Analusis, **1998**, *26*, 317-325

Characterization of multi-way spectral data using Factorial Correspondence Regression

N. Gouti[1], M.-F. Devaux[2], B. Novales[2], D.N. Rutledge[1] and M.H. Feinberg[1,*]

[1] *Institut National de la Recherche Agronomique, Laboratoire de Chimie Analytique, 16 rue Claude Bernard, 75231 Paris Cedex 05, France*

[2] *Institut National de la Recherche Agronomique, Laboratoire de Technologie Appliquée à la Nutrition, Rue de la Géraudière, 44072 Nantes Cedex 03, France*

Abstract. The principal advantage in Factorial Correspondence Analysis, where rows and columns are processed symmetrically, is the possibility to have in the same factorial space observation (row) and variable (column) projections. For sequence of spectra, the joint plot is composed of projections of wavelengths and of spectra. In the reported study, the analyzed data set consisted in fluorescence emission spectra recorded on animal feed samples. Samples were composed of eight raw materials (4 cereals and 4 oilcakes) and 48 mixtures of one cereal and one oilcake. For each sample, several specific excitation wavelengths were used leading to a 3-dimensional or 3-way data set: one way for samples, one for emission wavelengths and one for excitation wavelengths. After reorganization of the data set, FCA was applied and the resulting joint plot allowed finding similarities between excitation-emission wavelength couples and samples. Furthermore, the association of the Partial Least-Squares regression (PLS) with the FCA method led to the selection of some wavelength couples characteristic of the eight raw materials. The mathematical procedure of this new regression technique, called Factorial Correspondence Partial Least Squares regression (FCR-PLS), is developed and the model validation, which is based on a cross-validation procedure to choose independent variables entering the regression equation, is reported. All computations were done with Matlab® and programming examples are given.

Key words. Multivariate data analysis – factorial correspondence analysis – factorial correspondence regression – partial least squares regression – cross-validation.

Introduction

Recent advances in instrument technology has given rise to new techniques that tend to produce very large data sets with two or more dimensions and of several thousands of measurements. For example, time dependent spectroscopy, such infrared or fluorescence spectroscopy, can produce multi-way data sets. These are built by stacking individual spectra of the same sample recorded at different wavelengths or time [1,2]. The amount of data becomes so large that conventional processing methods used routinely in spectroscopy, such as examination and processing of individual spectra, cannot be used efficiently. More sophisticated techniques must be proposed in order to automatically and globally extract pertinent information from the total data set.

Among these techniques, Multivariate Statistical Analysis (MSA) gives very interesting results. MSA techniques are applied to data sets that are two-way tables with individuals for lines and variables for columns [3]. Principal Component Analysis (PCA) is an important MSA tool for studying multi-way data sets in the variable space. PCA is often used to eliminate variable redundancy and highlight meaningful structures. Its basic principle consists in transforming the original variables into a set of independent Principal Components or Factors, which are orthogonal in the multivariate space. PCA has also been applied to images or image sequences where each image number plays the role of an individual and the pixels the role of variables. Several recent applications have been published in Nuclear Magnetic Resonance imaging [4], in microscopic Raman line-imaging [5], or in Secondary Ion Mass Spectroscopy [6,7].

A derived technique, called Factorial Correspondence Analysis (FCA), can also be applied successfully to multivariate images or series of data sets. The goal of FCA is also to eliminate redundant information, but it differs from PCA in that individuals and variables are considered as equivalent and processed symmetrically. FCA has been first applied in chemical application to process large chromatographic data set [8]. During the last decade, it was used more regularly to study geochemistry data set [9]. More recently, FCA was also used as a preprocessing technique for background noise elimination in multivariate images, which were obtained by elemental mapping using Electron Energy Loss Spectroscopy [10]. FCA can be applied without any *a priori* hypothesis about the noise as it merely analyzes the variance between the pixel intensities of a collection of images. Furthermore, it distinguishes between variance due to actual signal and variance induced by background and noise. However, the main advantage of FCA in MSA is to simultaneously project onto the same score plot both matrix rows and columns [11]. This joint plot can be very useful to detect relationships between rows and columns.

The goal of our study is to select the excitation-emission wavelength couples which are the most efficient to simply measure by fluorescence spectroscopy the origin of a cereal and/or an oilcake used in an animal feed. Samples are pure cereals and oilcakes commonly used for animal feeds and

* Correspondence and reprints.
Received February 10, 1998; revised July 3, 1998; accepted July 10, 1998.

the mixtures of both of these raw materials. The data set consists in the fluorescence emission spectra recorded at several excitation wavelengths in order to form a three-way data set. The aims of the study are mainly to describe spectral similarity between samples, compare excitation conditions and select discriminatory excitation-emission wavelengths. It will be then possible to develop an analytical method, which can be used to quantitatively determine the type of cereal, and/or oilcake that was used to prepare a given animal feed. This goal is rather important for animal nutritionists or for regulation bodies.

Description of the method

Review of the FCA technique

The following notation is used in this paper: bold upper case for matrices or sub-matrices (**X**, **Y** or Φ_r); bold lower case for vectors (**x**, **y** or ψ); italic upper case for constants (*N*, *P* or *J*) and italic lower case for indices and matrix elements (*i*, *j* or x_{np}). The mathematical background of FCA has been extensively described in the literature [12,13] and will not be detailed here. Only a few characteristic features of the technique will be exposed.

Let **X** be a data matrix of general element x_{np} (n = 1, 2, …, N; p = 1, 2, …, P) where n describes individual or sample number and p variable number. The PCA algorithm makes a clear distinction between individuals and variables: the metric used to measure the distance between two points in the multivariate space is an Euclidean distance. On the contrary, FCA maintains symmetry between individuals (rows) and variables (columns) and uses a metric that has been demonstrated to be a χ^2 metric. The preprocessing step in FCA, instead of standardizing variables as in PCA, transforms the **X** matrix into a new matrix **Z** such that the χ^2 metric used on **X** is equivalent to an Euclidean metric on **Z**.

X is transformed into a frequency matrix **F**; this consists in a simple change in units using the following formula:

$$f_{np} = \frac{x_{np}}{\sum_{n=1}^{N}\sum_{p=1}^{P} x_{np}} \quad \text{then} \quad \sum_{n=1}^{N}\sum_{p=1}^{P} f_{np} = 1.$$

Let $\mathbf{D}_r$ and $\mathbf{D}_c$ be the diagonal matrices containing, respectively, the sum of rows (r_{nn}) and sum of columns (c_{pp}) of the matrix **F**.

$$r_{nn} = \sum_{p=1}^{P} f_{np} \quad \text{and} \quad c_{pp} = \sum_{n=1}^{N} f_{np}\ .$$

The matrix **Z** is obtained as follows:

$$\mathbf{Z} = \mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-1/2} \tag{1}$$

Z represents the new row coordinates in the vectorial space defined by the columns.

Following these modifications, the coordinates of the centroid of the data projections in multivariate space are the square roots of the diagonal elements of $\mathbf{D}_c$. If we take this centroid as a new origin, resulting matrix **V**, also called the inertia matrix, can be written in matrix form as:

$$\mathbf{V} = \mathbf{W}'\mathbf{D}_r\mathbf{W} = \left(\mathbf{D}_r^{1/2}\mathbf{W}\right)\left(\mathbf{D}_r^{1/2}\mathbf{W}\right). \tag{2}$$

Where elements of **W** are:

$$w_{np} = z_{np} - \sqrt{c_{pp}}\ .$$

The inertia matrix differs from the covariance matrix by the weighting and the translation of each element. The rank of **V** is $Q = \min(P, N) - 1$ therefore, the smallest latent root is zero. The transformation of the original data set **X** into the matrix **W** allows writing the inertia matrix **V** as a covariance-variance matrix through a cross-product operation. This made the eigenvalue and eigenvector extraction easier.

The analysis of the variability contained in the data set can be performed through the diagonalization of matrix **V**. This operation is performed in order to decompose the global variability into orthogonal sources. The eigenvector matrix **U** defines these sources while the associated eigenvalue diagonal matrix **L** defines their relative importance.

After the eigenvalue decomposition, row projections onto Q factorial axes (or eigenvectors) are obtained from equation (1) as:

$$\Phi_r = \mathbf{Z}\mathbf{U} = \mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-1/2}\mathbf{U}\,. \tag{3}$$

Furthermore, column projections can be deduced from (3) by the following relation:

$$\Phi_c = \mathbf{D}_c^{-1}\mathbf{F}'\Phi_r\mathbf{L}^{-1/2}. \tag{4}$$

In FCA, rows and columns can be interchanged as they are processed in a symmetrical manner. In this study, we defined Φ_r as the individual projections or individual scores in the factorial space, and Φ_c as the variable projections or variable scores.

All computations were performed using Matlab®. Eigenvalue and eigenvector extraction and the FCA algorithms are reported in annex as Matlab functions.

Partial least squares regression applied to correspondence factors

While PCA can often be considered as a descriptive technique useful to understand data redundancy, when combined with multivariate regression techniques it can be used to build predictive models for one or more observed responses. The numerous applications of Principal Component Regression (PCR) and Partial Least Squares regression (PLS) in analytical chemistry illustrate the success of these combined techniques [14,15].

As explained earlier, FCA is also based on a factor extraction technique resulting in two sets of projection coordinates, Φ_r and Φ_c. Thus, a regression technique may be used in order to predict the values of a variable as a function of Φ_r. The principle of such a regression technique has already been described elsewhere by Gouti et al. [16] as Factorial Correspondence Regression (FCR). When associated with PLS regression we shall refer to the FCR-PLS technique.

In this study the Φ_r matrix used as predictor variables (N individuals × Q columns) and **Y** as a set of response variables (N individuals × R responses). Compared with other

regression techniques, when there are more than one response variable the PLS algorithm efficiently uses the correlation among response variables **Y**. Several iterative PLS regression algorithms are described in the literature based either on the classical [17] or the kernel algorithm [18].

The PLS regression model is written as:

$$\mathbf{Y} = \Phi_r\mathbf{B} + \mathbf{E} \quad (5)$$

where **B** represents the matrix of regression coefficients and **E** the error matrix. Before computation, a unit vector **1** was added to Φ_r in order to determine the coefficients of the constants of the model.

As underlined earlier, the advantage in using FCA is that row scores Φ_r and column scores Φ_c are in the same factorial space. Thus, it is possible to apply regression coefficients to Φ_c, as shown in the figure 1, leading to the following regression equation:

$$\Phi_c\mathbf{B} = \Psi. \quad (6)$$

The resulting matrix Ψ ($P \times R$ elements) is expressed in the same units as **Y**. Each element ψ_{pr} of the vector ψ_r can be interpreted as the influence of variable p on response vector $\mathbf{y}_r$.

In order to highlight the most significant variables for each response vector $\mathbf{y}_r$, Ψ is studentized by columns as follows:

$$\boldsymbol{\psi}_r^{(t)} = \frac{\boldsymbol{\psi}_r - \overline{\psi}_r}{s_{\boldsymbol{\psi}_r}} \quad (7)$$

where $\overline{\psi}_r$ is the mean value and $s_{\psi r}$ is the standard deviation of the vector ψ_r.

Using this transformation, higher absolute values indicate that the corresponding columns significantly influence the corresponding response vector, following a Student's t distribution law. Two critical t-values were used corresponding respectively to two bilateral risk levels of 25% and 50%:

$$\alpha = 25\% \Rightarrow |t_1| = 1.150$$
$$\alpha = 50\% \Rightarrow |t_2| = 0.674.$$

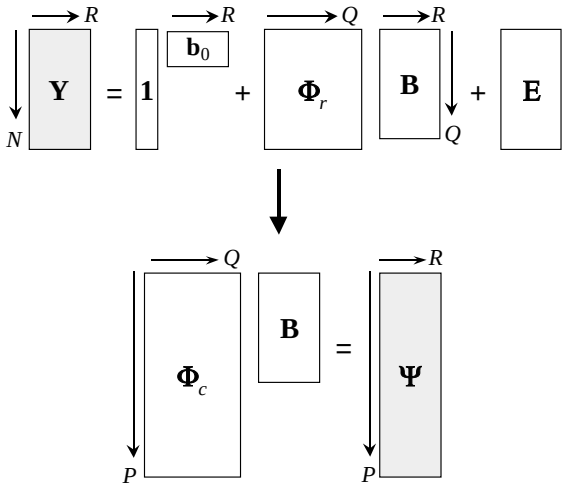

Fig. 1. Principle of the Factorial Correspondence Regression.

The interpretation rule is as follows: when a positive t-value is greater than the critical value, the associated column is highly correlated to the corresponding response vector; if negative, column and response are anti-correlated. It is then possible to select columns (i.e. wavelengths) which most influence a response vector.

Application of cross-validation technique to model validation

When dealing with predictive regression model, it is necessary to check the consistency of predicted values with observed values: this process is called model validation. The most effective way to validate a regression model consists in comparing new response values from a new data set to response values predicted with model built with the old data set. In many situations, collecting new data for validation purposes is not possible. The cross-validation algorithm [19] is a technique that can be applied when no new data are available. It consists in randomly discarding about 20% of observations from the data set using this left out group as a "new" data set in order to validate model predictive ability. This process is repeated until each observation has been discarded at least once.

Let $\mathbf{Z}^{(-G)}$ be the submatrix of **Z** in which a group of n_G observations have been discarded and $\mathbf{Y}^{(-G)}$ the corresponding response submatrix. Let $\mathbf{Z}^{(G)}$ and $\mathbf{Y}^{(G)}$ be their respective complements, i.e. submatrices containing discarded values. An inertia matrix $\mathbf{V}^{(-G)}$ is computed, as in equation (2), for each $\mathbf{Z}^{(-G)}$ giving a new eigenvector matrix $\mathbf{U}^{(-G)}$. Then a new row projection matrix $\Phi_r^{(-G)}$ is defined as in equation (3) and estimated PLS-coefficients $\hat{\mathbf{B}}^{(-G)}$ are obtained as in equation (5). Predicted values of group G are then obtained in matrix form:

$$\hat{\mathbf{Y}}^{(G)} = \mathbf{Z}^{(G)}\left(\mathbf{U}^{(-G)}\hat{\mathbf{B}}^{(-G)}\right).$$

Using this technique, it is possible to compute an error of prediction for each run. However, it is preferable to compute this error as a function of the number of factors entered in the model. This is called the Predictive Error Sum of Squares (PRESS) which is defined as follows, for a specific number of components A (dimension of the PLS model) [15,20]:

$$\mathrm{PRESS}_A = \sum_{k=1}^{K}\sum_{g=1}^{n_G}\sum_{r=1}^{R}\left(y_{gr} - \hat{y}_{gr}\right)^2.$$

where y_{gr} and $\hat{y}_{gr}$ are the elements of, respectively, matrices $\mathbf{Y}^{(G)}$ and $\hat{\mathbf{Y}}^{(G)}$. R is the number of response variables, K the number of iterations performed and n_G the number of discarded observations. In order to have a better estimate of expected variation when the model is applied, the product $K \times n_G$ is defined in such a way that all observations have been discarded the same number of times. The Matlab program that gives the K subsets of calibration and prediction is available from the authors. Instead of directly using the PRESS criterion, it is possible to calculate an average cross-validation error defined as the Root Mean Square Error of Cross-Validation (RMSECV) [14]:

$$\mathrm{RMSECV}_A = \sqrt{\frac{1}{RKn_G}\mathrm{PRESS}_A}. \quad (8)$$

Both of these criteria can be used to select the optimal number of components A that gives the best prediction. It is now classical to show that too large a value for A leads to overfitting and bad prediction due to the introduction of "noise", while too small a value causes underfitting due to systematic error. In our case, the model that gives the first minimal RMSECV is considered as the optimal model. The FCR-PLS cross-validation procedure is transcribed in Matlab code in the annex.

Experimental

Samples consisted of eight raw materials and 48 mixture samples, giving 56 samples in all. The raw materials were four cereals (wheat, barley, maize and triticale) and four oilcakes (soy, sunflower, rapeseed and groundnut). Each mixture consisted in combining one cereal and one oilcake, according to the following proportions: 75% – 25%, 50% – 50% and 25% – 75%.

Fluorescence spectra were recorded for each sample at eight different excitation wavelengths using a SLM 4800 spectrofluorometer, with a light incidence angle of 56°. Samples were reduced to powders and were directly analyzed, without others treatments, in a quartz tub of 1×1 cm. Each spectrum is an average of two spectra. All 56×8 spectra were put together giving a 3-way data table, as illustrated in figure 2. Each matrix $\mathbf{X}_j$ consisted in fluorescence spectra for N samples, recorded at a given excitation wavelength j over an I_j emission wavelength range. The emission wavelength ranges are not always exactly of the same width for all $\mathbf{X}_j$ since they depend on excitation wavelengths.

Before starting computation, the 3-way data array is reorganized and unfolded as usual for MSA techniques. The resulting data matrix $\mathbf{X}$ is composed of N rows and P columns. Reorganization of fluorescence data table lead to a matrix with $N = 56$ rows and $P = 884$ columns, each representing a specific excitation-emission wavelength couple as reported in the table I.

Table I. Excitation and emission wavelengths selected for recording multi-way spectral data.

Excitation wavelengths (nm)	*Emission wavelength ranges (nm)*
290	290 – 538
310	360 – 568
330	354 – 588
360	370 – 618
390	410 – 648
400	430 – 628
420	440 – 640
440	470 – 648

Results and discussion

FCA of the fluorescence spectral data set

Figure 3 presents the row projections in the space of the two first Correspondence Factors (CF) after submitting data to FCA. This projection plane summarizes 86.2% of the total variance: 60.7% for CF 1 and 25.5% for CF 2. Raw materials are presented in a rounded-corner box, as a regular character for the four cereals, such as wheat (W), triticale (T), barley (B), maize (M), and an italic character for the oilcakes, such as sunflower (*F*), groundnut (*G*), rapeseed (*R*) and soy (*S*). It is obvious that cereals and oilcakes form two distinct groups along CF 1. Mixtures are represented by a symbol according to the following rules. A symbol type is associated to each cereal (square for maize, triangle for triticale, …) and a gray level to each oilcake, from white for sunflower until black for soy. The size of the symbol is proportional to the amount of cereal into the mixture: for 25% of cereal the symbol is the smallest while it is the largest when there is 75% of cereal.

In figure 3, the distance between two points can be interpreted as a qualitative estimation of the similarity between

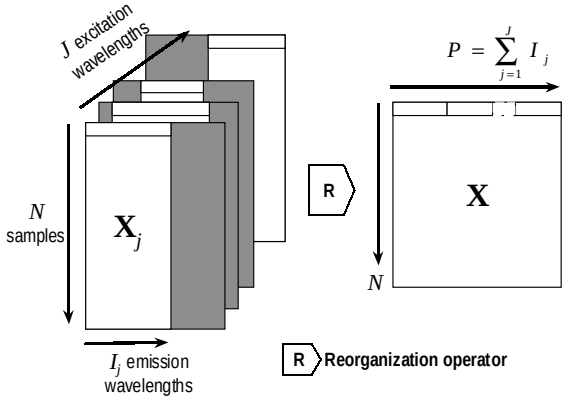

Fig. 2. Structure of multi-way spectral data. 3D-data set (N samples, J excitation wavelengths and I_j emission wavelengths) is reorganized into a 2D-matrix where rows are the samples and columns the emission wavelengths. For a given excitation wavelength, not all emission wavelengths were recorded.

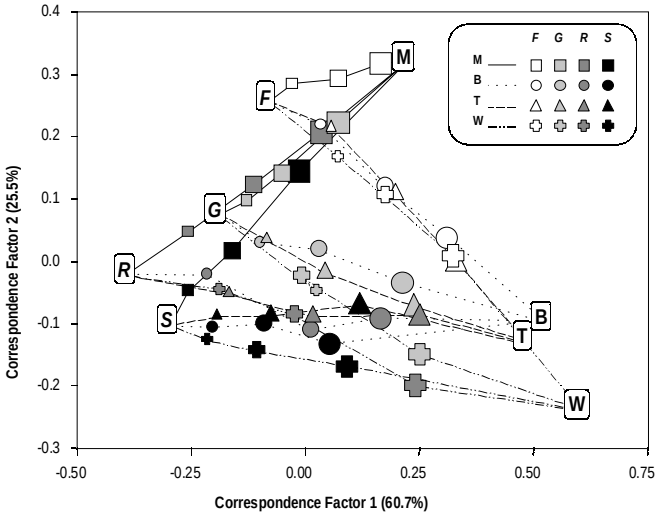

Fig. 3. Row projections in the two first factor space using Factorial Correspondence Analysis (for code abbreviation, see the text).

the corresponding rows. This means that two samples, which are close to each other, such as barley and triticale, have comparable spectral properties. For cereals, wheat, barley and triticale can be considered as having very similar properties, while maize seems to be different. The same situation occurs for sunflower compared to other cakes.

Thanks to the used graphical symbolism, it is possible to clearly establish that there is a linear relationship between each couple cereal/oilcake and the associated mixtures. These different relationships are shown in figure 3 as broken lines of various styles according to the cereal. There is proportionality between the distance of a mixture sample to the corresponding raw materials and its composition. That means, for example, that a 50% – 50% mixture is located at about equal distances from the cereal and the oilcake. Nevertheless, this proportionality is not observed for mixtures using soy where the projections are closer to the soy sample than to the cereals. This implies that soy has a greater effect on the fluorescence spectra and so should be easier to detect.

The column projections are reported in figure 4 as a function of emission wavelengths for the first two CFs. This representation was used in order to have simpler figures and easily highlight interesting columns, i.e. excitation wavelengths. For instance, figure 4 shows that high values on CF 1 are obtained when the emission wavelengths are at around 320 nm for an excitation wavelength of 290 nm. This wavelength couple, that we shall note {Ex: 290/Em: 310 – 330}, is strongly related to the cereals since in figure 3, high positive values correspond mainly to cereals. Using the same reasoning with high negative values along CF 1, it is possible to assign excitation-emission wavelength couples to rapeseed and soy, which can be characterized by couples {Ex: 420/Em: 470 – 500} and {Ex: 440/Em: 490 – 510}. These results are confirmed by the negative values of the column and the sample projections on CF 2. Moreover, the negative peaks at {Ex: 290/Em: 325} on CF 2, figure 4, is certainly related to the wheat sample. In fact, the wheat projection in figure 3 has the highest negative value. Finally, since sunflower and maize projections in figure 3 have the highest positive values on CF 2 and small absolute values on CF 1, it is possible to associate these samples to the following couples: {Ex: 310/Em: 520 – 560}, {Ex: 330/Em: 540 – 570} and {Ex: 360/Em: 530 – 600}.

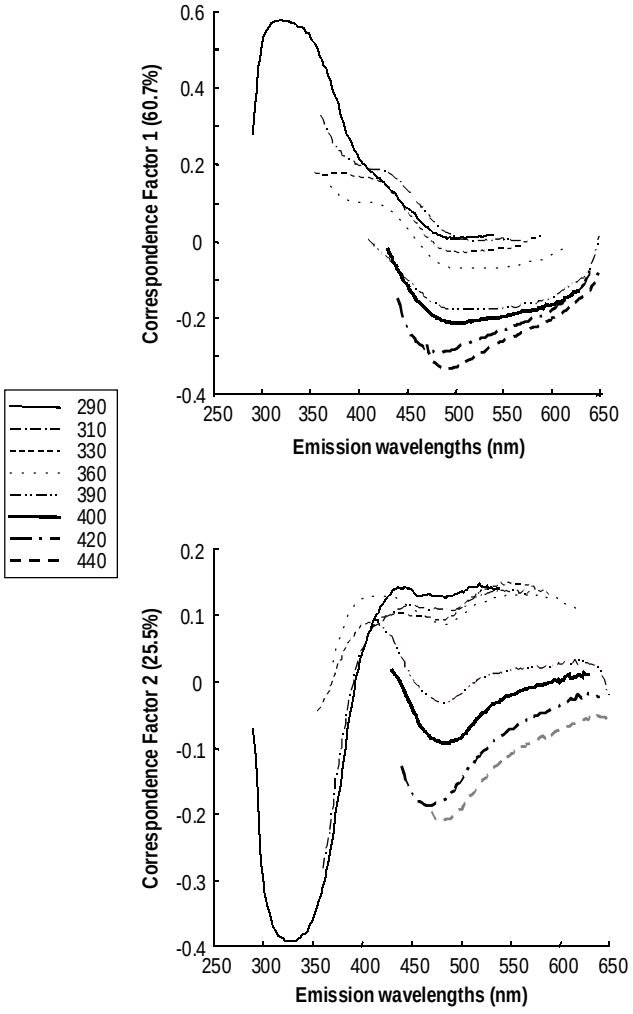

Fig. 4. Column projections as a function of emission wavelengths for Correspondence Factors 1 and 2.

However, the assignment of an excitation-emission wavelength couple to each sample type remains difficult even when using other CFs. At that time, it is difficult to characterize grapeseed sample and it is quite impossible to assign different wavelength couples to the barley and triticale samples.

FCR-PLS on mixture samples

Knowing the relative proportion of each cereal and oilcake in a mixture, it was possible to create 8 new variables which could be used to describe the content of the sample and thus build a matrix **Y** (R = 8 response vectors). Applying FCR-PLS to this matrix, it is possible to calculate the coefficients of the 8 models which best predict responses as a function of row projections Φ_r. These coefficients were then used on column projections Φ_c in order to relate excitation-emission wavelength couples to mixture composition.

The number of factors entering the best PLS regression model was defined by the cross-validation procedure described above. In order to have a good estimate of the RMSECV criterion, K = 56 iterations were used and n_G = 11 samples were left out at each iteration. These samples were chosen in such a way that after a cycle of 56 iterations each sample was discarded the same number of time, in the circumstances 11 times. The resulting RMSECV values are reported in figure 5. The first local minimal value was found for 16 entered CFs with $RMSECV_{16}$ = 11.02.

The **B** coefficient matrix was then computed and the matrix Ψ was calculated using equations (5) and (6). The studentized t-values obtained for the soy and noted $\psi^{(t)}_{soy}$ are reported in figure 6a *versus* emission wavelengths. Dashed lines correspond to both critical values t_1 and $-t_1$ and interesting points were detected above and below these lines. For instance, figure 6a shows that excitation-emission wavelength couples most significantly correlated to the proportion of soy in a mixture are {Ex: 440/Em: 470 – 500} and {Ex: 420/Em: 440 – 475} because the model coefficients are above the critical values in these regions of the spectrum. Similarly, emission wavelengths between 600 nm and 648 nm are significantly anti-correlated and could be defined as concurrent couples. Nevertheless, this region of the soy

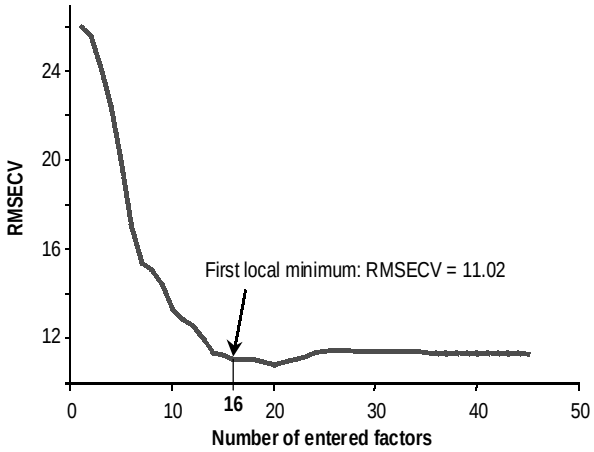

Fig. 5. Evolution of RMSECV during cross-validation process of PLS regression model for Φ_r. Local minimum is for 16 factors.

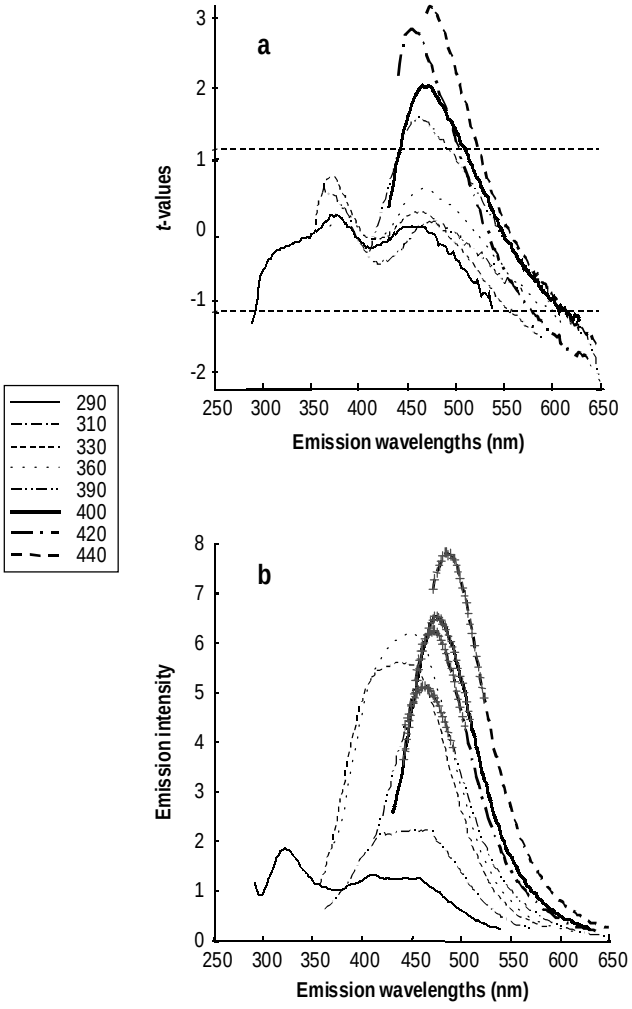

Fig. 6. Typical results obtained with soy and soy-based mixtures. a) Studentized $\psi^{(i)}_{soy}$ *versus* emission wavelengths (nm). Dashed lines correspond to critical values t_1 and $-t_1$. Raw soy emission spectra with most significant excitation-emission wavelength couples (+).

emission spectra usually corresponds to the low intensity peaks, as illustrated in figure 6b, which would not be well adapted to quantification. This means that a compromise must be found between couples, which are emphasized by FCR-PLS technique, and couples, which can actually be used for analytical purposes.

The same selection technique was applied to all other raw and mixture samples. All significant excitation-emission wavelength couples for each raw material are reported in table II. Figure 7 presents a graphical summary of optimal working ranges, which have been selected according to the described technique. On the vertical scale, eight monochromatic excitation wavelengths are presented while wide wavelength ranges represent the observed emission domains on the horizontal scale. Valuable excitation-emission wavelength couples are presented as boxes with the sample abbreviation code. It clearly appears that for each sample type several interesting excitation and/or emission domains can be used. As a rule, when available, couples located in the central part of the emission domain must be preferred.

The less valuable situation occurs with wheat and triticale as no specific domain can be found for wheat. Both these species are very close, present comparable biological properties and can be characterized by the couple {Ex:290/Em:320}. Nevertheless, only triticale has interesting wavelengths around 370 nm when excited at 330 nm. For maize, two couples {Ex: 360/Em: 420} and {Ex: 330/Em: 410} can be proposed. Concerning barley, the situation is not very good as no significant emission wavelengths were found above the critical value t_1. Therefore, a smaller critical value t_2 was applied and some interesting couples were selected at {Ex: 290/Em: 360 – 380}. For soy, couples {Ex: 400/Em: 475} and {Ex: 420/Em: 470} are preferred to couple {Ex: 440/Em: 485}, which were located at the beginning of the emission range and may be biased by instrumental artifact. For rapeseed, significant couples were present at the end of

Table II. Most valuable excitation-emission wavelength couples for the different sample types obtained by the FCR-PLS method.

	Samples	*Proposed wavelengths couples (nm) Excitation wavelength / Emission wavelength range*
Oilcakes	Soy	400 / 460–490 420 / 460–480 440 / 470–500 390 / 450–480
	Rapeseed	440 / 540–580 420 / 540–560
	Groundnut	400 / 430–440 390 / 410–420
	Sunflower	330 / 500–515 290 / 450–480 310 / 505–525
Cereals	Maize	360 / 400–440 330 / 385–420 310 / 400–440
	Triticale	290 / 300–325 330 / 360–370
	Barley	290 / 360–390
	Wheat	290 / 310–340

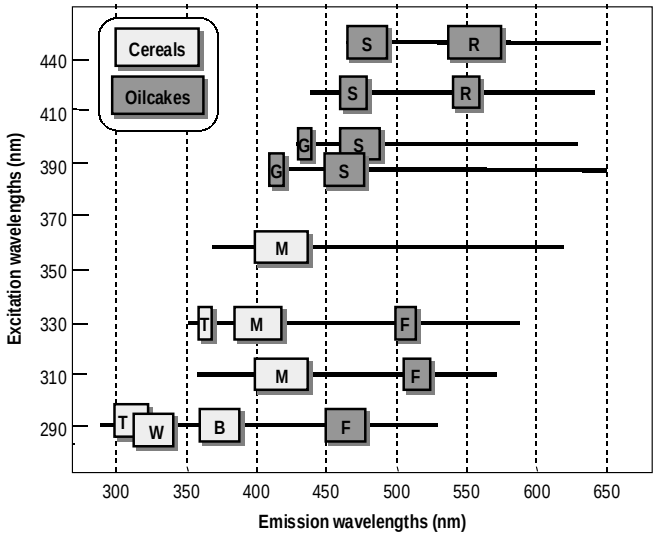

Fig. 7. Optimal excitation-emission wavelength couples selectable for quantification.

the emission range for the two excitation wavelengths at 420 nm and 440 nm. The highest *t*-values were found for the later emission wavelengths, but the corresponding intensities are minimal. The same is found for sunflower and groundnut spectra. However, groundnut could be characterize by the couples {Ex: 400/Em: 440} and {Ex: 390/Em: 415}, and sunflower by the couples {Ex: 290/Em: 460} and {Ex: 330/Em: 510}.

Conclusion

In the present work, the proposed FCR-PLS technique was demonstrated to be efficient in interpreting complex multiway fluorescence spectral data. The major advantage of this technique is to yield a simultaneous projection of samples and variables in a unique factorial space. Thus, any regression model developed on the sample score matrix can be directly applied to variable scores. Moreover, FCR and FCR-PLS methods highlight spectral variations only due to response vector(s) used for the regression; others variations being neglected.

In conclusion, it is possible to propose to use fluorescence spectroscopy as an analytical method to measure the proportion of a cereal and/or an oilcake in an animal feed and authenticate a composition. As a general guideline, short excitation wavelengths (from 290 nm to 330 nm) are to be selected for cereals, while high excitation wavelengths (330 nm to 440 nm) are to be preferred to identify oilcakes.

As stated by Benzécri (cited by Mellinger [11]), FCA, although it is mainly used as a descriptive tool, can also be an effective means of implementing discriminant and regression procedures. In other words, FCA can be used as an exploratory tool (in order to obtain some general knowledge about the relationships between the rows and the columns of a table), as a direct synthesis tool (the calculated factors are used as new variables to describe the data), or/and as a multistage synthesis tool (the factor characteristics are used for further multivariate statistical analysis, such as multivariate regression).

Annex

All computations were carried out with Matlab 4.2.c for Windows [21]. Three customs functions developed for this publication are presented. In order to facilitate code reading, all variable names start with mt for a matrix, vt for a vector, sg for a real scalar and in for an integer. Matlab internal functions and keywords are in boldface characters.

Intrinsic functions

mtX = **diag**(vtV)	creates a square matrix mtX with elements of vtV on the diagonal.
vtV = **diag**(mtX)	returns the main diagonal vtV of the matrix mtX.
[mtV, mtL] = **eig**(mtX)	computes eigenvectors mtV and eigenvalues mtL of the matrix mtX, so that mtX × mtV = mtV × mtL.
vtI = **find**(vtV > 10)	returns value indices of vtV satisfying the constraint.
mtY = **fliplr**(mtX)	returns mtX with row preserved and columns flipped in the left/right direction.
mtY = **flipud**(mtX)	returns mtX with columns preserved and rows flipped in the up/down direction.
mtY = **inv**(mtX)	mtY is the inverse matrix of mtX.
vtV = **linspace** (sgA, sgB, inN)	returns a row vector vtV (inN elements) linearly spaced between sgA and sgB.
mtI = **ones**(inR, inC)	returns a matrix (inR rows and inC columns) with all elements set to ones.
[inR,inC] = **size**(mtX)	returns row and column numbers of the matrix mtX.
[vtR,vtI] = **sort**(vtV)	sorts vtV in ascending order: vtR is the ranged vector and vtI the index vector.
mtY = **sqrt**(mtX)	elements of mtY are the square root of the elements of mtX.
vtV = **sum**(mtX)	returns a row vector with the sum over each column of mtX.
sgS = **sum**(vtV)	returns the sum of the elements of vector vtV.
mtZ = **zeros**(inR, inC)	returns a matrix with inR rows and inC columns with all elements set to zeros.

Custom functions

Function [mtB] = **pls2**(mtX, mtY, inA)

% returns the PLS coefficients mtB for the regression of mtX on mtY within A components.

% The Matlab code of this function was obtained from reference [21].

% The PLS algorithm is a kernel algorithm for data sets with many variables.

```
Function [mtvec, mtval] = eigorder(mtX)
% Returns eigenvalues in decreasing order and corresponding eigenvectors.
% INPUT ARGUMENT
%     mtX       square matrix.
% OUTPUT ARGUMENTS
%     mtval     diagonal matrix of eigenvalues in decreasing order.
%     mtvec     matrix of eigenvectors.
% LOCAL DECLARATIONS
%     inC       number of columns.
%     vtorder   index vector.
%     vtval     vector of eigenvalues.
%     vtrem     index vector of zero eigenvalues.
%-------------------------------------------------------------------
inC = size(mtX, 2);
[mtvec, mtval] = eig(mtX);
% Sorting eigenvalues and corresponding eigenvectors:
[vtval, vtorder] = sort(diag(mtval));
vtval = flipud(vtval);
mtvec(:, linspace(1, inC, inC)) = mtvec(:, vtorder);
mtvec = fliplr(mtvec);
% Removing zero eigenvalues and corresponding eigenvectors:
   vtrem = find(vtval < 1e-10);
   vtval(vtrem) = [ ];
   mtval = diag(vtval);
   mtvec(:, vtrem) = [ ];
```

```
Function [mtPr,mtPc,mtZ,mtW,mtDr,mtvec,mtval] = FCA(mtX)
% Factorial Correspondence Analysis.
% INPUT ARGUMENT
%     mtX       raw observation matrix.
% OUTPUT ARGUMENTS
%     mtPr      matrix of the row projections.
%     mtPc      matrix of the column projections.
%     mtZ       new row coordinates as Chi2 metric used on mtX is equivalent to a Euclidean metric on mtZ.
%     mtW       translation of mtZ so that centroid is new origin.
%     mtDr      diagonal matrix of the sum of rows.
%     mtvec     matrix of eigenvectors.
%     mtval     diagonal matrix of eigenvalues.
% LOCAL DECLARATIONS
%     inR       number of rows.
%     inC       number of columns.
%     mtF       frequency matrix.
%     mtDc      diagonal matrix of the sum of columns.
%     mttmp     temporary matrix.
%-------------------------------------------------------------------
[inR, inC] = size(mtX);
mtF = mtX ./ sum(sum(mtX));
mtDr = diag(sum(mtF'));
mtDc = diag(sum(mtF));
mtZ = inv(mtDr) × mtF × inv(sqrt(mtDc));
mtW = mtZ – ones(inR, inC) × sqrt(mtDc);
% Eigenvalues and eigenvectors of the inertia matrix:
mttmp = sqrt(mtDr) × mtW;
[mtvec, mtval] = eigorder(mttmp' × mtmp);
% Row and column projections on the factorial axes:
mtPr = mtZ × mtvec;
mtPc = inv(mtDc) × mtF' × mtPr × inv(sqrt(mtval));
```

Comments on FCA: when the number of variables is greater than the number of observations, the kernel eigenvalue decomposition algorithm may be used [22]:

```
[mtvec, mtval] = eigorder(mttmp × mttmp');
mtvec = mttmp' × mtvec × inv(sqrt(mtval));
```

When dealing with large problems, it is also recommended to use sparse matrix functions in order to avoid memory constraints. For example, the diagonal matrix of the sum of columns could be written as follows:

```
mtDc = sparse(1:inC, 1:inC, sum(mtF));
```

```
Function vtcv = CV_FCPLS(mtX, mtY, mtCset, mtPset)
% Returns Cross-Validation Root Mean Square Error (RMSECV) values for all possible PLS models.
% INPUT ARGUMENTS
%     mtX       raw observation matrix.
%     mtY       response matrix.
%     mtCset    matrix of the calibration subsets.
%     mtPset    matrix of the prediction subsets.
% OUTPUT ARGUMENTS
%     vtcv      RMSECV vector computed for PLS models of different dimension.
% LOCAL DECLARATIONS
%     mtPr, mtPc, mtZ, mtW, mtDr, mtvec, mtval
%               similar to the FCA algorithm.
%     inR       number of rows (observations).
%     inCF      number of Correspondence Factors.
%     inrun     number of iterations: times each PLS model is computed.
%     inout     number of left out observations at each run.
%     vtpress   vector of the Predictive Error Sum of Square values.
%     inloop    cross-validation loop index.
```

```
% vtpred   index vector of the current predictive set.
% vtcal    index vector of the current calibration set.
% mttmp    temporary matrix.
% indim    loop index denoting the dimension of the PLS model.
% mtb      regression coefficients matrix associated to Correspondence Factors.
% mtg      matrix of the regression coefficients associated to the initial variables.
%-------------------------------------------------------------
% Factorial Correspondence Analysis:
[mtPr, mtPc, mtZ, mtW, mtDr, mtvec, mtval] = FCA(mtX);
inR = size(mtX, 1);
inCF = size(mtPr, 2);
[inrun, inout] = size(mtPset);
% Initialization of PRESS vector:
vtpress = zeros(1, inCF);
% Cross-validation loop, at each loop the calibration and predictions sets are new ones:
for inloop = 1:inrun
   vtpred = mtPset(inloop, :);
   vtcal = mtCset(inloop, :);
   % Eigenvalues and eigenvectors of the new inertia matrix:
   mttmp = sqrt(mtDr(vtcal, vtcal)) × mtW(vtcal, :);
   [mtvec,mtval] = eigorder(mttmp' × mttmp);
   % Row projections:
   mtPr = mtZ(vtcal, :) × mtvec;
   % Adding a constant term for the regression:
   mtPr = [ones(inR – inout, 1), mtPr];
   % Loop for the dimension of the PLS model:
   for indim = 1:inCF
      mtb = pls2(mtPr, mtY(vtcal,:), indim);
      mtg = [mtb(1,:); mtvec × mtb(2:size(mtg, 1), :)];
      % Computation of the PRESS value according to the predictive set and the dimension of the model:
      vtpress(indim) = vtpress(indim) + sum(sum((mtY(vtpred, :) – [ones(inout, 1), mtZ(vtpred, :)] × mtg).^2));
   end
end
% Vectors of the RMSECV values, the best PLS model is the one which gives the first local minimum of vtcv:
vtcv = sqrt(vtpress./(inrun × inout × size(mtY, 2)));
```

References

1. Rutledge, D. N.; Barros, A. S.; Gaudard, F. *Mag. Res. Chem.* **1997**, *35*, S13-S21.
2. Bonnet, N.; Simova, E.; Thomas, X. *Microsc. Microanal. Microstruct.* **1991**, *2*, 129-142.
3. Basilevsky, A. Statistical Factor Analysis and Related Methods. Theory and Applications, John Wiley & Sons Ltd, 1993.
4. Geladi, P.; Grahn, H. Multivariate Image Analysis, John Wiley & Sons, Chichester, 1996.
5. Drumm, C. A.; Morris, M. D. *Appl. Spectro.* **1995**, *49*, 1331-1337.
6. Gouti, N.; Van Espen, P.; Feinberg, M. H. *Chemometr. Intell. Lab. Syst.*, in press.
7. Van Espen, P.; Janssens, G.; Vanhoolst, W.; Geladi, P. *Analusis* **1992**, *20*, 81-90.
8. Hirsch, R. F..; Gaydosch, R. J.; Chrétien, J. R. *Anal. Chem.* **1980**, 723-728.
9. Mellinger, M.. *Chemometr. Intell. Lab. Syst.* **1987**, *2*, 93-108.
10. Trebbia, P.; Bonnet, N. *Ultramicroscopy* **1990**, *34*, 165-178.
11. Mellinger, M. *Chemometr. Intell. Lab. Syst.* **1987**, *2*, 61-77.
12. Benzécri, J. P. L'Analyse des Données: 2. L'Analyse des Correspondances, Dunod, Paris, 1st ed. 1973, 2nd ed 1980.
13. Lefebvre, J. Introduction aux analyses statistiques multidimensionnelles, Masson, Paris, 1980.
14. Casal, V.; Martin-Alvarez P. J.; Herraiz, T. *Anal. Chim. Acta* **1996**, *326*, 77-84.
15. Kowalski, K. G. *Chemometr. Intell. Lab. Syst.* **1990**, *9*, 177-184.
16. Gouti, N.; Rutledge, D. N.; Feinberg, M. H. *Analyst* **1998**, *123*, 1783–1790.
17. Höskuldsson, A. *Chemometr. Intell. Lab. Syst.* **1988**, *2*, 211-228.
18. Rännar, S.; Lindgren, F.; Geladi, P.; Wold, S. *J. Chemom.* **1994**, *8*, 111-125.
19. Wold, S. *Technometrics* **1978**, *20*(4), 397-405.
20. Höskuldsson, A. *Chemometr. Intell. Lab. Syst.* **1996**, *32*, 37-55.
21. MATLAB® for Windows Version 4.2.c, (The MathWorks Inc., 1994).
22. Wu, W.; Massart, D. L.; De Jong, S. *Chemometr. Intell. Lab. Syst.* **1997**, *36*, 165-172.