Characterisation and classification of vegetable oils by combining near and mid infrared signals

Nathalie Estephan[a], Antonio Barros[b], Ivonne Delgadillo[b], Douglas N. Rutledge[a]

[a] *Laboratoire de Chimie Analytique, Institut National Agronomique INA P-G, 16 rue Claude-Bernard, 75231 Paris Cedex 05, France*

[b]*Departemento de Quimica, Universidade de Aveiro, 3810 Aveiro, Portugal*

Introduction

Infrared spectroscopy is a powerful analytical tool for determining various constituents in agricultural and food products because it is a fast, non-destructive, multi-analytical technique and it is not dependent on highly skilled personnel to operate the instrumentation.[1]

Chemometric methods, such as point-by-point analysis of variance (ppANOVA), principal component analysis (PCA), discriminant factor analysis (DFA) and partial least squares regression (PLS) can be used to detect and extract the information contained in infrared signals, and then optimise the use of that information for a particular application.[2,3]

The signals observed in the NIR region correspond to combinations and harmonics of the fundamental molecular vibrations observed in the Mid infrared (MIR) spectra. Therefore, it should be possible to have a better assignment of the NIR wavelengths by studying the two signals conjointly. The combination of the two signals should also increase the predictive ability of regression or discriminant models created using multivariate chemometric methods. This simultaneous analysis may be done by outer product analysis (OPA) which can reveal how the signals vary simultaneously as a function of some property, such as physico–chemical parameters.

The OPA calculates for each sample the product of intensities at all combinations of frequencies in the two domains to produce an outer product matrix. The complete set of OP matrices may then be analysed using chemometric techniques such as PCA, FDA or PLS. Plots of loadings, B coefficient and discriminant function are drawn to visualise the simultaneous variations in the two domains as a function of the predicted property or classification criterion.

In this study, this method is used to characterise nine groups of vegetable oils and to classify them as a function of their physico-chemical properties.

Materials and methods

Data set

- Vegetable oil samples : olive oil (ov, ovv), sunflower (tb), soy bean (sr, sh), peanut (ar), grapeseed (pr), sesame (se) and a mixture of refined oils (mr).
- Samples were provided by Bipea "Bureau inter-professionnel d'études analytiques".
- The values for a range of physico-chemical properties of the samples were obtained from inter-laboratory analyses.

Apparatus

A Fourier transform NIR and MIR spectrometer (Bruker Vector 33) with spectral acquisition software 'Opus' was used to do the measurements.

Acquisition parameters :

	NIR	**MIR**
Mode of acquisition	Diffuse reflection (Thoma cell)	Attenuated total reflection (ATR)—ZnSe
Detector	Integration sphere	DTGS
Spectral range	4000–9500 cm^{-1}	700–4000 cm^{-1}
Resolution	4 cm^{-1}	4 cm^{-1}
Sample scans	64	64
Reference scans	64	64
Reference	Air	Air

Chemometrics

- **ppANOVA** - calculates for each variable, one after the other, the part of the total variability due to the samples belonging to particular groups. It also calculates the residual variability not due to the groups.[6]
- **PCA -** A set of n linear combinations (PC1 - PCn) of the n original variables (X1 - Xn) :

$$PC_i = X_1 \,.\, u_{i1} + X_2 \,.\, u_{i2} + ... + X_n \,.\, u_{in} \qquad (1)$$

$$PC_i = X \,.\, u^T \qquad (2)$$

Calculated so that the first PCs point in the direction of greatest dispersion of the samples in the variable space. These PCs may be viewed as a set of new axes in the multidimensional space of the original variables.[6]

- **PLS -** models the relationship between a set of predictor variables X (n objects x k variables) and a set of response variables Y (n objects x m responses).[4]

In this study, there is only one response (physico-chemical parameter), so Y has dimensions (*n* objects × 1 response).

The PLS regression procedure may be written as :

$$Y = XB + E \qquad (3)$$

The regression model is generated by calculating the B coefficients matrix that minimises the error matrix E.

- **OPA** – For each sample, the Outer Product (OP) calculates the product of the intensities for all combinations of frequencies in the two domains. The two vector-signals of each sample thus give an OP matrix.

Therefore, for each sample, the values of one signal are weighted by the values of the other.
For each sample i, if we have two initial signal-vectors **x** and **y**, the vector $\mathbf{x}^{T}_{i}$ (1,m) in the X domain is multiplied by the vector $\mathbf{y}^{T}_{i}$ (1,p) in the Y domain to obtain an OP matrix for the sample i with a size (m, p).
The OP matrix of sample i can be unfolded to give an OP vector of size (1,m*p).
For the n samples, the n OP vectors are concatenated to give a matrix K (n,m*p).
The analysis of the matrix K gives us as results vectors such as **b** coefficient, loadings and discriminant factors of sizes (1,m × p).
These vectors are then folded back to give outer product (OP) matrices which facilitate the detection of relations between the variables in the two domains (Figure 1).

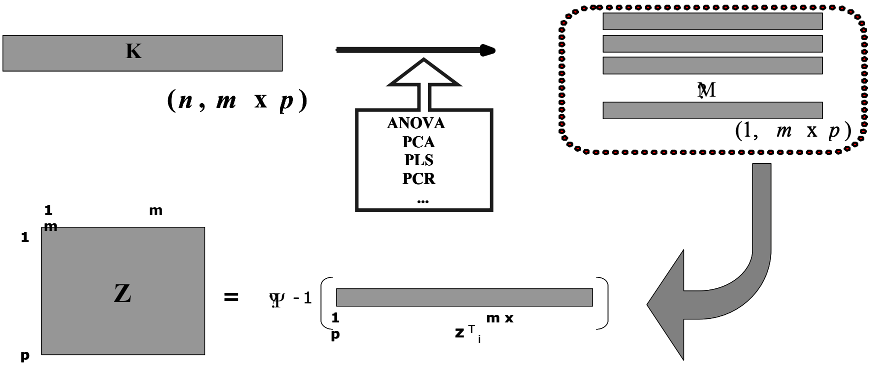

Figure 1. Principle of outer product analysis.

Using techniques like ppANOVA, PLS, PCR, it is possible to selectively highlight a particular source of variation.
By selecting profiles through the resulting folded OPA matrices, it is possible to artificially increase the resolution of signals.

Results and discussions

FDA on MIR and NIR

To see the distribution of the samples and to maximise the separation of the predefined groups, an FDA was applied on the matrix created by performing OP between the MIR and NIR spectra of the oils after centring and reducing the columns in the matrix, corresponding to the variables.

The results show that the greatest separation of the samples is along the first discrimant factor (DF1) where the olive and the peanut oil groups are separated from the other oil groups (Figure 2). The rate of correct classification of the samples into the oil groups is higher with the OP (MIR⊗NIR) than with the two sets of spectra taken separately.

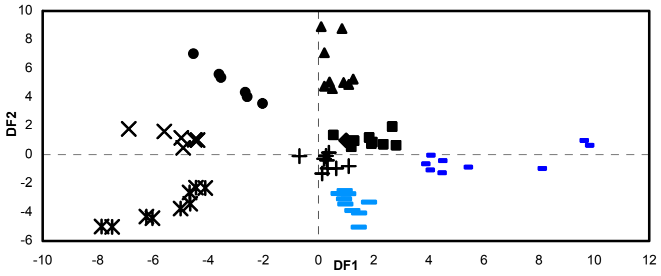

Figure 2. Scatter plot of DF1 vs. DF2. [♦ar x ov ✱ ovv • pr – sr — sh + se ■ tb ▲ mr].

PLS on MIR and NIR

Regression models were built as a function of the two physico-chemical parameters: unsaponifiables (INSAP) and the sterols (STER). These models were created by using partial least square regression (PLS) on the standard normal variant (SNV) pre-treated MIR and NIR matrices.

The number of the latent variables (LV) used to build each predictive model for each parameter was determined by *internal cross validation (leave-3-out)*.

The folded b vector was plotted to see how the spectral regions behave with the variation of each parameter in the samples (Figures 3 and 4).

In Figure 3, the regions of the folded b vector which are positive are correlated with the evolution of the INSAP parameter in the samples. These regions correspond to the following simultaneous variations of two sorts of correlation:

1) the 1^{st} overtone of –C–H of saturated hydrocarbons in NIR at 5683 and 5810 cm^{-1} with: 725, 1039, 1081, 1120, 1137, 1166, 1187 and 1242 cm^{-1} in MIR.

2) the 1^{st} overtone of =C–H of unsaturated hydrocarbons in NIR at 5995 and 6040 cm^{-1} with: 916, 950, 968, 985, 1068, 1106, 1153 and 1390 cm^{-1} in MIR.

For the regions which are negatively correlated, the simultaneous variations evolve in the opposite way to the INSAP. These regions correspond again to two sorts of correlation:

1) the 1^{st} overtone (ov) of –C–H of saturated hydrocarbons in NIR at 5683 and 5810 cm^{-1} with: 916, 950, 968, 985, 1068, 1106, 1153 and 1390 cm^{-1} in MIR.

2) the 1^{st} ov. of =C–H of unsaturated hydrocarbons in NIR at 5995 and 6040 cm^{-1} with: 725, 1039, 1081, 1120, 1137, 1166, 1187 and 1242 cm^{-1} in MIR.

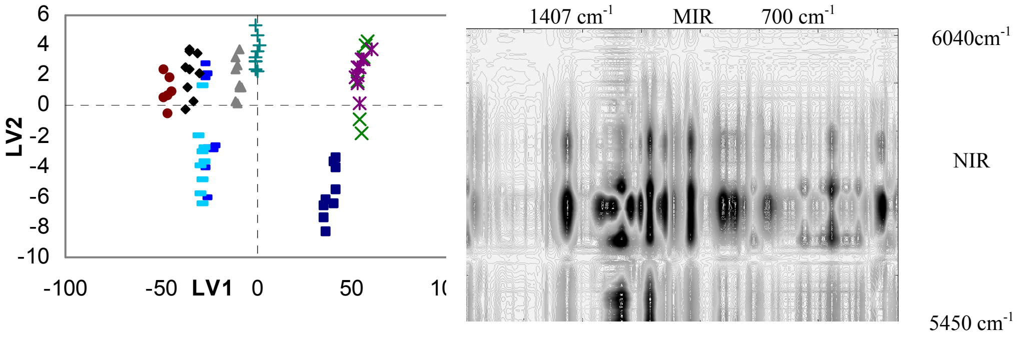

Figure 3. Comparison of the scatter plot of LV1 vs. LV2 with the folded b vector for the prediction model for INSAP (2LVs, RMSEC 16.5%). [■ ar x ov ✱ ovv • pr – sr — sh + se ♦ tb ▲ mr]

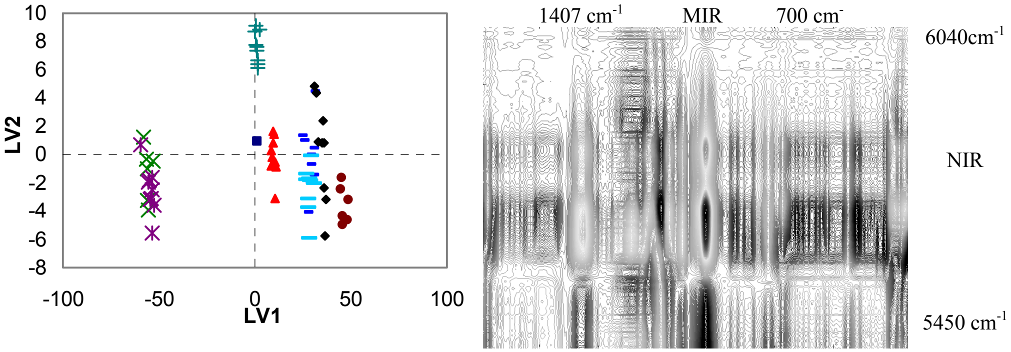

Figure 4. Comparison of the scatter plot of LV1 vs. LV2 with the folded b vector for the prediction model for STER (2LVs, RMSEC 21.18%). [■ ar x ov ∗ ovv • pr – sr — sh + se ♦ tb ▲ mr].

In the literature,[5,8] the attributions of the peaks in the fingerprint region of MIR is still somewhat ambiguous but these results show, through the correlation with the NIR bands, that one series of peaks is clearly associated with unsaturation and another series with saturation. We can also see that the peaks in MIR which are positively related to the bands in NIR corresponding to the unsaturated bonds, are negatively related to bands in NIR corresponding to saturated ones, and vice versa.

In figure 4, we can also notice two sort of correlation in the folded b vector,:

1) positive correlation (related to simultaneous variation evolving in the same direction as STER) for the 1st ov. –C–H of saturated hydrocarbons in NIR at 5681 and 5828 cm^{-1} and negative correlation (related to the simultaneous variations evolving oppositely to STER) for the 1st ov. of =C–H of unsaturated hydrocarbons in NIR at 6040 cm^{-1} with:

713, 732, 916, 939, 1039, 1081 and 1242 cm^{-1} in MIR.

2) positive correlation for the 1st overtone –C–H of saturated hydrocarbons in NIR at 5681 cm^{-1} and 5828 cm^{-1} and for the 1st overtone of =C–H of unsaturated hydrocarbons in NIR at 6026 cm^{-1} with:

1114, 1147 and 1166 cm^{-1} in MIR.

By comparing the correlations for the two parameters (INSAP and STER), it can be seen that most of the MIR peaks are the same and behave in the same way in relation to saturation. This can be explained by the fact that sterols are an important part of the unsaponifiables.

Conclusion

PCA and PLS applied to the OP matrices highlighted simultaneous variations in the two domains as a function of the classification criteria INSAP and STER. By examining the relations between bands in the two domains related to the property being predicted, they also allow us to have a better understanding on how and why the variables contribute to the model.

The combination of the two domains allows the extraction of complementary information about spectral characteristics of the oils.

OPA facilitates the interpretation of differences between samples. It has the advantage of being able to produce predictive and discriminant models, something which is not possible with other methods of combining spectra, such as classical 2D Correlation Spectroscopy.

References

1. R. Wilson, *Spectroscopic Techniques for Food Analysis*. VCH Publishers, Weinheim, Germany, pp. 13-52, (1994).
2. D.N. Rutledge, A.S. Barros and F. Gaudard, *Magn. Res. Chem*. **35,** S13 (1997).
3. D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte and L. Kaufman, *Chemometrics : A Textbook*. Elsevier, Amsterdam, The Netherlands (1998).
4. D.N. Rutledge, A.S. Barros and R. Giangiacomo, *Interpreting Near Infrared Spectra of Solutions by Outer Product Analysis With Time Domaine-NMR, Magnetic Resonance in Food Science*. The Royal Society of Chemistry, Cambridge, UK, pp. 180–192, (2001).
5. J. Harwood and R. Aparicio, *Hanbook of olive oil- Analysis and Properties*. Aspen Publications, Maryland, USA, pp. 209–242 (2000).
6. D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi and J. Smeyers-Verbeke, *Data Handling in Science and Technology, Handbook of Chemometrics and Qualimetrics*: *Part A*. Elsevier, Amsterdam, The Netherlands, pp. 20A, 121–150, 519–556 (1997).
7. D. Boskou, *Olive oil: Chemistry and Technology*. AOCS Press, Illinois, USA, pp. 52–83 (1996).
8. D. Bertrand and E. Dufour, *La spectroscopie infrarouge et ses applications analytiques*, TEC & DOC, Paris, France, pp. 147–157 (2000).