Analusis, **1998**, *26*, 310-316

Generalised Canonical Correlation Analysis for the interpretation of fluorescence spectral data

M.-F. Devaux[1,*], P. Courcoux[2], E. Vigneau[2] and B. Novales[1]

[1] *INRA – LTAN, BP. 71627, 44316 Nantes Cedex 03, France*
[2] *ENITIAA/INRA, Statistique Appliquée à la Caractérisation des Aliments, rue de la Géraudière, 44072 Nantes, France*

Abstract. The paper reports an application of Generalised Canonical Correlation Analysis to fluorescence spectral data. Emission fluorescence spectra can be recorded for several excitation wavelengths and can be presented as 3-way data tables. The objectives of the data treatment are to describe and compare the samples by taking into account all the emission spectra, and to reveal characteristic excitation and emission wavelengths. Generalised Canonical Correlation Analysis has been tested on fluorescence emission spectra acquired for binary mixtures of raw materials in the food domain. The application of the method within the context of spectral data is presented.

Key words. Generalised Canonical Correlation Analysis – fluorescence spectroscopy – multi-way data table.

Introduction

Fluorescence analysis is a developing method in the food domain for the purpose of quality control. Fluorescence spectroscopy has been shown to be a selective and sensitive technique which makes it possible to identify, characterise and quantify chemical compounds in a variety of samples [1]. Fluorescence refers to the light emitted by molecules during the period they are excited by photons in the ultra violet or visible range. The spectral data are acquired by lighting the samples at a given wavelength called “excitation wavelength” and by recording emitted light for a range of wavelengths called “emission wavelengths”. Food products are generally complex and the differentiation of products may require to analyse together several emission intensities recorded for several excitation wavelengths. As spectral data depends both on excitation and emission wavelengths, several sets of emission spectra can be acquired for a same set of samples and fluorescence data can be arranged according to a 3-way structure (Fig. 1). The 3 dimensions correspond

* Correspondence and reprints.
Received January 29, 1998; revised August 25, 1998; accepted August 31, 1998.

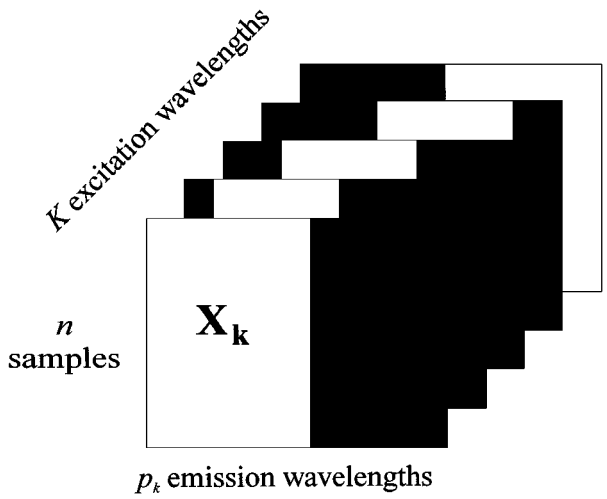

Fig. 1. Three-way structure of fluorescence spectral data.

to the samples, the excitation wavelengths and the emission wavelengths, respectively. Each data table X_k describes the fluorescence intensities obtained for the n samples at the k^{st} excitation wavelength. The spectral range of the p_k emission wavelengths generally differ from one table to the other.

The objective of the present work was to find a suitable data treatment that takes into account the 3-way structure of the tables for analysing fluorescence spectral data. The approach was focused on the description of the data and on the highlight of notable points in each of the 3 dimensions. In particular, specific excitation or emission wavelengths were searched for and a global comparison of the samples from all the spectral data was expected. Moreover, fluorescence data treatment had to take into account the collinearity among the emission wavelengths caused by the digitisation of the continuous spectra.

Usual multidimensional data treatments widely applied on spectral data such as Principal Component Analysis or PLS are designed to analyse single data tables [2]. The use of multi-way methods in chemistry, spectroscopy and more generally in chemometrics is growing. Procedures such as multilinear PLS [3], PARAFAC [4] or more general Tucker models [5] are discussed by several authors. In the present work, Generalised Canonical Correlation Analysis was studied [6]. The method, though not new, has been little applied because of some difficulties to interpret the results obtained. Canonical correlation analysis, which makes it possible to jointly study two data tables, has been successfully applied to mid and near infrared spectral data [7]. The generalisation of the method proposed by Carrol was applied in the present paper to fluorescence spectral data.

Generalised Canonical Correlation Analysis

The objective of generalised canonical correlation analysis (GCCA) is to describe the correlations between K data tables observed for the same n samples. The method is a generalisation of canonical correlation analysis which study the correlations between 2 groups of variables by assessing, in each group, linear combinations that are as correlated as possible. The generalisation of the method to more than 2 data tables requires to choose a criterion for measuring the correlation between K data tables. The most usual criterion consists in assessing canonical variates z^j so that the sum of the squared correlation coefficients between them and each data table is maximum [6]. Considering the K centred tables X_k, the criterion corresponding to the j^{st} canonical variate can be written as [8]:

$$m^j = \sum_{k=1}^{K} z^{j\prime} X_k (X_k' X_k)^{-1} X_k' z^j \,. \qquad (1)$$

The variates z^j are determined in order to maximise m^j with the constraints of being normalised and orthogonal to each other:

$$z^{i\prime} z^j = 1 \text{ and } z^{i\prime} z^j = 0 \text{ for } i \neq j. \qquad (2)$$

The criteria m^j are maximum when the canonical variates are eigenvectors of:

$$P = \sum_{k=1}^{K} X_k (X_k' X_k)^{-1} X_k' \,. \qquad (3)$$

As the spectral data are highly correlated, assessing $(X_k'X_k)^{-1}$ is generally difficult. A solution consists in replacing the raw data by their principal components. Principal components are separately determined for each original data table by:

$$C_k = X_k U_k \qquad (4)$$

where U_k are the eigenvectors of the variance covariance matrix $(X_k'X_k)$. Canonical variates are therefore assessed from the principal components C_k as eigenvectors of:

$$Q = \sum_{k=1}^{K} C_k L_k^{-1} C_k' = C L^{-1} C' \qquad (5)$$

where $L_k = C_k'C_k$ is the diagonal matrix of the eigenvalues of $(X_k'X_k)$, L is the diagonal matrix which diagonal elements are the eigenvalues L_1, ..., L_k, ..., L_K and C is the data table obtained by merging the K data tables C_k:

$$C = \left[C_1 \middle| \ldots \middle| C_k \middle| \ldots \middle| C_K \right]. \qquad (6)$$

Generalised Canonical Correlation Analysis applied on principal components therefore corresponds to the principal component analysis of the normalised components.

The results of the analysis are the eigenvalues λ^j of Q, the canonical variates $Z = (z^j)$ and the eigenvectors A of $L^{-1}C'C$. The eigenvalues correspond to the maximised values of the criteria m^j. The canonical variates makes it possible to draw similarity maps of the samples. The canonical eigenvectors A can be used to relate the canonical variates and the original variables corresponding to the emission wavelengths. A relation between X_k, the canonical variates Z and the sub-matrix of A, noted A_k, corresponding to C_k can be written:

$$X_k = Z \Lambda^{1/2} A_k' L_k U_k' \qquad (7)$$

where Λ is the diagonal matrix of eigenvalues λ^j.

The columns of

$$S_k = A_k' L_k U_k' \qquad (8)$$

are linear combinations of the eigenvectors U_k. When principal component analysis is applied to spectral data, the

eigenvectors can be drawn as "*spectral patterns*" and interpreted in a spectroscopic way [9]. Linear combinations of eigenvectors can still be interpreted as spectral patterns [7]. In the case of canonical correlation analysis, they are called "*canonical spectral pattern*" and reveal the contribution of each emission wavelengths to the canonical variates.

The canonical variates and the spectral pattern make it possible to compare the samples and the emission wavelengths, respectively. A comparison of the excitation wavelengths can be obtained by examining each individual multiple correlation coefficients between the canonical variates and the data tables. These values correspond to the contribution of each table to the criterion m^j.

Application

The method was applied to fluorescence spectra measured for a sample set composed of binary mixtures of raw materials.

Samples

Eight raw materials have been collected: 4 cereals (wheat, barley, triticale and maize) and 4 oilcakes (soya, rapeseed, sunflower and groundnuts). Each material was ground in a hammer mill fitted with a 1 mm grid. Binary mixtures were obtained by mixing one cereal with one oilcake in the proportion 25% – 75%, 50% – 50% and 75% – 25%. The complete sample set contained 48 mixtures and 8 raw materials.

Fluorescence spectra

Emission spectra were recorded with a spectrofluorimeter SLM 4800C (SLM Instruments, Illinois). The powdered samples were placed in a 1 × 1 cm quartz cell without any other preparation. The spectral acquisition was performed with a front face fluorescence device at a 56° angle in order to minimise the specular reflection of the excitation light. The absorption spectrum of rhodamin B in pure ethanol solution was used as reference spectrum. For each sample, the emission spectrum was obtained by assessing the ratio between the sample and the reference spectra.

Emission spectra were recorded for 8 excitation wavelengths varying between 290 and 440 nm. Table I give the excitation wavelengths and the range of the corresponding emission wavelengths. The spectra were not recorded for the same range of emission wavelengths in order to avoid the recording of the excitation peaks together with the emission peaks. The spectra were digitised each 2 nm.

Number of principal components

The 8 excitation wavelengths lead to 8 different spectral sets described in 8 data tables. Each table was first transformed by principal component analysis before being submitted to generalised canonical correlation analysis. Components with a variance equal to 0 had to be discarded. Principal components being ranged in decreasing order of variance, only the first ones were considered. In the present work, the number of components was chosen identical for each data table. In order to select a small number of significative components, the choice was based on the criterion:

$$c(p,g) = \sum_{j=1}^{g} m^j(p) \qquad (9)$$

where p is the number of principal components considered in each data tables, g is the number of canonical variates calculated and $m^j(p)$ is the criterion obtained for the j^{st} canonical variate when p components are included in the analysis. The evolution of $c(p,g)$ according to p and g was examined.

Results and discussion

Spectra

Figure 2 shows the spectra of soya, wheat, rapeseed and binary mixture composed of 50% of wheat and rapeseed. The spectra of wheat were very different from the 2 other raw materials. The wheat spectra exhibited a strong emission peak around 320 nm after an excitation at 290 nm. A peak was also observed around 420 nm for excitation wavelengths 310, 330 and 360 nm. In the case of the soya and rapeseed sample, emission peaks were observed around 500 nm after an excitation between 390 and 440 nm, and around 410 and 470 nm after an excitation at 330 and 360 nm. The mixture exhibited emission peaks from the 2 raw materials. Large variations in the intensity of the spectra could also be observed.

Number of principal components

Figure 3 shows the evolution of the criterion $c(p,g)$ according to the number of principal components considered in the data tables and to the number of canonical variates calculated. Each curve represents the criterion obtained for a same number of components. The criterion did not much increase after taking 5 principal components into account. Few additional correlation could be found by including more principal components in the analysis. Canonical variates describe the correlation between the data tables. The curves in figure 3 indicated that 5 canonical variates were sufficient to describe almost all the correlation. The other variates did not participate much to the criterion. The following results were therefore obtained by choosing the first 5 principal components in each data table and by examining the first 5 canonical variates.

Table I. Excitation and emission wavelengths for spectral acquisition.

Excitation wavelengths (nm)	*Emission wavelengths (nm)*	*Number of wavelengths*
290	290 – 540	125
310	360 – 570	105
330	354 – 590	118
360	370 – 620	125
390	410 – 650	120
400	430 – 630	100
420	440 – 640	101
440	470 – 650	90

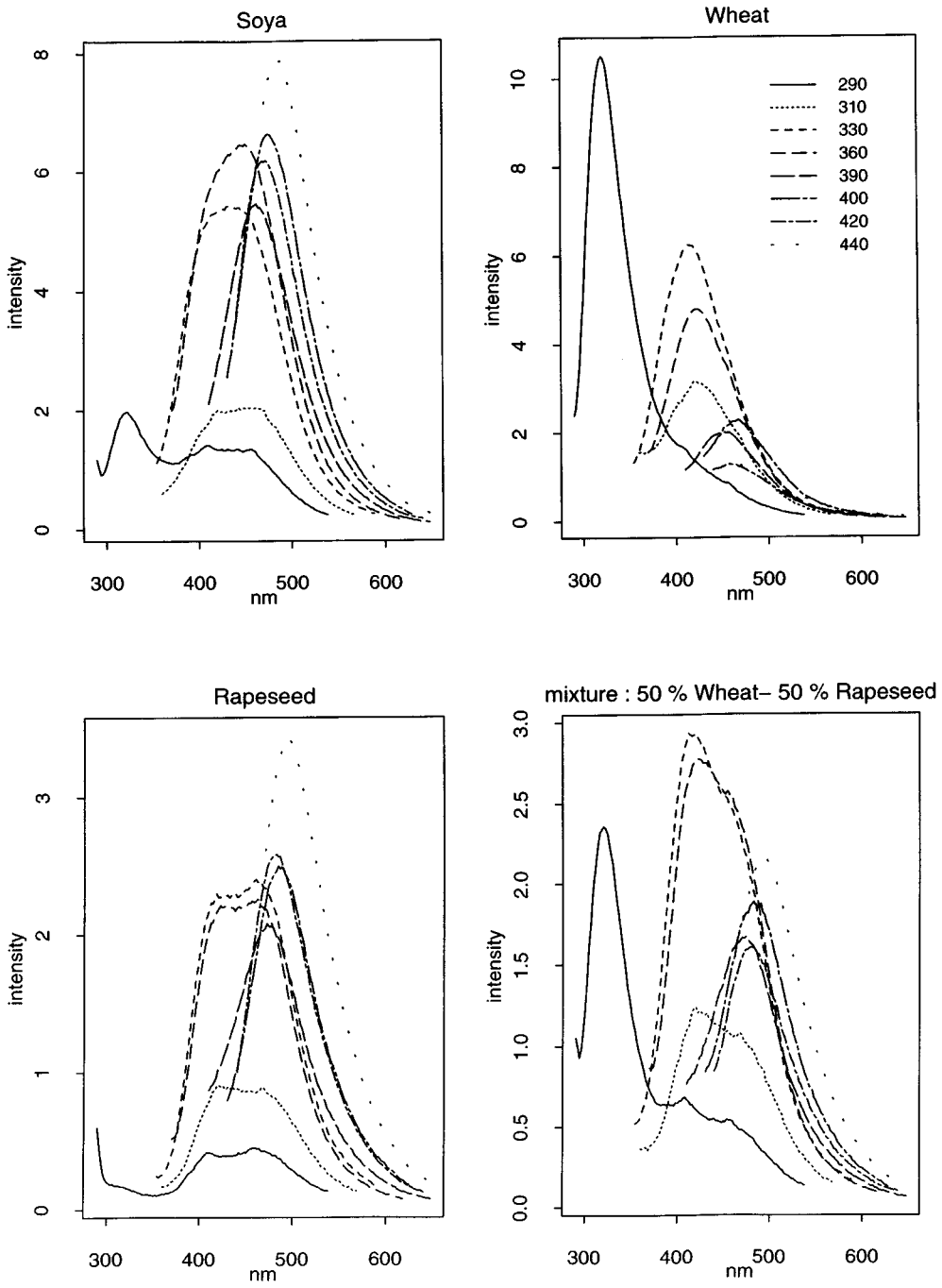

Fig. 2. Fluorescence spectra of soya, wheat, rapeseed and a 50% – 50% mixture of wheat and rapeseed. The 8 emission spectra corresponding to the 8 excitation wavelengths are drawn on the same figure.

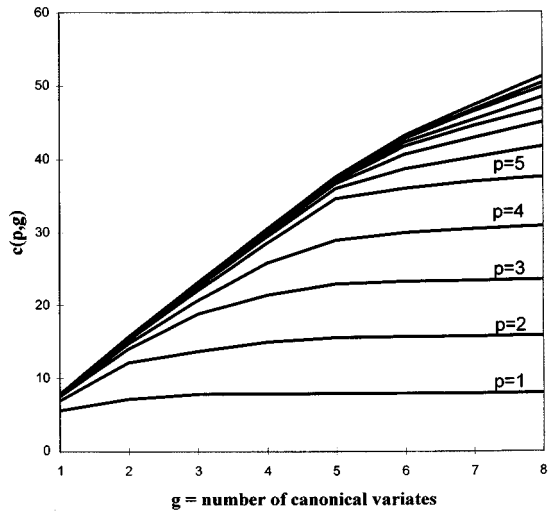

Fig. 3. Choice of the number of principal components to consider in generalised canonical correlation analysis. Criterion *c(p,g)* according to the number *p* of principal components considered and the number *g* of canonical variates calculated.

Comparison of the data tables

Table II gives the values of the squared multiple correlation coefficients observed between each data tables and the first 5 canonical variates. The criteria m^j corresponding to each canonical variate are also given. The m^j values are the sum of the squared multiple correlation coefficients and were therefore at most equal to 8. The m^j values observed for the first two canonical variates were higher than 7 indicating that all the data tables highly contributed to the information they described. The values observed for z^1 were higher than 0.95 for each data tables and those observed for z^2 were higher than 0.92 excepted for the table corresponding to an excitation wavelength of 290 nm. Canonical variate z^3 mainly described a strong correlation observed for the excitation wavelengths between 360 and 420 nm. Generalised canonical correlation analysis therefore revealed the data tables highly correlated in the first 5 canonical variates. The values in table II showed that the information obtained by an excitation at 290 nm was the most different in comparison to the other tables.

Comparison of the samples

Canonical similarity maps are obtained from the canonical variates taken by pairs. Figure 4 shows the similarity maps of canonical variates 1 and 2. Lines have been drawn that join some of the raw materials and their binary mixtures. Soya was clearly identified by canonical variate z^1 and appears very different from all the other raw materials. As the first canonical variate described an information common to all data tables (Tab. II), variate z^1 indicated that soya showed characteristics fluorescence properties for all the excitation wavelengths. Rapeseed was opposed to the 4 cereals according to the second variate. This second characterisation could also be found in all the data tables as revealed by the multiple correlation coefficients. The mixtures were found in intermediate position between the raw materials. Figure 5 shows the canonical similarity map of variates 3 and 4. Variate z^3 made it possible to separate maize from the 3 other cereals. Table II indicated that the most correlated excitation wavelengths were 360, 390, 400, 420 and 290 nm. Sunflower and groundnuts were found in opposite position according to variate z^4 though not as isolated than previously observed for soya, rapeseed or maize. Excitation wavelengths 310, 330, 390, 420 and 440 were the most involved in this variate (Tab. II). Sunflower and groundnuts were separated from the 4 cereals according to canonical variate z^5 (not shown).

The similarity maps drawn from the first 5 variates revealed the raw materials characterised by several excitation wavelengths. Soya, rapeseed and maize were the most characteristic products. Sunflower and groundnuts could be identified by taking several canonical variates into account. Wheat, barley and triticale were however always found together. These materials are very similar in comparison to

Table II. Squared multiple correlation coefficients between the canonical variates and the data tables.

	z^1	z^2	z^3	z^4	z^5
290	0.95	0.80	0.82	0.49	0.46
310	0.96	0.97	0.74	0.89	0.86
330	0.98	0.97	0.77	0.89	0.91
360	0.97	0.98	0.93	0.72	0.93
390	0.98	0.96	0.94	0.88	0.83
400	0.98	0.96	0.95	0.77	0.70
420	0.99	0.92	0.88	0.91	0.65
440	0.99	0.94	0.77	0.91	0.75
m^j	7.8	7.5	6.8	6.5	6.1

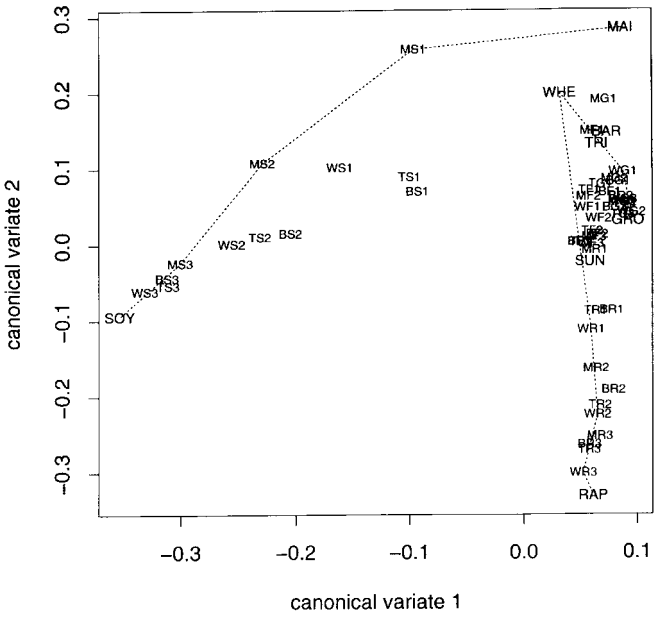

Fig. 4. Canonical similarity maps of variates z^1 and z^2. Codes of raw materials: wheat WHE, barley BAR, triticale TRI, maize MAI, soya SOY, rapeseed RAP, sunflower SUN and groundnuts GRO. Codes of mixtures: first letter W, B, T, M for the 4 cereals; second letter S, R, F, G for the 4 oilcakes; third letter 1, 2, 3 for proportion 25, 50, 75 of oilcakes, respectively.

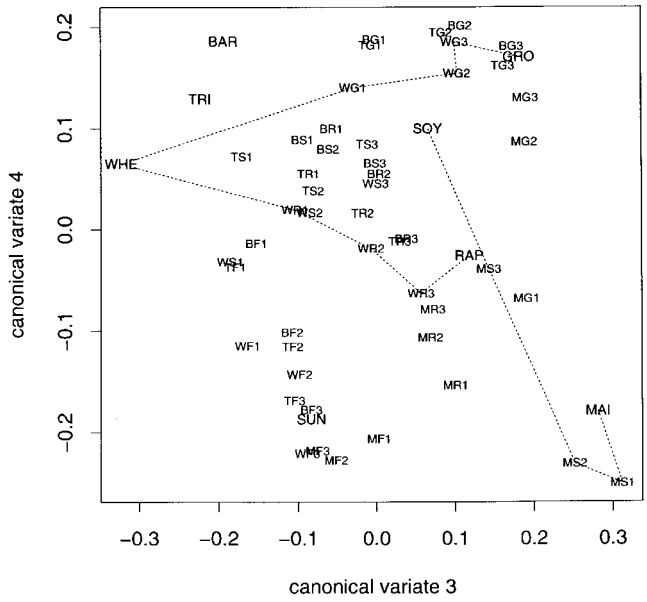

Fig. 5. Canonical similarity maps of variates z^3 and z^4. Same codes as figure 4.

the other products. Moreover, fluorescence may fail to identify them because of similar spectral properties towards the method. Another reason may be that generalised canonical correlation analysis firstly reveals information common to several data table. An identification of a product found in only one data table would not be highlighted in the first canonical variates.

Canonical spectral patterns

Figure 6 shows the spectral patterns associated to canonical variates z^1, z^2, z^3 and z^4. Spectral patterns can exhibit positive and negative peaks that may be associated to the positive and negative values of variates. The first pattern only showed negative peaks that could therefore be related to soya. This pattern was, in fact, very similar to the fluorescence spectra of soya (Fig. 2). The peak around 470 – 480 nm observed for excitation wavelengths 390, 400 and 420 nm was revealed with a higher intensity on the pattern in comparison to the spectra. This showed that the corresponding pairs of excitation-emission wavelengths were more characteristic of soya than 330 – 410 or 360 – 410 pairs also observed for the raw spectra in figure 2. The second canonical pattern contrasted positive peaks at 320 and 420 nm to negative peaks around 490 nm. Positive peaks could be related to cereals while the negative ones were associated to rapeseed. In comparison to the raw spectra in figure 1, spectral pattern highlighted some specific peaks that can be associated to the raw materials in binary mixtures of cereals and oilcakes. The third pattern made it possible to specifically associate the emission peak observed at 320 nm after an excitation at 290 nm to wheat, barley and triticale. This peak was not found in maize. Spectral pattern

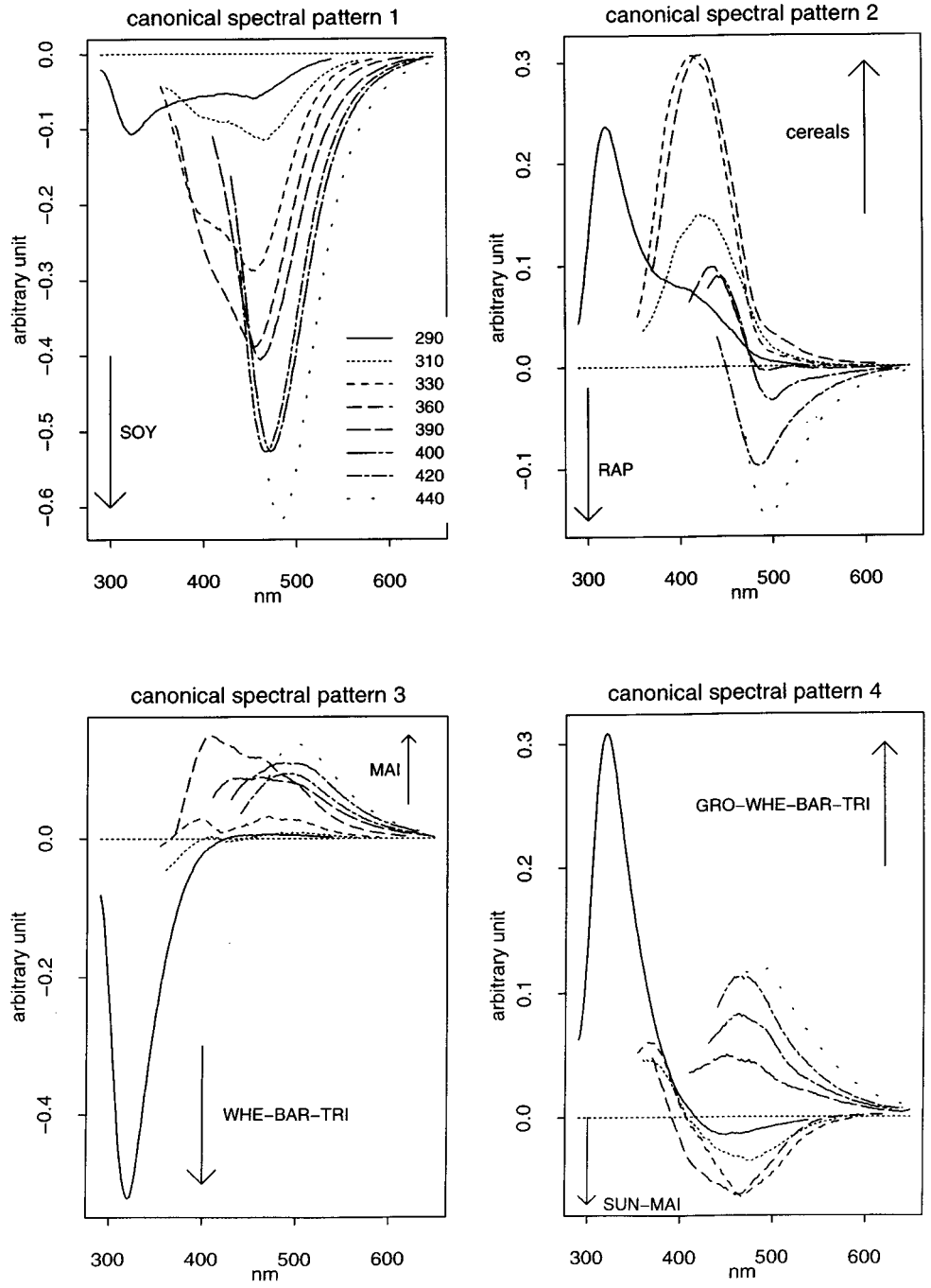

Fig. 6. Canonical spectral patterns of variates z^1, z^2, z^3 and z^4. The 8 patterns corresponding to the 8 excitation wavelengths are drawn on the same figure.

4 revealed emission peaks around 465 nm both in the positive and negative side of the figure. In this case, the excitation wavelengths, i.e. the multiway nature of fluorescence, made it possible to distinguish between sunflower and groundnuts.

Table III summarises the results obtained from the 5 canonical variates. The use of the similarity maps together with the spectral pattern and squared multiple correlation coefficient made it possible to extract pairs of excitation-emission wavelengths specific of some raw materials. The raw materials the most characterised by the emission fluorescence spectra could be pointed out by generalised canonical correlation analysis.

Conclusion

Generalised canonical correlation analysis could be successfully applied to sets of fluorescence spectral data. The spectral data sets had to be preliminary transformed by principal component analysis in order to eliminate the high collinearity among the emission wavelengths. In such a way, the analysis is the principal component analysis of the normalised components of each original data table merged in a single table. This procedure made the application of generalised canonical correlation analysis very simple to achieve and provided interpretable results. This advantage was previously outlined by Muller [10] and Devaux et al. [7] for canonical correlation analysis when applied to 2 groups of variables. Generalised canonical correlation analysis therefore provides a description of the samples by taking into account the correlation between the data tables after elimination of the collinearities within each table.

Generalised canonical correlation analysis highlights the common information to the data tables and do not focus on the description of each data table. Other techniques have been designed for the same general purpose aiming to describe the data tables together with revealing correlated information. In co-inertia analysis, an orthogonal decomposition of each data table is performed [11]. The application of co-inertia analysis on the same fluorescence data also proved to be of interest [12].

Table III. Assignments of pairs of excitation-emission wavelengths to raw materials.

Canonical variate	*raw material*	*pairs of excitation-emission wavelengths*
z^1	soya	440–485 420–470 400–470
z^2	rapeseed	440–495 420–485
	cereals	360–420 330–410 290–320
z^3	wheat barley triticale	290–320
	maize	360–420
z^4	sunflower	330–465 360–465
	peanut	440–485 420–465
	wheat barley triticale	290–320
z^5	cereals	330–410 360–420 290–320
	sunflower groundnuts	390–450 400–450 420–450

References

1. Munck, L. Fluorescence analysis in foods, Longman Scientific & Technical, London, 1989.
2. Martens, H.; Naes, T. Multivariate calibration, John Wiley, New York, 1989.
3. Bro, R. *J. Chemometr.* **1996**, *10*, 47-61.
4. Harshman, R. A.; Lundy, M. E. *Comp. Stat. Data Anal.* **1994**, *18*, 39-72.
5. Smilde, A. K.; Wang, Y.; Kowalski, B. R. *J. Chemomet.* **1994**, *8*, 21-36.
6. Carroll, J. D. *Proc. Amer. Psy. Ass*. **1968**, 227-228.
7. Devaux, M. F.; Robert, P.; Qannari, A.; Safar, M.; Vigneau, E. *Appl. Spectrosc.* **1993**, *47*, 1024-1029.
8. Saporta, G. Probabilité, analyse des données et statistique, Éditions Technip, Paris, 1990.
9. Robert, P.; Devaux, M. F.; Bertrand, D. *J. Near Inf. Spectrosc.* **1996**, *4*, 75-84.
10. Muller, K. E. *Amer. Statist.* **1982**, *36*, 342-354.
11. Chessel, D.; Hanafi, M. *Rev. Stat. Appl.* **1996**, *XLIV*(2), 35-60.
12. Courcoux, P.; Devaux, M. F.; Vigneau, E.; Novales, B. Decomposition of multiway fluorescence spectral data, TRICAP, Washington (USA), May 4-9, 1997.