Norwegian University
of Life Sciences

Master's Thesis 2023 30 ECTS
Faculty of Science and Technology

Optimization and assessment of Omega-3 fatty acids from high-dimensional spectroscopic data in the Atlantic salmon breeding programs

Jisoo Park
Master of Science in Data Science

Optimization and assessment of predicted Omega-3 fatty acids from high dimensional spectroscopic data in Atlantic salmon breeding programs

Key words: PLSR, NIR, RAMAN, Markov-Blanket, sample selection, variable selection

Jisoo Park

Master Thesis
submitted in fulfilment of the requirements for Master of Science in Data Science, at the Norwegian University of Life Sciences in Ås, Norway

Main supervisor
Dr Gareth Frank Difford
Faculty of Biosciences, Institute of Animal and Aquaculture Sciences
Norwegian University of Life Sciences (NMBU), Ås

Co-supervisors
Dr Nils Kristian Afseth
Raw Materials and Process Optimization
NOFIMA, Ås

Acknowledgements

This thesis would not have been possible without these wonderful people who have contributed beyond words.

First and foremost, I would like to thank my supervisor, Dr. Gareth Frank Difford, for your dedicated support, invaluable guidance, feedback, and imparting knowledge throughout this entire thesis. I always consider myself lucky to have taken your class (BIN302) in the first semester and to have you as my supervisor. You helped me with new ideas for developing further analysis, and your passion made my work complete.

I would also like to thank my co-supervisor, Dr. Nils Kristian Afseth, for his feedback, vast knowledge, and guidance on a successful thesis. Thank you for allowing me to use the room at Nofima, where I can fully concentrate on my thesis work and study, and the canteen that provides nutrition to save time for taking care of myself.

I sincerely enjoyed discussing and developing my thesis with both of you during this journey.

Lastly, I would like to deeply appreciate my parents and my brother, and all my old friends in Korea. Without your support and love, I would never have accomplished this journey.

Abstract

Atlantic salmon is well known as a rich source of Omega-3 fatty acids (in particular, ALA, DHA, and EPA), and these fatty acids are heritable, thus, the selection of parents will influence their levels in the offspring. The gold standard method of recording Omega-3 fatty acids in Atlantic salmon is costly, time-consuming, and destructive to the sample. For selective breeding purposes, a more affordable method is needed to measure Omega-3 fatty acids in thousands of related salmon. In many breeding programs, vibrational spectroscopy is primarily used with the Partial Least Squares Regression (PLSR) model to measure and predict phenotypes such as Omega-3 fatty acids. However, there is a knowledge gap in estimating heritability using vibrational spectroscopy and finding the effect of sample selection and variable selection methods in the data analysis process perspective. Hence, we optimized and assessed the predicted Omega-3 fatty acids from the high-dimensional spectroscopy data (NIR and Raman spectroscopy data) in the breeding program according to the multiple scenarios combined with sample selection (Kennard-Stone and Random Sampling) and variable selection (with or without Markov Blanket). We found that the PLSR model accuracy and the resulting heritability estimates generally increase with adopting the variable selection method. We also found that NIR spectroscopy has a good affinity with Kennard-Stone sampling, while Raman spectroscopy showed stable performance regardless of sample selection. It is possible to achieve improved PLSR model accuracy and heritability estimates by utilizing the Markov Blanket approach.

Table of Contents

List of Abbreviations

ALA: Alpha(α)-linolenic acid

BLUE: Best Linear Unbiased Estimator

BLUP: Best Linear Unbiased Prediction

DNA: Deoxyribonucleic acid

DHA: Docosahexaenoic acid

EBV: Estimated Breeding Value

EDA: Exploratory Data Analysis

EMSC: Extended Multiplicative Signal Correction

EPA: Eicosapentaenoic acid

FT-IR: Fourier-Transform Infrared

GC: Gas Chromatography

IAMB: Incremental Association Markov Blanket

MSC: Multiplicative Signal Correction

MSE: Mean Square Error

RS: Random Sampling (Sample)

KS: Kennard-Stone

MB: Markov Blanket

PCA: Principal Component Analysis

PCR: Principal Component Regression

PLSR: Partial Least Squares Regression

PRESS: Prediction Residual Error Sum of Squares

RMSE: Rooted Mean Square Error

RMSECV: Rooted Mean Square Error Cross Validation

RMSEP: Root Mean Squared Error of Prediction

SNP: Single Nucleotide Polymorphism

SNV: Standard Normal Variate

List of Figures

List of Tables

1. Introduction

Global fisheries and aquaculture production tend to increase continuously, and the global consumption of marine food has also increased in recent years. In addition, as the world population continues to grow, aquaculture has the potential to provide food sources and nutrients [1]. Aquatic foods are considered an important element of healthy diets because they are not only valued as a rich source of animal protein but also contain Omega-3 fatty acids and micronutrients essential for improving nutrition and health outcomes. Research shows that fish consumption helps to reduce the risk of cardiovascular disease because of the high quantities of Omega-3 fatty acids in some fish species [2]. It is well-known that Atlantic salmon is a rich source of Omega-3 fatty acids. Atlantic salmon farming began with the first breeding program used in the late 1960s in Norway, and it is the dominant species in aquaculture production with the fact that Norway is the biggest producer of farmed Atlantic salmon [1,3]. The primary feed ingredients of Atlantic salmon were traditionally marine-based oils and protein sources until around the 1990s; where concerns over environmental impacts and supply-demand linked with the aquafeed industry using wild fish as feed sources, most feed resources were replaced to some extent with plant-based materials over the past 30 years [1,4]. This change in feed ingredients and feed composition has resulted in a reduction in levels of Omega-3 content in the diets of Atlantic salmon, and therefore, also in their fillets. Of particular importance is eicosatetraenoic acid (EPA; C20-5n3) and docosahexaenoic acid (DHA; C22-6n3) which are reduced, but a concurrent increase in α-linolenic acid (ALA; 18:3n-3) commonly found in plant-based ingredients [5]. This has also led to increasing concerns related to lower fish growth rates, and impaired fish health alongside the decreasing nutritional benefit to the human consumer. Hence, research into ways to maintain healthy fish husbandry and provide high-level Omega-3 content as a method of ensuring sustainable production has gained impetus in recent years.

Selective breeding is one such approach that can be used to alter the characteristics of plants and animals, including fish, in a desired way. Provided a trait or characteristic is heritable, artificially selecting individuals with the most desirable characteristics from parents to the next generation results in permanent and cumulative changes over generations [6]. Currently, Atlantic salmon selection programs in Norway include increased growth rate, increased survival (or resistance to particular diseases), fillet yield, fillet fat percentage, and fillet color as selection criteria [7,8]. In order to select traits, it is necessary to employ statistical methodologies which estimate the genetic parameters of the population. These include estimating heritability, the genetic correlations among traits, and estimated breeding values (see Section 2 below). These statistical models combine two sources of information: namely the phenotypes recorded for thousands of related individuals under the conditions they are expected to perform and how these individuals are related through a pedigree [7,8].

Despite evidence that Omega-3 fatty acids like ALA, EPA, and DHA content in the fillet of Atlantic salmon are significantly heritable (h^2 ranging from 0.09 – 0.26), no breeding programs currently include these in their breeding objectives. The primary barrier to including Omega-3 fatty acids in a breeding program is the prohibitive cost of recording the phenotypes on large numbers of Atlantic salmon. In instances where the phenotypes are expensive or difficult to measure, researchers investigate the use of easier or cheaper alternatives of measurement. In the case of quantitative genetics, an alternative phenotype which is cheaper or easier can replace the expensive gold standard or “true” phenotype, if their heritability estimates are approximately equal

and their genetic correlation is above 0.80 [9]. The alternative phenotype can also be employed as an “indicator trait” provided it is still heritable and genetically correlated to the true phenotype below 0.80.

Fatty acids are quantified using several biochemical steps, which include extracting all lipids from the tissues by hydrolysis using acids and organic solvents, then the fatty acids from triglycerides or free fatty acid forms are converted to methyl esters. Gas chromatography (GC) is then used to separate the methyl esters on the basis of molecular weights and then identified and quantified using analytical tools (e.g., flame ionization detectors, mass spectrometry) [10]. Results show that one measurement of all fatty acids in an Atlantic salmon fillet costs more than 1500 Norwegian Kroner (~150 United States Dollars). Apart from the high cost, the disadvantages of this method are that it is destructive to both fish and the fillet because they are sacrificed in the measurement process, and it is time-consuming since samples must be transported to laboratories for measurement [11]. Sampling and extraction steps are also always prone to errors, contributing to the overall measurement uncertainties. But still, as the gold standard method is the most accurate and precise method, it is used to benchmark and compare with other methods such as spectroscopy. Non-destructive, affordable, and rapid alternative measurements are therefore needed to measure phenotypic ALA, EPA, and DHA for selection purposes.

Vibrational spectroscopy is a collection of techniques used to measure the chemical bonds or functional groups in molecules to determine their chemical composition. It is the techniques are based on how chemical bonds interact with sources of electromagnetic radiation such as light or lasers. Near-infrared (NIR) and Raman spectroscopy are two types of vibrational spectroscopy currently used in food research, including the analysis of Atlantic Salmon fillets [12]. Vibrational spectroscopy techniques have powerful benefits when used, including non-destructive, fast analysis, reasonable cost, and minimal sample preparation requirements [12,13]. The techniques are widely applied to measure the chemical components in the food industry due to these characteristics. Furthermore, the use of vibrational spectroscopy data heavily relies on various statistical analyses, such as Exploratory Data Analysis (EDA), regression analysis for chemical content prediction, and classification [14]. The vibrational spectroscopic dataset is usually high-dimensional, consisting of many often-collinear variables. In addition, in order to obtain quantitative information from a spectrum, a subset of samples is always required to predict the chemistry of new samples through calibration/training on a reference sample. To overcome these challenges, Partial Least Squares Regression (PLSR) is frequently used to create new latent variables from the existing data and employ them to build a regression model. Some papers have already shown the successful application for the prediction of Omega-3 fatty acids using PLSR with vibrational spectroscopic measurement. For example, Afseth et al., 2022, [15] applied PLSR on vibrational spectroscopy data to predict a wide selection of fatty acids in Atlantic salmon and evaluated the potential usages. Interestingly, there are few studies that deal with how PLSR can be used to effectively predict fatty acid features from vibrational spectroscopic data for breeding purposes and in breeding programs.

In general, a data analysis procedure consists of five stages: data collection, data processing, exploratory data analysis, model fitting, and model evaluation. Data collection (acquisition) defines the research question and gathers the required data. Data processing extracts or converts meaningful information from the data and contains the data cleaning process in case of missing data. Exploratory data analysis is the process of understanding the data, and it includes visualization to investigate any specific patterns. Model fitting is often referred to as a modeling

process applied to machine learning through sample selection (data partitioning) and variable (feature) selection, and one or more machine learning models possibly applied to find the best approach. Lastly, model evaluation assesses the fitted models based on the statistical criteria. (See Figure 1)

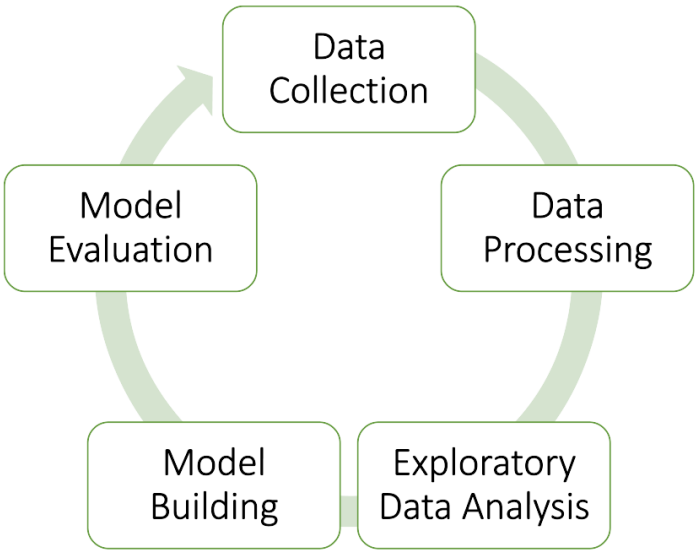


Figure 1. General Process of Data Analysis

As vibrational spectroscopy requires a subset of samples with both the expensive reference data and concurrent vibrational spectra, a key decision is which samples should be sent for reference analysis. From a modeling and prediction perspective, these data are further partitioned into the training and test data sets which are required to build the models in data science. The sample selection methods are used to split the dataset into two subsets. The training set is used to build the model, while the test set is used to test the prediction performance and generalization of the model. The most common method in the sampling technique is random sampling (RS) selection which is the probabilistic methodology. The other method alternatively used for sample selection is the Kennard-stone (KS), which chooses samples using the Euclidean distance in predictors space. It is difficult to say which is the better option for sample selection, and it depends on the characteristics of the data set and the deployed modeling processes. For example, Nawar et al., 2018 [16] reported that the KS and RS produced similarly but it is depending on the training data set size in NIR data, and Ferreira et al., 2022 [17] eported that the KS method outperforms the RS practically in NIR data. Although these two sampling methods are widely utilized for analysis, they have not been directly compared in both NIR and Raman spectroscopy data.

Furthermore, variable (feature) selection is a technique that chooses variables among data with a large number of variables and relatively few samples. It is useful in terms of data reduction and improving interpretability, as well as may help to increase the model prediction performance. There are three categories in variable selection methods: filter, wrapper, and embedded method [18]. T. Mehmood et al., 2012, [19] explained the variable selection methods only in PLSR based on loading weights, scores, and loadings. Instead of the variable selection methods introduced above, we would like to apply the Markov blanket (MB) as a variable selection method in PLSR. MB is a variable selection method in a Bayesian network to select the optimal subset of data by

a probability distribution. The MB in the Bayesian Network gained attention to providing a flexible analysis tool of biological data such as Single Nucleotide Polymorphism (SNP) or gene expression [20]. Felipe et al., 2014 [21], and Dorea et al., 2018 [22] have both recently applied an MB for variable selection to predict feed intake in dairy cattle from milk Mid-Infrared (MIR) spectra recorded in milk and reported MB as improving prediction accuracy, although it is not a classical methodology. The effect of MB variable selection in combination with PLSR predictions has not been tested in Atlantic salmon or otherwise.

In the present study, we have access to a large dataset of 613 samples with both NIR and Raman spectra as well as reference omega-3 fatty acid content (in particular, EPA, DHA, and ALA) in a pedigreed population of adult Atlantic salmon. The aim of this thesis is to optimize and evaluate the PLSR model predictive performance of EPA, DHA, and ALA content using NIR and Raman spectroscopy in Atlantic salmon fillets for measurement purposes in breeding programs. Several specific objectives are tested to achieve this aim; (1) The effect of random sampling versus Kenard-Stone sampling on model prediction, (2) The effect of PLSR predictions with or without a Markov-Blanket as a variable selection adoption, (3) Comparison of NIR and Raman spectra for prediction, and (4) Genetic evaluation of NIR and Raman predictive phenotypes against the true phenotypes.

2. Background and Theory

This section will introduce the theoretical background needed to understand the multidisciplinary concepts and experimental approaches used in this thesis. This chapter also covers topics across multiple research areas and lays the groundwork for the materials and methods section that follows thereafter.

2.1. Introduction to Quantitative Genetics and Animal Breeding

Selective breeding, also known as “artificial selection”, is a process where humans choose individual animals based on desired characteristics to be mated and produce offspring with those same desired characteristics. Selective breeding predates modern science and has been successfully employed by humans for thousands of years, resulting in domesticated livestock breeds and crops. Before the fields of genetics and statistics existed, humans had observed that mating the best individuals based on certain characteristics tended to increase or improve those characteristics in their offspring [23]. It is now known that animals pass on 50% of their DNA (genes) in their sperm and oocytes to the next generation. The division of 50% of the DNA into the oocyte or sperm follows meiosis and is a random division of DNA. After fertilization, the resulting offspring have half their chromosomes for each of their parents. Thus, the genetic relationship between offspring and each parent is expected to be 0.5 and this is called the additive genetic relationship [23]. Genes located on the chromosomes encode proteins and other transcription factors, which provide the blueprint for an organism. Some traits are termed qualitative traits that are controlled by the inheritance of relatively few genes, whilst some traits are termed quantitative traits that can be controlled by many thousands of genes, many of which have very small effects [24].

2.1.1. Phenotypes and their heritability

Modern selective breeding requires the combination of two sources of information: 1) phenotypic information and 2) the relationship between individuals.

Definition box

- Gene: the set of instructions in DNA that codes for proteins that produce a trait
- Genotype: the genetic code of the individual that includes all the information can be found inside the individual's cells
- Phenotype: the set of observable physical traits (i.e., eye colors)

The question of which traits or phenotypes are desirable to improve highly varies between terrestrial and aquatic livestock. Examples of phenotypes in Atlantic salmon are growth rate, fillet fat content, disease resistance, et cetera. The phenotype is the result of the genotype and environment, as shown in Equation (1). Examples of environmental factors which can influence a phenotype include age, sex, temperature, and feed nutrition. The environmental effect can also interact with the genotype to determine its phenotype. In other words, phenotype refers to the observable physical or behavioral characteristics that are affected by both genotype and environmental factors [7]. The relationship among all concepts can be described as follows:

$$\text{Phenotype} = \text{Genotype} + \text{Environment} \quad (1)$$

In quantitative genetics, these effects are resolved by looking at the phenotypic differences between individuals as the phenotypic variance and then partitioning the phenotypic variance into genetic and environmental contributions. An important note is that the genotype can be divided into three different effects in terms of variation: additive genetic variance, dominance, and epistatic interaction. Assuming there is no interaction between genotype and environmental factors under the same environmental conditions, the formula (1) above can be expanded as follows:

$$\text{Phenotype} = \text{Additive Genotype} + \text{Dominance} + \text{Epistasis} + \text{Environment} \quad (2)$$

Both dominance and interaction, which are called non-additive genetic variance, are less favorable as they require specialized breeding designs to estimate their variance components, while additive genetic variance is most often used in selective breeding [8]. The complete genotype of an individual is unobservable, but rather the expected genetic covariance between individuals can be estimated from pedigree relationship matrices i.e., the genetic covariance between full-sibs (having the same male (sire) and female (dam) parents) is 0.5, and between half-sibs (having one parent in common) is 0.25. Note, more recently DNA marker information using commercial Single Nucleotide Polymorphism (SNP) arrays has been used to better estimate the realized genetic relationships between individuals [7]. For the purpose of this thesis, pedigree relationships will be used to model additive genetics relationships. Hence, the final formula that we are interested in to link phenotypes and additive genetic relationships between individuals is as follows:

$$\text{Phenotype} = \text{Additive Genotype} + \text{Environment} \quad (3)$$

Heritability is one of the major parameters in animal breeding which provides the proportion of the phenotypic variation caused by genetic variation. It plays a pivotal role in estimating the relative magnitude of variation when planning a breeding program and predicts the response to selection or individuals' breeding values. There are two definitions of heritability in the broad or narrow sense. The broad sense of heritability is defined by dividing genotypic variance by phenotype variance: ($h_b = h^2 = V_G/V_P$); however, it is not often used due to the genetic variance containing dominance and interaction effects. The narrow sense meaning of heritability uses only the additive genetic variance: ($h_n = h^2 = V_A/V_P$). It is commonly used to show the proportion of genetic variance that is possibly transmitted to the next generation. Heritability varies based on the population, traits, and environment. (See Figure 2 for the simple graphical division of the heritability.)

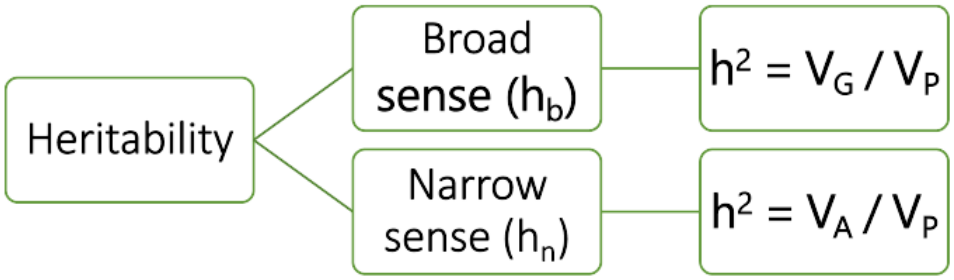


Figure 2. The difference between the broad and narrow sense of heritability is whether using additive genetic variance or genotypic variance.

2.1.2. Linear Mixed Model to estimate variance components, predict breeding values, and genetic correlations

As the complete genotype of an individual is never fully known, animal breeders make use of specialized linear mixed models. Firstly, it includes "fixed effects", which are non-genetic factors or regressors that contribute to phenotyping differences in the population, and their effects must be estimated or taken into account to prevent unintentional confounding with the genetic effects. Secondly, “random effects” which are estimated using a variance-covariance structure (e.g., the expected relationship between individuals based on a pedigree to estimate additive genetic effects). As a result, linear mixed models provide variance components and random solutions for the random effects, while simultaneously estimating the non-genetic fixed effects. In the special case where the random effect is modeled using the additive genetic variance-covariance matrix, the random solutions are called Estimated breeding values (EBV). EBVs are used to rank individuals for selection purposes. If variance components have been estimated previously, it is also possible to predict EBVs using the Best Linear Unbiased Prediction (BLUP) is a methodology designed by Henderson in 1984 [8] to estimate the fixed and random effects simultaneously. The equation for a mixed linear model is the following equation:

$$y = Xb + Za + e \quad (4)$$

, where y indicates the individual phenotype vectors, X is the design matrix that relates the element of y to those of b and b is the fixed effect vector corresponding to the X, Z is the design matrix that relates the element of y to those of a, a is the random effect solutions which represent

the estimated breeding values (additive genetic variance) vector, and e is the residuals under a normal distribution [7,8].

2.1.3. Genetic correlations between traits

Multiple different phenotypes can be measured in individuals or their relatives. In the case that multiple phenotypes are recorded on the same individuals, it is possible to estimate Pearson's correlation, which is referred to as the phenotypic correlation in animal breeding and genetics. It is the identical concept of Pearson's correlation (r), which measures the linear correlation between the two data sets; however, in quantitative genetics, it includes both the additive genetic effects, as well as the non-genetic fixed effects, and the residual error. Breeders are interested in the additive genetic covariance shared between different phenotypes, as it can inform on shared genetic effects between the two phenotypes and how selection for one of the phenotypes will result in a change in the other. In the case of two phenotypes x and y, their genetic correlation (r_G) is given by:

$$r_G = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad (5)$$

, where σ_{xy} represents the covariance between the phenotypes x and y , σ_x and σ_y represent the additive genetic standard deviations. In order to estimate the genetic correlation between two or more traits, substantial amounts of individuals with phenotypes are required 10^2-10^4 [25].

2.1.4. Predicting response to selection from genetic parameters

Animal breeding is centered around achieving desired genetic gain in the following generations. A key equation called the "Breeders Equation" is used to predict genetic gain or response to selection for a given population. Artificial selection is the selection of individuals for transmitting or improving the economically or socially desirable traits by generation. Assuming there is no change in environmental conditions, the effect appears as a change in the population average, which is called a response to selection. The response to selection refers to the difference between the average of the selected parent generation and the average of its offspring, and it plays a pivotal role in animal and plant breeding. It can represent as a formula as follows:

$$R = S \times h^2 \quad (6)$$

, where S is called selection differential, which shows the magnitude of selection. S can be divided into phenotypic standard deviation ($\sigma_P = \sqrt{V_P}$), and selection intensity denoted by i , the above formula can be described as follows:

$$R = i \times \sigma_P \times h^2 \quad (7)$$

The response to selection shows the genetic improvement by one generation, and it accumulates as the generation number increases since the average increases by the generation.

2.1.5. The role of alternative phenotypes

When the traditional or “true phenotypes” are expensive or difficult to measure, it is important to evaluate potential alternative methods of recording or predicting phenotypes. Assuming one case of artificial selection, animal breeders want to gauge the relative response of the true phenotype y when selection is based on the alternative phenotype x. In this case, researchers need to estimate the heritability of both phenotypes and their genetic correlation. Traditionally, if the heritability estimates are of a similar magnitude and the genetic correlation is >= 0.80, the alternative phenotype can replace the “true” phenotype [9]. Importantly, even if the alternative phenotype is less genetically correlated to the “true” phenotype < 0.80 it can still be used as an indicator trait to increase the accuracy of the EBVs for the “true” phenotype, particularly if the alternative phenotype can be more readily recorded on large numbers of related individuals under commercial conditions [26].

2.2. Vibrational Spectroscopy

Vibrational spectroscopy is known as one class of spectroscopic techniques that utilizes the interactions between matter (i.e., molecules) and electromagnetic radiation. This class of techniques includes near-infrared (NIR) and Raman spectroscopy. According to Sultanbawa et al., 2021 [27] and Cozzolino, 2021 [28], there are several main characteristics of vibrational spectroscopy; 1) they are non-destructive techniques that require minimal sample pretreatment [29], 2) they provide information on chemical components as a vibrational "fingerprint" (see vibrational spectroscopy Table 1.1 in [29]), and 3) the spectral data available is usually used in combination with multivariate analysis or machine learning techniques to extract useful qualitative or quantitative information of the sample. However, a pre-processing step is usually required to remove non-relevant physical information in the spectra.

Three parameters are important to understand the basic principles of Spectroscopy; wavelength, frequency, and wavenumbers. The wavelength is defined as the distance between the peak point, and it is represented as lambda (λ). The frequency (υ) is the cycle of waves that represents how many times repeating per unit of time. Lastly, the wavenumbers ($\overline{\upsilon}$) are defined as the number of waves per unit distance, typically centimeters (cm^{-1}). The relationship between parameters is followed by the expression:

$$\overline{\upsilon} = \frac{\upsilon}{\left(\frac{c}{n}\right)} = \frac{1}{\lambda} \tag{8}$$

, where c is the speed of light and n is the refractive index of the medium it is passing through.

When light interacts with materials, photons (a particle of light that carry energy proportional to the radiation frequency) may be absorbed, dispersed or passed through the sample without interaction. In vibrational spectroscopy, the molecule is exposed to an external energy source, such as infrared (IR) radiation or visible or near-infrared light (Raman scattering). Each molecule has a unique level of vibrational energy that depends on the molecular chemical bonds and the geometrical arrangement. Both NIR and Raman spectroscopy identify specific molecules through the interaction with light, but each has a different way of photon energy transfer (i.e., NIR is absorption, Raman is scattering) that changes its vibrational state [29,30].

Molecular vibrations can be explained by using quantum theory (harmonic oscillator model) [29,31]. Molecular vibrations can be explained by using quantum theory (harmonic oscillator model). The molecules can only exist in a quantized energy state in quantum mechanics, and the vibrational energy has certain discrete values in quantum mechanics. The potential energy can be described as a harmonic oscillator model that has a probability distribution by each state, and the differences between the states are constant. The molecules can only change between the adjacent level in the harmonic oscillator model (i.e., fundamental transitions). However, the anharmonic oscillator is a more realistic approach because it provides different energy distances as the vibrational quantum number increases. The distance between each level becomes smaller as the level increases until dissociation is reached. Thus, the transitions in a harmonic oscillator model result in $\triangle v = \pm 1$, while the anharmonic oscillator model results in overtones (energy transition such as v = 0 to v = 2 or v = 0 to v = 3) and combination bands (when two or more fundamental vibrations occur simultaneously). A simple graphic description of the harmonic oscillator and anharmonic oscillator model is as in Figure 3.

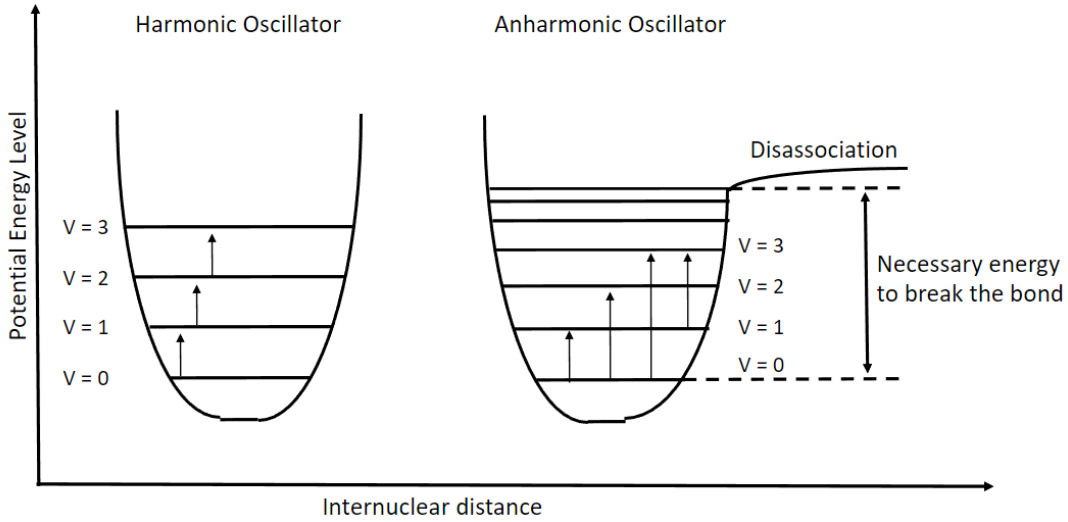


Figure 3. Harmonic oscillator and anharmonic oscillator**.**

2.2.1. Near-Infrared (NIR) spectroscopy

NIR spectroscopy is one type of vibrational spectroscopy that uses light in the electromagnetic spectrum region that corresponds to the wavelength approximately in the range of 750-2500 nm (wavenumbers: 13,300-4,000 cm^{-1}) [30,31]. NIR absorption empirically follows the Lambert-Beer law [14], which indicates there is a linear relationship between the spectrum absorbance and a particular analyte concentration.

The absorption process in NIR spectroscopy requires molecular vibrations that result in changes in the vibrational energy level by dipole moments. The molecular vibration changes the dipole moment of the molecule to transfer the energy from the NIR light. The dipole moments generate the forces from the separation of positive and negative electrical charges in the opposite direction during the vibration [29]. The absorption intensity differs from the sample or matter and is also related to the magnitude of the dipole change during the vibration [30,31]. With these characteristics, NIR spectroscopy is related to the combination between molecular functional groups consisting of e.g., C (carbon), N (nitrogen), O(oxygen), and S (sulfur) [29].

NIR spectroscopy has a wide range of applications within the food and agricultural area; for example, Pizarro et al., 2007 [32], predicted the caffeine content and roasting color using NIR spectroscopy, Holroyd, 2013 [33], showed the applications on milk and milk products by using NIR spectroscopy, Cozzolino et at., 2005 [28], applied to identification of fish species used to make the fishmeal. These examples show that NIR spectroscopy is suitable to use for predicting or classifying food quality and production.

2.2.2. Raman spectroscopy

Raman scattering was first observed by two Indian scientists, Raman and Krishnan, in 1928, but Raman spectroscopy was established as an analytical technology much later due to a lack of technical solutions to deal with this weak scattering phenomenon [34]. According to Orlando et al., 2018 [35] and Lohumi et al., 2014 [36], Raman spectroscopy has several benefits over infrared spectroscopy, such as less influence from water and the use of fiber optic probes.

The basic theory of Raman spectroscopy is described in [29]. Like NIR spectroscopy, Raman spectroscopy also provides chemical information (e.g., chemical structure, spectral patterns) by detecting vibrations in molecules based on Raman scattering. Raman spectroscopy is a scattering technique, and both Rayleigh scattering and Raman scattering can occur when photons reach a molecule. Scattering occurs when light deviates from the path and proceeds in different directions. Rayleigh scattering is when the scattered light holds the original energy, and Raman scattering is when it loses or gains energy. Rayleigh scattering is the most dominant event, while Raman scattering is much less likely to occur. Both events create a virtual state that causes the molecule to distort (polarizes) from its transient status, which is instantaneously absorbed because of the transition from the bottom state to the virtual state, resulting in the creation and scattering of new photons. Raman scattering can be divided into two: Stokes scattering and Anti-Stokes scattering. Stokes Raman scattering occurs when the energy of the scattered light is lower than that of the incident light, and it is more commonly observed than Anti-Strokes Raman scattering. On the other hand, Anti-Stokes Raman is the opposite and less common. Anti-Stokes scattering is a transition from the virtual state to the ground state and involves energy transfer to the scattered light. Raman scattering, such as Stokes scattering, is related to a molecule vibration that changes the polarizability of a chemical bond or functional group. Therefore, Raman spectroscopy measures how much energy light has decreased or increased compared to Rayleigh scattering. The difference is expressed as Raman shift, which means the frequency of the molecule's vibration [29,34].

Raman spectroscopy is also widely used in pharmaceuticals/biological applications such as detecting drugs and distinguishing cancerous [36], and it is also used in aquaculture, such as measuring and predicting the fatty acids of Atlantic salmon [15]. Like NIR spectroscopy, pre-processing is an important part of the analysis of Raman spectroscopy data.

2.2.3. Preprocessing of spectral data

Pre-processing is commonly performed to remove unwanted physical information and noise from the raw spectra, obscuring the actual chemical information. It is carried out before applying the modeling process, and it affects the following analysis process, including the modeling and model assessment, depending on the proper choice of pre-processing. According to Rinnan et al., 2009 [37], the pre-processing techniques have two broad categories: scatter corrections and spectral

derivatives. The former techniques are suggested to decrease the variability due to scattering by adjusting baseline shifts and intensity variations between samples, and the latter is used to avoid noise inflation by applying a derivative or smoothing. In this thesis, pre-processing methods based on scatter correction, including Standard Normal Variate (SNV) and Extended Multiplicative Scatter Correction (EMSC), have been used.

2.2.3.1. Standard Normal Variate (SNV)

SNV is a widely applied method in NIR, and it is designed to reduce the effects of baseline shift and intensity variations between spectra. It is obtained by subtracting the overall mean from the spectrum and dividing the value by the standard deviation. It transforms the original spectrum into an adjusted spectrum to have a similar effect to normalization by making the mean close to zero and the standard deviation close to one. However, as already pointed out by Rinnan et al., 2009 [37] SNV has a disadvantage in that it is sensitive to outliers. Guo et al., 1999 [38] proposed a robust method that uses the percentile median or means as an alternative method to get reasonable results.

2.2.3.2. Extended Multiplicative Scatter Correction (EMSC)

Another spectroscopy preprocessing technique is called Multiplicative Scatter Correction (MSC). This approach aims to remove the undesirable scatter effect in the observed spectra. Extended MSC (EMSC) is an expanded version of MSC to account for additional variability in the spectra. Like MSC, it aims to correct the scattering effect in the measured spectra. However, it is more flexible in dealing with the individual variations of baseline and considers *a priori* knowledge from an analyte of interest [37]. EMSC models are flexible and can be extended by polynomial extensions, configuration spectra, and orthogonal subspaces. EMSC with a polynomial model estimates the parameters, including an arbitrary slope, a quadratic term, or terms of higher polynomial order, and is used to correct the additive or multiplicative effects. The full rank of the spectrum matrix is required to consider the linear independence between spectra in the polynomial EMSC model. The parameter estimation from the polynomial model provides many quantitative features from the spectrum. The polynomial extension model of EMSC helps to get better the correction of spectra and to prevent over-fitting problems in the EMSC modeling. The EMSC model under constituent spectra reduces the disturbance challenge by adding the configuration spectrum around the mean spectrum to the model, which is required to make the least squares estimation of the parameters. The EMSC model by orthogonal subspace extension takes a general approach, like adding a physical model that estimates undesirable features from the spectra [39].

2.3. Sample selection methods – choosing the samples for model training

Sampling is a statistical process to obtain a subset (pre-determined number of observations) from a large population, and it is utilized to estimate an unknown population through a well-representative sample dataset. For prediction purposes, sampling is used as a fundamental step to split data into training and test data sets. It allows us to work with a subset of the dataset instead of the entire dataset. This can be helpful in several ways: for instance, increasing speed and computational efficiency of model building and validation. In this step, we decide how many samples are assigned to training data and select which sampling technique is used. There are several diverse types of sampling methods, including simple random sampling, stratified sampling,

and cluster sampling, among others. There is no golden rule on which sampling method should be used, rather it depends on the requirements and goals of the analysis and the data's characteristics. In this thesis, we chose Random Sampling (RS) and Kennard-Stone (KS; often referred to as Computer-Aided Design of Experiment; CADEX) algorithms as sampling methods to assess the differences between the selected sampling methods.

2.3.1. Random Sampling Selection

Random Sampling (RS, which is often referred to as Simple Random Sampling; SRS) is one of the probability-based sampling methods, and it is the most used sampling technique because of its easier implementation by computer software. It randomly selects the observations without replacement from a finite list of samples, so that each observation has an equal chance of being selected.

2.3.2. Kennard-Stone Sampling Algorithm

The KS algorithm was proposed by Kennard and Stone [40] and it is a commonly used sampling method in the Chemometric fields. It is based on the distance over the X variable space to capture as much diversity in the original data set. To select a representative subset of n samples from a given set of Nsamples ($n < N$), the KS algorithm follows stepwise three procedures; [41]

1. Find two samples that have the farthest distance between data points, assign them to the training dataset
2. Find the next sample that has the longest distance from the already selected samples.
3. Repeat step 2 until it reaches the desired number of samples

By repeating step 2, the training dataset is filled with the minimum distance to all already included samples. A graphical representation of the KS algorithm is depicted in Figure 4, which elucidates the first and second procedures of the algorithm. The classical KS algorithm suggested the Euclidean distance to compute; however, the modified KS algorithm proposed the Mahalanobis distance [42]. The Euclidean distance is a straightforward measure of the distance between two points. The Mahalanobis distance is a measure of distance that considers the covariance between variables. Both are commonly used measures in data science; however, the Euclidean distance is a simpler and more intuitive measure, while the Mahalanobis distance is more appropriate when the variables are not independent.

While RS is based on statistical randomness, the KS sampling algorithm determines the training data set based on the mathematical distance metric between data points. Thus, repeated RS will generate multiple different subsets of data, whilst repeated sampling of KS will generate the sample subset unless additional data is added to the selection pool.

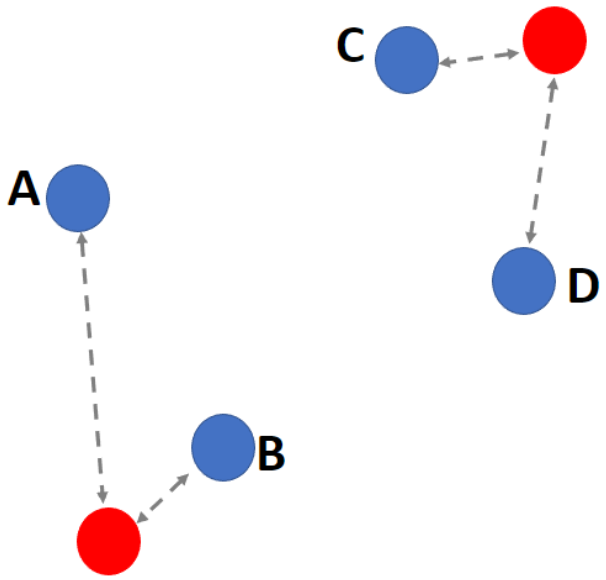


Figure 4. A graphical representation of the KS algorithm. Two red points imply the longest distance among the given data and select as a starting point for the training data set. When comparing the distances of each blue dot and the red dot, A, the farthest distance between the two, is chosen. The process then re-iterates until the required number of subsamples is reached.

2.4. Variable selection – choosing the variables used for training models

Variable selection, also known as feature selection in machine learning, is the process of selecting a subset of related variables (also referred to as features or predictors) for building a statistical model. The general purpose of variable selection is to identify the most relevant and informative variables to predict a target variable and to reduce the number of variables used in the model by removing or dropping less relevant variables. It can be expected to improve model accuracy (when the number of predictors is large) because the model is trained on only the most valuable information by selecting the most relevant variables. Furthermore, models with fewer variables increase interpretability, as there are fewer variables to consider. Lastly, it reduces the overfitting problem by reducing the number of variables used in the model. There are several different techniques for performing variable selection, including filtering methods, wrapper methods, and embedded methods [18]. Here, a Markov Blanket (MB) structure learning algorithm is used as a variable selection method to achieve the minimal subsets as this has recently been shown to perform well in livestock datasets ultimately used for animal breeding purposes.

2.4.1. Markov blanket – Bayesian Network structure

Bayesian network (BN) is a graphical model that can be described as a directed acyclic graph (DAG) $G = (V, E)$, where the nodes $v_i \in V$ correspond to each random variable $X = \{X_1, X_2, \cdots, X_N\}$, and the edges (or arcs) represent the dependent relationship between the linked nodes by conditional probabilities. It needs a few more terminologies to elucidate the BN; directed separation criterion (D-separation), which considers the directionality of arrows in the graph, and an independency map (I-map), which is the probabilistic dependence structure between the random variables in X and the nodes V of G. The property of BN (Markov property) enables factorization that can calculate the joint probability distribution of the random variable X(Global

distribution) as a product of the conditional probability distributions associated with each variable X_i (Local distribution) [20,43]. In other words, the key concept of BN is DAG which represents the probabilistic dependencies split into each variable in a given random variable set X, and it also can be used to infer the causal effect only under certain assumptions [44]. BN may be hard to illustrate as an image if there are many variables, but it enables us to obtain a Markov blanket (MB) that is a minimal set of the given random variables (Figure 5). The MB is the set of node A that consists of the direct parents of A, the children of A, and all the other nodes which share a child with A by following D-separation. In other words, the collection of variables conditioned on which all other variables are probabilistically independent.

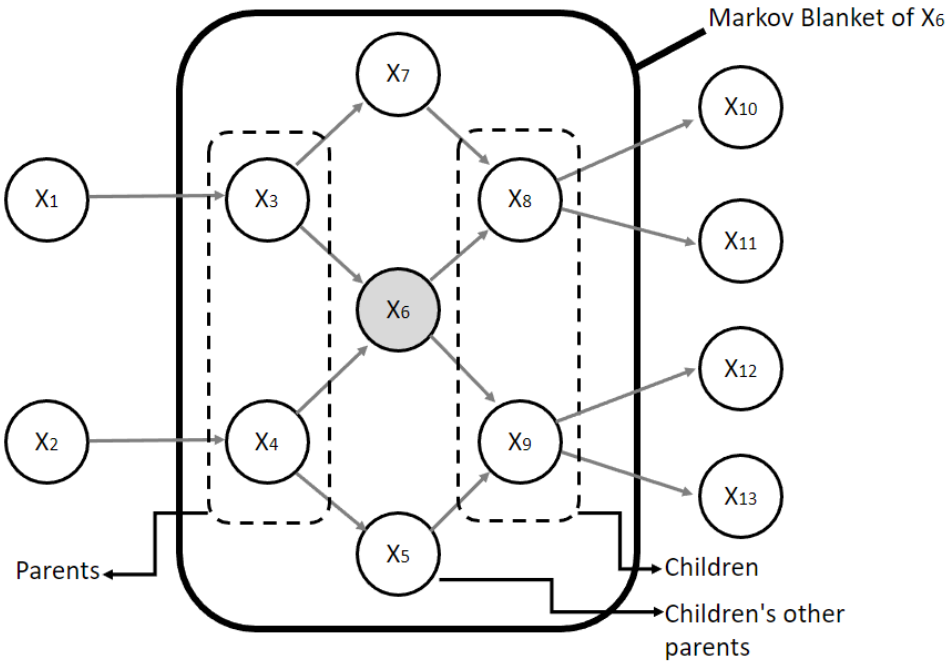


Figure 5. An example of a graphical sketch of MB. The bold line represents the MB of X_6 in a given random variable X_1 to X_{13}, and it consists of its parent's node (X_3, X_4), children's node (X_8, X_9), and children's other parents (X_5, X_7)

2.4.2. Constraint-based structure learning algorithm – Incremental Association Markov Blanket (IAMB)

Constructing a BN structure is called learning, including structure learning and parameter learning. The former aims to identify the graph structure of BN to have the minimal MB of the data, and the latter is for the parameter estimation of the global distribution. There are three broad approaches to finding a BN structure learning, including constraint-based, score-based, and ensemble (hybrid), and several algorithms have been proposed. The Inductive Causation (IC) algorithm proposed by Verma and Pearl in 1991 [45], which offers a theoretical framework for understanding the BN structure, serves as the basis for all constraint-based algorithms; however, it is unfeasible in the real world due to the exponentially increasing computation problem. Several algorithms have been developed to improve this issue, and we will discuss the Incremental Association Markov Blanket (IAMB) algorithm among them.

The IAMB algorithm is a constraint-based structure learning algorithm for BN, and it consists of two phases that have a forward selection followed by a backward selection. It starts from the empty set and adds the nodes (variables) to make the maximum MB set, including the false positives. Then, it removes the unrelated nodes one by one by conditional independence test in the backward phase. It repeats until the BN structure converges to a stable. Since the performance of IAMB relies on the number of conditional independence and its associated calculations, IAMB has $O(MB \times N)$ an average time complexity. It also has $O(N^2)$ in the worst case if the number of MB is equal to N. (See the pseudocode provided in Figure 2 of [46].)

2.5. Partial Least Squares Regression (PLSR)

Partial Least Squares (PLS, often called a projection to latent structures) is a statistical method used for analyzing the relationship between two data sets (X and Y), and it was first designed by the Sweden statistician Herman Wold in the late 1960s. It became popular in chemometrics as a standard analysis tool for chemical analytics because of the works of Svante et al., 2001 [47].

PLS has the advantage of processing data in the following two cases: (1) the wide matrix, where there are a small number of observations (n; rows) and the relatively large number of predictors (p; variables), and (2) multicollinearity of predictors, meaning the presence of high correlations between the predictors. In the first case ($n \ll p$), the problem of having more predictors than degrees of freedom (common in the least squares methods) is overcome by regressing on a number of latent variables for which there are adequate degrees of freedom. However, it can cause an overfitting problem that is not generalized to the new data set when we build the model using high-dimensional data. In this case, it leads to poor model performance and gives consequently inaccurate prediction values than what is expected based on the test data performances. In the second case, multicollinearity, which is due to highly correlated predictors, results in sensitive and inconsistent coefficient estimation in regression. In other words, multicollinearity causes an unstable regression model, so it is difficult to estimate the regression coefficients. Hence, it is one realistic approach to use the PLS method in these cases.

An understanding of principal component analysis and principal component regression is needed to know how PLSR overcomes the two problems which are previously mentioned. PLS solves these problems through a variant of the principal component analysis (PCA), which is the dimensional reduction technique. PCA is an unsupervised method that extracts the latent variables (components) based on the variation that exists in the predictors (X) using singular value decomposition (SVD). The principal components have the favorable property of being orthogonal to each other, and this property is the solution for the multicollinearity problem because orthogonal sets are automatically linearly independent. The number of principal components is the same as the number of predictor variables; however, the proportion of variation they explain in the predictor variables is from the most to the least. This results in another favorable property where a subset of principal components can explain a high percentage of variation in the original predictor variables. These principal components can be used as the variable instead of the original predictor variables and reduce the number of variables to a subset less than n. This approach can be extended to regression called Principal Components Regression (PCR). Similarly, PLS-Regression (PLSR) is a supervised method in machine learning and a type of multivariate regression that makes predictions by extracting the orthogonal factors known as PLS components. Both methods are used for dimension reduction, but different results are obtained because of the different approaches. In terms of PCR, it only takes into account the predictors (X) to perform the

Principal Component Analysis (PCA) and find the maximum variance, fitting the model based on the principal components. However, there is no guarantee that the latent variables explain the target variable Y since it only relies on the given X data. On the other hand, PLSR implements the same by considering the relationship that exists between predictors (X) and response variables (Y).[48]. PLSR decomposed the given X and Ysimultaneously as followed:

$$X = TP^T + \epsilon_1 \quad (9)$$

$$Y = QP^T + \epsilon_2 \quad (10)$$

In the equation (9) and (10), T and U are called the scores matrix, which is orthogonal to each other and contains the latent variables, and P and Q are the loadings (weights). The PLSR model can be simplified the two relations: outer relations between X and Y and inner relations between T and U.

Choosing the appropriate number of components is an essential task for the complexity and accuracy of the PLSR model. It may lead to an underfitting problem with fewer components, while an overfitting problem results from too many components. Cross-validation (see Section 2.6) using the calculation of the prediction residual error sum of squares (PRESS) is one way the selection of model complexity. A minimum PRESS value is calculated differently by the given data set, at least, empirically selecting the number of components that have the minimum PRESS is desirable [49].

2.6. Model Assessment – Metrics and cross-validation

Model evaluation is the last step in the analysis process to assess how generally applicable the fitted model is and to measure the explainability of the model. The most common evaluation metric is R^2 (R-squared, often called the coefficient of determination in Statistics), which shows how much of the variance for a dependent variable in a regression model is explained by one or more independent variables. The mathematical formula is followed as:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \quad (11)$$

It is calculated by subtracting the proportion of the Sum of Squared Errors (SSE) divided by the Sum of Squared Total (SST) from 1. SSE measures the discrepancies between data and predicted data by the model, and SST measures the differences between the data and its mean. Thus, the ratio of SSE and SST has represented the overall variation percentage that cannot be explained by the model. Theoretically, we can say that the model fits well when R^2 is close to 1. But in principle, the R^2 has the possibility to take a negative value when the SSE is greater than SST in Formula 11. It means either the model did not explain enough of the given data, or the data contains the non-linearity characteristic that results in very large errors.

Another widely used evaluation metric is Root Mean Square Error (RMSE), which is the square root of the Mean Square Error (MSE). The MSE measures the average of the differences between the actual data and the predicted data by the linear model, and it is often used in machine learning as a cost function for model performance. It is easy to apply and implement, while it takes the

disadvantage of being affected by outliers. The RMSE alleviates this disadvantage by taking the root.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N_P}(\hat{y}_i - y_i)^2}{N_P}} \quad (12)$$

, where $\hat{y}_i$ are the predicted values, y_i represents the actual value of individual fatty acids for the number of N_P samples.

Cross-validation is a statistical method used to estimate the prediction error (e.g., R^2 and or RMSE), and it is commonly used in machine learning to compare and select the best model because it is easy to implement and understand. In the case of PLSR and many other methods, increasing the number of latent variables will typically result in an increase in R^2 and a decrease in RMSE. However, when such models are tested with independent datasets, they usually show a drastic drop in performance and generalization. To combat this, cross-validation is used to determine model performance in a proportion of the data that has not been used to train the model. The k-fold cross-validation is one type of cross-validation test that randomly divides the data into k groups (folds) of almost equal size. The first fold is assigned as a validation set, and the rest (k-1 groups) parts are used to build the model in a test set. The MSE is then calculated on the first fold. This process is repeated k times until all k-folds are treated as a validation set, it results in the k number of MSE, and k-fold cross-validation is calculated by taking the average of these (Figure 6). The k in the k-fold cross-validation uses empirically 5 to 10 to take advantage of computational time, but there is no golden rule. When k is equal to the sample size, then it is known as leave-one-out cross-validation [50].

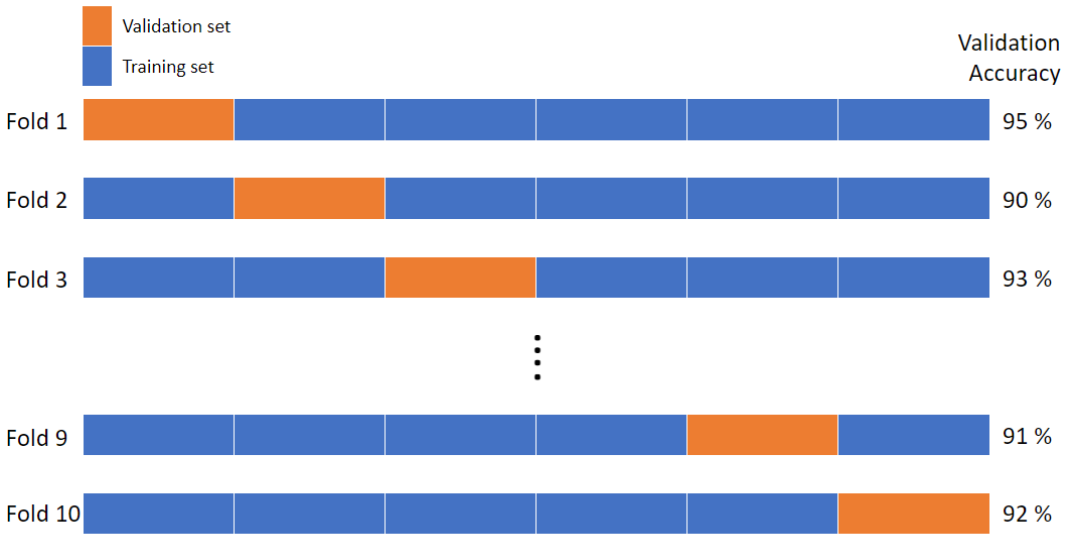


Figure 6. A graphical representation of k-fold cross-validation. Each fold selected a different validation set and calculated the validation accuracy. MSE by k-fold cross-validation calculated the average validation accuracy of each fold.

3. Materials and Methods

3.1. Sample data acquisition

The processing of biological salmon samples in the current dataset has been described previously [15,51,52]. Briefly, the salmon originated from Benchmark Genetics AS as part of their commercial breeding practice and comprised 194 full-sibling families, which means the same set of biological parents, from 92 sires (father; male) and 194 dams (mother; female). The pedigree dataset contains all genetic relationships traced back to a total of 5 generations of information (F0 to F4) and includes a total of 1685 individuals. All fish were transferred to the sea when they reached an average body weight of 113.1g and were reared under the same conditions until they reached an average body weight of 3605g. They were fed a diet with high fish oil content and fasted about 2 weeks before slaughter. After slaughter, a total of 668 sample data were obtained by recording characteristics such as weight (g), length (cm), age, and sex. Salmon muscle fillet was followed by the Norwegian Quality Cut (NQC) standards and was stored in freezing conditions of -20 degrees after harvesting. Total lipids were extracted from muscle tissues following the method described by Folch et al., 1957 [53], and the fatty acids composition of the total lipids was analyzed as the method described by Mason and Waller [54]. Each fatty acid was expressed as a percentage of total fatty acid contents; for the purpose of this thesis, we only further considered three Omega-3 fatty acids; ALA, EPA, and DHA.

3.2. Acquisition of spectroscopy measurements

Both spectroscopic techniques were measured on ground salmon samples taken from the Atlantic salmon fillets. NIR spectra were measured using FOSS NIR Systems XDS Rapid ContentTM Analyzer (FOSS Analytical A/S, Hillerød, Denmark) with a range from 400 nm to 2500 nm in 0.5 nm increments. The NIR spectra were converted to absorbance (A) units using the reflectance (R); A = log10(1/R). The NIR spectra selected in a range from 1150 nm to 2500 nm were used because they contained the relative Omega-3 fatty acid information [12], and SNV was applied as a pre-processed method for a baseline correction in a range from 1150 nm to 2450 nm. All the replicated measurements of NIR spectra were averaged for further analysis.

Raman spectra were measured using a Kaiser Raman-RXN2™ Multi-channel Raman analyzer (Kaiser Optical Systems, Inc., Ann Arbor, MI, USA) with a range from 300 cm^{-1} to 1890 cm^{-1} in 0.3 increments. Raman spectra were selected between 500 cm^{-1} and 1800 cm^{-1}, and EMSC with a sixth-order polynomial extension was used as a pre-processing method for scatter correction. A sixth-order polynomial extension showed the best performance and it related to constructing a more parsimonious prediction model by choosing fewer components to reach minimum RMSE cross-validation (RMSECV) [55]. All the triplicated Raman spectra were averaged to the nearest whole number for further analysis. In short, SNV was applied to NIR spectra for the pre-processed method, while EMSC with a sixth-order polynomial extension was applied to Raman spectra [15,52]. After calculating the average of each spectrum wavelength, both have 1300 equal numbers of spectral wavelengths (Raman shift) and are used as predictors (variables, features) in regression models.

3.3 Experimental design and data partitioning

A total of 613 data were used in the subsequent analysis after all necessary data were merged and cleaned. All analyses were performed with R version 4.1.2 (R core team 2021), which is an extensively used statistical software. As described in Section 2.2, splitting the sample into training and testing datasets is a base stage for building models and validation. In data science, the data can be divided into two parts: training and test data; the training data set is empirically allocated from 60% up to 80% of sample data, while the rest is automatically assigned as a test data set. In this thesis, 100 samples (16%) are assigned as a training data set by both RS and KS, which is a sampling selection method, while the test data set has 513 observations (84%). This particular approach of allocating a relatively smaller number of training data samples was employed for two reasons. Firstly, in practice, due to the high cost of reference method analysis and logistical constraints, only 20-100 samples have concurrent spectral measurements in practice, and reference measurements are used for model building and validation. Secondly, the genetic models require adequately large data (10^2-10^3) to converge, and we wished to estimate quantitative genetic parameters on data not used for model building. The reasoning was to ensure that quantitative genetic analyses avoid overfitting problems and overly optimistic results. According to Difford et al., 2021 [52], in commercial and research of Atlantic salmon, the number of fish with spectroscopic measurements can vary from 19 to more than two thousand samples. Afseth et al., 2006 [56], obtained and utilized 50 salmon samples of Atlantic salmon for the potential of Raman spectroscopy using PLSR. As an extra quality control step, two further steps were conducted; 1) the model performance varied the sample sizes from 50 to 500 where run, and samples sizes above 100 showed diminishing returns (results presented in Appendix A as this was outside of the scope of this thesis) and 2) the model performance using the full data (n = 613) showed similar results by using fewer number of components than published studies (results presented in Appendix B).

The KS algorithm implementation in R software used the kenStone function of the prospectr package [41], and it can choose either the Euclidean or the Mahalanobis distance to compute the distance between two data. The Mahalanobis distance was selected as a distance metric because each wavelength is highly correlated to the others. RS sampling method used the sample function in the base package, which is a built-in function in R, and it was performed without replacement to keep the training data size consistent and prevent observations from being selected more than once. To overcome the small training sample size and to ensure concrete results by comparison of sample selection methods, 10-fold cross-validation was applied. In addition, since RS generates a different 100 samples and KS always results in the identical 100 samples, it was decided to run the models with 10 repetitions of RS to approximate the variation in R^2 and RMSE to be expected in practice when using the RS approach. Effectively 10-fold cross-validation is conducted within each RS repetition (See in Figure 7).

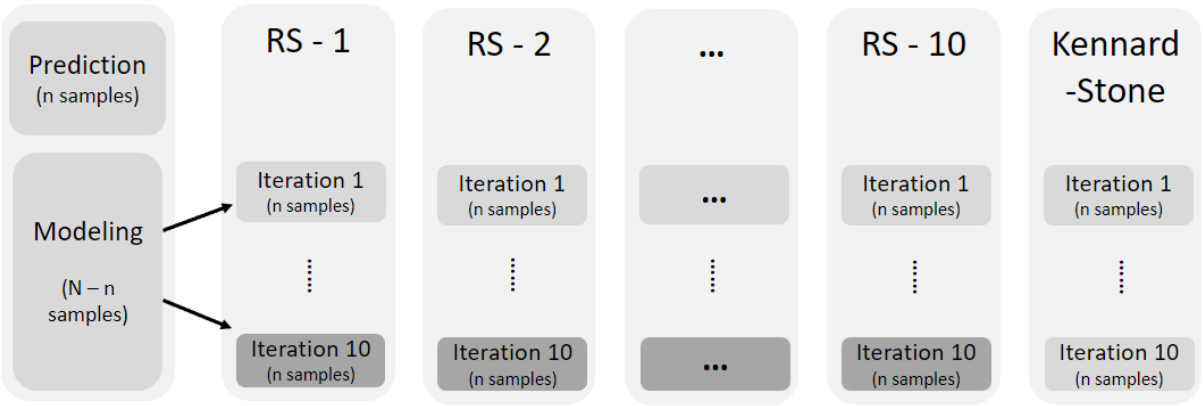


Figure 7. Graphical representation of 10-fold cross-validation by the sample selection methods. RS iterated 10 times enables to get the range of the PLSR model accuracy, while KS operated only one time due to the designed algorithm.

The MB approach was implemented as a variable selection method to get the minimal subset of variables from vibrational spectroscopy data. Several algorithms have been developed to implement the BN learning structure, but the IAMB algorithm was performed for detecting MB in the given vibrational spectroscopy data because it shows good performance in terms of computation time and relatively high accuracy [57]. The learn.mb function of the bnlearn package in R [58] was used to detect the MB of each fatty acid using all wavelengths in both spectroscopy (n = 1300).

Therefore, four scenarios were defined for each of the vibrational spectra types based on different combinations of sample selections and variable selections;

- RS using full variable (without variable selection)
- RS using MB
- KS using full variable (without variable selection)
- KS using MB

These scenarios, illustrated by different colors in Figure 8, included the 100 training samples that were selected based on RS or KS, and vibrational spectra were full (e.g., 1300 variables) or a reduced subset using MB.

PLSR was built to predict each of the respective Omega-3 fatty acids of interest (EPA, DHA & ALA) using either NIR or Raman spectroscopy under the four scenarios, and both spectroscopy data were preprocessed identically to have the same number of variables before regression analysis. The fitted model was constructed by using the plsr function in R pls package with a maximum of 15 principal components. The appropriate number of components was chosen based on the smallest RMSEP (Root Mean Squared Error of Prediction) values within the validation set to avoid the over-or underfitting problem in the PLRS model. The Kernel algorithm was chosen instead of the NIPALS algorithm, which is the classic orthogonal scores algorithm as default in the plsr function, because it yields the same results as NIPALS, and the kernel algorithm is fast for most cases as well [59].

Based on the findings from the PLSR models tested, it was decided that the genetic evaluations would be conducted only on KS scenarios, as these results are inherently more stable and tended to result in the lowest RMSE. As a result, 12 different models (3 fatty acids x 2 vibration spectra

x 2 scenarios based on the variable selection) were used to predict the respective fatty acid content in the remaining 513 samples that were not used in the model training. The Rdmu package in R was used to interface with the DMU version 6, which was developed for analyzing multivariate mixed models in quantitative genetics. Quantitative genetic parameters were estimated by employing the Rdmuai function using multivariate mixed models in DMU version 6 [60]. Using equation (6) above, where y indicates the individual phenotype vectors (reference phenotypes for EPA, DHA, and ALA, or predicted phenotypes for EPA, DHA, and ALA), X is the fixed effect matrix, which only included sex (three levels; male, female, and unknown), b is the coefficient vector corresponding to the X, and Z is the random effect matrix of genetic effects using the pedigree-derived numerator relationship matrix A, a is the vector of estimated breeding values $a \sim N(0, A_\sigma)$, and e is the residuals under a normal distribution. Bivariate models of equation (6) were run between paired reference and predicted phenotypes (e.g., reference EPA and predicted EPA). The DMU output files then contained the phenotypic and genotypic correlations with their associated standard errors and heritability estimates.

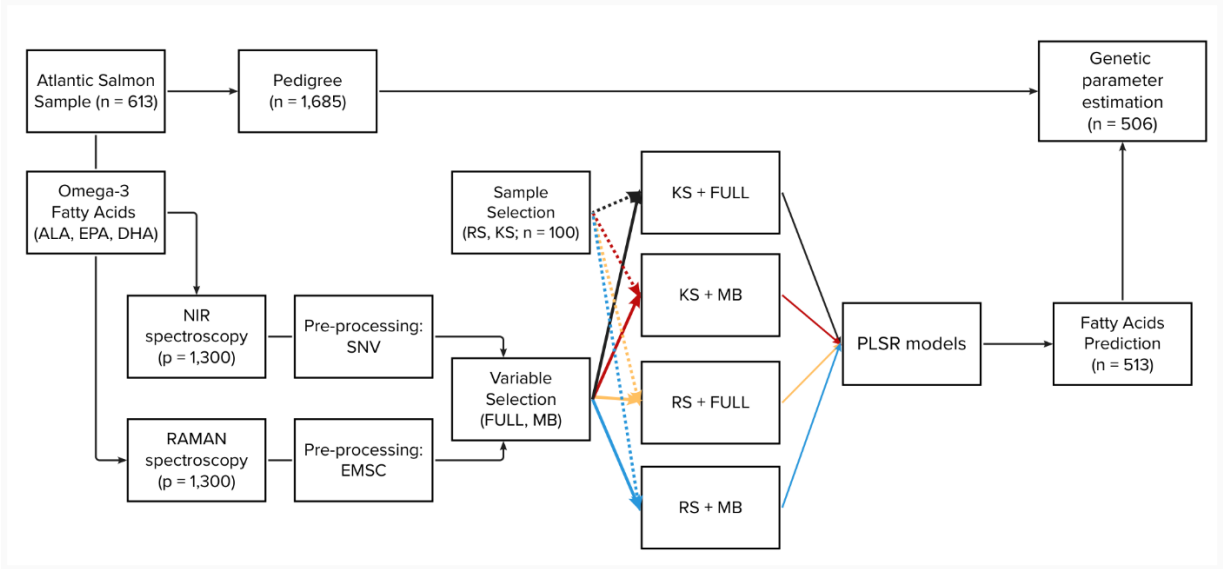


Figure 8. A graphical description of the methodology of this thesis for Atlantic salmon. Four different colors indicate different scenarios by sample selection and variable selection methods.

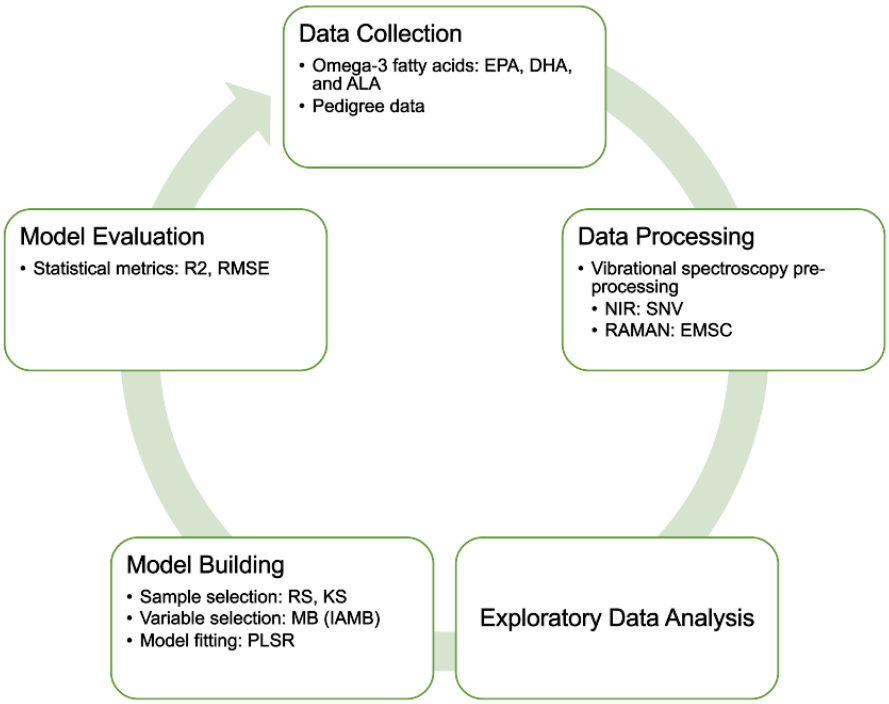


Figure 9. The summary of the data analysis process for this thesis according to Figure 1**.**

In short, Figures 8 and 9 illustrate the graphical description of the methodology flows of this thesis and a graphical summary of the data analysis process, which correspond to Figure 1. The four distinguished colors in Figure 8 suggest the different scenarios by the combination of sample selection and variable selection. The twelve respective PLSR models are compared in the following result section, encompassing the individual fatty acids, vibrational spectroscopy, and sample and variable selection methods.

4. Results

Table 1 shows the descriptive statistics of Omega-3 fatty acids using the reference data, which contained 613 samples. DHA has the highest percentage of fatty acids among the total Omega-3 fatty acids, whilst ALA was the lowest in value. The proportional content of DHA has a wide range (4.34-9.04), while ALA has a narrow range (2.86-4.35).

Table 1. Descriptive statistics of proportional content (%) of Omega-3 fatty acids

Fatty Acids[1]	Mean	Min	Max	SD[2]
EPA	5.137	2.57	6.47	0.462
DHA	6.758	4.34	9.04	0.523
ALA	3.447	2.86	4.35	0.226
Total Sum	15.342	12.33	17.94	0.702

[1] EPA = (eicosatetraenoic acid, C20-5n3), DHA = (docosahexaenoic acid, C22-6n3), and ALA = (α-linolenic acid, 18:3n-3)
[2] SD = standard deviation

4.1 NIR and Raman Spectra and IAMB selection

The NIR and Raman spectra for all samples are visually described in Figure 10 and Figure 11. As described in Section 2.1 and Section 2.2, each spectroscopy has a unique feature to identify chemical compounds as a fingerprint. In the literature [12], the NIR spectra characterized wider peaks than the Raman spectra due to the water absorption, which are found at 1190, 1930 nm (O – H stretch; the molecule vibration motion of the oxygen and hydrogen), and 1450 nm (O – H stretch overtones). The molecular vibration of the first overtones (v = 0 to v = 2) are located around 1715 and 1760 nm, and the second overtones (v = 0 to v = 3) are located at about 1200 nm. On the other hand, the Raman spectroscopy showed very sharp peaks, where the first two highest peaks are located around 1665 (C = C stretch; which refers to the carbon-carbon double bonds) and 1440 (CH_2 scissoring; one of bending vibrations of CH_2 groups) cm^{-1}, and the second highest peaks are found around 1266 (C – H rock; the molecule rocking motion of the carbon-hydrogen) cm^{-1} and 1300 (CH_2 twist; one of bending vibrations of CH_2 groups) cm^{-1} [15,61].

The selected wavelengths for each Omega-3 fatty acid of interest (EPA, DHA, and ALA) using the MB in NIR spectroscopy (Figure 10) and Raman spectroscopy (Figure 11) are contrasted using color-coded vertical dotted lines. All the list of selected wavelengths (Raman shifts) by MB are in Appendix C. The IAMB algorithm used to get the optimal sets for each fatty acid performed well, but it only returned two wavelengths (at 1450 and 2427 nm) for EPA in NIR spectroscopy. Hence, this particular scenario could not be applied to the PLSR model for EPA estimation. In the case of NIR and ALA, and DHA, thirty-seven wavelengths were selected from the total 1300 wavelengths, of which three wavelengths (1665, 1751, and 1762 nm) overlapped.

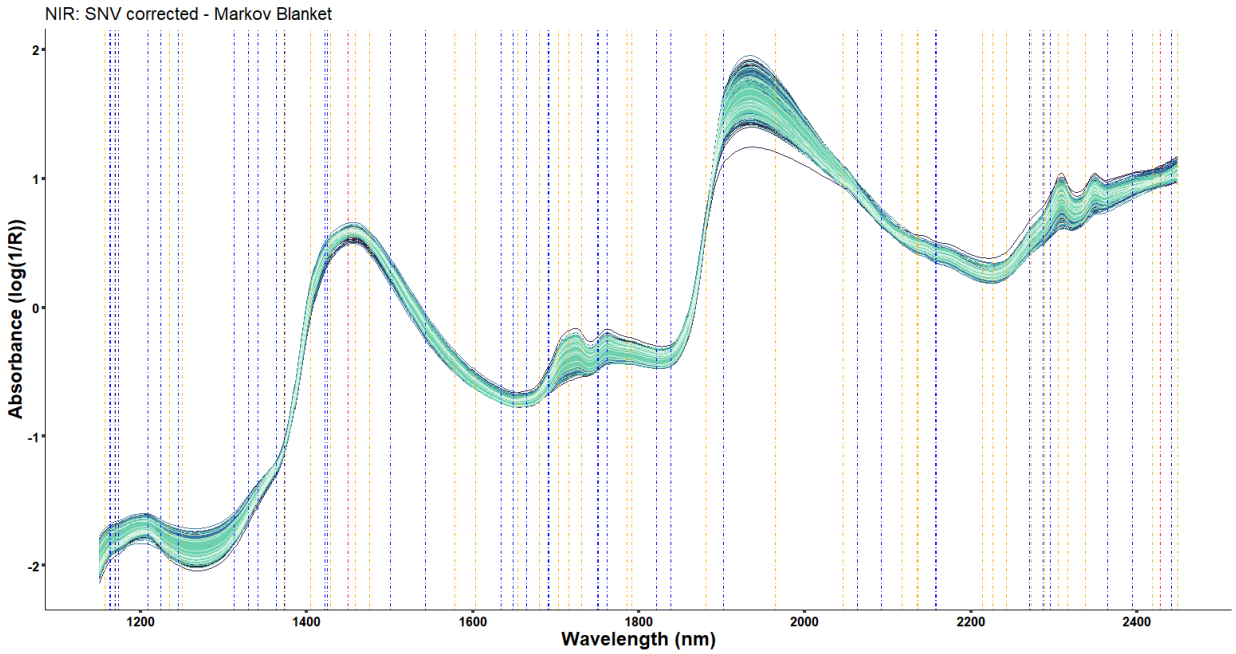


Figure 10. A solid vertical line indicates the selected wavelengths using the MB in NIR spectroscopy for each fatty acid; eicosatetraenoic acid (EPA; Red dotted line), docosahexaenoic acid (DHA; Blue dotted line), and α-linolenic acid (ALA; Orange dotted line).

In the case of Raman, the IAMB algorithm selected seventy-four Raman shifts for EPA, thirty-three for DHA, and seventy for ALA. There are no common variables across Raman spectroscopy for all fatty acids. Interestingly, each pair has several variables in common; five Raman shifts (1371, 1453, 1469, 1610, and 1692 cm^{-1}) between EPA and DHA, four shifts (822, 835, 1648, and 1679 cm^{-1}) between EPA and ALA, and two shifts (929 and 1666 cm^{-1}) between DHA and ALA. Figure 11 shows several unselected ranges for all fatty acids, including the range of 770-821 cm^{-1}.

Some of the wavelengths (Raman shifts) selected by MB are exact matches or very close to the fingerprints in the literature. For example, the MB picked at 1450 nm in EPA, around 1200 nm, and 1750 nm in DHA and ALA in NIR spectra, while it selected at exactly 1665 cm^{-1} in EPA, 1666 cm^{-1} in DHA and ALA, which is close to 1665 cm^{-1}, and some wavelengths (Raman shifts) around 1400 and 1300 cm^{-1} across Omega-3 fatty acids in Raman spectra.

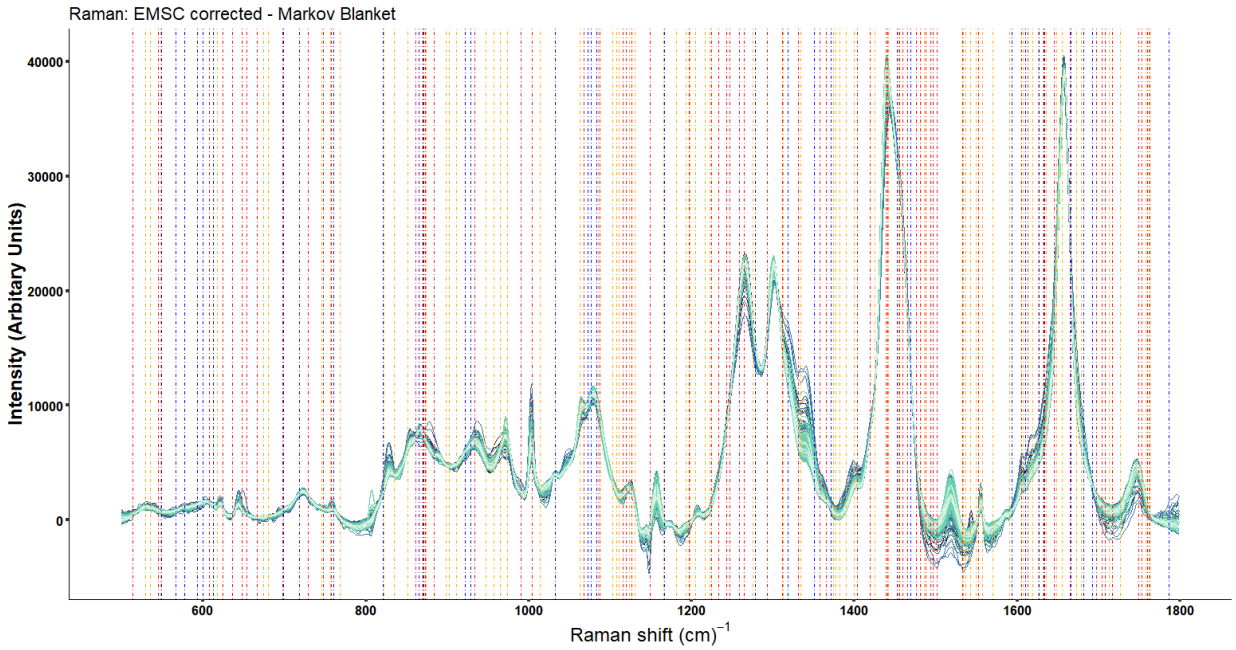


Figure 11. A solid vertical line indicates the selected wavelengths using the MB in RAMAN spectroscopy for each fatty acid; eicosatetraenoic acid (EPA; Red line), docosahexaenoic acid (DHA; Blue line), and α-linolenic acid (ALA; Orange line)

4.2. Prediction of Omega-3 fatty acids across scenarios

The predictive performance of PLSR models across all four scenarios contrasting different combinations of RS and KS with or without MB variable selection for NIR and Raman are presented in Table 2 for comparison. To ensure robust results under RS scenarios, 10 repetitions were conducted to approximate the full variability RS can result in. Note that KS chose identical samples due to the algorithms, while RS selected non-identical training data samples for each iteration. In the case of different RS iterations, the optimal PLSR model can have different numbers of components, as such the range of components is included in Table 2. Additionally, a further PLSR result for the entire dataset using both spectroscopy is in Appendix B, in which the overall results showed a similar or slightly less model accuracy but used less number of components than Afseth et al., 2022 [15].

The overall results in Table 2 show that the R^2 and RMSE varied according to the fatty acids, vibrational spectroscopy (NIR and Raman spectroscopy), sample selection (KS and RS), and variable selection methods (with or without MB adoption). In general, KS showed the best affinity in NIR spectroscopy rather than Raman spectroscopy, while RS showed better performance in Raman spectroscopy. Raman spectroscopy showed robustness rather than NIR spectroscopy. The model performance is overall improved with the MB adoption rather than using all 1300 variables. Graphical results are presented in Figure 12 for EPA and NIR where MB could not be utilized and in Figure 13 for NIR with ALA and DHA. Figure 14 graphically presents all three fatty acids of interest with Raman spectra.

Table 2. PLSR results of ten times repetition for each fatty acid across scenarios of sample and variable selection for NIR and Raman spectroscopy

Fatty Acids[1]	Spectroscopy[2]	Sample Selection[3]	Variable Selection[4]	Variable Number	Component Number[5]	Mean R^2 (Min-Max)	Mean RMSE (Min-Max)
EPA	NIR	RS	FULL	1300	12(14)	0.5 (0.38-0.58)	0.33 (0.3-0.39)
			MB	2	-	-	-
		KS	FULL	1300	14	0.57	0.31
			MB	2	-	-	-
	Raman	RS	FULL	1300	4(8)	0.66 (0.54-0.73)	0.27 (0.24-0.32)
			MB	74	4(9)	0.71 (0.61-0.8)	0.25 (0.22-0.29)
		KS	FULL	1300	11	0.59	0.25
			MB	74	15	0.64	0.23
DHA	NIR	RS	FULL	1300	11(13)	0.51 (0.38-0.59)	0.38 (0.34-0.44)
			MB	37	11(15)	0.55 (0.43-0.62)	0.36 (0.32-0.41)
		KS	FULL	1300	12	0.56	0.34
			MB	37	11	0.67	0.36
	Raman	RS	FULL	1300	5(9)	0.62 (0.54-0.69)	0.33 (0.3-0.38)
			MB	33	5(7)	0.69 (0.56-0.74)	0.3 (0.27-0.36)
		KS	FULL	1300	5	0.61	0.29
			MB	33	7	0.67	0.27
ALA	NIR	RS	FULL	1300	12(15)	0.46 (0.22-0.56)	0.17 (0.16-0.22)
			MB	37	13(15)	0.58 (0.42-0.69)	0.15 (0.13-0.17)
		KS	FULL	1300	15	0.54	0.16
			MB	37	15	0.65	0.15
	Raman	RS	FULL	1300	6(10)	0.62 (0.47-0.69)	0.14 (0.13-0.17)
			MB	70	6(11)	0.73 (0.6-0.8)	0.12 (0.1-0.15)

		KS	FULL	1300	9	0.53	0.13
			MB	70	10	0.65	0.11

[1] EPA = (eicosatetraenoic acid, C20-5n3), DHA = (docosahexaenoic acid, C22-6n3) and ALA = (α-linolenic acid, 18:3n-3), [2] NIR = Near-Infrared Spectroscopy, Raman = Raman Spectroscopy,
[3] RS = Random Sample, KS = Kennard Stone, [4] FULL = complete spectra, MB = Markov Blanket,
[5] Range of Number of Components = first interquartile range to the third interquartile range because of the variety of ranges based on the RS.

The PLSR models using NIR spectroscopy yielded lower estimates in RS scenarios; EPA has an average R^2 of 0.5 (in a range of 0.38-0.58) across components 12 to 14, DHA has an R^2 of 0.51 (in a range of 0.38-0.59) across components 11 to 13, and ALA has an R^2 of 0.46 (in a range of 0.22-0.56) across components 12 to 15. The identical model but using the KS method provided a better estimation for each fatty acid (EPA, DHA, and ALA) with estimates R^2 around 0.55 (0.57, 0.56, and 0.54) when the model used 14, 12, and 15 components, respectively. Furthermore, the R^2 is generally increased when the MB is applied to the PLSR models. The MB opted for an identical number of wavelengths for each vibrational spectroscopy depending on the individual fatty acid, and the PLSR models were constructed using different numbers of latent variables according to the spectroscopy data, sample selections, and variable selection. The MB is not employed in the case of the EPA in the PLSR models using NIR spectroscopy because only two wavelengths are selected in the IAMB algorithm. For DHA estimation, the PLSR model using NIR spectroscopy combined with RS and MB shows an average R^2 of 0.55 (in a range of 0.43-0.62) across the number of components 11 to 15 among 37 wavelengths. ALA estimation shows an average R^2 of 0.58 (in a range of 0.42 to 0.69) across the number of components 13 to 15 among 37 wavelengths. Besides, the R^2 showed the best performance when the MB is combined with the KS in the PLSR model; DHA estimates an average R^2 of 0.67 with 11 latent variables among 37 selected wavelengths, while ALA estimates an average R^2 of 0.65 using 15 components within 37 selected wavelengths. Note that the MB chose the same total number of wavelengths in DHA and ALA, but they are mostly different wavelengths.

The PLSR models using Raman spectroscopy yielded better estimates than those using NIR spectroscopy, regardless of employing sample selection and variable selection. It built the models using relatively fewer components than using NIR spectroscopy. The PLSR models using Raman spectroscopy in RS provide an average R^2 of 0.66 (in a range of 0.54-0.73) by components between 4 and 8 for EPA, while it generates an average R^2 of 0.62 (in a range of 0.54-0.69) using between 5 and 9 latent variables for DHA, and an average R^2 of 0.62 (in a range of 0.47-0.69) using 6 to 10 components for ALA. Interestingly, the model used KS yielded a lower prediction range than RS. The PLSR models constructed with Raman spectroscopy using KS provide an average R^2 of 0.59 with 11 components for EPA, an average R^2 of 0.61 with 5 components for DHA, and an average R^2 of 0.53 with 9 components for ALA. As with the PLSR model using NIR spectroscopy, employing the MB also helped to increase the R^2 no matter which sample selection methods were used. The PLSR models constructed with Raman spectroscopy using RS adding the MB as a variable selection provide an average R^2 of 0.71 (in a range of 0.61-0.8) using between 4 and 9 components for EPA in 74 selected wavelengths through MB, while it provides an average R^2 of 0.69 (in a range of 0.56-0.74) using 5 to 7 latent variables at 33 selected wavelengths for DHA, and an average R^2 of 0.73 (in a range of 0.6-0.8) using 6 to 11 latent variables at 70 selected wavelengths for ALA. The PLSR models constructed with Raman spectroscopy using KS adding the MB generates an average R^2 of 0.64 using 15 components for

EPA, an average R^2 of 0.67 using 7 latent variables for DHA, and an average R^2 of 0.65 using 10 latent variables for ALA.

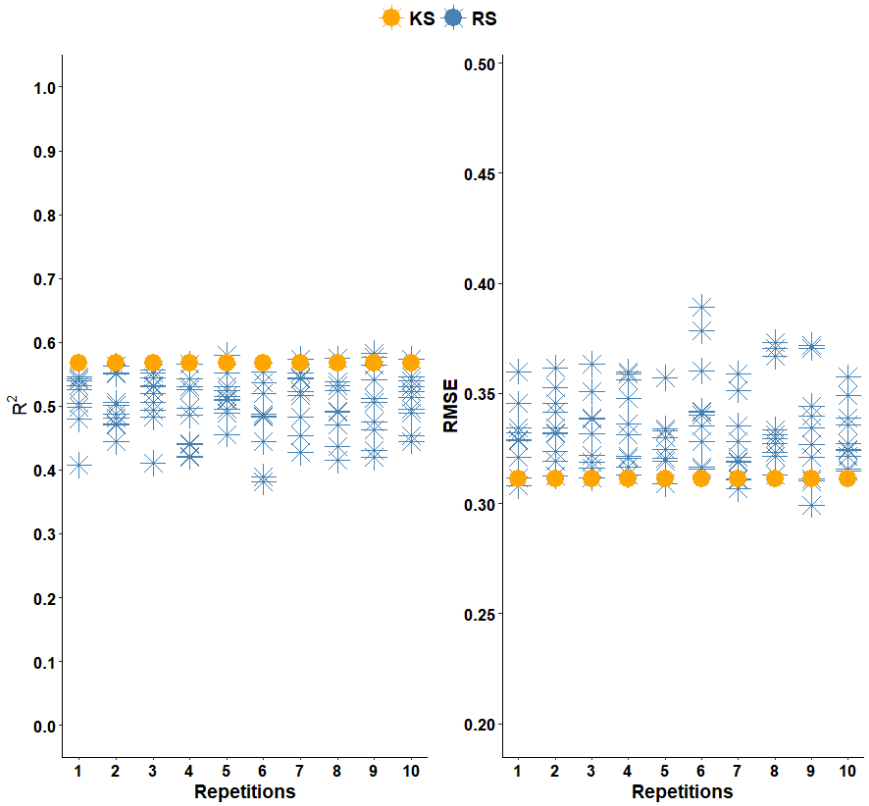


Figure 12. Comparison of KS and RS repetitions of EPA by the PLSR model using NIR spectroscopy. The left side of the figure shows R^2 of ten times repetitions by the sampling method and the right side shows RMSE. In each repetition, the orange dot represents the value obtained by the fitted model on the dataset partition following the KS, and the blue asterisks denote the corresponding values on the n = 100 dataset partitions provided by the RS method.

Figure 12 shows a range of each model assessment value (R^2 and RMSE) of EPA by the fitted PLSR model using NIR spectroscopy depending on the selected sample by the RS or KS. In Figure 12, R^2 has the lowest value (R^2 = 0.38) in repetition 6 of the RS, which shows poor performance even compared with KS. The KS method gives identical R^2 values (R^2 = 0.57) regardless of the iteration and enables us to obtain similar or better R^2 values than the RS method, and the RMSE values also consistently show a similar pattern. Figure 13 and Figure 14 describe the ten times iteration results of the PLSR using the vibrational spectroscopy combined with the sample selection methods and variable selection. Figure 13-A and Figure 13-B show the comparison results for DHA of the PLSR models using NIR spectroscopy based on the sample selection methods, while Figure 13-C and Figure 13-D show the comparison results for ALA. One noticeable point is that sample selection methods combined with MB showed an improvement in R^2 values rather than without MB adoption, whilst it didn't reduce the RMSE values as much as it did. Figures 14-A, C, and E show the ten times repetition results of R^2 comparison according to the sample selection methods combined with MB for the individual fatty acids of the PLSR models using Raman spectroscopy, and Figures 14-B, D, and F show ten times repetition results of RMSE.

Similarly, the iteration results show that applying MB to the sample selection methods boosts R^2 and reduces the RMSE. An interesting point of Figure 14 is that RS with MB shows the highest range of RS among the other scenarios. Some repetitions in Figure 14-A and C show the highest value (R^2 = 0.8) that can be yielded for the RS combined with MB.

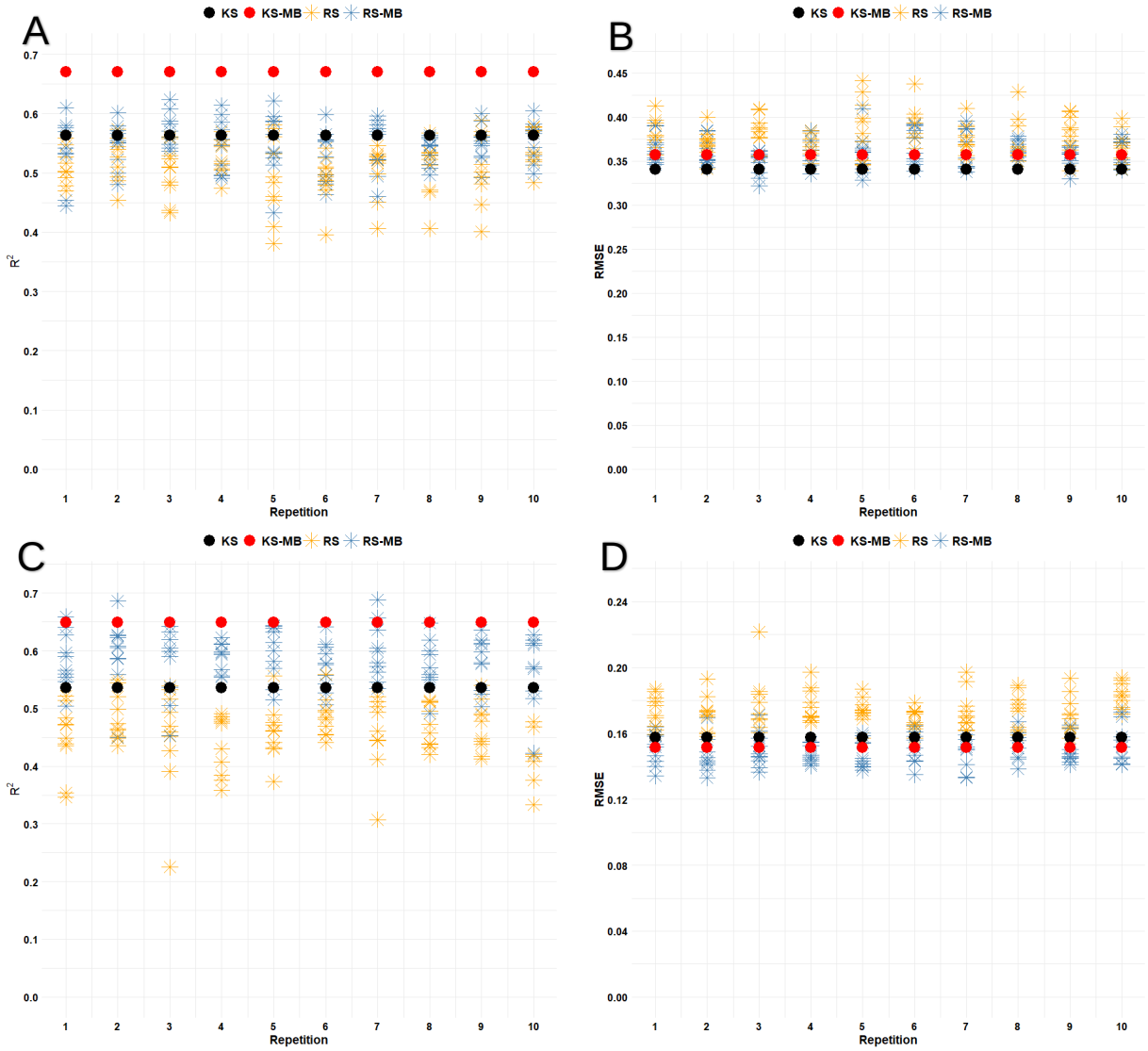
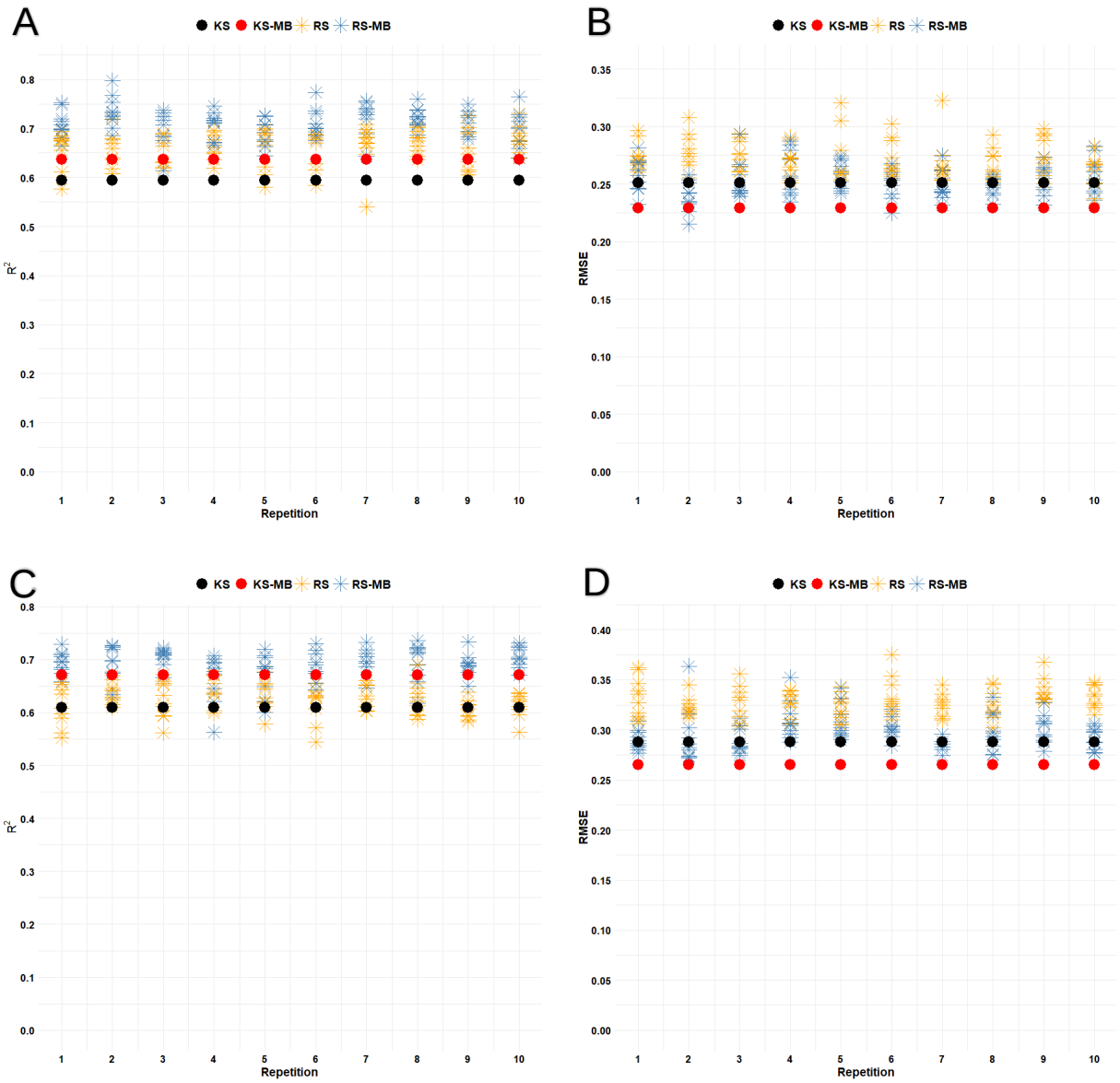


Figure 13. Comparison R^2 and RMSE for DHA (A-B) and ALA (C-D) using NIR spectral according to scenarios combining sample selection and variable selection. Each repetition contains random sampling (RS) or Kennard-Stone sampling (KS) with Markov Blanket MB variable selection or not. Note, MB was not successful for EPA and is not presented (See Figure 12).


A
KS
KS-MB
RS
RS-MB
R^2
Repetition
B
RMSE
Repetition
C
R^2
Repetition
D
RMSE
Repetition

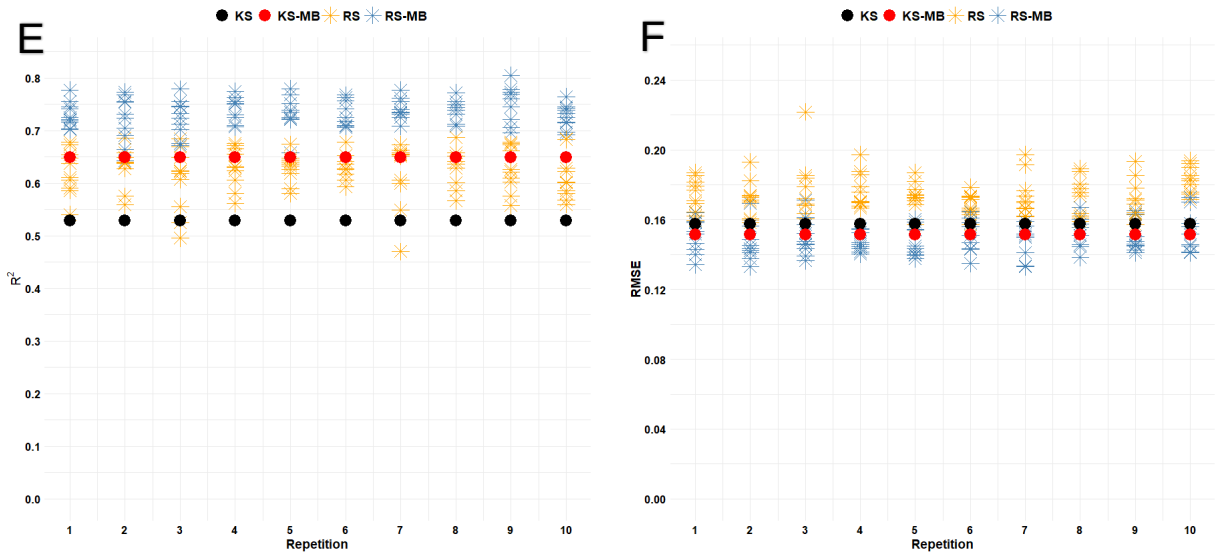


Figure 14. Comparison R2 and RMSE for EPA (A-B), DHA (C-D) and ALA (E-F) using Raman spectral according to scenarios combining sample selection and variable selection. Each repetition contains random sampling (RS) or Kennard-Stone sampling (KS) with Markov Blanket MB variable selection or not. In each iteration, the black dot represents the value from the fitted model on the data set following the KS and the red dot describes the value from the fitted model of the KS combined with MB, while the orange asterisks denote each iteration value corresponding on the n = 100 data set selected by the RS and the blue asterisks describe each fitted values by the RS combined with MB.

4.3. Quantitative Genetic parameters

The quantitative genetic parameters of the predicted EPA, DHA, and ALA from the KS sampling method with or without MB variable selection are presented in Table 3. The genetic parameters for the reference method individual fatty acids and those predicted from the PLSR were generally very similar. For instance, all predicted fatty acids were significantly heritable ranging from h^2 = 0.22 - 0.44 with standard errors in the range of 0.10 - 0.11. Importantly, the heritability estimates are remarkably similar between the reference phenotypes and the predicted phenotypes for the fatty acids, except for ALA using NIR with MB variable selection, which was approximately half (h^2 = 0.22) of the respective reference phenotype heritability (h^2 = 0.49).

The genetic correlations between the reference phenotype and the predicted phenotype for all fatty acids were generally very strong and positive ranging from r_G = 0.69 - 0.99, with standard errors ranging from 0.004 to 0.08. There was a tendency for higher genetic correlations in Raman spectroscopy rather than in NIR spectroscopy and a tendency to yield higher genetic correlations when the MB is applied instead of the full spectral data. The Raman spectroscopy applied to MB for ALA shows the highest genetic correlation (r_G = 0.99). Regardless of the spectroscopy data, the genetic correlation for DHA applied to the MB provides a value greater than 0.9 and, albeit marginally, gives a higher value in Raman spectroscopy (r_G = 0.94) than in NIR spectroscopy (r_G = 0.93). The phenotypic correlation also tends to increase when the MB is applied; however, there are cases in which the phenotypic correlation does not increase even if the MB is applied. The phenotypic correlations reduce by approximately 4% (r_P = 0.76 to r_P = 0.73) for DHA and 3% (r_P = 0.58 to r_P = 0.55) for ALA in NIR spectroscopy.

Table 3. Genetic parameter estimation results by RMDU.

Fatty Acid[1]	Spectra[2]	Variable Selection[3]	Genetic correlation (r_G) ± S.E*.	Phenotypic correlation (r_P) ± S.E.	True phenotype h^2 ± S.E.	Predicted phenotype h^2 ± S.E.
EPA	NIR	FULL	0.69 (0.04)	0.70 (0.11)	0.35 (0.10)	0.41 (0.11)
		MB	-	-	-	
	Raman	FULL	0.84 (0.04)	0.75 (0.08)	0.43 (0.11)	0.38 (0.11)
		MB	0.86 (0.04)	0.79 (0.10)	0.39 (0.11)	0.35 (0.11)
DHA	NIR	FULL	0.87 (0.08)	0.76 (0.16)	0.45 (0.12)	0.41 (0.11)
		MB	0.93 (0.08)	0.73 (0.16)	0.39 (0.11)	0.43 (0.11)
	Raman	FULL	0.83 (0.05)	0.77 (0.12)	0.36 (0.10)	0.39 (0.11)
		MB	0.94 (0.08)	0.83 (0.17)	0.39 (0.11)	0.44 (0.11)
ALA	NIR	FULL	0.66 (0.01)	0.58 (0.02)	0.34 (0.10)	0.35 (0.11)
		MB	0.81 (0.01)	0.55 (0.02)	0.49 (0.12)	0.22 (0.10)
	Raman	FULL	0.75 (0.00)*	0.63 (0.01)	0.35 (0.10)	0.31 (0.10)
		MB	0.99 (0.01)	0.78 (0.10)	0.49 (0.11)	0.36 (0.10)

[1] EPA = (eicosatetraenoic acid, C20-5n3), DHA = (docosahexaenoic acid, C22-6n3) and ALA = (α-linolenic acid, 18:3n-3),
[2] NIR = Near-Infrared Spectroscopy, Raman = Raman Spectroscopy,
[3] MB = Markov Blanket, FULL = complete spectra, *S.E. = standard error, note these are rounded to the nearest digit two points after the decimal point, thus <= 0.004 is set to 0.00. h^2 = heritability.

5. Discussion

In this thesis, we estimated individual Omega-3 fatty acids predicted by vibrational spectroscopy and their heritability based on the estimated predicted phenotypes. We have identified the potential for further research and evaluated the results across all combinations per scenario.

5.1. Small datasets and repetitions in PLSR models

As we mentioned in Section 3.3, the PLSR models were restricted to training on 100 samples which were selected according to the sample selection methods RS and KS. This choice was motivated by mimicking practical circumstances where difficulties in obtaining enough reference data are the reason researchers look into vibrational spectra and PLSR models. The second motivation was to have enough independent data to estimate genetic parameters like heritability and genetic correlations with sufficiently small standard errors. The fitted PLSR model performed well utilizing the small number of data sets; however, the k-fold cross-validations only used randomly selected 10 samples among the training data. Since reasonable doubts may arise as to whether this is a reliable number of training data for fitting the model or whether the cross-validations work well, the model fitting and assessment procedure was repeated ten times using a small number of training data based on the sample and variable selection methods by individual fatty acids. This approach enabled us to see the distribution of R^2 and RMSE by the RS sampling method and helped to find a possible experiment in a given situation. We acknowledge that increasing the training sample sizes will further increase the R^2 and decrease the RMSE of prediction as demonstrated by Afseth et al., 2022 [15], but the current work demonstrated what

can be achieved with training sample sizes more closely aligned to commercial practices Difford et al., 2021 [52] for Atlantic salmon. Interpretation of the PLSR model using the full dataset (Appendix B) was out of scope for this thesis, and it is not recommended to be used to train data on the testing dataset from a Machine Learning perspective due to the objective of testing the model is generally estimating how well predictions performs to unseen data. However, considering the commercial practice, which is hard to obtain enough samples as expected in data science, and comparing to previous results by Afseth et al., 2022 [15], it can be noted that comparable results by using fewer components as a positive control.

5.2. Comparison of the sample selection methods

It is hard to say which sampling method is best because the RS and KS are unique methods; RS chooses randomly based on the uniform distribution, while KS selects a sample according to the distance between the two observations based on the predictor variable space. Repeated random sampling results suggest that the sample selection may lead to contrasting consequences depending on the vibrational spectra data and the fatty acid to be predicted. Most of the estimated R^2 provides better performance when the KS is applied for sample selection. It generates identical R^2 and RMSE values no matter how many iterations due to its algorithm. Some specific cases in each iteration by RS give slightly above or even worse R^2 than estimated by KS with respect to the NIR spectroscopy; For example, at least one of each repetition (5) and (6) for ALA achieves over the R^2 of 0.55, which is slightly above than KS repetitions, while one of the repetitions (3) reaches the lowest R^2, and all of the repetitions (4) show the lower R^2 in RS than KS. It may indicate that KS worked well in NIR spectroscopy for most situations, but RS may give comparable performance by the coincidence of selecting the optimal sample. As Nawar et al., 2018 [16] and Ferreira et al., 2022 [17] reported the KS performed similarly or exceeded the RS practically in NIR data, and our results somewhat supported those points. Interestingly, it shows the opposite results when it comes to Raman spectroscopy. In most cases of EPA and ALA and some in DHA, it attains higher estimated values of R^2 by the RS; as shown in Figure 14, one value in repetition 7 shows a noticeably lower value among the RS iterations using Raman spectroscopy, while the KS shows an invariant value (R^2 = 0.59) over iteration. Although the R^2 ranges of the RS repeats in each fatty acid are different, utilizing RS is more likely to choose a better sample in Raman spectroscopy compared to NIR spectroscopy.

5.3. Variable selection effects of Markov Blanket

When the MB is applied as a variable selection method in addition to the sample selection method, each fatty acid selects a different number of variables (wavelengths, Raman shifts) by spectroscopic data, as shown in Table 2. If the PLSR model is constructed with the selected variables through the MB from the spectroscopic data, it shows an overall pattern of increasing the R^2 and decreasing the RMSE. However, it is not in all cases; the PLSR model could not apply for EPA in NIR spectroscopy because it selected only two wavelengths through the MB, and the PLSR model for DHA in Raman spectroscopy increased the RMSE. Likewise, the PLSR models required similar or more latent variables to build. This may imply that the MB, as a variable selection method, helps to increase prediction accuracy to some extent without affecting PLSR.

In addition, as shown in Figure 10 and Figure 11, it is possible to distinguish which part of each fatty acid is concentrated in the spectroscopy data by different colors, and some regions in vibrational spectroscopy data were not considered as a selected variable at all. Among selected

variables according to the fatty acids and the spectroscopy data, a few common variables exist depending on the combination; In NIR spectroscopy, the three wavelengths (1665, 1751, and 1762 nm) between DHA and ALA were commonly selected by MB, while five Raman shifts (1371, 1454, 1469, 1610, and 1692 cm^{-1}) in EPA and DHA, two (929, 1666 cm^{-1}) in DHA and ALA, and four (822, 835, 1648, 1679 cm^{-1}) between EPA and ALA in Raman spectroscopy. A few wavelengths (Raman shifts) selected by MB were exactly matched or closed with the chemical characteristics of Omega-3 fatty acids, as described in the literature [15,61]; In EPA, 1450 nm in NIR spectra and around 1665 cm^{-1} across fatty acids in Raman spectra. Nonetheless, it is hard to check how these selected variables via MB are related to some extent to the individual fatty acid because the MB can be used for causal analysis only if certain assumptions are met.

In this thesis, the IAMB was employed as a variable selection method to find the best subset because it is reasonably fast, accurate, and widely applicable according to the result from [62]. As discussed earlier, the IAMB worked well on the given data set and had an impact on improving the prediction accuracy in the fitted PLSR model. Apart from that, one limitation of this paper is only one algorithm (IAMB) was tested and implemented for MB as a variable selection. There are several different algorithms to implement MB, but there is no guarantee that the other algorithms perform well on the given data set or return identical variables like the IAMB. Nevertheless, given the overall improvement MB had on PLSR model prediction and the uncertainty as to why MB improves PLSR prediction, it is recommended that other variable selection algorithms are also investigated to shed light on the causes and limitations of variable selection in conjunction with PLSR prediction.

5.4. Comparison of Vibrational spectroscopy methods NIR and Raman

As mentioned in Section 3.3, we set the maximum number of principal components as 15 to construct the PLSR models. Some PLSR models required the highest latent variables, as described in Table 2. One notable thing is the PLSR models easily reach the maximum number of components (e.g.,15 here) when it comes to using NIR spectroscopy whilst Raman spectroscopy had consistently lower numbers. In general, lower numbers of latent variables are favored because it is an indicator that the vibrational spectroscopy method is capturing chemical information from the fatty acid of interest, and it makes the PLSR model interpretability easier. It may raise interest in how many maximum latent variables are required when constructing a PLSR model using NIR spectroscopy. However, it may indicate that NIR spectroscopy has more complexity than Raman spectroscopy in terms of interpretability, which is required more latent variables to construct the PLSR models.

The principle of each spectroscopy is different, but it is widely used to quantify the measurement of chemical components. In this paper, we utilized NIR and Raman spectroscopy to estimate the fatty acid and to compare the prediction ability. As a result, it showed two general points: 1) NIR spectroscopy showed a lower prediction accuracy than Raman spectroscopy regardless of sample and variable selection, and 2) NIR required more components than Raman spectroscopy in the PLSR models for measuring individual fatty acids. This shows similar results to those reported by Afseth et al., 2022 [15] that Raman spectroscopy has more possibility to achieve high and optimal performance. Hence, it indicates that Raman spectroscopy can be more robust and stabilities to measure fatty acid compared to NIR spectroscopy.

One limitation of this thesis is that vibrational spectroscopy is generally measured on one specific surface of the sample, which cannot cover the whole sample and cannot get spatial information. Although our results showed that vibrational spectroscopy is a useful tool to measure and capture the chemical information that we desired, further studies may consider employing hyperspectral images (i.e., high-throughput and spatial information across the fillet) to overcome this drawback, which can be used to maximize the information extraction from the data and can apply more complex models such as neural networks.

5.5. Genetic parameters estimation and validation of predicted phenotypes

Genetic parameters of individual fatty acids were estimated in 506 samples using a training dataset of 100 samples selected solely by KS solely but with or without MB variable selection. This choice was motivated by the stability of KS samples across vibrational spectra types and the Omega-3 fatty acids and the fact that MB improved predictions in NIR and was generally as good as RS in Raman on average. Furthermore, KS is inherently a completely repeatable sample selection method which eliminates the extra variability added when employing RS. The heritability estimates for EPA, DHA, and ALA ranged from (h^2 = 0.34 - 0.49, which is the “True Phenotype” heritability from Table 3) using the reference method, which was far larger than those of Horn et al., 2018, whose estimates ranged from (h^2 = 0.22 - 0.26) on 668 samples, of which our 506 samples are a subset of the research by Horn et al., 2018. A careful review by Horn et al., 2018 found that body weight was used as a covariate, and sex was included as a fixed effect. It is one of the different points since we only considered sex as a fixed effect in all genetic regression models. Hence, it may have reduced their heritability because body weight is correlated with fatty acids, and it is highly hereditary.

According to Equations (6) and (7), heritability is a major factor to be considered in estimating the EBV and the response to selection for a given variable. Higher heritability results in higher response to selection or improvements in the response variables in the next generation, which are the desired goal for breeders. The results from Table 3 showed that heritability tended to increase after adopting the MB as a variable selection method, and the predicted heritability tended to have a higher heritability when the genetic correlation (r_G) and phenotypic correlation (r_P) were close to each other. All genetic and phenotypic correlations achieved strong positive values whether or not the adoption of MB, and the correlations improved when the MB was applied, which is a similar pattern to the results of individual estimation in Omega-3 fatty acids. In particular, the genetic correlations for DHA in the adoption of the MB are over 0.9, and the genetic correlation of ALA reached the highest value (r_G = 0.986) after the adoption of the MB. One potential point of issue is that the MB approach does not always work well to predict fatty acids and their heritability. For example, the predicted heritability for ALA using the MB approach was recorded as the lowest estimation (h^2 = 0.22). It may result from the huge gap between the genetic and phenotypic correlation (r_G = 0.81 and r_P = 0.55) and may lead to undesirable consequences, despite the advantage of MB adoption.

To the best of our knowledge, this is the first time that predicted phenotypes from vibrational spectra have been validated using genetic correlations between the true and predicted phenotypes for Omega-3 fatty acids. Difford et al., 2021 [52], used the current dataset to estimate genetic correlations between true and predicted total fat, and also found a genetic correlation of 0.93 using NIR spectroscopy and 0.98 using Raman spectroscopy, although they did not investigate the use of MB. In the vast majority of cases, geneticists simply estimate the R^2 and or RMSE in small sample sizes and assume that this would result in high genetic correlations. This

is not always the case. For example, Poulsen et al., 2014 [63] predicted fatty acid content in the milk of dairy cows using gas chromatography (GC, which is known as a more accurate than vibrational spectrometry but expensive and time-consuming method) and Fourier-Transform Infrared (FT-IR). The results showed that four out of eight fatty acids had high genetic correlations (> 0.94), two had moderate (0.45-0.52), and one had a highly negative genetic correlation (-0.86) in Stearic Acid (C18:0, which is a saturated fatty acid). If the authors had not conducted a genetic validation and proceeded to select based on spectral predicted phenotypes, Stearic Acid would have resulted in a negative unwanted response to selection in the opposite direction intended as seen by the strong negative genetic correlation (-0.86) between the true and predicted phenotype.

The estimated heritability also tended to have a higher estimate when it was used by Raman spectroscopy. Although using a different species, Blay et al., 2021 [64] estimated the fatty acids and their heritability using Raman spectroscopy in the adipose viscera tissues (organs, not muscle as we have done) from Rainbow trout and reported very low heritability estimates regarding EPA, DHA, and ALA. They trained on ridge regression models using 268 samples from an experimental facility and then predicted phenotypes in a separate population of 1400 samples from commercially selected fish for estimating the genetic variability. Surprisingly, the genetic and phenotypic correlations of Omega-3 fatty acids that we are interested in recorded strongly negative values amongst each other, while their heritability recorded low and varied (ALA= 0.02, EPA= 0.16, and DHA= 0.03, respectively). It is hard to compare directly with the current study as we had different species, different tissues, and different predictive models, but since the studies show such different heritability estimates to our own, it is worth testing our approach in an independent commercial dataset as Blay et al., 2021 [64] have done to ensure our results are reproducible in different populations of Atlantic salmon.

6. Conclusion

This paper showed the optimization possibility by combining appropriate methods to achieve competitive output from the data acquisition to the model evaluation. To estimate the Omega-3 fatty acids and their heritability, we used vibrational spectroscopy combined with sample selection methods and the MB as a variable selection. Although we used a small number of training data to train the PLSR model, both NIR and Raman spectroscopy worked well in general, but Raman showed overall better prediction accuracy and estimation of genetic parameters across the fatty acids. When it comes to choosing the sample selection methods in the given spectroscopic data, NIR spectroscopy achieved higher accuracy in the KS rather than the RS. Raman spectroscopy works well in both sample selection methods since it shows analogous results; however, it could achieve higher accuracy in the RS only if it selected a good representative sample to elucidate the fatty acid. It is recommendable to use KS as sample selection to obtain solid and reproducible results in general, but it needs to be noted that RS also shows superior results by the good representative selection in Raman spectra. The MB performed well in most cases of Omega-3 prediction and heritability estimation, and there is no doubt that it helped to increase the prediction accuracy and reduce the variable. However, which variable selection methods perform the best or to what extent the selected variables are related to the individual fatty acids are not in the scope of this thesis and warrant further investigation. Most importantly, the prediction of Omega-3 fatty acids using Raman spectroscopy in the muscle of Atlantic salmon is highly feasible and has the possibility to break the bottleneck in acquiring suitable numbers of phenotypes for breeding purposes.

References

1. FAO. The State of World Fisheries and Aquaculture 2022. FAO; 2022 Jun.

2. Ahern M, Thilsted S, Oenema S, Kühnhold H. The role of aquatic foods in sustainable healthy diets. UN Nutrition Discussion Paper. 2021;

3. Janssen K, Chavanne H, Berentsen P, Komen H. Impact of selective breeding on European aquaculture. Aquaculture. Elsevier B.V.; 2017;472:8–16.

4. Ytrestøyl T, Aas TS, Åsgård T. Utilisation of feed resources in production of Atlantic salmon (Salmo salar) in Norway. Aquaculture. Elsevier; 2015;448:365–74.

5. Sprague M, Dick JR, Tocher DR. Impact of sustainable feeds on omega-3 long-chain fatty acid levels in farmed Atlantic salmon, 2006-2015. Sci Rep. Nature Publishing Group; 2016;6.

6. Hill WG, Mackay TFC. D. S. Falconer and Introduction to Quantitative Genetics. Genetics. 2004;167:1529–36.

7. Xu S. Quantitative Genetics. Quantitative Genetics. Springer International Publishing; 2022.

8. Trygve G. Selection and breeding programs in aquaculture. Dordrecht: Springer; 2005.

9. Robertson A. The Sampling Variance of the Genetic Correlation Coefficient. Biometrics. 1959;15:469.

10. Chiu HH, Kuo CH. Gas chromatography-mass spectrometry-based analytical strategies for fatty acid analysis in biological samples. J Food Drug Anal. No longer published by Elsevier; 2020;28:60–73.

11. Jens Petter Wold, Tiril Aurora Lintvedt. Fast and effective measurements of fatty acid composition in salmon fillets [Internet]. 2023 [cited 2023 Apr 17]. Available from: https://nofima.com/results/fast-and-effective-measurements-of-fatty-acid-composition-in-salmon-fillets/

12. Afseth NK, Segtnan VH, Marquardt BJ, Wold JP. Raman and Near-Infrared Spectroscopy for Quantification of Fat Composition in a Complex Food Model System. Appl Spectrosc. 2005.

13. Pasquini C. Near Infrared Spectroscopy: Fundamentals, Practical Aspects and Analytical Applications. J. Braz. Chem. Soc. 2003.

14. Beć KB, Huck CW. Breakthrough potential in near-infrared spectroscopy: Spectra simulation. A review of recent developments. Front Chem. Frontiers Media S.A.; 2019.

15. Afseth NK, Dankel K, Andersen PV, Difford GF, Horn SS, Sonesson A, et al. Raman and near Infrared Spectroscopy for Quantification of Fatty Acids in Muscle Tissue—A Salmon Case Study. Foods. MDPI AG; 2022;11:962.

16. Nawar S, Mouazen AM. Optimal sample selection for measurement of soil organic carbon using on-line vis-NIR spectroscopy. Comput Electron Agric. Elsevier B.V.; 2018;151:469–77.

17. Ferreira R de A, Teixeira G, Peternelli LA. Kennard-Stone method outperforms the Random Sampling in the selection of calibration samples in SNPs and NIR data. Ciência Rural. FapUNIFESP (SciELO); 2022;52.

18. Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. J Mach Learn Res. JMLR.org; 2003;3:1157–82.

19. Mehmood T, Liland KH, Snipen L, Sæbø S. A review of variable selection methods in Partial Least Squares Regression. Chemometrics and Intelligent Laboratory Systems. 2012. p. 62–9.

20. Nagarajan R, Scutari M, Lèbre S. UseR ! Bayesian Networks in R with Applications in Systems Biology [Internet]. Available from: http://www.springer.com/series/6991

21. Felipe VPS, Silva MA, Valente BD, Rosa GJM. Using multiple regression, Bayesian networks and artificial neural networks for prediction of total egg production in European quails based on earlier expressed phenotypes. Poult Sci. Oxford University Press; 2014;94:772–80.

22. Dórea JRR, Rosa GJM, Weld KA, Armentano LE. Mining data from milk infrared spectroscopy to improve feed intake predictions in lactating dairy cows. J Dairy Sci. Elsevier Inc.; 2018;101:5878–89.

23. Hutu I, Oldenbroek K, Waaij L. Animal breeding and husbandry. first. Timisoara: Agroprint; 2020.

24. Johannsen W. The genotype conception of heredity1. Int J Epidemiol. Oxford University Press; 2014;43:989–1000.

25. Visscher PM. On the Sampling Variance of Intraclass Correlations and Genetic Correlations. Genetics. 1998;149:1605–14.

26. Coffey M. Dairy cows: in the age of the genotype, #phenotypeisking. Animal Frontiers. 2020;10:19–22.

27. Sultanbawa Y, Smyth HE, Truong K, Chapman J, Cozzolino D. Insights on the role of chemometrics and vibrational spectroscopy in fruit metabolite analysis. Food Chemistry: Molecular Sciences. 2021;3:100033.

28. Cozzolino D. An Overview of the Successful Application of Vibrational Spectroscopy Techniques to Quantify Nutraceuticals in Fruits and Plants. Foods. 2022;11:315.

29. Larkin P. Infrared and raman spectroscopy: principles and spectral interpretation. Elsevier; 2011.

30. Blanco M, Villarroya I. NIR spectroscopy: a rapid-response analytical tool. TrAC Trends in Analytical Chemistry. 2002;21:240–50.

31. Pasquini C. Near infrared spectroscopy: A mature analytical technique with new perspectives – A review. Anal Chim Acta. 2018;1026:8–36.

32. Pizarro C, Esteban-Díez I, González-Sáiz J-M, Forina M. Use of Near-Infrared Spectroscopy and Feature Selection Techniques for Predicting the Caffeine Content and Roasting Color in Roasted Coffees. J Agric Food Chem. 2007;55:7477–88.

33. Cattaneo TMP, Holroyd SE. The Use of near Infrared Spectroscopy for Determination of Adulteration and Contamination in Milk and Milk Powder: Updating Knowledge. J Near Infrared Spectrosc. 2013;21:341–9.

34. Smith E, Dent G. Modern Raman Spectroscopy-A Practical Approach. John Wiley & Sons, Inc.; 2005.

35. Orlando A, Franceschini F, Muscas C, Pidkova S, Bartoli M, Rovere M, et al. A Comprehensive Review on Raman Spectroscopy Applications. Chemosensors. 2021;9:262.

36. Lohumi S, Lee S, Lee H, Cho B-K. A review of vibrational spectroscopic techniques for the detection of food authenticity and adulteration. Trends Food Sci Technol. 2015;46:85–98.

37. Rinnan Å, Berg F van den, Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. TrAC - Trends in Analytical Chemistry. 2009. p. 1201–22.

38. Guo Q, Wu W, Massart DL. The robust normal variate transform for pattern recognition with near-infrared data. Anal Chim Acta. 1999;382:87–103.

39. Afseth NK, Kohler A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. Chemometrics and Intelligent Laboratory Systems. 2012;117:92–9.

40. Kennard RW, Stone LA. Computer Aided Design of Experiments. Technometrics. 1969;11:137–48.

41. Stevens A, Ramirez-Lopez L. An introduction to the prospectr package [Internet]. 2022 [cited 2023 Feb 2]. Available from: https://cran.r-project.org/web/packages/prospectr/vignettes/prospectr.html

42. Saptoro A, Tadé MO, Vuthaluru H. A modified Kennard-Stone algorithm for optimal division of data for developing artificial neural network models. Chemical Product and Process Modeling. 2012;7.

43. Pearl J. PROBABILISTIC REASONING IN INTELLIGENT SYSTEMS: Networks of Plausible Inference REVISED SECOND PRINTING. 1988.

44. Koller D, Friedman N. Probabilistic graphical models: principles and techniques. MIT Press, Cambridge; 2009.

45. Verma TS, Pearl J. Equivalence and Synthesis of Causal Models. Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence. 1990;220–7.

46. Tsamardinos I, Aliferis CF, Statnikov A. Algorithms for Large Scale Markov Blanket Discovery [Internet]. 2003. Available from: www.aaai.org

47. Wold S, Sjostrom M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems [Internet]. 2001;58:109–30. Available from: www.elsevier.comrlocaterchemometrics

48. Mateos-Aparicio G. Partial least squares (PLS) methods: Origins, evolution, and application to social sciences. Commun Stat Theory Methods. 2011;40:2305–17.

49. Geladi P, Kowalski BR. PARTIAL LEAST-SQUARES REGRESSION: A TUTORIAL. Anal Chim Acta. Elsevier Science Publishers B.V; 1986.

50. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York, NY: Springer New York; 2009.

51. Horn SS, Ruyter B, Meuwissen TH, Hillestad B, Sonesson AK. Genetic effects of fatty acid composition in muscle of Atlantic salmon. Genetics Selection Evolution. BioMed Central Ltd.; 2018;50:1–12.

52. Difford GF, Horn SS, Dankel KR, Ruyter B, Dagnachew BS, Hillestad B, et al. The heritable landscape of near-infrared and Raman spectroscopic measurements to improve lipid content in Atlantic salmon fillets. Genetics Selection Evolution. BioMed Central Ltd; 2021;53.

53. FOLCH J, LEES M, SLOANE STANLEY GH. A simple method for the isolation and purification of total lipides from animal tissues. J Biol Chem. 1957;226:497–509.

54. Mason ME, Waller GR. Dimethoxypropane Induced Transesterification of Fats and Oils in Preparation of Methyl Esters for Gas Chromatographic Analysis. Anal Chem. 1964;36:583–6.

55. Liland KH, Kohler A, Afseth NK. Model-based pre-processing in Raman spectroscopy of biological samples. Journal of Raman Spectroscopy. John Wiley and Sons Ltd; 2016;47:643–50.

56. Afseth NK, Wold JP, Segtnan VH. The potential of Raman spectroscopy for characterisation of the fatty acid unsaturation of salmon. Anal Chim Acta. 2006;572:85–92.

57. Zhang Y, Zhang Z, Liu K, Qian G. An improved IAMB algorithm for Markov blanket discovery. J Comput (Taipei). 2010;5:1755–61.

58. Marco Scutari. bnlearn: Bayesian Network Structure Learning, Parameter Learning andInference. 2022.

59. Mevik B-H. Introduction to the pls Package. 2022.

60. Maia RP, Madsen P, Labouriau R. Multivariate survival mixed models for genetic analysis of longevity traits. J Appl Stat. Routledge; 2014;41:1286–306.

61. Czamara K, Majzner K, Pacia MZ, Kochan K, Kaczor A, Baranska M. Raman spectroscopy of lipids: a review. Journal of Raman Spectroscopy. 2015;46:4–20.

62. Chang K, Lee J, Jun CH, Chung H. Interleaved incremental association Markov blanket as a potential feature selection method for improving accuracy in near-infrared spectroscopic analysis. Talanta. Elsevier B.V.; 2018;178:348–54.

63. Aagaard Poulsen N, Eskildsen C, Skov T, Larsen L, Buitenhuis A. Comparison of Genetic Parameters Estimation of Fatty Acids from Gas Chromatography and FT-IR in Holsteins. 2014.

64. Blay C, Haffray P, D'Ambrosio J, Prado E, Dechamp N, Nazabal V, et al. Genetic architecture and genomic selection of fatty acid composition predicted by Raman spectroscopy in rainbow trout. BMC Genomics. BioMed Central Ltd; 2021;22.

Appendix A

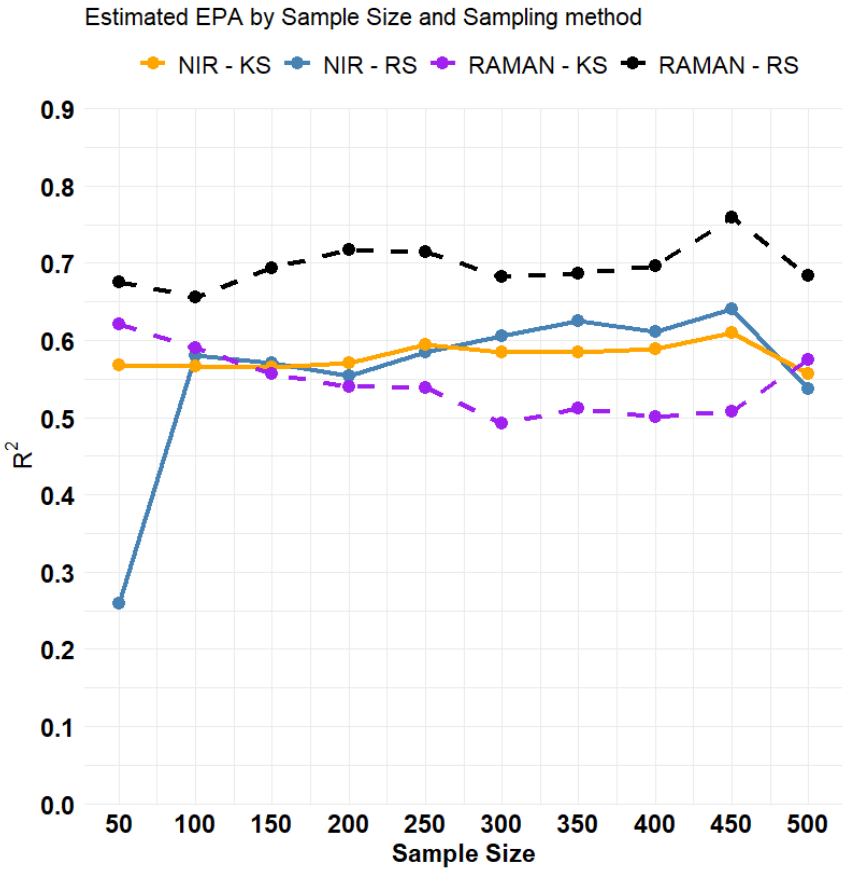


Figure A 1. The predictive performance of PLSR models with increasing training sizes from 50 to 500 samples for EPA under different scenarios of sample selection and variable selection for EPA.

Appendix B

Table B 1. PLSR results for the whole data set (n = 613) using NIR and Raman spectroscopy

	NIR spectroscopy[1]			**Raman spectroscopy**[1]		
Fatty Acids[2]	PCs[3]	R^2	RMSE	PCs[3]	R^2	RMSE
EPA	11	0.62	0.28	5	0.73	0.24
DHA	11	0.63	0.32	7	0.77	0.25
ALA	15	0.69	0.13	8	0.81	0.1

[1] Both spectroscopy data have the identical number of variables (p = 1300),
[2] EPA = (eicosatetraenoic acid, C20-5n3), DHA = (docosahexaenoic acid, C22-6n3) and ALA = (α-linolenic acid, 18:3n-3),
[3] PCs = Number of Components (maximum = 15)

Appendix C

Table C 1. All selected list of spectroscopy wavelengths/Raman shifts by Markov Blanket

Fatty Acids[1]	NIR spectroscopy[2]	Raman spectroscopy[3]
EPA	1450, 2427	514, 545, 608, 624, 636, 648, 654, 666, 698, 729, 748, 757, 822, 835, 861, 870, 873, 884, 934, 990, 1004, 1068, 1087, 1116, 1120, 1126, 1149, 1197, 1224, 1233, 1243, 1247, 1259, 1265, 1293, 1312, 1331, 1358, 1365, 1371, 1404, 1419, 1439, 1441, 1453, 1455, 1459, 1465, 1469, 1476, 1481, 1488, 1494, 1497, 1501, 1533, 1552, 1605, 1610, 1613, 1634, 1645, 1648, 1665, 1679, 1692, 1697, 1704, 1708, 1717, 1749, 1753, 1760, 1763
DHA	1163, 1164, 1170, 1173, 1209, 1224, 1245, 1313, 1330, 1341, 1363, 1373, 1422, 1425, 1501, 1543, 1634, 1648. 1665, 1691, 1692, 1750, 1751, 1762, 1821, 1838, 1902, 2063, 2092, 2157, 2158, 2270, 2286, 2295, 2364, 2394, 2441	549, 566, 577, 593, 600, 612, 699, 718, 760, 821, 865, 871, 922, 929, 1032, 1073, 1077, 1083, 1167, 1278, 1318, 1351, 1371, 1453, 1469, 1594, 1610, 1627, 1633, 1666, 1682, 1692, 1786
ALA	1157, 1169, 1235, 1250, 1374, 1405, 1429, 1458, 1476, 1578, 1603, 1654, 1665, 1680, 1703, 1716, 1731, 1751, 1762, 1786, 1791, 1881, 1964, 2046, 2117, 2135, 2136, 2213, 2226, 2242, 2272, 2288, 2305, 2316, 2337, 2418, 2449	529, 535, 548, 617, 674, 680, 746, 756, 768, 822, 835, 851, 869, 874, 898, 902, 911, 929, 947, 957, 965, 974, 1014, 1063, 1085, 1103, 1108, 1111, 1124, 1130, 1166, 1181, 1193, 1198, 1205, 1217, 1222, 1275, 1313, 1334, 1374, 1377, 1382, 1390, 1400, 1425, 1442, 1454, 1486, 1534, 1537, 1543, 1549, 1557, 1570, 1591, 1607, 1619, 1626, 1632, 1636, 1648, 1655, 1666, 1673, 1679, 1713, 1727, 1758, 1762

[1] EPA = (eicosatetraenoic acid, C20-5n3), DHA = (docosahexaenoic acid, C22-6n3) and ALA = (α-linolenic acid, 18:3n-3)

[2] selected NIR spectroscopy (nm) by MB

[3] selected RAMAN spectroscopy (cm^{-1}) by MB

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway