".*-","",rownames(X))),unique(trim(gsub
    (".*-","",rownames(X)))))
          plugs<-1:length(unique(trim(gsub(".*-","",rownames(X)))))
          cal.lines<-round(length(plugs)*size)
          elimin<-numbers<=cal.lines
     }
     if(Test=="cal"){
          elimin<-!elimin
          #browser()
     }
     if(ncom=="no"){
          ncom<-dim(X)[2]-1
          }
     if(val=="pc"){
          pc.res<-PcCv(Y[elimin],X[elimin,],ncom,val.cv.meth)
          #browser()
          if(remo=="T"){
          res$remov<-plott(Y[elimin],pc.res$y_hat,X[elimin,])
          }
          res$RsqCal<-cor(pc.res$y_hat[,1],Y[elimin])^2
          res$pc.res.cal<-pc.res$rmsep
          best<-which.min(pc.res$rmsep)
          E<-eigen(t(scale(X[elimin,],scale=F))%*%scale(X[elimin,],scale=F))
    $vectors[,1:best]
```

```
        b_hat<-E%*%solve(t(E)%*%t(scale(X[elimin,],scale=F))%*%scale(X[elimin
    ,],scale=F)%*%E)%*%t(E)%*%t(scale(X[elimin,],scale=F))%*%scale(Y[elimin],
    scale=F)
        }
  if(val=="pls"){
        pls.res<-PlsCv(Y[elimin],X[elimin,],ncom,val.cv.meth)
        if(remo=="T"){
            res$remov<-plott(Y[elimin],pls.res$y_hat,X[elimin,])
        }
        res$pls.res.cal<-pls.res$rmsep
        res$RsqCal<-cor(pls.res$y_hat[,1],Y[elimin])^2
        best<-which.min(pls.res$rmsep)
        out_x2<-scale(X[elimin,],scale=F)
        out_y2<-scale(Y[elimin],scale=F)
        P<-W<-matrix(nrow=dim(X)[2],ncol=best)
        T<-matrix(nrow=length(Y[elimin]),ncol=best)
        Qa<-c()
        for(b in 1:best){
            w_a<-t(out_x2)%*%out_y2
            W[,b]<-w_a
            t_a<-out_x2%*%w_a
            T[,b]<-t_a
            p_a<-t(out_x2)%*%t_a%*%(1/(t(t_a)%*%t_a))
            P[,b]<-p_a
            q_a<-t(out_y2)%*%t_a%*%(1/(t(t_a)%*%t_a))
            Qa[b]<-q_a
            out_x2<-out_x2-t_a%*%t(p_a)
            out_y2<-out_y2-t_a%*%q_a
        }
            b_hat<-W%*%solve(t(P)%*%W)%*%Qa
        }
        for(i in 1:length(Y[!elimin])){
            y_hat[i]<-mean(Y[elimin])+t(b_hat)%*%(X[!elimin,][i,]-colMeans(X[
    elimin,]))
        }
        res$Test.beta.hat<-b_hat
        res$y_hat.from.Test<-y_hat
        res$y_from.Test.set<-Y[!elimin]
        res$RsqTest<-cor(y_hat,Y[!elimin])^2
        res$coeffic<-lm(y_hat~Y[!elimin])
        res$correl<-FirstAnalyses(X[elimin,],Y[elimin])
        res$ncom.from.testset<-best
        res$new.rmsep<-sqrt(mean((Y[!elimin]-y_hat)^2))
        res$elimin<-elimin
  return(res)
}
```

```
#------ Selection of best additional responses in CPLSR-------------------------
#  Warning: The function needs long time for dataset with
#              many additional responses !!
#
# Function which calculates RMSEP for all combinations
# of additional responses and finds the lowest rmsep
# for each nr of additinoal responses. The best combinations
# are retained, the Beta, RMSEP values are retained
# from that combination. Parameters taken in:
# plug:         The dataset used
# maxcom: Maximum nr of components
# met:          Values fish, line, plug. Defines what
#               is storred in the CV segments.
#-------------------------------------------------------------------------------

# Example CpplsOnAllComp(new.plug1,maxcom=3,met="line")

CpplsOnAllComp<-function(plug,maxcom,met){
      all.data<-names(plug)[3:dim(plug)[2]]
      res<-list()
      Best.Rmsep<-list()
      Where.Best.Rmsep<-list()
      Best.Comb<-list()
      all.com<-list()
      Best.Coeff<-list()
      All.Coeff<-list()
      all.Rmsep<-list()
      for(i in 1:length(all.data)){all.com[[i]]<-combn(names(plug)[3:dim(plug)
    [2]],i)}
      for(l in 1:length(all.com)){
            all.Rmsep[[l]] <- matrix(0,nrow=maxcom+1,ncol=dim(all.com[[l]])[2])
            for( b in 1:dim(all.com[[l]])[2]){
                  plug$addY<-as.matrix(plug[all.com[[l]][,b]])

                  modell<-cppls(Fat~NIR,data=plug,segments=segme(plug,met),ncom=
    maxcom,Y.add=addY,validation="CV")
                  all.Rmsep[[l]][,b]<-RMSEP(modell)$val[1,1,]
                  }
            }
            for( i in 1:length(all.Rmsep)){
                  Best.Rmsep[[i]]<-min(all.Rmsep[[i]])
                  pos.of.best<-which(all.Rmsep[[i]]==min(all.Rmsep[[i]]),arr.ind=
    TRUE)
                  Where.Best.Rmsep[[i]]<-pos.of.best
                  Best.Comb[[i]]<-all.com[[i]][,pos.of.best[,2]]
            }
            for ( i in 1:length(Where.Best.Rmsep)){
```

```
                plug$addY2<-as.matrix(plug[Best.Comb[[i]]])
                Best.Coeff[[i]] <- matrix(0,nrow=15,ncol=1)
                All.Coeff[[i]] <- matrix(0,nrow=15,ncol=maxcom)
                ExtraModel<-cppls(Fat~NIR,data=plug,segments=segme(plug,met),ncom=
    maxcom,Y.add=addY2,validation="CV")
                Best.Coeff[[i]]<-ExtraModel$coefficients[,,Where.Best.Rmsep[[i
    ]][1]-1]
                All.Coeff[[i]]<-ExtraModel$coefficients
            }

            res$allRmsep<-all.Rmsep
            res$BestCoeff<-Best.Coeff
            res$AllCoeff<-All.Coeff
            res$all.comb<-all.com
            res$Where.Best.Rmsep<-Where.Best.Rmsep
            res$Best.Rmsep<-Best.Rmsep
            res$Best.Comb<-Best.Comb
     return(res)
}


#-- Segments definition function used in CpplsOnAllComp function-----------
# Function used in CpplsOnAllComp which divided the data into
# segments where segments are defined by fish,
# line and plug. Parameters are:
# plug:         The dataset as a dataframe
# met:          The segmentation definition.
#-------------------------------------------------------------------------

segme<-function(plug,met="fish"){
     segments<-list()
            if(met=="line"){
            nn<-dim(plug)[1]
            numbers<-1:nn
            for( i in 1:nn){
            segments[[i]]<-which(numbers==i)
            }
     }
     if(met=="fish"){
            numbers<-match(gsub(".*","",rownames(plug)),unique(gsub(".*","",
    rownames(plug))))
            fish<-unique(numbers)
            sequ<-1:length(numbers)
                  for(i in 1:length(fish)){
                        segments[[i]]<-which(numbers==i)
                        }
            }
     if(met=="plug"){
```

```
            numbers<-match(trim(gsub(".*-","",rownames(plug))),unique(trim(gsub
    (".*-","",rownames(plug)))))
            plug<-unique(numbers)
            sequ<-1:length(plug)
            for(i in 1:length(plug)){
                  segments[[i]]<-which(numbers==i)
                  }
            }
     return(segments)
}

#--------- Example of the plotting scripts used ----------------------------------
# ggplot2 package was used for plotting. This is one
# example out of all scripts used to plot the results in the thesis.
#---------------------------------------------------------------------------------

# Cal Test pls line

umm<-c(0.25,0.5,0.75)

TestCalPlsLine<-data.frame()
newRmsep<-data.frame()
names<-c("RMSEP","Dataset","CalSize","NrofCompoentsRetained")
named<-c("RMSEP","NrofCompoentsRetained","Dataset","CalSize")

for( i in umm){
temp<-data.frame()
temp2<-data.frame()
temp<-data.frame(Test(new.plug1$Fat,new.plug1$NIR,ncom=5,size=i,val="pls",val.cv.
    meth="line",test.meth="part",Test="test",remo="F")$pls.res.cal,rep("Dataset1
    ",5),paste(rep(i*100,5),"%",sep=""),c(1:5))
temp2<-data.frame(Test(new.plug1$Fat,new.plug1$NIR,ncom=5,size=i,val="pls",val.cv
    .meth="line",test.meth="part",Test="test",remo="F")$new.rmsep,Test(new.
    plug1$Fat,new.plug1$NIR,ncom=5,size=i,val="pls",val.cv.meth="line",test.meth
    ="part",Test="test",remo="F")$ncom.from.testset,"Dataset1",paste(i*100,"%",
    sep=""))
names(temp2)<-named
names(temp)<-names
TestCalPlsLine<-rbind(TestCalPlsLine,temp)
newRmsep<-rbind<-rbind(newRmsep,temp2)
}
colnames(newRmsep)<-named
colnames(TestCalPlsLine)<-names

for( i in umm){
temp<-data.frame()
temp2<-data.frame()
```

```
temp<-data.frame(Test(new.plug2$Fat,new.plug2$NIR,ncom=5,size=i,val="pls",val.cv.
    meth="line",test.meth="part",Test="test",remo="F")$pls.res.cal,rep("Dataset2
    ",5),paste(rep(i*100,5),"%",sep=""),c(1:5))
temp2<-data.frame(Test(new.plug2$Fat,new.plug2$NIR,ncom=5,size=i,val="pls",val.cv
    .meth="line",test.meth="part",Test="test",remo="F")$new.rmsep,Test(new.
    plug2$Fat,new.plug2$NIR,ncom=5,size=i,val="pls",val.cv.meth="line",test.meth
    ="part",Test="test",remo="F")$ncom.from.testset,"Dataset2",paste(i*100,"%",
    sep=""))
names(temp2)<-named
names(temp)<-names
TestCalPlsLine<-rbind(TestCalPlsLine,temp)
newRmsep<-rbind<-rbind(newRmsep,temp2)
}
colnames(newRmsep)<-named
colnames(TestCalPlsLine)<-names

for( i in umm){
temp<-data.frame()
temp2<-data.frame()
temp<-data.frame(Test(new.plug3$Fat,new.plug3$NIR,ncom=5,size=i,val="pls",val.cv.
    meth="line",test.meth="part",Test="test",remo="F")$pls.res.cal,rep("Dataset3
    ",5),paste(rep(i*100,5),"%",sep=""),c(1:5))
temp2<-data.frame(Test(new.plug3$Fat,new.plug3$NIR,ncom=5,size=i,val="pls",val.cv
    .meth="line",test.meth="part",Test="test",remo="F")$new.rmsep,Test(new.
    plug3$Fat,new.plug3$NIR,ncom=5,size=i,val="pls",val.cv.meth="line",test.meth
    ="part",Test="test",remo="F")$ncom.from.testset,"Dataset3",paste(i*100,"%",
    sep=""))
names(temp2)<-named
names(temp)<-names
TestCalPlsLine<-rbind(TestCalPlsLine,temp)
newRmsep<-rbind<-rbind(newRmsep,temp2)
}
colnames(TestCalPlsLine)<-names
colnames(newRmsep)<-named


windows(10,4)
ok<-ggplot(TestCalPlsLine,aes(x=NrofCompoentsRetained,y=RMSEP,colour=CalSize))+
    geom_line()+geom_point()+facet_wrap(~Dataset)+opts(title="Partial Least
    Square Regression Cross Validated By Line with 25-75% of the data for
    calibratoin")+xlab("Nr of components retained")
ok + geom_point(data = newRmsep,size = 2,label=c('25,50,75'))+geom_text(data=
    newRmsep,label="NewRmsep",size=3,hjust=1.1)

ggsave(file="CalTestPlsLine.pdf")
```

Bibliography

[1] Valdimar Ingi Gunnarsson. Selective breeding-the key to sustainability in aquaculture, 2012. URL http://www.fisheries.is/aquaculture/slective-breeding/.

[2] Stofnfiskur. History of stofnfiskur, 2012. URL http://www.stofnfiskur.is/?modID=1&id=69.

[3] T. Gjedrem and M. Baranski. *Selective breeding in aquaculture: An introduction(1th ed.)*, volume 10 of *Reviews: Methods and Technologies in Fish Biology and Fisheries*. Dordrecht, Netherlands : Springer-Verlag, 2009. ISBN 9789048127726. URL http://books.google.is/books?id=2ec7Txf3KZkC.

[4] T. Gjedrem. *Selection and breeding programs in aquaculture (1th ed.)*. Dordrecht, Netherlands : Springer-Verlag, 2005. ISBN 0471754951. URL http://books.google.is/books?id=y1uJcgAACAAJ.

[5] Trygve Gjedrem. *AKVAFORSK i nasjonal og internatsjonal akvakultur*. AKVAFORSK, 2007.

[6] The Research Council of Norway. *Aquaculture Research: From Cage to Consumption*. The Research Counchil of Norway, 2005. URL http://www.forskningsradet.no/en/Newsarticle/New_book_on_the_knowledge_status_of_Norwegian_aquaculture_research/1236685398480?lang=en.

[7] Martin Kermit, Jens Petter Wold, and Astrid Woll. On-line quality classification of crabs. In *Innovation in Efficient Food Processing*, 2010. URL http://www.fiskerifond.no/files/projects/attach/martin_kermit-qvision-krabbekonferansen-2010-10022010.pdf.

[8] Are Folkestad, Jens Petter Wold, Kjell-Arne Rørvik, Jon Tschudi, Karl Henrik Haugholt, Kari Kolstad, and Turid Mørkøre. Rapid and non-invasive measurements

of fat and pigment concentrations in live and slaughtered atlantic salmon (salmo salar l.). *Aquaculture*, 280(1–4):129 – 135, 2008. ISSN 0044-8486. URL http://www.sciencedirect.com/science/article/pii/S0044848608003372.

[9] Distell. Distell fish fatmeter, 2012. URL http://www.distell.com/index.php/products/prd-fish-fatmeter/introduction/ffm-general-description.

[10] Emil Veliyulin, Claas van der Zwaag, Wolfgang Burk, and Ulf Erikson. In vivo determination of fat content in atlantic salmon (salmo salar) with a mobile nmr spectrometer. *Journal of the Science of Food and Agriculture*, 85(8):1299–1304, 2005. ISSN 1097-0010. URL http://dx.doi.org/10.1002/jsfa.2117.

[11] Vegard H Segtnan, Martin Høy, Oddvin Sørheim, Achim Kohler, Frank Lundby, Jens Petter Wold, and Ragni Ofstad. Noncontact salt and fat distributional analysis in salted and smoked salmon fillets using x-ray computed tomography and nir interactance imaging. *Journal of Agricultural and Food Chemistry*, 57(5):1705–1710, 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19256551.

[12] Vegard Segtnan, Martin Høy, Frank Lundby, Bjørg Narum, and Jens Wold. Fat distribution analysis in salmon fillets using non-contact near infrared interactance imaging: a sampling and calibration strategy. *Journal Of Near Infrared Spectroscopy*, 17(5):247, 2009.

[13] F Alfnes, Ag Guttormsen, Gro Steine, and Kari Kolstad. Consumers willingness to pay for the color of salmon: A choice experiment with real economic incentives. *American Journal of Agricultural Economics*, 88(4):1050–1061, 2006. URL http://www3.interscience.wiley.com/journal/118558683/abstract.

[14] Cozzolini D, Cynkar W.U., Dambergs R.G., Janik L., and Gishen M. Near infrared spectroscopy in the australian grape & wine industry, 2007. URL http://www.crcv.com.au/viticare/vitinotes/Viti-Notes/near%20infrared%20spectroscopy%20technology%20for%20grape%20and%20wine%20quality%20measurement/Near%20infrared%20spectroscopy%20in%20the%20Australian%20grape%20and%20wine%20industry.pdf.

[15] Norway Odenberg. Qvision 500 analyzer, 2012. URL http://www.odenberg.com/wp-content/uploads/2011/05/Odenb_QV_500_8p_IR_UK_2011_1k.pdf.

[16] John Verzani. *Using R for Introductory Statistics*. Chapman & Hall/CRC, 2005. URL http://wiener.math.csi.cuny.edu/UsingR/. ISBN 1-584-88450-9.

[17] P.J. Bickel and K.A. Doksum. *Mathematical statistics: Basic ideas and selected topics*. Upper Saddle River, NJ: Prentice Hall, 2001. ISBN 0132306379. URL http://www.amazon.com/Mathematical-Statistics-Basic-Selected-Topics/dp/013850363X.

[18] Douglas C. Montgomery, Elizabeth A. Peck, and Geoffrey G. Vining. *Introduction to Linear Regression Analysis (4th ed.)*. Chichester, UK: John Wiley & Sons Inc, Hoboken, July 2006. ISBN 0471754951. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0471754951.

[19] Trevor Hastie, Robert Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations (1th ed.)*. Springer Series in Statistics. New York: Springer-Verlag, New York, NY, USA, 2001. ISBN 0387952845. URL http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

[20] H. Martens and T. Naes. *Multivariate calibration (1th ed.)*. Chichester, UK: John Wiley & Sons Inc, Hoboken, 1989. ISBN 047909793. URL http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471930474.html.

[21] Jens Petter Wold, Johansen Ib-Rune, Haugholt Karl Henrik, Tschudi Jon, Thielemann Jens T, Vegard H Segtnan, Bjø rg Narum, and Wold Erik. Non-contact transflectance near infrared imaging for representative on-line sampling of dried salted coalfish (bacalao). *Journal Of Near Infrared Spectroscopy*, 14(1):8, 2006. ISSN 0967-0335. URL http://www.sintef.no/Publikasjoner-SINTEF/Publikasjon/?pubid=SINTEF+S101.

[22] Jens Petter Wold, Martin Kermit, and Astrid Woll. Rapid nondestructive determination of edible meat content in crabs (cancer pagurus) by near-infrared imaging spectroscopy. *Appl. Spectrosc.*, 64(7):691–699, Jul 2010. URL http://as.osa.org/abstract.cfm?URI=as-64-7-691.

[23] CAMO. *Theory used by UNSCRAMBLER*, 2011. URL http://www.camo.com/downloads/U9.6%20pdf%20manual/The%20Unscrambler%20Method%20References.pdf.

[24] R.A. Johnson and D.W. Wichern. *Applied multivariate statistical analysis (6th ed.)*. Pearson Prentice Hall, Upper Saddle River, NJ, 2007. ISBN 9780131877153. URL http://books.google.is/books?id=gFWcQgAACAAJ.

[25] Kristian Hovde Liland. *Multivariate Analysis- Method Development And Novel Applications In Spectrometry*. PhD thesis, Norwegian University of Life Sciences, 2009.

[26] Ulf G. Indahl, Kristian Hovde Liland, and Tormod Næs. Canonical partial least squares—a unified pls approach to classification and regression problems. *Journal of Chemometrics*, 23(9):495–504, 2009. ISSN 1099-128X. URL http://dx.doi.org/10.1002/cem.1243.

[27] Silje Ottestad, M Hy, Astrid Stevik, and Jp Wold. Prediction of ice fraction and fat content in superchilled salmon by non-contact interactance near infrared imaging. *Journal Of Near Infrared Spectroscopy*, 17(2):77–87, 2009.

[28] Geir H. Sørland, Per M. Larsen, Frank Lundby, Alf-Petter Rudi, and Thierry Guiheneuf. Determination of total fat and moisture content in meat using low field nmr. *Meat Science*, 66(3):543 – 550, 2004. ISSN 0309-1740. URL http://www.sciencedirect.com/science/article/pii/S0309174003001578.

[29] Davies, T. The history of near infrared spectroscopic analysis: Past, present and future "from sleeping technique to the morning star of spectroscopy". *Analusis*, 26 (4):17–19, 1998. URL http://dx.doi.org/10.1051/analusis:199826040017.

[30] Center for Natural Resource Information Technology Texas AgriLife Research. Implement nirs fecal sampling technology. In *Lecture,notes*, 2011. URL http://cnrit.tamu.edu/elearning/1_history.html.

[31] Brian G. Osborne. *Near-Infrared Spectroscopy in Food Analysis*, pages 1–13. John Wiley & Sons, Ltd, 2006. ISBN 9780470027318. URL http://dx.doi.org/10.1002/9780470027318.a1018.

[32] Matforsk. Analytical methods development study, subtast 4.2.2. In *Rapid and non-destructive fat and salt distrubution analysis*, 2007.

[33] Martin Høy. personal communication, 2012.

[34] R. J. Barnes, M. S. Dhanoa, and Susan J. Lister. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.*, 43(5):772–777, May 1989. URL `http://as.osa.org/abstract.cfm?URI=as-43-5-772`.