



Special Issue: Proceedings of the 13<sup>th</sup> International Conference on Near Infrared Spectroscopy—NIR-2007

# Incorporating chemical band-assignment in near infrared spectroscopy regression models

Frank Westad,<sup>a,b</sup> Angela Schmidt<sup>b</sup> and Martin Kermit<sup>b</sup>

<sup>a</sup>M-Analyse, Oslo, Norway. E-mail: frank.westad@online.no

<sup>b</sup>Camo Software AS, Oslo, Norway

In this paper, we present an approach for incorporating chemical band assignment information in regression models between spectra and constituents. It is shown how the matrices in this L-shaped data structure can be combined and give direct information of the relationships between theoretical chemical band assignment, spectral wavelengths and the responses. The chosen application is NIR spectroscopic measurements of canola seeds. Variable selection based on partial least squares regression using jack-knifing within a cross-model validation (CMV) framework is applied for removing non-relevant spectral regions. Extended multiplicative scatter correction was applied as a spectral pre-treatment to remove physical scatter effects in the spectra. The results show a high degree of correspondence between the objectively found wavelength bands from CMV and the reported chemical interpretation found in the literature.

**Keywords:** PLS regression, cross-model validation, band assignment, extended multiplicative scatter correction, fatty acids, interpretive spectroscopy

## Introduction

NIR spectroscopy data is often applied in building quantitative models, typically with various regression methods in a multivariate calibration context. As the NIR spectra emerge from overtones and combinations of fundamental mid infrared (IR) absorptions, the NIR bands are naturally broad and overlapping and, thus, not selective (accompaniment is that their intensity is 10–50 times less than their corresponding mid-IR bands). This is one of the reasons why NIR applications in general focus on empirical approaches rather than spectral interpretations when compared to classical IR and Raman spectroscopy.

There is always an aspect of interpretation in multivariate regression models and it is important to understand the cause and effect rather than spurious correlations in the observed data. Also, it is known that regression coefficients themselves should be interpreted with care. Two reasons for this are:

1. The size and sign of the regression coefficients may not be interpretable because a given least square solution (one of many) does not “know” if  $2+2=4$  or  $-8+12=4$ . This effect becomes more prominent in latent variable models when the model dimensionality approaches the MLR solution. However, abrupt changes in the regression coefficient dimensionality occur long before there are numerical problems due to collinearity.
2. The regression model itself may be computed from samples where constituents are correlated. One such example is seen by the high negative correlation ( $> -0.9$ ) between fat and water in red meat. In this case, it is important to interpret the relevant wavelengths from a regression model for water and to assess whether it is the chemical band from water or fat, or both, which is being modelled.

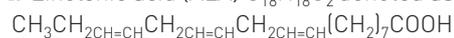
This work presents how to incorporate chemical information by representing regions for different functional groups as

a matrix from the table of absorption bands in the literature. Thus, the spectral data is denoted  $\mathbf{X}$ , the responses  $\mathbf{Y}$ , and the band-specific information as  $\mathbf{Z}$ . A two-step procedure, where the relationship between  $\mathbf{Z}$  and  $\mathbf{X}$  is modelled prior to modelling  $\mathbf{Y}$ , is presented. Cross-model validation (CMV)<sup>1-4</sup> is applied as a robust variable selection method to remove the irrelevant wavelengths.

## Materials and methods

One hundred and sixty seven samples from Canola seeds have been measured in reflectance mode with a dispersive spectrometer model 5000 (NIRSystems Inc., Silver Springs, MD, USA) equipped with a spinning ring cup module (diameter 50 mm) in a wavelength range from 1100 nm to 2500 nm in 2 nm increments. The reference used was the standard ceramic plate of the instrument. Canola is a trademarked brand name for a group of cultivars from rapeseed variants, initially bred in Canada, from which edible oil with especially low amounts of cancer-causing erucic acid (C22:1, <2%) and toxic glycosinolates (<30  $\mu\text{mol}$ ) is obtained. However, canola oil is healthy due to its low content of saturated fatty acids (palmitic acid C16:0, stearic acid C18:0), high level of mono-unsaturated oleic acid (MUFA) C18:1 (nearly 60%) and an intermediate level of poly-unsaturated fatty acids [(PUFAs) such as linoleic acid C18:2 and  $\alpha$ -linolenic acid C18:3 (ALA)]. In particular the PUFAs represent omega-6 (linoleic) and beneficial omega-3 (ALA, an essential nutrient) fatty acids, which are in a good 2:1 balance. Dry canola seeds contain about 10% seed moisture, 20–30% protein and 40–50% oil. In this sample set of whole canola seeds, only the ALA contents have been determined as fatty acid methyl ester by gas chromatography (FAME-GC), ranging between 4.99–11.12%.

$\alpha$ -Linolenic acid (ALA)  $\text{C}_{18}\text{H}_{32}\text{O}_2$  denoted as C18:3



## Significance testing and cross-model validation

So-called CMV<sup>1-4</sup> has been applied in other applications to reduce the chance of spurious correlations being interpreted as valuable information. This problem arises when variable selection procedures are employed in situations where  $\mathbf{X}$  and/or  $\mathbf{Y}$  have relatively few objects and many variables (typically >500). It has been shown that cross-validation is too optimistic when thousands of variable combinations are evaluated in a search for the “best” model. CMV is based on the simple idea of performing a two-layer cross-validation. One informative way to visualise the results is to count the number of times every variable is found to be significant, and present it as “frequency of significance”.<sup>4</sup> This procedure is generic for any method which aims at finding a sub-set of significant variables among many. Combined with jack-knifing for uncertainty estimates in partial least squares (PLS) regression,<sup>5</sup> the procedure can be performed as follows:

1. leave out sample(s).

- do cross-validated PLS regression with jack-knifing and collect significant variables.
- repeat step 2 with new sample(s) left out until all samples have been left out.
- count the number of times a variable has been found significant in step 2. The number of times a variable was found significant relative to the total number of cross-validated sub-models calculated is called frequency of significance.
- select variables with a frequency of significance higher than a predefined threshold for further analysis.

In this paper, we selected the variables with a frequency of significance equal to 100%, thus found to be significant in all sub-models, for subsequent analyses.

## Analysis of L-shaped data

Let the constituent data be denoted as  $\mathbf{Y}$  ( $N \times J$ ), spectra as  $\mathbf{X}$  ( $N \times K$ ) and chemical band assignment for a number of chemical groups ( $L$ ) as  $\mathbf{Z}$  ( $L \times K$ ). Figure 1 depicts the structure of the data. It is worth mentioning that  $\mathbf{Y}$  and  $\mathbf{Z}$  share no common dimension, but are linked through the spectral data,  $\mathbf{X}$ . There are several ways to analyse these three matrices.<sup>6</sup> The procedure applied in this paper is:

- calculate  $\mathbf{G}$  = correlation matrix ( $\mathbf{Z}$ ,  $\mathbf{X}$ )
- use  $\mathbf{G}$  ( $N \times L$ ), as the new matrix linking  $\mathbf{Z}$  and  $\mathbf{Y}$

Within this framework,  $\mathbf{Y}$  may be modelled from  $\mathbf{X}$  (spectra) only or from  $\mathbf{G}$  alone. In addition, it is possible to apply the augmented matrix  $[\mathbf{X}|\mathbf{G}]$  since  $\mathbf{G}$  can now be positioned together with the other matrices with  $N$  rows. One interesting aspect of modelling  $\mathbf{Y}$  from  $\mathbf{G}$  is that the regression model is based on the inherent link between the actual spectra and theoretical band assignment, giving direct “chemical” interpretation. Naturally, with broad bands as in NIR, the assignments are rather crude and one should always interpret the results in the light of the chemical background knowledge. By applying this procedure in

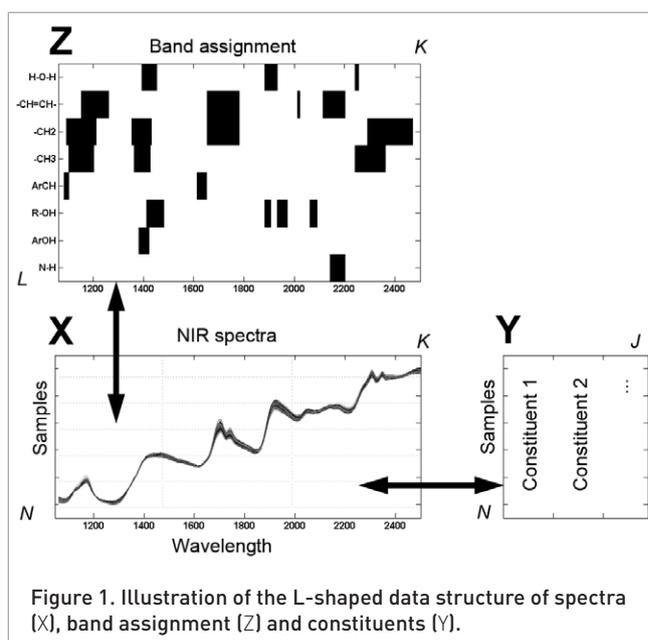


Figure 1. Illustration of the L-shaped data structure of spectra ( $\mathbf{X}$ ), band assignment ( $\mathbf{Z}$ ) and constituents ( $\mathbf{Y}$ ).

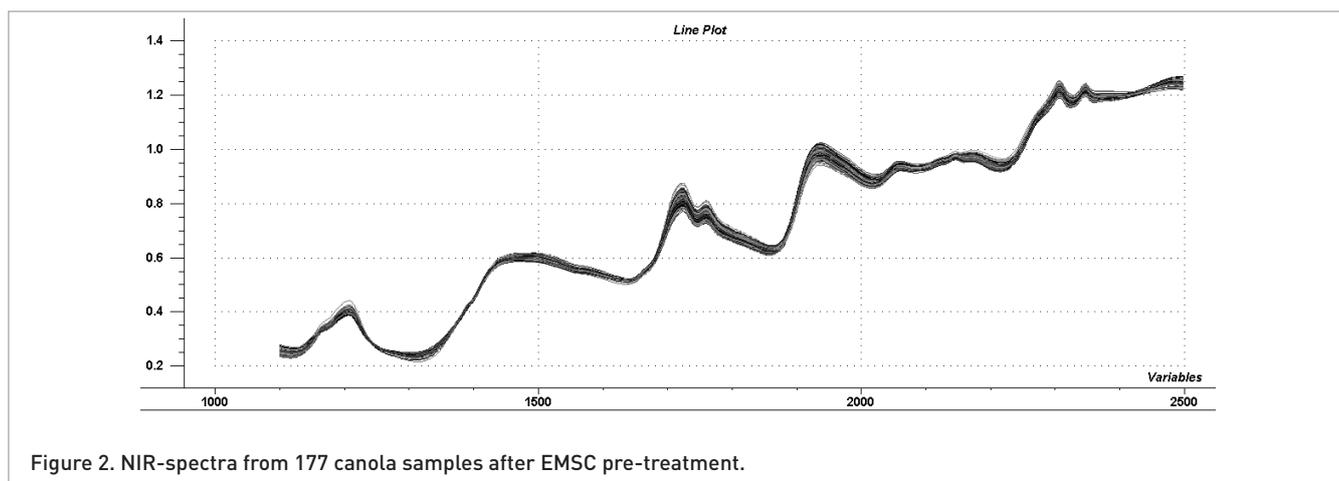


Figure 2. NIR-spectra from 177 canola samples after EMSC pre-treatment.

e.g. IR or Raman spectroscopy, a more detailed interpretation compared to NIR will then be possible.

### Analysis software

The Unscrambler 9.7 (Camo Software AS, Oslo, Norway) was used in order to perform pre-treatment and PLS regressions. CMV with jack-knifing was implemented in Matlab ver. 7.2 (MathWorks Inc., Natick, MA, USA).

## Results

### Spectral pre-treatment

The 177 sample spectra were pre-treated with extended multiplicative scatter correction (EMSC).<sup>7</sup> The EMSC-model was developed using a historical data set of similar canola seed samples, non-overlapping with this set during investigation. Figure 2 shows the spectra after application of the EMSC pre-treatment.

There were several reasons for choosing a scatter correction method for pre-treatment of the spectra. Firstly, the spectra were collected directly from the canola seeds, which will give various degrees of scatter. Secondly, all samples were lying inside the Hotelling  $T^2$  confidence ellipse after EMSC pre-treatment, and no outlier had to be reported. This was not the case for the raw data, where several samples were indicated as outliers. Thirdly, the correlation matrix, **G**, (see text below and Figure 5) had higher absolute values due to removal of the scatter effects which caused a dominant baseline offset in the spectral data.

### Interpretive spectroscopy<sup>8</sup> of the NIR region 1100–2500 nm

NIR spectra are dominated by overtone and combination bands of C–H, O–H and N–H functionalities,<sup>9</sup> whereas the corresponding overtones of the most intense MIR fundamental absorptions (from polar groups like C–O, C=O etc.) are rarely represented. They usually absorb at wavelengths above 6667 nm ( $< 1500\text{ cm}^{-1}$ , fingerprint region). As such, their first

overtones still occur in the mid-IR region. However, most fundamentals that involve light hydrogen absorb at wavelengths smaller than 6667 nm ( $> 1500\text{ cm}^{-1}$ ). Thus, their first overtones already appear in the NIR range.

Since Canola seeds have around 50% oil which consists of fatty acids, the absorptions observed in its NIR spectrum are primarily due to vibrational modes from C–H functional groups. Moreover, C–H group frequencies<sup>8</sup> can be attributed to three main functional groups:<sup>10</sup> (1)  $-\text{CH}_2$  methylene, (2)  $-\text{CH}_3$  methyl and (3)  $-\text{CH}=\text{CH}-$  ethenyl which, in turn, can be assigned to different regions in the canola seed spectrum according to Table 1 and Figure 3.

Figure 3 shows a mean spectrum for canola seed with the six regions, A–F, depicted.

The regions are interpreted as follows:

*Region A:* 2<sup>nd</sup> overtone from C–H stretching mode has mostly information on conjugation<sup>11</sup> and shows an overlapping peak

Table 1. Assignments of major NIR absorption bands for oils and fats.

Region	Wavelength nm	Molecule	Vibration
A	1090–1180	$-\text{CH}_2$	2 <sup>nd</sup> overtone
	1100–1200	$-\text{CH}_3$	
	1150–1260	$-\text{CH}=\text{CH}-$	
B	1350–1430	$-\text{CH}_2$	combination
	1360–1420	$-\text{CH}_3$	
	1390–1450	$\text{H}_2\text{O}$	
C	1650–1850	$-\text{CH}_2$	1 <sup>st</sup> overtone
		$-\text{CH}_3$	
		$-\text{CH}=\text{CH}-$	
D	1880–1930	$\text{H}_2\text{O}$	combination
	2010–2020	$-\text{CH}=\text{CH}-$	
E	2100–2200	$-\text{CH}=\text{CH}-$	combination
F	2240–2360	$-\text{CH}_3$	combination
	2290–2470	$-\text{CH}_2$	

Source: Hourant *et al.*<sup>10</sup> (2000)

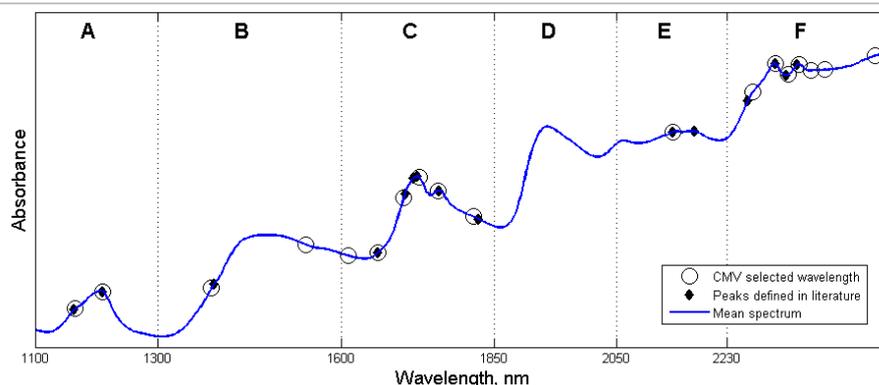


Figure 3. Mean spectrum of canola seed samples with indicated regions A–F<sup>10</sup> (see text and Table 1 and 2 for details).

with a maximum around 1210 nm and a shoulder at 1164 nm. The latter is used for the spectroscopic determination of the Iodine Value (IV)<sup>12</sup> indicating the degree of unsaturation, or equivalently, the number of double bonds.

*Region B:* Combination of C–H stretching and other vibrational modes of the molecule concerned resulting in a peak centred near 1392 nm, perturbed by the 1<sup>st</sup> overtone from moisture at 1440 nm.

*Region C:* 1<sup>st</sup> overtone from C–H stretching shows two prominent bands around 1725 and 1760 nm, characteristically for the 1<sup>st</sup> overtone of C–H stretching vibration of methyl, methylene, and ethenyl groups. According to Hourant *et al.*,<sup>10</sup> oils rich in MUFA have a peak centred near 1725 nm that is referring to oleic acid (C18:1, highly concentrated in canola seed and oil). The assignment of the peak at 1760 nm to the saturated components (–CH<sub>2</sub>, –CH<sub>3</sub>) can be based on two reasonings: (1) the degree of triglyceride unsaturation decreases along the wavelength axis,<sup>10</sup> (2) the less intense 1760 nm absorption corresponds to the lower amount of saturated fatty acids (SFA), such as palmitic acid C16:0 and stearic acid C18:0, that are relatively low (total SFA in

canola seeds is far below 10%). Moreover, Hourant *et al.*<sup>10</sup> mention another maximum at 1824 nm, negatively correlated with both, total amount of PUFA and iodine value and used to discriminate between PUFA and MUFA categories of edible oils and fats.

*Region D* is dominated by the combination band from moisture, but has poor information for the characterisation of oil and fats.

*Region E:* This area includes peaks at 2142<sup>13</sup> and 2176<sup>10</sup> nm that are related to the absorptions of fatty acids having cis double bonds<sup>10,13</sup> from ethenyl –CH=CH–: in canola these are oleic, linoleic and linolenic acid. These bands are derived from combinations of C–H stretching with other vibrational modes.

*Region F:* Contains the most intense absorption bands located in vicinity of 2264, 2308, 2326, and 2344 nm, due to combination of C–H stretching and bending modes<sup>10</sup> of methyl (–CH<sub>3</sub>) and also methylene (–CH<sub>2</sub>) functional groups.<sup>8</sup> The first three peaks mentioned before were used by Hourant *et al.*<sup>10</sup> to classify canola oil, distinguishing it from other types of edible oils.

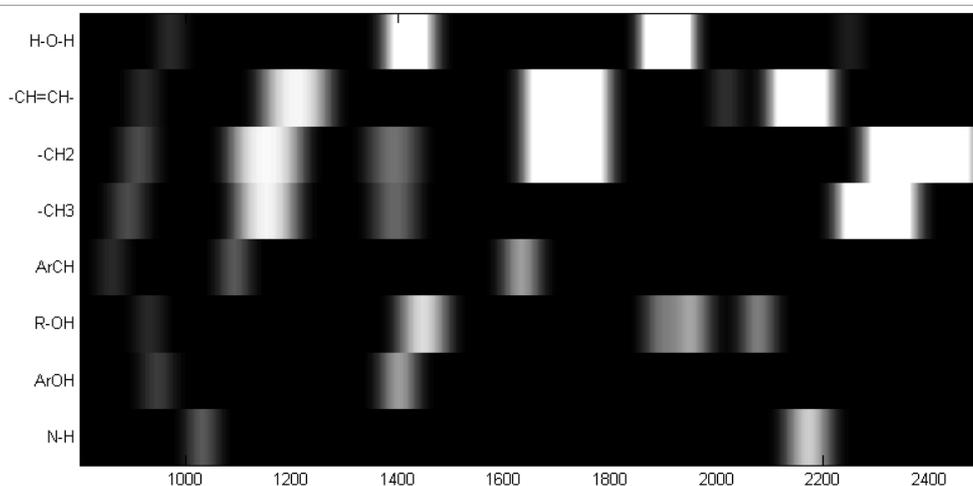


Figure 4. Band assignments after convolution with a Gaussian filter.

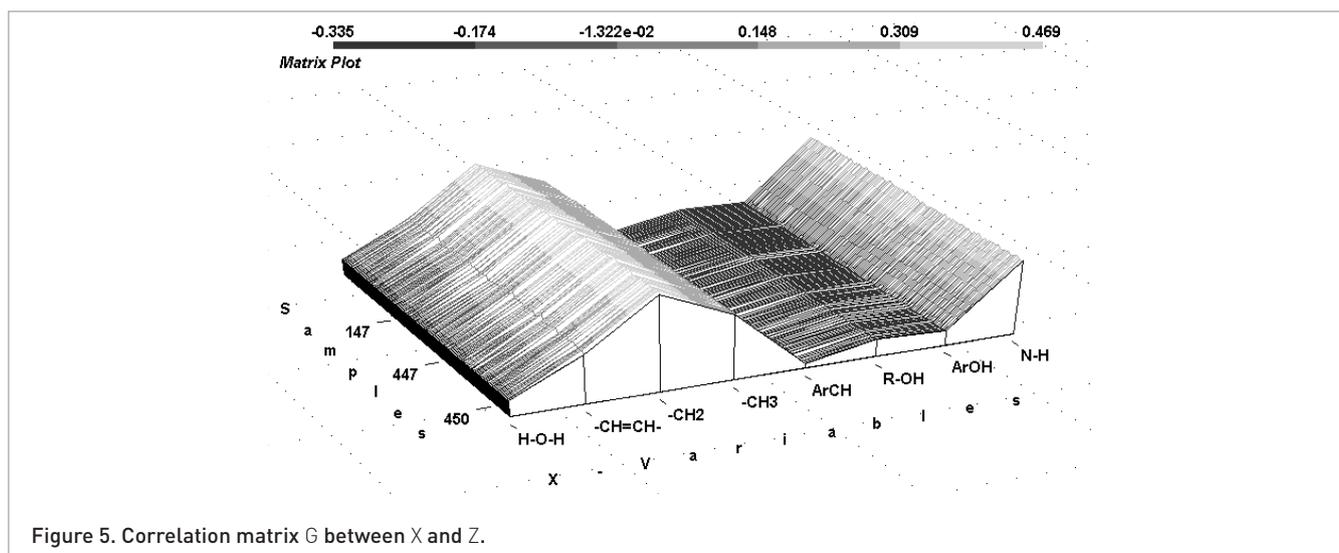


Figure 5. Correlation matrix  $G$  between  $X$  and  $Z$ .

### Band assignment and correlation matrix

Eight group frequencies were assigned to their respective vibrational regions from an NIR-chart,<sup>14</sup> and an initial binary matrix was built using 1 to indicate peaks and 0 elsewhere. In addition, peaks were weighted either as weak (0.5), normal (1), strong (2) or very strong (4). The band assignment matrix was then convolved with a Gaussian filter of size  $N=30$  [corresponding to 60 nm when the spectral resolution is equal to 2 nm] for each group. The resulting band assignment matrix  $Z$  is shown in Figure 4.

The samples from the EMSC pre-treated spectra at these selected wavelengths ( $X$ ) were then correlated with the band assignment matrix  $Z$ , giving the correlation matrix  $G$  shown in Figure 5.

Below follows the interpretation of the correlation matrix  $G$  for the functional groups:

#### H–O–H

The first molecular group for the band assignment is moisture, which was found negatively correlated across this sample set. Moisture is present even in dry canola seeds, having around 10% seed moisture. In the NIR spectrum, broad absorptions from water can clearly be identified around 1440 nm (region B) and 1900 nm (region D).

#### –CH=CH–

The group frequency from ethenyl is correctly positive correlated with the samples because it presents the functional group expressing the degree of unsaturation in the unsaturated oleic C18:1, linoleic C18:2 and especially linolenic C18:3 acids having 1, 2 or 3 *cis*-double bonds, respectively. All of them are present in canola seeds as described above.

#### –CH<sub>2</sub>–, –CH<sub>3</sub>

Methylen together with methyl show the highest correlation, because they are the main functional groups in the fatty acid chains, with the general formula  $\text{CH}_3(\text{CH}_2)_n\text{COOH}$  ( $n$  typically an even number between 12 and 22).

#### Ar–CH, Ar–OH

Aromatic –OH is negatively correlated because it is close to the moisture vibration H–O–H when referring to the correlation loadings plot in Figure 7. The same applies to Ar–CH, which can be found in proximity to –CH<sub>3</sub> very close to the 1<sup>st</sup> principal component (PC). Thus two dimensions or PCs are not enough to distinguish between these types of C–H: aromatic C–H or normal “chain” C–H. However, the influence of Ar–OH and Ar–CH is not at all significant, when referring to the small regression coefficients after nine PCs in Figure 8.

#### R–OH

Because all kinds of fatty acids in oils and fats are forming ester bonds with a glycerol backbone, resulting in triglycerides, due to their breakdown free fatty acids or glycerol itself can also be present.

#### N–H

The N–H group frequencies are positively correlated, and can be explained by the considerable protein content in canola seeds.

### Wavelength selection and regression modelling

The first regression model was run with EMSC pre-treated spectral data using all wavelengths as descriptors ( $X$ ) and concentration of ALA as the response variable ( $Y$ ). The initial model revealed no outlier, and all models in this stage are made on the 177 samples, using full cross-validation. Nine PLS-components were chosen as the optimal number. Variable selection in the inner models in the CMV procedure was performed with uncertainty estimates from jackknifing.<sup>5</sup> The main purpose was to remove non-significant wavelengths to ease interpretation in the following analysis. In total, 182 out of the 700 original wavelengths were found to be significant in CMV. These could further be reduced to 18 frequency regions with connected significant wavelengths, as shown at the bottom of Figure 6, where these 18 frequency

Table 2. Results from initial PLS-regression models using the full wavelength range as well as wavelengths selected by using CMV.

Predictors (X)	Data region	Model rank	RMSECV (Full cross-val.)
Raw data	All wavelengths (700)	10	0.82
EMSC pretreated	All wavelengths (700)	10	0.79
Raw data	CMV selected (182)	8	0.81
EMSC pretreated	CMV selected (182)	7	0.82
Raw data	Peaks from CMV (18)	9	0.81
EMSC pretreated	Peaks from CMV (18)	6	0.81

regions have reached the 100% frequency of significance line. Table 2 indicates the effect of wavelength selection after using CMV and the pre-treatment by EMSC. As an objective criterion to compare the models, the model rank suggested by the software program is shown here, as well as the *RMSECV* from full cross-validation (*RMSECV*). The *RMSECV* may be held against the reference values of ALA varying between 5.0% and 11.1%. As can be seen, the EMSC pre-treatment improved the model by reducing the optimal number of PCs by three when CMV selected peaks are being used.

To simplify the interpretation, only the centre wavelengths of these 18 frequency regions found by CMV shown in Figure 6 (bottom) were selected for further analysis. As indicated by Table 2, a reduced model rank was also achieved when only these wavelengths were used in regression. The 18 centre wavelengths found were then compared with the regions/

wavelengths given in References 10–13. Some were redundant, but most of the wavelengths (apart from a weak band at 2176 nm and one within a cluster of peaks at 1717 nm) known to be important for fat and oils were found by the CMV method. Table 3 shows a list of the wavelengths selected with the two approaches and Figure 3 illustrates this comparison.

In the final step, the 18 significant wavelengths from CMV were combined in an augmented matrix with the correlation matrix **G** shown in Figure 5, constituting an L-model as described in the Method section above (Figure 1). For this model, variable standardisation was necessary due to the difference in value range found in the correlation matrix **G** and the wavelengths **X**. By the use of full cross-validation, the best model rank was found to be nine. With this approach, eight samples were treated as outliers and thus removed. It should be noted that results presented for this L-model in Table 4 are

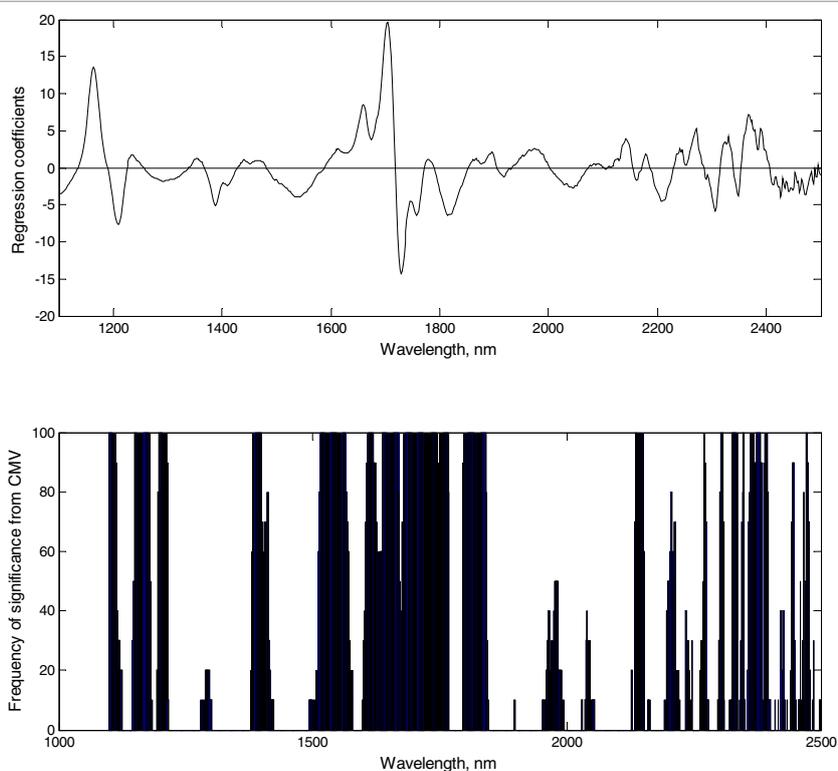


Figure 6. Regression coefficients for a nine-component PLSR model (top) and frequency of significance from cross-model validation (bottom).

Table 3. List of wavelengths selected by CMV compared to wavelengths defined in relevant literature.<sup>10–14</sup>

Region	Description	Important wave-lengths for fat and oils (nm)	Significant wavelengths from CMV, peak maximum position (nm)
A	Degree of unsaturation <sup>10–12</sup>	1164	1166
	-CH 2 <sup>nd</sup> overtone <sup>10</sup>	1210	1210
B	C-H combination <sup>10</sup>	1392	1388
			1542
C	C-H 1 <sup>st</sup> overtone <sup>14</sup> from CH=CH- <i>cis</i> -CH <sub>2</sub> -CH <sub>3</sub>	1660	1660
		1704 C18:3 <sup>10</sup>	1702
		1717 C18:2 <sup>10</sup>	
		1725 C18:1 <sup>10</sup>	1728
		1760 SFA <sup>10</sup>	1758
		1824 <sup>10</sup>	1816
E	CH combination CH=CH- <i>cis</i> <sup>10,13</sup>	2142	2142
		2176	
F	CH combination from -CH <sub>3</sub> <sup>10</sup> 2240–2360 nm	2264	2272
		2308	2308
		2326	2330
		2344	2348
	CH combination from -CH <sub>2</sub> <sup>10</sup>	2290–2470	2366
			2390
			2472

not directly comparable with the results from the initial model (Table 2) used in the estimation of the **G** matrix, due to outlier removal and different weight settings.

A combined correlation loadings plot showing the centre wavelengths from CMV, the functional groups from the NIR chart and the constituent ALA is displayed in Figure 7. In this

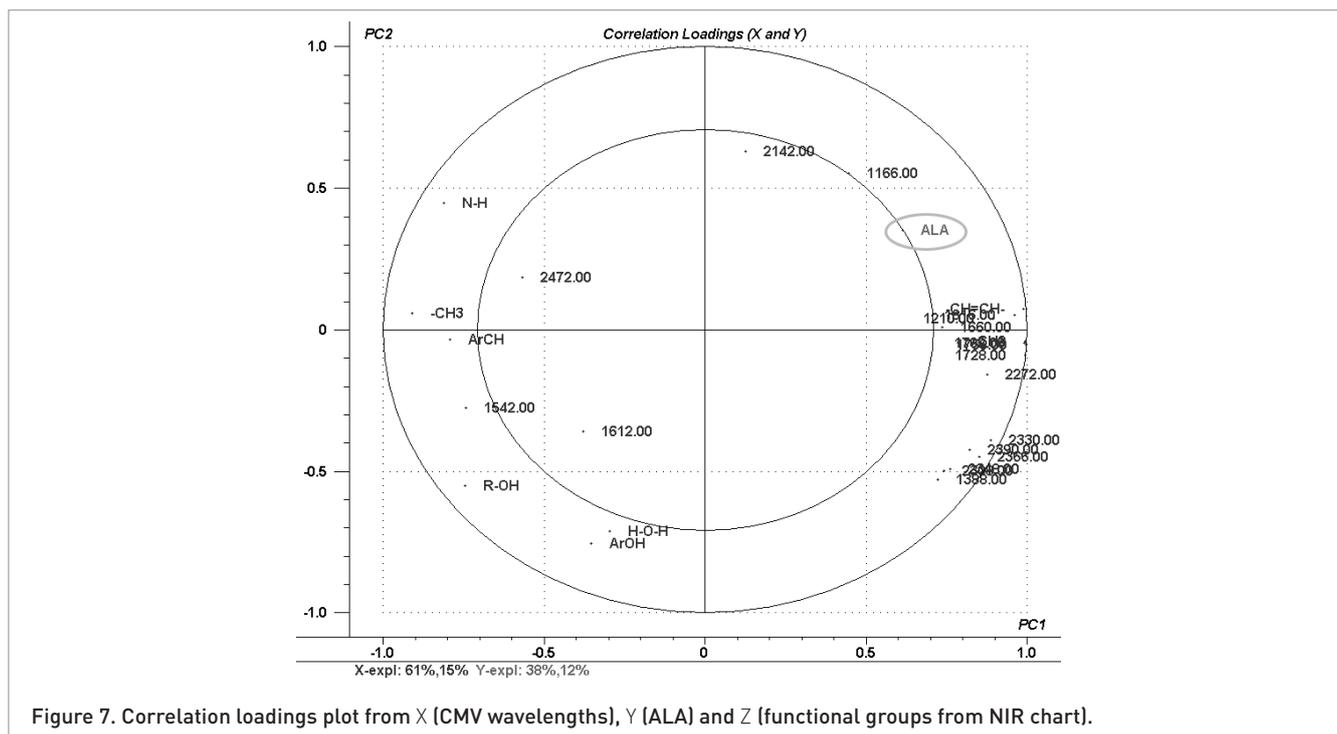


Table 4. Results from different PLS-regression models with EMSC pre-treatment and inclusion of correlation matrix G.

Predictors (X)	NIR data region	Model rank	RMSECV (full cross-val.)
EMSC pretreated+G	All wavelengths (700)	9	0.63
EMSC pretreated+G	CMV selected (182)	9	0.62
EMSC pretreated+G	Peaks from CMV (18)	9	0.62

two-dimensional plot (two PCs) the explained variance is already considerable: 76% in **X** explain 50% of **Y**. Some functional groups are close to relevant wavelengths, for example, C-H<sub>2</sub> and -CH=CH- to the cluster of wavelengths in region C and 1210 (region A). As can be seen, dominant NIR spectral C-H stretching features like methyl C-H<sub>3</sub>, methylene C-H<sub>2</sub> and ethenyl =CH as well as aromatic CH are found to be very close to PC1, whereas combination bands in region B, E and F as well as Ar-OH, H-O-H and R-OH are elsewhere in this plot. However, as described previously, two PCs are not sufficient to distinguish between aromatic C-H and aliphatic C-H. Thus it will be referred to the regression coefficients at rank 9, shown in Figure 8, where light grey bars indicate coefficients corresponding to functional groups. Coefficients shown in black represent the 18 centre wavelengths selected by CMV. The relative importance and sign for these wavelengths are consistent with the regression coefficients from the initial model with all 700 wavelengths [Figure 6 (top)]. Clearly the -CH=CH- and the -CH<sub>2</sub> dominate among the functional groups, which is to be expected from high level oil in canola seed samples. Now the aromatic functional groups have small regression coefficients and, thus, low influence. Two redundant CMV wavelengths (1542nm and 1612nm), that could not be explained from literature as important for fats and oils, have low b-coefficients.

## Conclusions

Theoretical band assignment for various molecular groups can be incorporated with spectral data and constituents by use

of a correlation matrix between spectra and the band assignment table. Jack-knifing with cross-model validation acts as a filter to remove the wavelengths of no interest. The correlation matrix, **G**, can be combined with selected spectral wavelengths in the model so that chemical background information pertaining to the actual measured spectra is visualised. With 16 out of 18, the majority of the wavelengths found by the CMV selection procedure gave plausible explanations from the chemistry of canola seed. This enhances the interpretation of the regression model and existing chemical knowledge is confirmed from this empirical analysis. On the other hand, findings that at first glance are not consistent with established theory may give new application specific information, for example, about impurities in the samples. Further research is required to interpret data sets with multiple **Y** variables.

## References

1. U. Hjort, *Computer Intensive Statistical Methods*. Chapman and Hall, London, UK, p. 40 (1994).
2. E. Anderssen, K. Dyrstad, F. Westad and H. Martens, "Reducing over-optimism in variable selection by cross-model validation", *Chemometr. Intell. Lab. Syst.* **84(1/2)**, Special issue. p. 69 (2006).
3. L. Nørgaard and R. Bro, *Proceedings of International PLS 99, CISIA - CERESTA, France*, p. 187 (1999).
4. F. Westad, N.K. Afseth and R. Bro, "Finding relevant spectral regions between spectroscopic techniques by use of cross model validation and partial least squares

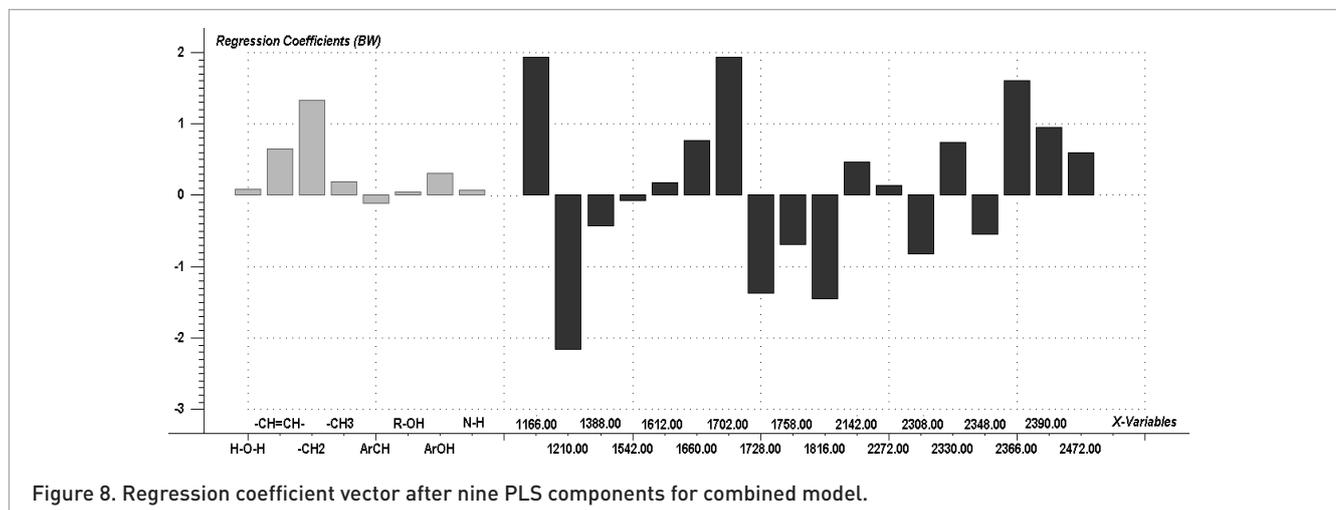


Figure 8. Regression coefficient vector after nine PLS components for combined model.

- regression", *Anal. Chim. Acta* **595**, 323 (2007). doi: [10.1016/j.aca.2007.02.015](https://doi.org/10.1016/j.aca.2007.02.015)
5. F. Westad and H. Martens, "Variable selection in near infrared spectroscopy based on significance testing in partial least squares regression", *J. Near Infrared Spectrosc.* **8**, 117 (2000).
  6. H. Martens, E. Anderssen, A. Flatberg, L.H. Gidskehaug, M. Høy, F. Westad, A. Thybo and M. Martens, "Regression of a data matrix on descriptors of both its rows and of its columns via latent variables: L-PLSR", *Comput. Stat. Data Anal.* **48(1)**, 103 (2005). doi: [10.1016/j.csda.2003.10.004](https://doi.org/10.1016/j.csda.2003.10.004)
  7. H. Martens, J. Pram Nielsen and S. Balling Ellingsen, "Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures", *Anal. Chem.* **75(3)**, 394 (2003). doi: [10.1021/ac020194w](https://doi.org/10.1021/ac020194w)
  8. J. Workman, "Interpretive spectroscopy for near infrared", *Near infrared spectroscopy: the future waves*, Ed by A.M.C. Davies and Phil Williams, NIR Publications, Chichester, West Sussex, UK, p. 6 (1996).
  9. H. Siesler, "Quality control and process monitoring by vibrational spectroscopy", *NIR news* **11(3)**, 9 (2000).
  10. P. Hourant, V. Baetten, M.T. Morales, M. Meurens, R. Aparicio, "Oil and fat classification by selected bands of NIRS", *Appl. Spectrosc.* **54(8)**, 1168 (2000). doi: [10.1366/0003702001950733](https://doi.org/10.1366/0003702001950733)
  11. M. Snieder, N.A.G. Dekker, S.C.C. Wiedemann and W.G. Hansen, "Determination of very low unsaturation levels in oleochemical by NIR", in *Near infrared spectroscopy: stretching the NIR-spectrum to the limit*, Ed by A.M.C. Davies and A. Garrido-Varo, NIR Publications, Chichester, West Sussex, UK, p. 999 (2004).
  12. *AOCS Standard Procedure for Iodine Value (IV)*, ABB Application Note (2004).
  13. D. Behmer, A. Montasell and C. Villar Pascual, "Applications with a Mediterranean flavour: quality control of olives and olive oil with FTNIR", in *Near infrared spectroscopy: NIR in action—making a difference*, Ed by G.R. Burling-Claridge, S.E. Holroyd and R.M.W. Sumner. New Zealand NIRS Society Inc., Hamilton, New Zealand, p. 330 (2007).
  14. Guide for Infrared Spectroscopy, p.4 <http://www.brueke-optics.de/downloads/index.html>.