Omega sample with NIR, RAMAN

Code

Report Bin302 High throughput phenotyping for smart farming 2021

Jisoo Park

Last compiled date is 2021-12-01

Introduction

During Fall 2021, we learned the High Throughput phenotype for smart farming that covered multiple areas-Image Analysis, Vibrational Spectroscopy, Machine and Deep Learning, Sensors in Time Series, and benchmark and validation.

Combine with the Vibrational Spectroscopy and Multivariate analysis enables a wide range of analyses in multidisciplinary research. What we select among the several topics is vibrational spectroscopy with near-infrared(NIR) and RAMAN using Omega3 data to assess and compare its prediction applying machine learning methods.

We assume that smoothing spectra performs well to predict the omega3 fatty acid. The main question is whether this assumption can apply to the data set. Thus, the purpose of this paper is to compare the prediction value and accuracy of omega3 fatty acids among NIR, RAMAN, smoothing methods by using Principal Components Analysis(PCA) and Principal Components Regression(PCR), Partial Least Square Regression(PLSR). The assessments are based on the Root Mean Squared Error(RMSE), R^2, and Concordance correlation coefficient(CCC). The mathematical definition of each method is as below;

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

$$CCC(\rho_c) = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

The data analysis process has a typical process regardless of the field. Simply speaking, it conducts data pre-processing, data modeling and completes the evaluation of the model. We follow these steps; First, we split the training and testing data with 60:40. As a general rule of thumb, the data split into 70:30; however, it could adjust depending on the number of data. Here, we randomly select the fish id and split it into the training and test data set with a 60:40 ratio due to the small number of the data. Applying PCR and PLS model to the training data gives us an overview of the model. We look for the number of principal components based on the lowest RMSE to explain the variance. Each model applies to the test data set for predicting the omega3 fatty acids and compares its RMSE, R^2. Lastly, we will calculate Lin's concordance correlation coefficient for consolidating the outcome.

All data analysis processes and statistical methods are implemented within R and R-studio. In R, pls packages provide PCR and PLS methods. The other choice for executing PCR and PLS is using caret, tidymodels packages.

Description of data

Code

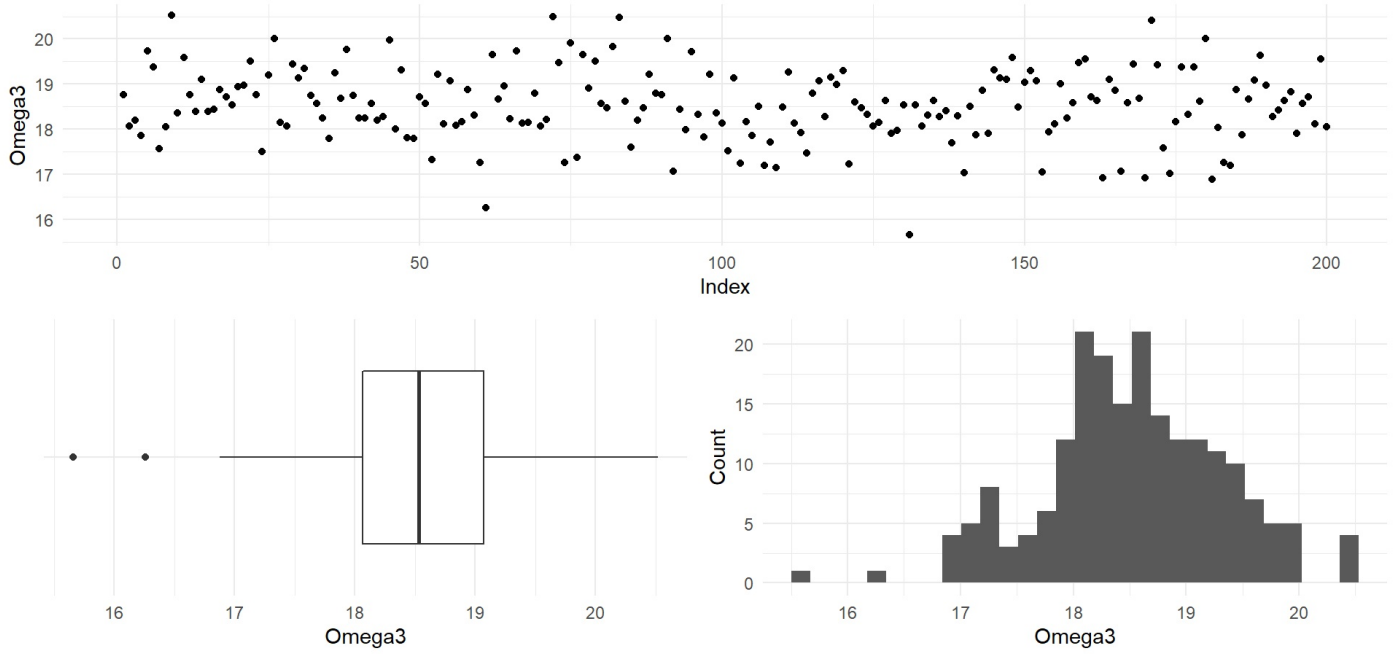
The given data set has comprised of three CSV files named `Omega Assistant`, `NIR Assistant`, and `RAMAN Assistant`. Each file has a 200 unique id of fish used to merge based on the purpose. Omega Assistant set contains 200 observations with four variables; unique id of the fish, the total contents of omega3 and omega6 fatty acids, pigment. It could be considered all three variables except id as the target variable in prediction. However, we would like to focus on omega3 that known as beneficial nutrition for human beings.

`RAMAN Assistant` and `NIR Assistant` contain their spectra information as a variable. Spectroscopy is widely applied in the fields of Physics or analytical chemistry to study the interaction between light and material depending on the wavelength. NIR and RAMAN are included in vibrational spectroscopy because they identify and quantify biomolecules by analyzing the spectra absorbed by a substance. Since both are the one way of vibrational spectroscopy techniques with different molecules analyses and have different wavelengths and noises, we generally need additional data processing accordingly before applying the machine learning methods. It employs the Savitzky-Golay filtering with the second derivatives to reduce the noise effect.

Most omega3 fatty acids are in a range of 17 and 20. However, we can see there are two outliers in the boxplot and index plot. It may need to take into account to delete or other treatment because it affects the prediction value. Omega3 and Omega 6 have a negative relationship holds the Pearson correlation -0.59. A salmon in the given data set has contents of high omega3 with relatively low omega6.

Code

Fig 1. Omega3 distribution



Code

Table 1. Omega3 content below 16.5

n	Omega3	Omega6	pigment
299	16.26	12.29	7.7
431	15.66	14.83	7.1

Code

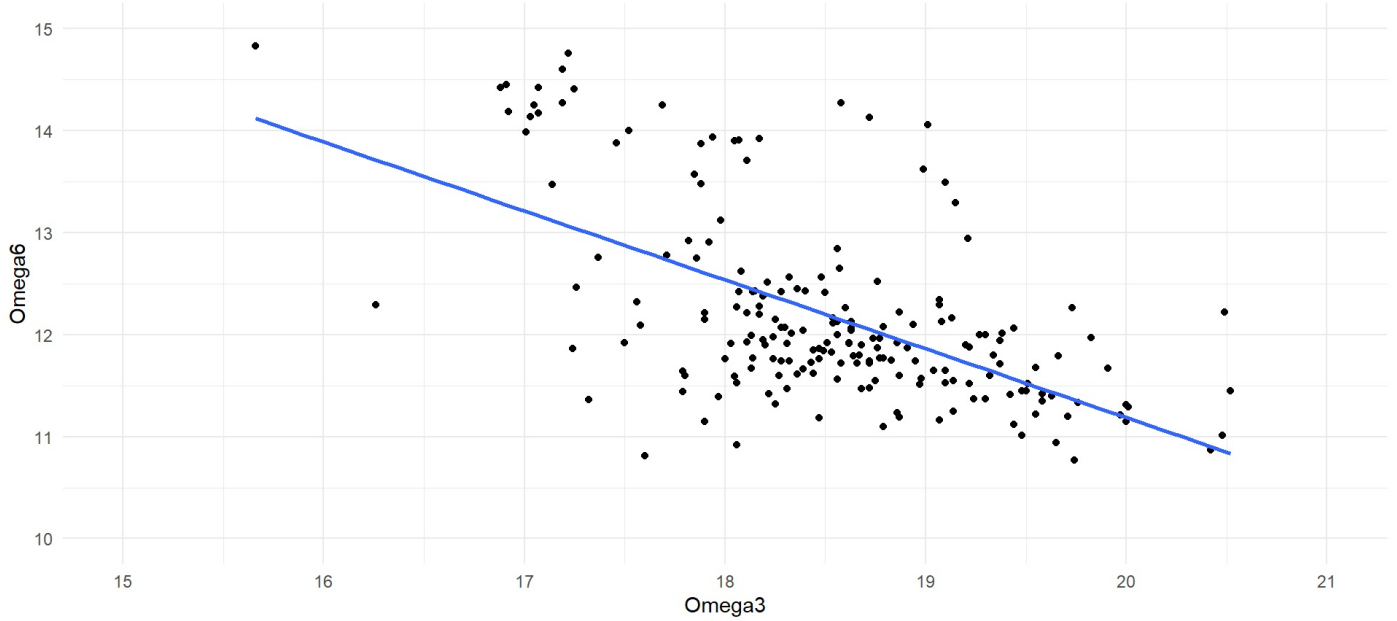
Table 2. Correlation table

	Omega3	Omega6	pigment
Omega3	1.0000000	-0.5941631	0.0560357
Omega6	-0.5941631	1.0000000	-0.3273615
pigment	0.0560357	-0.3273615	1.0000000

Code

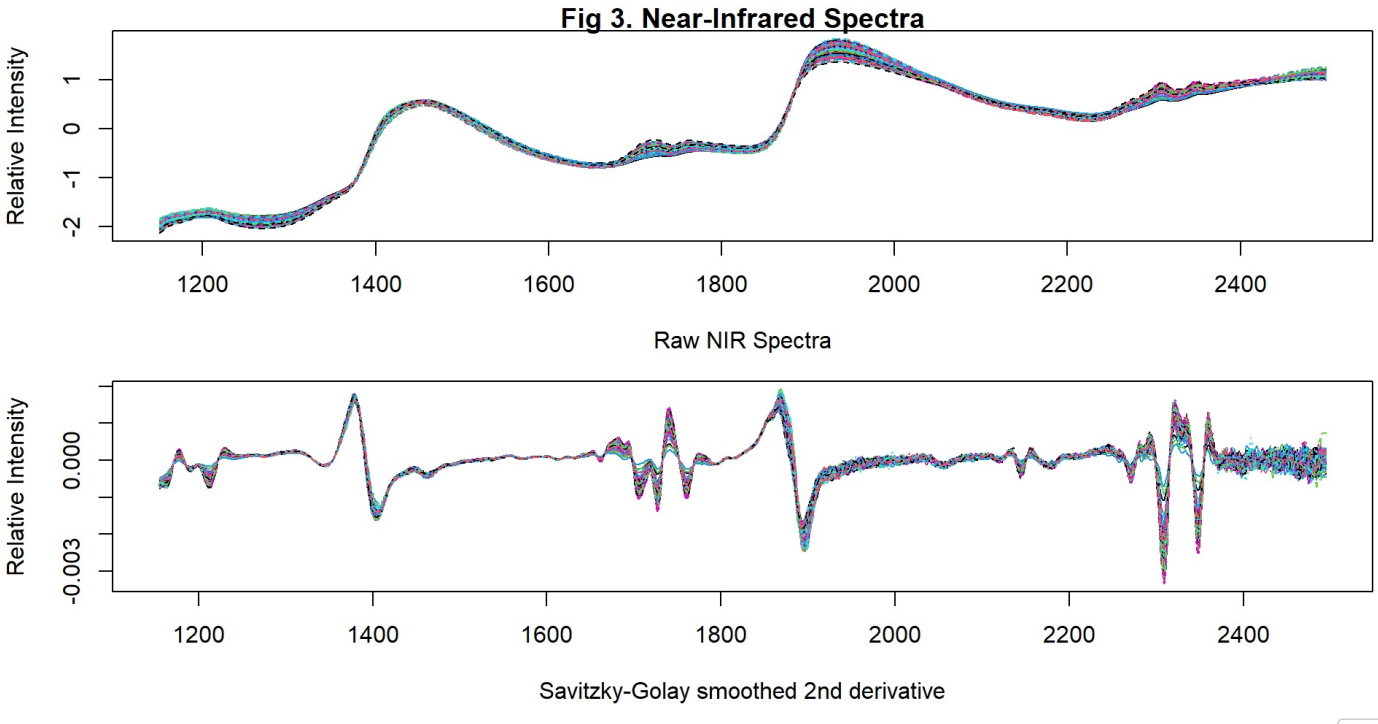
Fig 2. Correlation between Omega3 and Omega 6

corr: -0.5942

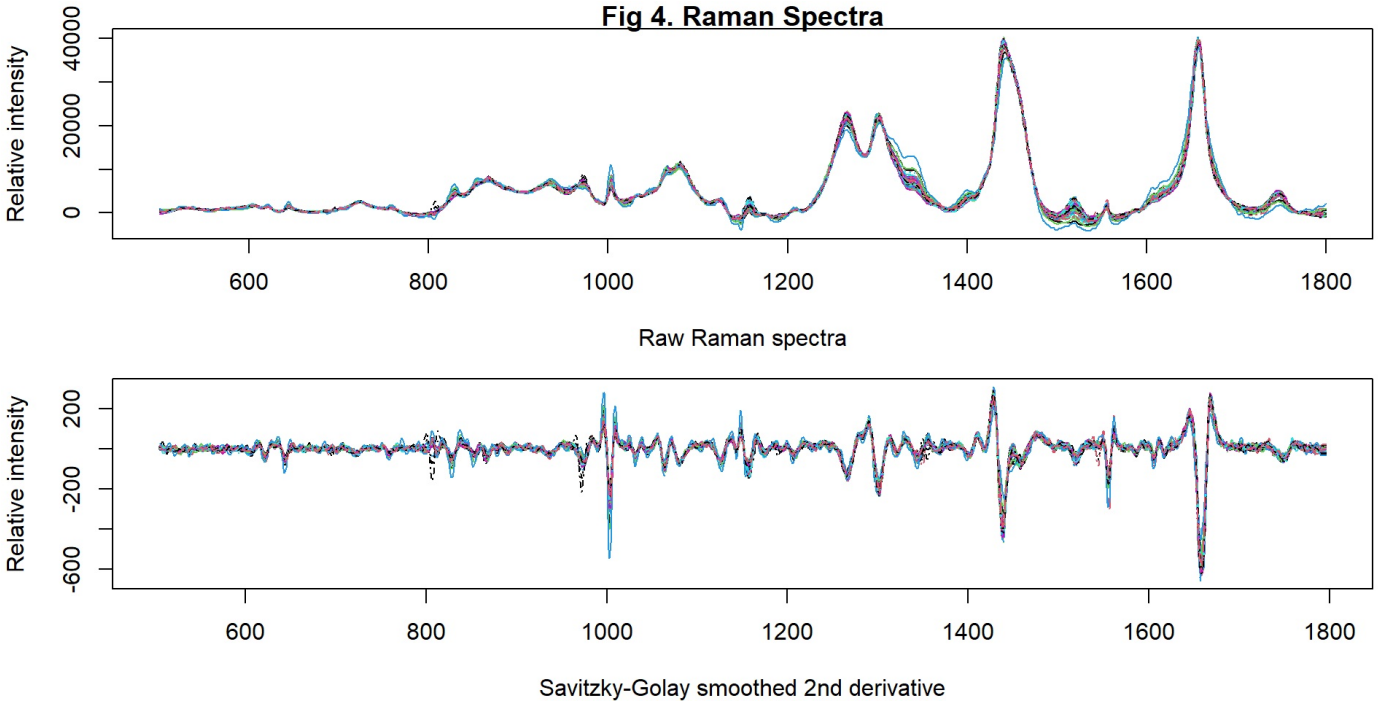


All spectra in the given NIR Assignment data set have 1350 variables, from 1150 to 2499. On the other hand, all spectra in the given RAMAN Assignment have 1301, from 500 to 1800. The second derivative of the Savizky-Golay filter flattens a spike of data, so most of the data points are located closely at zero. We expect that it has a similar normalization effect.

Code



Code



Code

Methods

Spectroscopy data is high-dimensional data comprised of spectra. PCA allows reducing its high dimension to a reasonable number of new variables. Mathematically, it uses a matrix that decomposes the value into eigenvector and eigenvalue to maximize its variance and create a new composition. Thus, it enables to explain the variance of data with a small number of principal components which is the main advantage. One drawback of PCA only considers the linear combination between the independent variables, so it does not explain the relationship between the target variable.

Principal component regression analysis performs by using the principal component as an explanatory variable instead of the existing variable. In addition, it is a regression analysis methodology used to solve the multicollinearity problem and the high-dimensional problem. Partial least square regression is similar to principal component regression in terms of using the principal components. It is a widely used technique in chemometrics, bioinformatics, and related fields rather than principal component regression. While principal component regression finds a linear combination for maximizing its variance within the independent variables, principal component regression finds maximum correlation with dependent variable considering the target variable. Thus, both PCR and PLSR is practical regression technique to use principal component.

Modeling

It is possible to divide the unique fish id into 60:40 and compare the predicted values by having the same id. We set the random seed for a reproducible experiment.

Note that there is a gap in specific terminology between phenotype and data science. In data science, the split data are called the training set and the testing set, while in chemometrics and related fields the data are called calibration and validation sets.

Code

Code

Table 3. Descriptive Statistics of splited data

Index	Omega3_count	Omega3_min	Omega3_max	Omega3_mean	Omega3_sd
Total	200	15.66	20.52	18.52258	0.8097732
Test	80	15.66	20.48	18.62333	0.8646881
Train	120	16.26	20.52	18.45542	0.7673505

PCA

In principle components analysis, we can see that the raw spectra could reduce by principal component. While the Nir spectra explain up to 85% by the second components, the Raman spectra explain only 61%. It means that the near-infrared could reduce as two principal components instead of using all. Raman spectra could select four compositions to contain over 70% of the variance. Despite performing well in each raw spectra, it doesn't work enough in Savitzky-Golay smoothed filter.

NIR

Code

```
## Importance of first k=10 (out of 200) components:
##                           PC1     PC2      PC3    PC4     PC5     PC6     PC7
## Standard deviation     31.3333 13.2327 10.65561 6.2028 4.98321 3.39715 1.33410
## Proportion of Variance  0.7272  0.1297  0.08411 0.0285 0.01839 0.00855 0.00132
## Cumulative Proportion   0.7272  0.8569  0.94105 0.9696 0.98795 0.99650 0.99781
##                           PC8     PC9    PC10
## Standard deviation     1.2181 0.63095 0.45890
## Proportion of Variance 0.0011 0.00029 0.00016
## Cumulative Proportion  0.9989 0.99921 0.99936
```

Code

```
## Importance of first k=10 (out of 200) components:
##                           PC1      PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     23.5371 10.92528 5.99043 5.29292 4.24029 4.01245 3.86935
## Proportion of Variance  0.4131  0.08901 0.02676 0.02089 0.01341 0.01201 0.01116
## Cumulative Proportion   0.4131  0.50213 0.52889 0.54978 0.56319 0.57520 0.58636
##                           PC8     PC9    PC10
## Standard deviation     3.84790 3.74698 3.66735
## Proportion of Variance 0.01104 0.01047 0.01003
## Cumulative Proportion  0.59740 0.60787 0.61790
```

Code

RAMAN

Code

```
## Importance of first k=10 (out of 200) components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     23.6976 15.3463 9.67431 7.93235 5.55923 5.19337 4.42346
## Proportion of Variance  0.4316  0.1810 0.07194 0.04836 0.02375 0.02073 0.01504
## Cumulative Proportion   0.4316  0.6127 0.68461 0.73297 0.75673 0.77746 0.79250
##                           PC8     PC9    PC10
## Standard deviation     3.97163 3.67375 3.49423
## Proportion of Variance 0.01212 0.01037 0.00938
## Cumulative Proportion  0.80462 0.81500 0.82438
```

Code

```
## Importance of first k=10 (out of 200) components:
##                           PC1     PC2     PC3     PC4     PC5     PC6    PC7
## Standard deviation     15.3226 9.68231 7.10036 5.61320 5.43067 4.79883 4.6588
## Proportion of Variance  0.1817 0.07256 0.03902 0.02439 0.02283 0.01782 0.0168
## Cumulative Proportion   0.1817 0.25428 0.29330 0.31769 0.34051 0.35834 0.3751
##                           PC8     PC9   PC10
## Standard deviation     4.60682 4.33780 4.2528
## Proportion of Variance 0.01643 0.01456 0.0140
## Cumulative Proportion  0.39156 0.40613 0.4201
```

Code

summary

Code

Fig 5-1. Principal Component Analysis - NIR

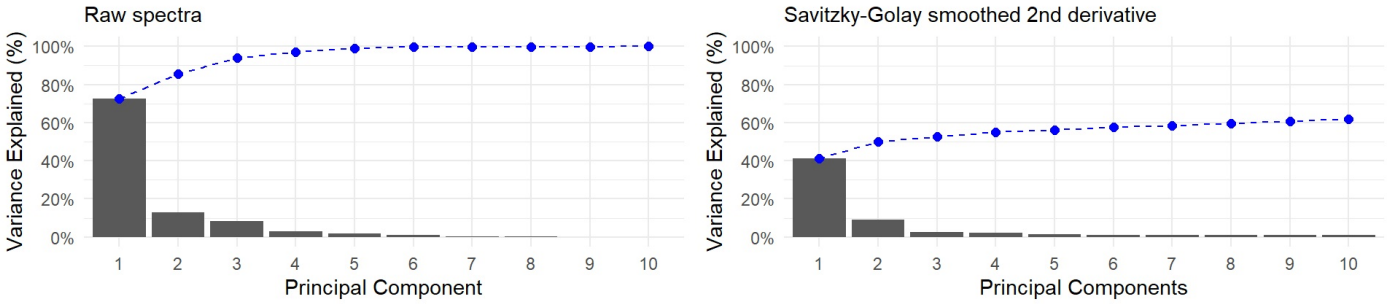
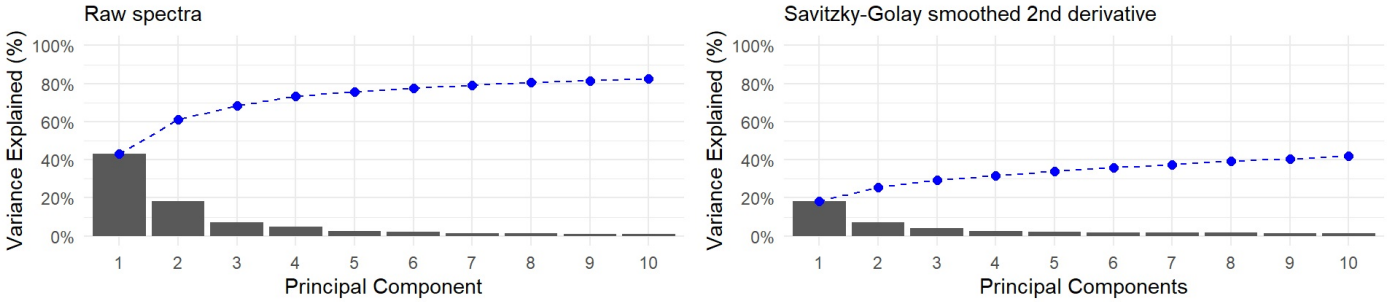


Fig 5-2. Principal Component Analysis - RAMAN



PCR

Principal component regression has a similar approach to principal component analysis. It has one more step before linear regression. As the first step, we conduct the principal component analysis to find the combination of maximizing its variance. Once we figure out what number of selections appropriately explain, we apply it to linear regression. The basic approach for determining the number of components is following the minimize adjusted CV point. We could not find other studies on if there exists the gold standard on the number of PCs. The minimized Adjusted Cross-Validation marked 23 in `PCR-NIR` , 78 in `PCR-NIR SG` , 13 in `PCR-RAMAN` , and 86 in `PCR-RAMAN SG` , as shown in Fig 7. It tends to choose a large number of components when the data is filtered.

NIR

Code

Code

RAMAN

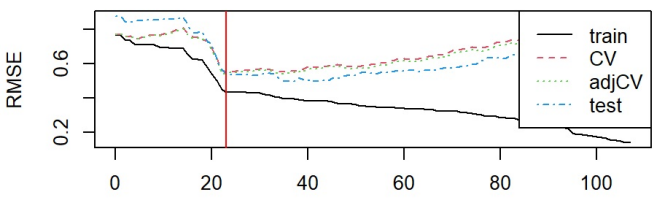
Code

Code

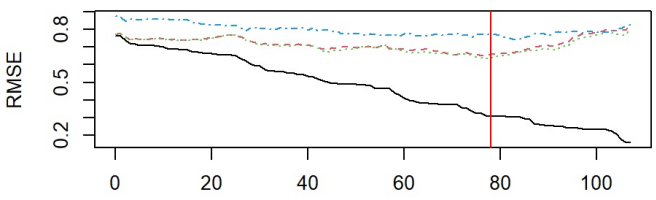
summary

Code

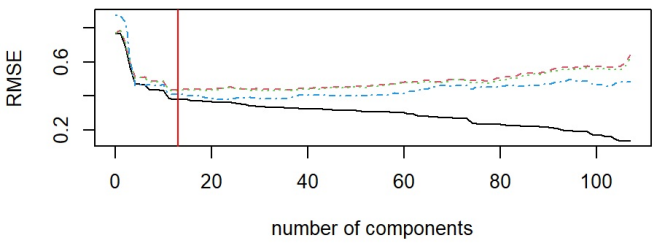
Fig 7. Cross-Validation: PCR - NIR



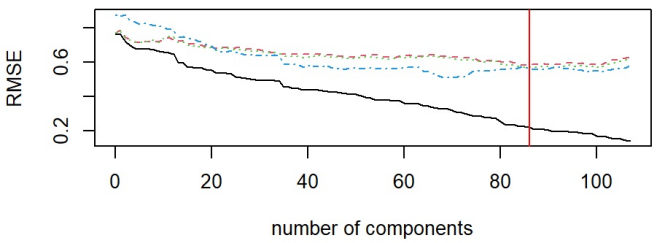
PCR - NIR SG



PCR - RAMAN



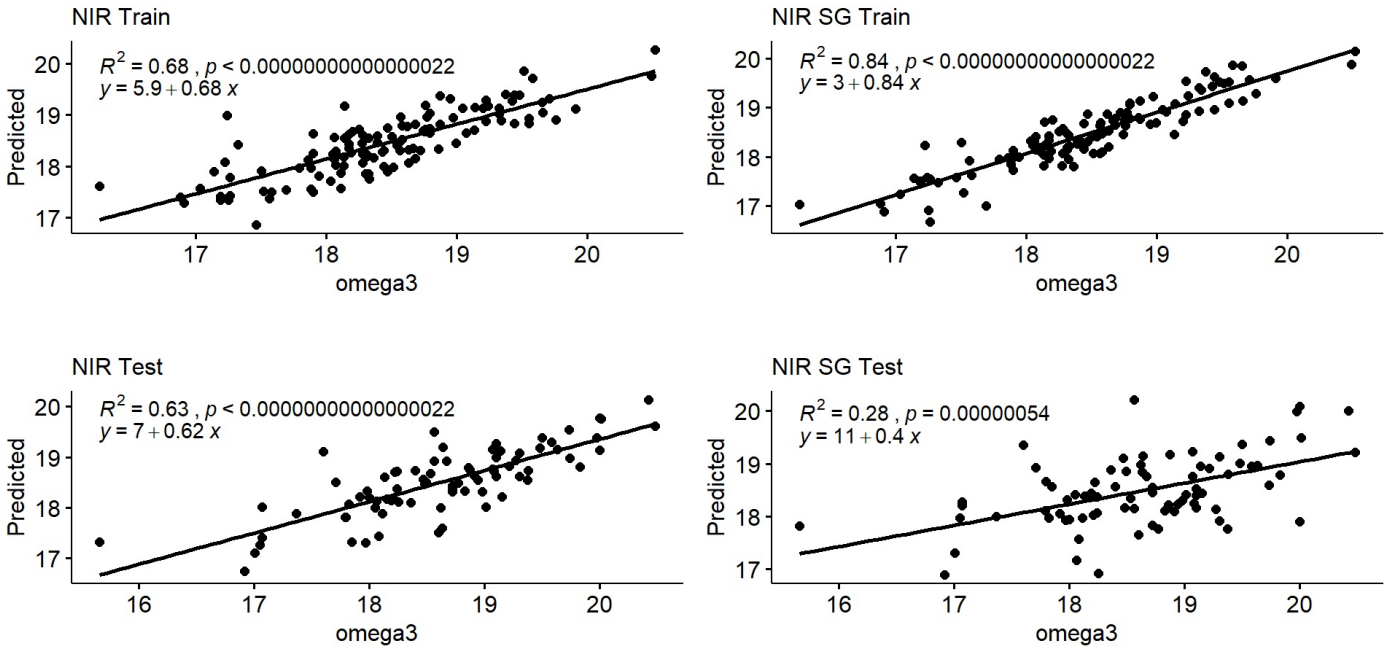
PCR - RAMAN SG



Code

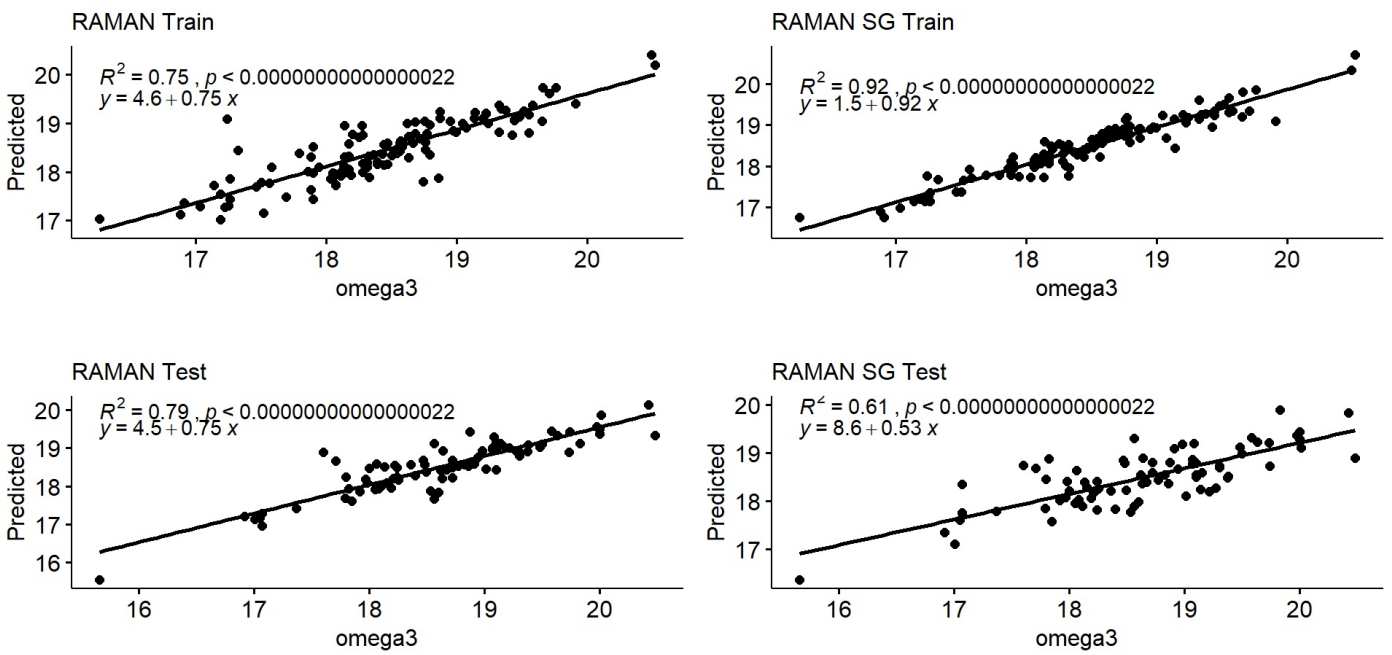
Code

Fig 8. PCR-NIR



Code

Fig 9. PCR-RAMAN



PLS

Partial Least Square Regression has a similar approach with principal component regression in terms of looking for a combination of components. One distinct difference is it considers the relationship between the Omega3 and spectra. As the same method, the minimized Adjusted Cross-Validation marked 13 in `PCR-Nir` , 6 in `PCR-NIR SG` , 6 in `PCR-RAMAN` , and 7 in `PCR-RAMAN SG` , as shown in Fig 11. One noticeable point is the Partial Least Square Regression needs fewer components than the Principal Component Regression.

NIR

Code

Code

RAMAN

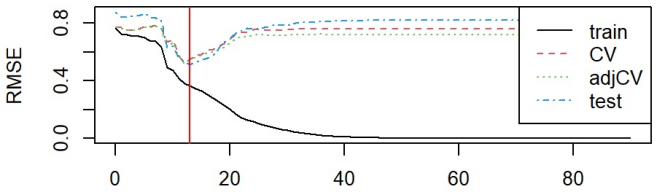
Code

Code

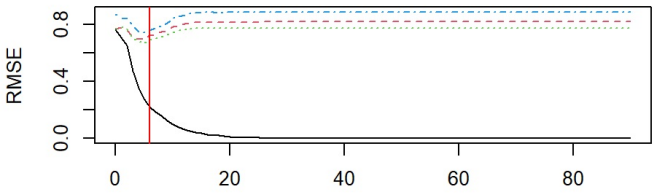
summary

Code

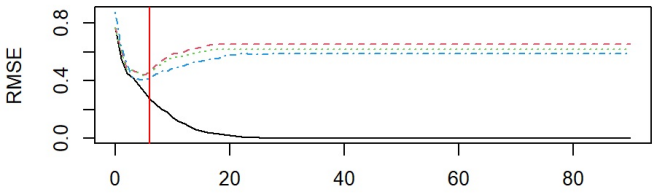
Fig 11. Cross-Validation; PLSR - NIR



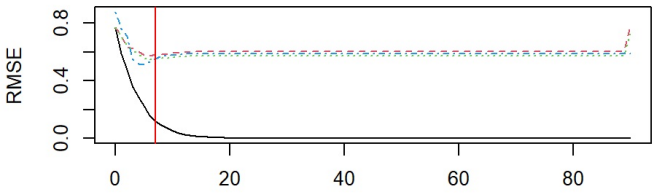
PLSR - NIR SG



PLSR - RAMAN



PLSR - RAMAN SG

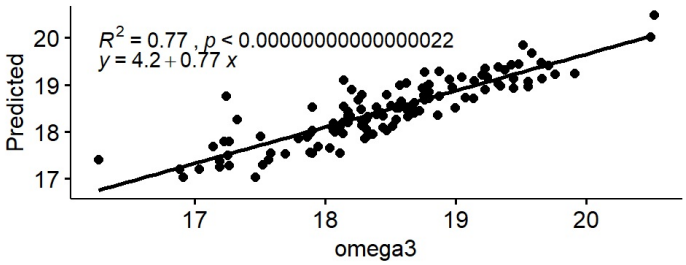


Code

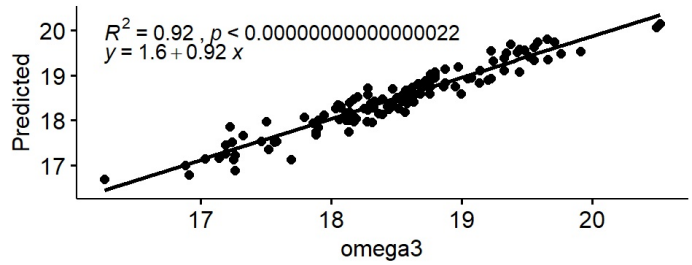
Code

Fig 12. PLS-NIR

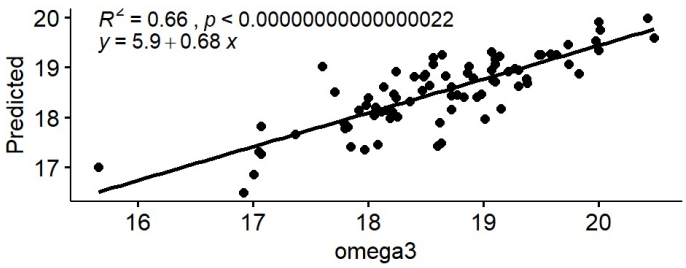
NIR Train



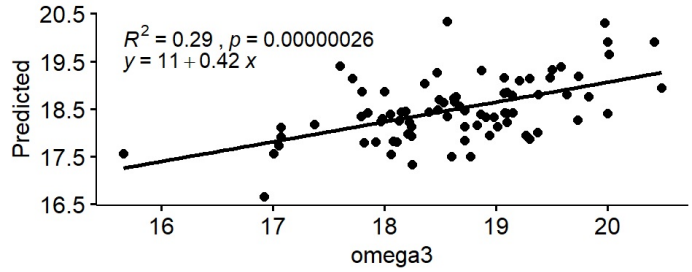
NIR SG Train



NIR Test

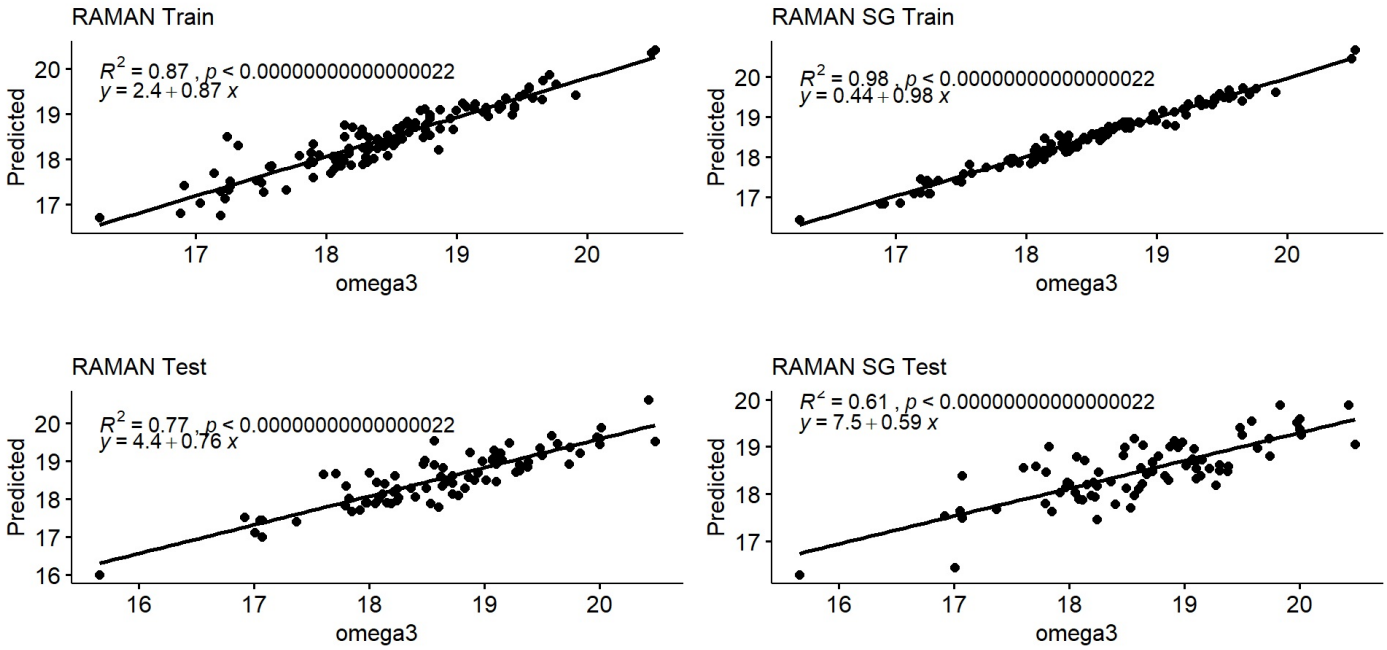


NIR SG Test



Code

Fig 13. PLS-RAMAN



Code

Results – comparisons

Code

Table 4. RMSE based on the model

model	RMSE
PCR-NIR	0.1193906
PCR-NIRSG	0.1396748
PCR-RAMAN	0.1160718
PCR-RAMANSG	0.1336575
PLSR-NIR	0.1124282
PLSR-NIRSG	0.1413866
PLSR-RAMAN	0.0863519
PLSR_RAMANSG	0.1361441

Code

Table 5. Concordance Correlation Coefficient

model	est	lower	upper
PCR-NIR	0.7591429	0.6584118	0.8331578
PCR-NIRSG	0.4995296	0.3276452	0.6394073
PCR-RAMAN	0.8695116	0.8089303	0.9118211
PCR-RAMANSG	0.7110124	0.6058470	0.7917569
PLSR-NIR	0.7915765	0.6992593	0.8579248
PLSR-NIRSG	0.5133775	0.3434394	0.6507647
PLSR-RAMAN	0.8635542	0.7990278	0.9084177
PLSR_RAMANSG	0.7413627	0.6375691	0.8187221

Table 6. Model summary of test data

Methods	# of Comps	formula	RMSE	R^2	Estimated CCC
PCR-NIR	23	$y = 7 + 0.62x$	0.119	0.63	0.75
PCR-NIRSG	78	$y = 11 + 0.4x$	0.139	0.28	0.49

PCR-RAMAN	13	$y = 4.5 + 0.75x$	0.116	0.79	0.86
PCR-RAMANSG	86	$y = 8.6 + 0.53$	0.133	0.61	0.71
——-:	—-:	——-:	———:	—-:	———:
PLS-NIR	13	$y = 5.9 + 0.68x$	0.112	0.66	0.79
PLS-NIRSG	6	$y = 11 + 0.42x$	0.141	0.29	0.51
PLS-RAMAN	6	$y = 4.4 + 0.76x$	0.086	0.77	0.86
PLS-RAMANSG	7	$y = 7.5 + 0.59x$	0.136	0.61	0.74

Discussion & Conclusion

We apply a multivariate statistical methodology to compare predicted Omega3 using different spectra. It shows different results by the methods and validation measurements.

In Raman spectra prediction, it shows the highest RMSE value(0.77-0.79) and CCC(0.86) regardless of principal component regression or partial least square regression. However, the partial least square regression has more efficiency to predict with the lower number of components within the same condition. Furthermore, Raman spectra in partial least square regression have the lowest RMSE value which indicates that accuracy.

The number of components to be selected varies depending on which method and spectra are applied. The R2 is slightly different between the training and the test data set. If there is a large gap between the train and test data, there may occur an overfitting or underfitting problem. These problems lead to less credibility of the model. Since there is no golden standard for determining the number of components, we believe there could be further development that suggests new methods.

In conclusion, Raman spectra with a partial least square regression could be used to predict the Omega3 fatty acids of salmon. In this paper, we only focused on predicting Omega3 fatty acids using spectra. It could be a further developed model with an effort of searching for the best balance among Omega3, Omega6, and pigment.

References

- Lecture Note Week39 (https://nmbu.instructure.com/courses/7475/files/1370004?module_item_id=148118)
- PLS packages in R (https://cran.r-project.org/web/packages/pls/vignettes/pls-manual.pdf)
- Process Improvement Using Data, Chapter 6 (https://learnche.org/pid/latent-variable-modelling/projection-to-latent-structures/advantages-of-projection-to-latent-structures)
- Hands-On Machine Learning with R, Chapter 4 (https://bradleyboehmke.github.io/HOML/)
- spectroscopy review (https://epjtechniquesandinstrumentation.springeropen.com/articles/10.1140/epjti/s40485-015-0018-6)
- Determination of fatty acids and lipid classes in salmon oil by near infrared spectroscopy (https://www.sciencedirect.com/science/article/pii/S0308814617311408)
- Investigation of NIR spectra pre-processing methods combined with multivariate regression for determination of moisture in powdered industrial egg (https://www.redalyc.org/journal/3032/303258327015/html/)