Available online at www.sciencedirect.com

SCIENCE DIRECT®

Chemometrics and Intelligent Laboratory Systems 79 (2005) 31 – 41

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Discriminating from highly multivariate data by Focal Eigen Function discriminant analysis; application to NIR spectra

J.M. Roger*, B. Palagos, S. Guillaume, V. Bellon-Maurel

Cemagref BP 5095-34033 Montpellier Cedex1, France

Received 15 September 2004; received in revised form 28 January 2005; accepted 7 March 2005
Available online 21 June 2005

Abstract

Discriminating between classes from spectra deals with an ill-conditioned problem, which is generally solved by means of dimension reduction, using principal component analysis or partial least squares regression. In this paper, a new method is presented, which aims at finding a parcimonious set of discriminant vectors, without reducing the dimension of the space. It acts by scanning a restricted number of scalar functions, called Focal Eigen Functions. These functions are theoretically defined and some of their interesting properties are proven. Three scanning algorithms, based on these properties, are given as examples. An application to real spectroscopic data shows the efficiency of that new method, compared to the Partial Least Squares Discriminant Analysis.

Keywords: Multivariate discriminant analysis; NIR spectroscopy

1. Introduction

Analytical chemistry and process monitoring involve more and more multivariate indirect sensors, like spectrometers. For example, Near Infra-Red (NIR) spectrometry is a powerful analytical tool, increasingly used in industry [1,2]. The signal of such devices (i.e. the spectrum) is made up of hundreds of intensity values measured with regards to wavelengths. In order to create calibration models, a (generally) linear relationship is sought between spectra and reference measurements (responses). Two mathematical problems arise in spectrometer calibration: The first one, related to dimensioning, is due to the fact that the calibration set generally contains more variables (wavelengths) than individuals (spectra); The second one, related to conditioning, is due to the huge intercorrelation of the measured spectral variables. For quantitative response, these two problems are generally solved by factorial methods, like principal component regression (PCR) or partial least square regression (PLSR) [3]. These methods build a restricted number of latent variables from the original ones, little or not correlated, thus revealing interpretable structure. Discrimination from NIR spectra is less common, in spite of a great number of potential applications: defect detection, object or product recognition, outlier detection, etc. In discrimination, the variable to predict is qualitative, i.e. it takes its values in a unordered discrete set. Except in the simple case of 2 classes, which is analogous to a quantitative response case, the factorial regression methods are unsuited. The discriminant methods which solve the issues of dimensioning and conditioning proceed similarly to factorial regression: A classical discriminant analysis (DA) is performed on latent variables, provided either by a principal component analysis (PCA-DA), or by a PLS between the spectra and the class membership (PLS-DA) [4]. As far as regression is concerned, PLSR is generally more powerful than PCR, since the latent variable design takes into account the relationship between the spectra variables and the responses. Due to the same reason, in the discrimination case, PLS-DA is generally more efficient than PCA-DA.

The method presented in this paper does not proceed by dimensionality reduction. It is based on scanning functions,

* Corresponding author. Tel.: +33 46 7046383; fax: +33 46 7046306.
E-mail address: roger@montpellier.cemagref.fr (J.M. Roger).

0169-7439/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2005.03.006

called Focal Eigen Functions, to seek the optimal discriminating vectors. This method is particularly dedicated to the ill-conditioned problems. But it yields the same result that the classical discriminant analysis for the well-conditioned ones.

In Section 2, the method is explained: the concept of Focal Eigen Function is introduced, some of their properties are described and finally three examples of scanning algorithms are given. A spectrometric data set is introduced in Section 3, as an ill-dimensioned and ill-conditioned problem illustration. The performances of the scanning algorithms, in comparison with the PLS-DA ones, are detailed in Section 4. Finally, the main conclusions are reminded in Section 5.

2. Theory

Let $\mathbf{X}_{(n\times p)}$ be a matrix containing a sample of n individuals, described by p variables. $\mathbf{X}$ is assumed to be centered, i.e. the mean value of each column is zero. Each individual of $\mathbf{X}$ belongs to one of the c classes $\{1,\ldots, c\}$, a priori known. Let $\mathbf{Y}_{(n\times c)}$ be the matrix containing the disjunctive encoding of the individuals, i.e. $y_{ij}=1$ if the individual i belongs to the class j and 0 if not.

Let $\mathbf{T}$ be the variance–covariance matrix:

$$\mathbf{T}=\frac{1}{n-1}\mathbf{X}^T\mathbf{X}$$

Let $\mathbf{B}$ be the between class variance–covariance matrix:

$$\mathbf{B}=\frac{1}{n-1}\mathbf{X}^T\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\mathbf{Y}^T\mathbf{X}$$

Let $\mathbf{W}$ be the within class variance–covariance matrix:

$$\mathbf{W}=\mathbf{T}-\mathbf{B}$$

It is assumed in the following that:

- The rank of $\mathbf{X}$ is maximal, i.e.: $r=\mathrm{rank}(\mathbf{X})=\min(n-1,p)$;
- The dimension of the space spanned by $\mathbf{X}$ is high enough to build a maximal discriminant space, i.e.: $r\geq c-1$;
- The rank of $\mathbf{B}$ is maximal, i.e.: $\mathrm{rank}(\mathbf{B})=c-1$. In other words, all the class centroids are distinct.

2.1. The discrimination issue

A discriminant model is a function from $\mathbb{R}^p$ to $\{1,\ldots, c\}$ which associates an individual $\mathbf{x}$ to a class i. Let us consider the functions which operate in three steps: (i) projection of $\mathbf{x}$ in a subspace Z of $\mathbb{R}^p$; (ii) calculation of the membership degrees for the c classes, according to the distance between the projected individual and the class centroids; (iii) assignment to the class i corresponding to the higher degree. Operation (i) is of major importance for the performances of discrimination. Hence, a lot of theoretical developments are carried out about the Z subspace identification, which generally involves a learning process using $\mathbf{X}$ and $\mathbf{Y}$.

The linear discriminant analysis thus consists first in determining a basis of Z, i.e. a matrix $\mathbf{U}$, such that $\mathbf{Z}=\mathbf{XU}$ optimally separates the classes. Since the earliest contribution of Fisher in 1936 [5], several alternatives have been proposed, review of which can be found in [6]. Discriminant analysis is also used in pattern recognition. A significant number of methods were produced there, a survey is available in [7–9]. As far as we know, no contribution is dedicated to ill-conditioned problems. Usually, the columns of $\mathbf{U}$ are the vectors that maximize the Wilks' Lambda criterion, defined as the between class variance to the within class variance ratio ($\mathbf{u}_i^T\mathbf{B}\mathbf{u}_i/\mathbf{u}_i^T\mathbf{W}\mathbf{u}_i$). Other authors also used the between class variance to the total variance ratio ($\mathbf{u}_i^T\mathbf{B}\mathbf{u}_i/\mathbf{u}_i^T\mathbf{T}\mathbf{u}_i$) [10]. The latter, varying between 0 and 1 is chosen in this paper.

Let $\mathbf{u}$ be an unitary vector. We search for the maximum of $\Lambda(\mathbf{u})$, which is a linear form of $\mathbb{R}^p$. Equating its gradient to zero yields:

$$(\mathbf{B}-\Lambda(\mathbf{u})\mathbf{T})\mathbf{u}=0 \quad (1)$$

On the other hand, let $z\in[0,\ 1]$ so that $(\mathbf{B}-z\mathbf{T})\mathbf{u}=0$. After left-multiplying this expression by $\mathbf{u}^T$, we get $z=\Lambda(\mathbf{u})$. Then, the space answering the problem is included into the union of the kernels of $\mathbf{B}-z\mathbf{T}$, with $z\in[0,\ 1]$.

2.2. Dimensioning and conditioning of the problem

The problem is known as *well-dimensioned* when $n>p$. The $\mathbf{T}$ matrix is invertible, because $\mathrm{rank}(\mathbf{T})=\dim(\mathbf{T})=p$. The values of z for which $E(z)$ is not reduced to $\mathbf{0}$ are given by the equation: $\det(\mathbf{T}^{-1}\mathbf{B}-z\mathbf{I})=0$, meaning that z is an eigenvalue of $\mathbf{T}^{-1}\mathbf{B}$. Since $\mathrm{rank}(\mathbf{T}^{-1}\mathbf{B})\leq c-1$, we get $c-1$ values for z, which correspond to the solutions of the Factorial Discriminant Analysis (FDA). The space E is splitted up into $c-1$ monodimensional subspaces $\{E(z_1),\ldots, E(z_{c-1})\}$ plus one subspace $E(0)$ of dimension $p-c+1$. The Z space is identified as $\cup_{i=1\ldots c-1}E(z_i)$.

The problem is known as *ill-dimensioned* when $n\leq p$. In this case, $\mathbf{T}$ is no more invertible, because $\mathrm{rank}(\mathbf{T})=n-1<\dim(\mathbf{T})=p$. In addition, as $\mathbf{B}-z\mathbf{T}$ is built by means of linear combinations of $\mathbf{X}$, $\mathrm{rank}(\mathbf{B}-z\mathbf{T})\leq\mathrm{rank}(\mathbf{X})<p$. Then, $\mathrm{rank}(\mathbf{B}-z\mathbf{T})<\dim(\mathbf{B}-z\mathbf{T})$ and any value of z between 0 and 1 satisfies $\det(\mathbf{B}-z\mathbf{T})=0$. $E(z)$ is never reduced to $\mathbf{0}$ and Z cannot be identified by solving the Eq. (1).

The problem is known as *ill-conditioned* when some columns of $\mathbf{X}$ are highly correlated one to the others. Ill-conditioned problems appear especially when the individuals of $\mathbf{X}$ are curves. In this case, some eigenvalues of $\mathbf{T}$ are close to 0. The $\mathbf{T}$ matrix inversion is thus unstable, even if the problem is well-dimensioned. Ill-conditioning is in fact very close to ill-dimensioning: In the first case, only a given number r^* of eigenvalues of $\mathbf{T}$ are significantly non-null and $p-r^*$ eigenvalues are close to 0; In the second case,

$r=n-1$ eigenvalues of $\mathbf{T}$ are non null and $p-r$ are null. We will thus treat in this paper the ill-conditioned case in the same way as the ill-dimensioned one.

2.3. Description of the FEF-DA

Our method aims at calculating a sequence of spaces with an increasing discriminant power. In well-dimensioned cases, this sequence converges to the FDA solution. For ill-dimensioned problems, the sequence provides a set of solutions, to be tested with regard to a learning criterion, like a validation error. The construction of this sequence is based on the analysis of the eigenvalues of $\mathbf{B}-z\mathbf{T}$, with $z \in [0, 1]$.

Let $F_1(z),\ldots, F_r(z)$ be the r largest eigenvalues of $\mathbf{B}-z\mathbf{T}$ in absolute value, with $z \in [0, 1]$, so that $F_1(z) \geq F_2(z) \geq \cdots \geq F_{r-1}(z) \geq F_r(z)$. This defines a family of r functions from [0, 1] to $\mathbb{R}$. As the calculus of $F_i(z)$ is based on continuous functions, it is possible to build the F_i functions so that they are continuous. To each $F_i(z)$ correspond two opposite unitary eigenvectors. Let $\mathbf{u}_i(z)$ be the vectorial function, from [0, 1] to $\mathbb{R}^p$, which relates z to the unitary eigenvector associated to $F_i(z)$ so that $\|\mathbf{u}_i(z+h)-\mathbf{u}_i(z)\|$ tends to 0 when h tends to 0. In other words, $\mathbf{u}_i$ is built to be continuous.

FEF-DA consists in using the $c-1$ first F_i functions, called Focal Eigen Fonctions (FEFs), as a support for an iterative scanning algorithm. At each iteration step, the associated vectorial functions $\mathbf{u}_i$ define a discriminant space, to be tested. Unlike the factorial methods, where the space is primarily reduced, our method identifies the space by completely and directly taking into account its discrimination abilities.

In the following of this section, some general properties of the functions F_i are examined, which allows us to define the focal eigen functions. Then, other properties are presented and examples of scanning algorithms, taking advantage of these properties, are proposed. All the proofs are reported in annex.

2.3.1. General properties of the Fi functions

Two properties of the F_i functions are used to define the focal eigen functions.

Property 1. *Each F_i function is derivable and strictly decreasing; its derivative is given by:*

$$F_i'(z) = -\mathbf{u}_i^T(z)\mathbf{T}\mathbf{u}_i(z). \quad (2)$$

Property 2. *Only the functions $F_1, \ldots, F_{c-1}$ can equate 0 on [0, 1].*

2.3.2. Definition and properties of the Focal Eigen Functions

Let $i \leq c-1$; Let z_i^* be the zero of the function F_i, if F_i can be zeroed. Otherwise, $z_i^*=1$. The restriction of F_i to $[0, z_i^*]$ is called the ith Focal Eigen Function. The vectorial function $\mathbf{u}_i$ associated to F_i is called the ith Focal Eigen Vector. The analytical form of F_i is given by:

$$F_i(z) = \mathbf{u}_i^T(z)\mathbf{B}\mathbf{u}_i(z) - z\mathbf{u}_i^T(z)\mathbf{T}\mathbf{u}_i(z) \quad (3)$$

Now, let us examine some properties of these functions:

Property 3. *In well-dimensioned cases ($n > p$), the zero of a focal eigen function belongs to [0, 1].*

Property 4. *In ill-dimensioned cases ($n \leq p$), the focal eigen functions are strictly positive on [0, 1] and null at 1.*

Property 5. *The curvature of the focal eigen functions is positive.*

Property 6. *The discriminant power of the focal eigen vectors increases with z, i.e. that the function $L_i(z)=\Lambda(\mathbf{u}i(z))$ is increasing.*

These properties will be employed to optimize the course of the eigen functions.

2.3.3. Implementation of the focal eigen functions

It has been shown that each focal eigen function is strictly positive at $z=0$, strictly decreasing and has a zero z_i* on]0; 1], as illustrated on Fig. 1. In the well-dimensioned case, $\{\mathbf{u}_1(z_1^*),\ldots, \mathbf{u}_{c-1}(z_{c-1}^*)\}$ is the solution given by FDA. In the ill-dimensioned case, we have $z_i^*=1$ and the solution is obviously over-fitted (it corresponds to a Wilks' Lambda of 1). In the well-dimensioned but ill-conditioned case, the zeros of F_i give also the FDA solution, which should be also considered as over-fitted.

As the well-dimensioned and well-conditioned cases are well managed by FDA, let us focus on the ill-conditioned or/and ill-dimensioned problems. The FEFDA consists in scanning the focal eigen functions to build a sequence of discriminant spaces, defined by the associated focal eigen vectors. Due to Property 6, this sequence has got an increasing discriminant power, provided that the functions are scanned with increasing sequences, converging to z_i^*. This increasing discriminant power is necessary to test the sequence with respect to an overlearning criterion.

2.4. Examples of scanning algorithms

Three scanning methods are given, as examples.

2.4.1. Vertical scanning

This method simultaneously scans all the focal eigen functions, vertically, from $F_i(0)$ downto 0. It uses the normalized functions F_i^*, defined as $F_i^*(z)=F_i(z)=F_i(0)$. From the above properties, F_i^* obviously is a bijection from $[0, z_i^*]$ to [0, 1]. Then, let β be a decreasing sequence, starting from a value smaller than 1 and converging to 0. Let S_i be the sequence defined by $S_i(k)=F_i^{*-1}(\beta(k))$. This sequence is increasing and converges to z_i^*. The discriminant space Z_k built at the kth step of this algorithm is

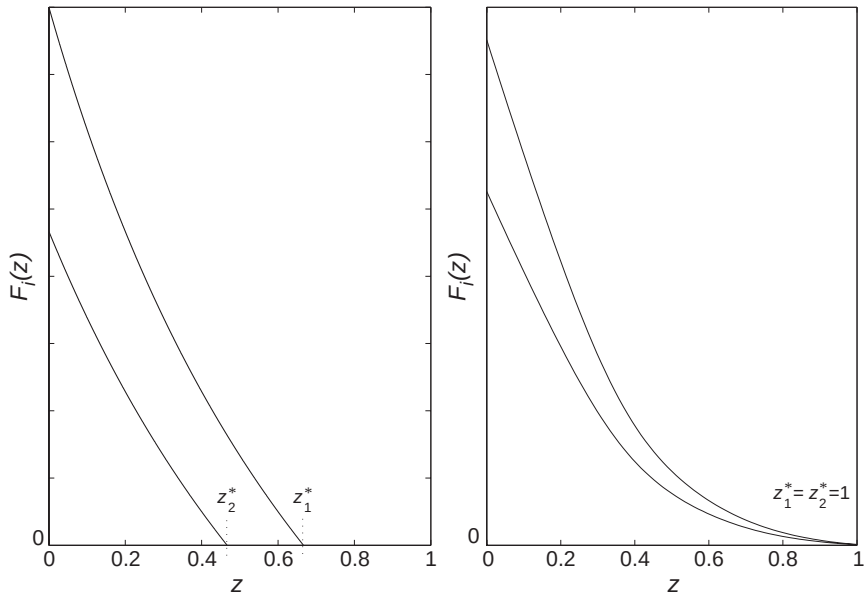

Fig. 1. A priori shape of the focal eigen functions, for 3 classes; on a well-dimensionned case (left) and on a ill-dimensioned case (right).

spanned by $\{\mathbf{u}_i(F_i^{*-1}(\beta(k)))\}_{i=1,\ldots,c-1}$. This method makes it possible to scan the whole functions. It has the advantage of proceeding with only one index. However, it supposes that the functions F_i behave similarly so that the same vertical cut corresponds to similar zones. Another drawback of this method is that it requires the inversion of $c-1$ functions at each step.

2.4.2. Asynchronous scanning

Unlike the former algorithm, asynchronous scanning processes each function F_i independently. Since each function F_i is positive, decreasing, and with positive curvature, applying the Newton-Raphson method to it produces a sequence s, which increasingly converges to z_i^*:

$$s(0) = 0$$

$$s(k+1) = s(k) - \frac{F_i(s(k))}{F_i'(s(k))}$$

Substituting $F_i'(s(k))$ by the form given by Eq. (2) and $F_i(s(k))$ by that given by Eq. (3) gives a very simple expression to s:

$$s(0) = 0$$

$$s(k+1) = \Lambda(\mathbf{u}_i(s(k))) = L_i(s(k))$$

The Newton Raphson method is known to converge quickly. To obtain a refined scanning, the following sequence s^* can be used:

$$s^*(0) = 0$$

$$s^*(k+1) = (1-\alpha_k) \times s^*(k) + \alpha_k \times L_i(s^*(k))$$

$$0 < \alpha_k \leq 1$$

The s^* sequence has the same convergence properties than s, but the lower α_k the lower the convergence speed. To adopt the same formalism as vertical scanning, rather than defining a sequence α_k, one can take an increasing sequence β, so that $\alpha_k = \beta(k+1) - \beta(k)$. As an advantage, the asynchronous scanning converges in a way adapted to each function, while using only one index.

2.4.3. Orthogonal scanning

Defining the discriminant space using orthogonal vectors offers some advantages: The space analysis becomes easier as the vectors convey complementary information one from the other. Moreover, the more independent the basis vectors, the more reliable the space. For these reasons, we propose a method for building orthonormal discriminant vectors, as an option of the two already described scanning algorithms.

A scanning (vertical, asynchronous or other) is performed on the first focal eigen function F_1, using a sequence s. For each $s(k)$, the scanning is run again on the $\mathbf{u}_1(s(k))$ orthogonal subspace, but always on the first function F_1, calculated on the new space. The procedure is summerized in the following recursive algorithm:

(1) $i = 1$
(2) For each $s(k)$, calculate $\mathbf{v} = \mathbf{u}_1(s(k))$
(3) If $i < c-1$:
  - Orthogonalisation of $\mathbf{X}$ with regard to $\mathbf{v}$: $\mathbf{X} = \mathbf{X}(\mathbf{I} - \mathbf{v}\mathbf{v}^T)$
  - $i = i+1$
  - Goto step 2
(4) Else, exit the algorithm.

The orthogonal scanning provides a tree of depth $c-1$. A node of depth k provided a orthonormal basis of a k-dimension discriminating space. This method thus gives a

significant number of solutions. But, the procedure requires a significant number of calculations, growing exponentially with respect to the number of classes.

3. Material and methods

3.1. Experimental data

The methods exposed in this article were applied on two data sets:

The first data set has been built from the Fisher's Iris famous data [5].[1] The raw data I_0 consisted of 150 individuals belonging to 3 groups of the same size ($n_1=n_2=n_3=50$) and described by 4 variables. These data has been transformed as following, in order to produce a ill-dimensioned and ill-conditioned problem (I_1): the initial matrix has been multiplied by the first 4 factors of a PCA carried out on a set of apple spectra, described by 176 variables. Thus, the new data was described by 176 variables, very correlated each to the others, since the rank of the new matrix was 4. Then, a uniform random noise was added, with an amplitude of 10^{-4} time the standard deviation of each variable, to finally obtain a matrix of rank 150 (149 after centering). Finally, a well-dimensioned but ill-conditioned problem was generated, by keeping the 40 first variables of I_1 data (I_2).

The second data set relied on varieties of wine grapes, to be discriminated by means of NIR and visible spectrometry. The spectra were measured in transmission on berries separated from the bunch, in laboratory conditions, with a ZEISS MMS1 spectrometer. The wavelengths ranged from 310 to 1100 nm. These data were collected within the framework of a project aiming at characterizing the sugar content and the acidity of wine grapes by NIR spectrometry. Thus, the berries were selected to span a great heterogeneity of maturity. Spectra were acquired by batches of 50 individuals. Each batch contained individuals of the same variety. The experimentation related to 3 varieties: *carignan* (crg), *grenache blanc* (grb) and *grenache noir* (grn). Only crg and grb varieties were measured on different batches, at various dates. From this base, a data base was created to constitute an example of ill-dimensioned and ill-conditioned problem. Its composition is given in Table 1. For crg and grb varieties, the training set and the test set are different batches, whereas for the grn variety, a batch of spectra was cut randomly in two equal parts. Thus, the calibration and test sets consisted of $n=125$ individuals described by $p=256$ variables.

[1] http://lib.stat.cmu.edu/DASL/Datafiles/Fisher'sIris.html.

Table 1
Constitution of calibration and test sets

Variety	Calibration		Test	
	Size	Date	Size	Date
crg	50	08 07 2003	50	08 26 2003
grb	50	08 07 2003	50	08 26 2003
grn	25	09 04 2003	25	09 04 2003

3.2. Model calibration

No model was tested on the first data set, which was used only to plot Fifunctions.

Four discriminant methods have been tested on the second data set: A PLS-DA (SIMPLS algorithm plus FDA on the scores), and the FEF-DA implemented with the three scanning algorithms detailed in Section 2. Whatever the method used, a leave-one-out cross-validation provided an error of cross-validation (CVE(%)), expressed as a percentage of bad classification. Its evolution has been monitored with regard to the following parameters:

- for PLS-DA, the number of latent variables: n_{LV};
- for the vertical and the asynchronous scannings: $\beta(k)$;
- for the orthogonal scanning: scanning of each level with the asynchronous method, using a sequence $\beta(k)$;
- for all the methods, the number of discriminant vectors: n_{DV}.

The best model of each method has been chosen by examining the evolution of CVE according to the parameters. Then, these models were applied to the test set. The results of this test were expressed with a prediction error (PE(%)) and a confusion matrix: $\mathbf{C}=\widehat{\mathbf{Y}}^t\mathbf{Y}$; c_{ij} is the number of individuals belonging to the class j and assigned to the class i. Assignment, during either the cross-validation or the test, was carried out by the minimum of the Mahalanobis distance in the discriminant space. Neither distance rejection nor ambiguity rejection was used.

4. Results and discussion

The graph of Fig. 2 (solid lines) shows the shape of the two eigen functions calculated on the data set I_0 (well-conditioned problem), with a logarithmic y-coordinate to enhance the small values. F_1 and F_2 equate 0, respectively, at 0.97 and 0.22, which correspond to the eigenvalues of $\mathbf{T}^{-1}\mathbf{B}$, solutions given by the FDA on the Fisher's data. On the same figure, but in dash-dot lines, the eigen functions of the data set I_1 (ill-dimensioned problem) are plotted. The zero of these functions is deffered to $z_1=z_2=1$. Inflection points remain at the location of the initial zeros (0.97 and 0.22). Finally, the eigen functions of the well-dimensioned but ill-conditioned problem (I_2) are drawn with dashed lines. These functions behave in an intermediate way

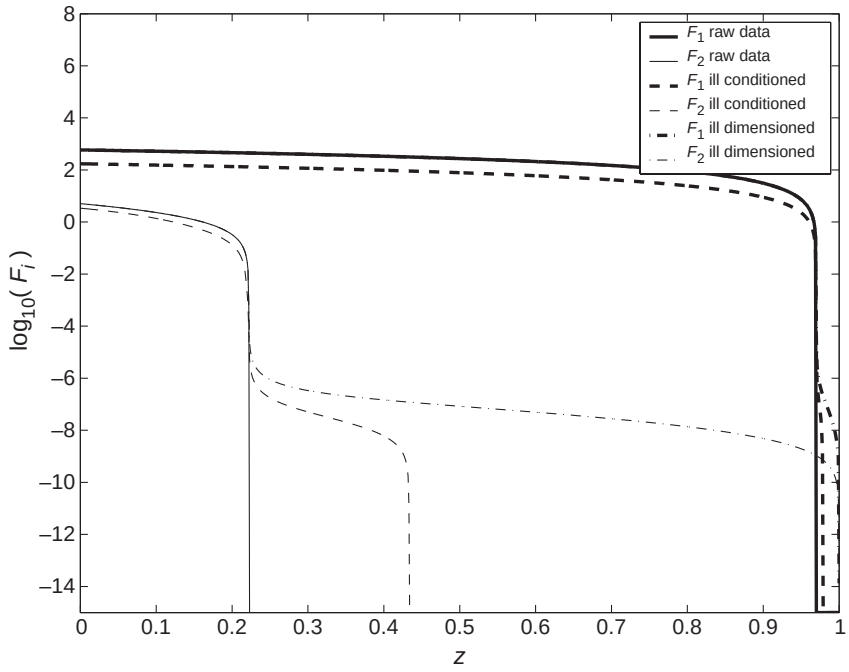

Fig. 2. Logarithm of F_i functions calculated on the Fisher's Irises (solid lines); on the same data artificially ill-conditioned (dashed line); and ill-dimensioned (dash-dot lines).

between the well-conditioned and the ill-dimensioned cases. They equate 0 for a value of z different from 1, but higher than the solution of the FDA carried out on the raw data. If a FDA was carried out on these data, the two discriminant axes would present Wilks' Lambda respectively equal to 0.98 and 0.43, instead of 0.97 and 0.22, which is obviously false. The shape of the functions is close to the ones corresponding to the ill-dimensioned case. This reinforces the idea to deal with the problem of conditioning in the same way as the dimensioning one.

The Fig. 3 shows a vertical scan on the data set I_2, with $\beta(k)=10^{-k}$. The Fig. 4 reports an example of asynchronous scanning, processed on the same data, with $\beta(k)=k$, i.e. using the standard Newton Raphson sequence. For both methods, it is noticeable that the scan is particularly refined at the location of the inflection points.

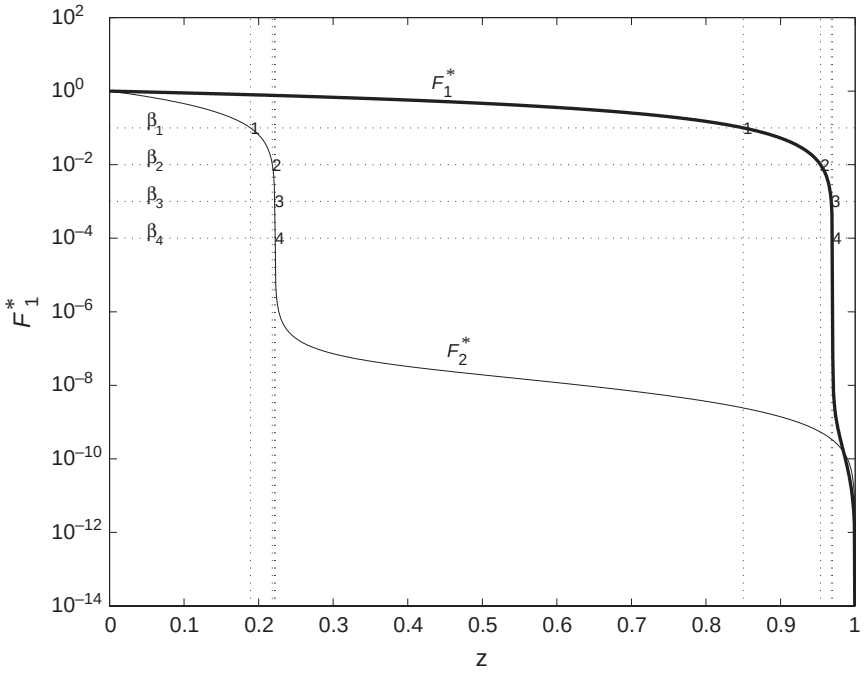

Fig. 3. Vertical scanning on the artificially ill-dimensioned Fisher's Iris, with $\beta(k)=10^{-k}$, $k \in \{1, 2, 3, 4\}$.

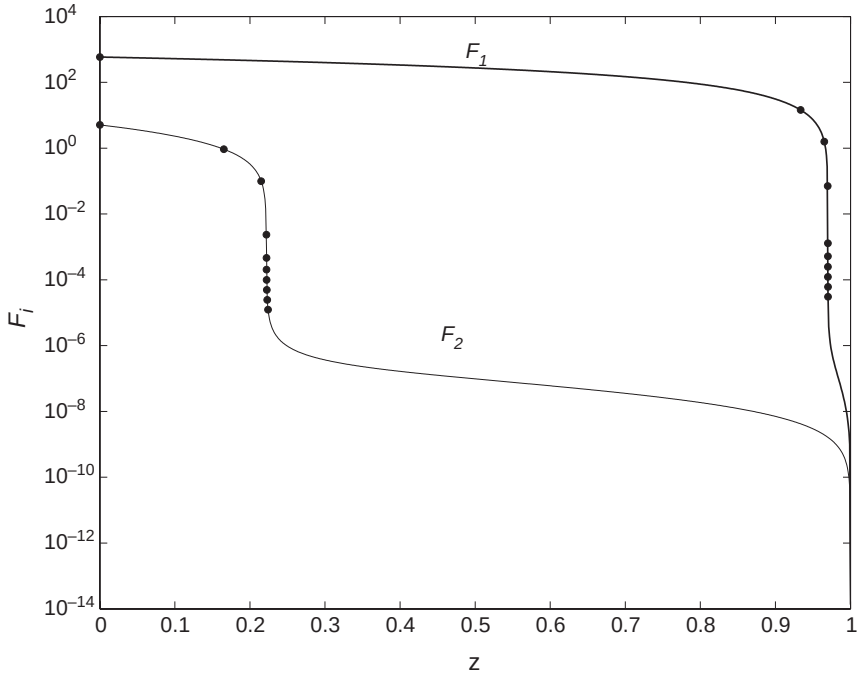

Fig. 4. Asynchronous scanning on the artificially ill-dimensioned Fisher's Iris, with $\beta(k)=1, 2,\ldots, 10$.

The following results address the second data set, devoted to wine variety discrimination.

The focal eigen functions F_1 and F_2 calculated on the calibration set are shown in Fig. 5a. They behave in accordance with the properties previously described. Fig. 5b uses a logarithmic y-coordinate, in order to magnify the low values.

The cross-validation error of the PLS-DA model, with regard to the number of PLS latent variables (n_{LV}) and of FDA discriminant vectors (n_{DV}) is plotted in Fig. 6a. The model performances are always better with $n_{DV}=2$. The model using 10 latent variables is retained; it corresponds to a cross-validation error (CVE) of 0.8%, i.e. 1 misclassified individual. On Fig. 6b, the same graph is reported for the vertical scanning. As above, using 2 discriminant vectors is more efficient. A value of $\beta=10^{-3.2}$ is kept as optimal. It corresponds to an error CVE=1.6%, i.e. 2 misclassified individuals. The cross-validation error for the asynchronous scanning is reported in Fig. 6c. The minimum value is CVE=0.8%, while using for β the sequence 1, 1.5, 2,. . .,

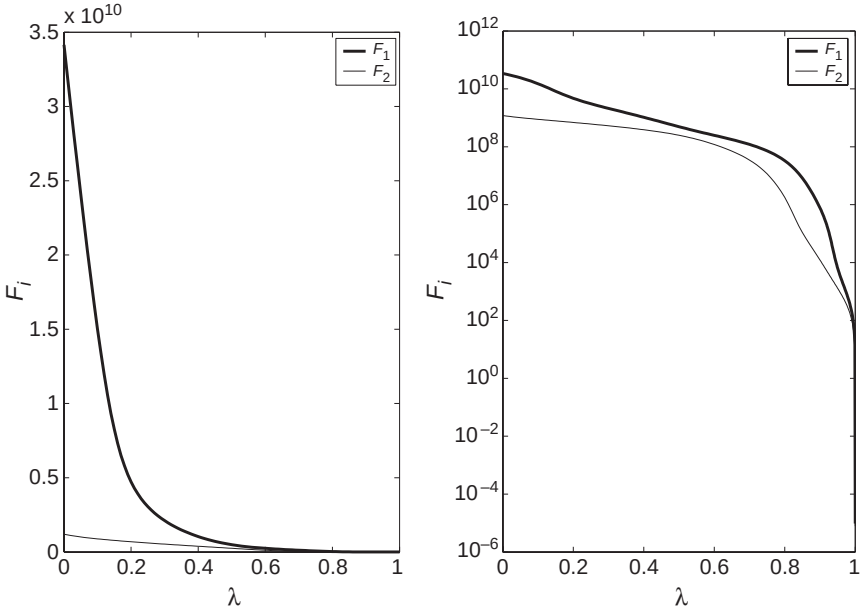

Fig. 5. Curves of F_1 and F_2; left (a), linear y-coordinate; right (b), logarithmic y-coordinate.

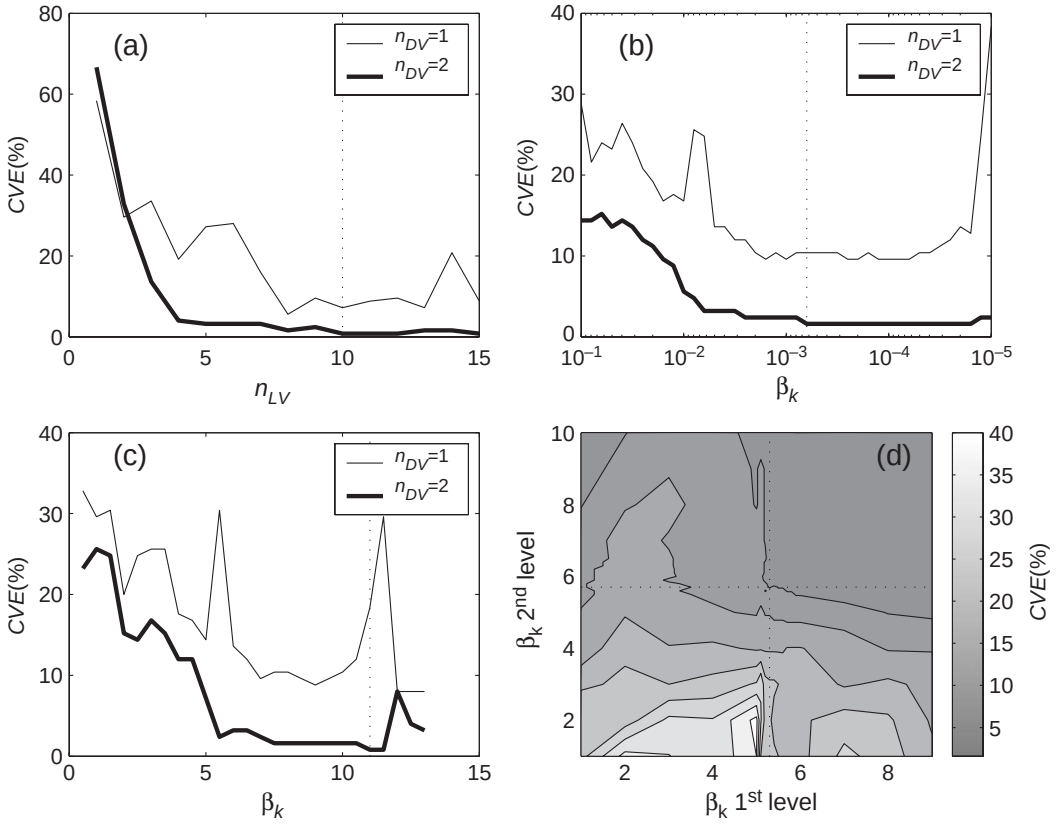

Fig. 6. Cross-validation error (CVE) evolution for: (a) PLS-DA, (b) vertical scanning, (c) asynchronous scanning and (d) orthogonal scanning.

10.5, 11, which has been retained. Last, Fig. 6d plots the cross-validation error for the orthogonal scanning, with regard to the two sequences β_1 and β_2 which were used at the two levels of the algorithm. A first trial was carried out with $\beta_1=\beta_2=1, 2,\ldots, 10$. A second one was processed, focussing on a refined zone around $\beta_1=5.5$ and $\beta_2=6$. The optimal value, finally retained is: $\beta_1=1, 2,\ldots, 5, 5.1, 5.2, 5.3$ and $\beta_2=1, 2,\ldots, 5, 5.5, 5.6,\ldots, 6$, corresponding to the left bottom corner of the plate zone of low errors. It corresponds to CVE=3.2%, i.e. 4 misclassified individuals.

The four models, parameterized as above and applied on the test set gave the results reported in Table 2. Firstly, it may be seen that all the models correctly predict the test sample, since the highest error is only 8%. The grn class is always perfectly discriminated; This may be due to the fact that, for this class, the calibration set and the test set were not really independent, since they were extracted of the same experiment, unlike the other classes. The grb class is very correctly predicted by the three FEF-DA methods (0–2% of error), but much less better by the PLS-DA (8% of error). The crg class is predicted almost perfectly by orthogonal scanning (2% of error), a little less better by asynchronous scanning (6% of error) and rather poorly by the two other methods (12% of error).

PLS-DA thus seems less efficient than FEF-DA, especially if the latter is implemented with orthogonal scanning. This first observation must nevertheless be moderated by the fact that the models, selected on the basis of the cross-validation error, can be over-fitted. Indeed, the best model in cross-validation (PLS-DA) appears the worst one in test, and conversely for the orthogonal scanning. Before concluding, it is thus advisable to examine the test of PLSDA for a smaller number of latent variables which is less likely to over-fitting. Fig. 6a shows two points in the CVE curve which could have been legitimately chosen: $n_{LV}=4$ or $n_{LV}=8$. The test results for the PLS-DA model with these new values are given by Table 3. The model with 4 latent variables is clearly under-fitted; it is not able to discriminate the class crg. The model

Table 2
Test results for the four discriminant models

PLS-DA				FEF-DA Vertical scanning			
$\widehat{Y}^T\mathbf{Y}$	crg	grb	grn	$\widehat{Y}^T\mathbf{Y}$	crg	grb	grn
crg	44	–	–	crg	44	–	–
grb	6	46	–	grb	6	49	–
grn	–	4	25	grn	–	1	25
PE=8.0%				PE=5.6%			
FEF-DA Asynchronous scanning				**FEF-DA Orthogonal scanning**			
$\widehat{Y}^T\mathbf{Y}$	crg	grb	grn	$\widehat{Y}^T\mathbf{Y}$	crg	grb	grn
crg	47	–	–	crg	49	–	–
grb	2	50	–	grb	–	50	–
grn	1	–	25	grn	1	–	25
PE=2.4%				PE=0.8%			

Table 3
Test result for the PLS-DA with 4 and 8 latent variables

$n_{LV}=4$				$n_{LV}=8$			
$\widehat{Y}^T\mathbf{Y}$	crg	grb	grn	$\widehat{Y}^T\mathbf{Y}$	crg	grb	grn
crg	37	–	–	crg	42	–	–
grb	12	50	–	grb	8	49	–
grn	1	0	25	grn	–	1	25
PE=10.4%				PE=7.2%			

with 8 latent variables is certainly a little better than the one with 10 latent variables, but always much less good than orthogonal scanning. In conclusion, on this test, the performances of PLS-DA are exceeded by those of the FEF-DA, especially when orthogonal scanning is used.

In Fig. 7 (top), the test set is projected into the discriminant space, for the FEF-DA model with othogonal scanning. The groups appear clearly separated. As a comparison, Fig. 8 (top) shows the same factorial map for the PLS-DA ($n_{LV}=10$). The crg and grb classes clearly overlap the grn class.

In Fig. 7 (bottom), the discriminant vectors of the orthogonal FEF-DA are drawn. In the 500–650 nm range, $\mathbf{u}_1$ presents a significant peak. As the spectra are in signal mode (and not in absorbance mode), this zone separates the berries according to their color and their transparency: The berries which are transparent and/or "green-yellow" gain in this area a positive coordinate on $\mathbf{u}_1$, unlike the more opaque and "purple" berries. This is in agreement with the factorial map: grb, which is a "white" grape variety (in fact rather yellow), is on the right while the other ones, which are "black" grapes varieties (in fact rather purple), are on the left. Similarly, the vector $\mathbf{u}_2$ separates the black grapes according to a finer decomposition from their color: we should find at the top of the map the reddest berries, because of the peak present at 650 nm and at the bottom, the greenest or yellowest berries, because of the negative peaks at 520 nm and 590 nm. Another zone also seems very discriminant, around 980 nm. It may be caused by the water absorption. It would seem that the berries of the class crg contain less water than those of the class grn (which certainly contain more dry matter, like sugars).

FEF-DA is a real alternative to factorial methods. With regard to PLS-DA, it owns some interesting properties:

- The discriminant vector calculations involve all the information. When a PLS-DA is calibrated, the choice of the latent variables cancels a certain amount of information. The factorial decomposition performed by the PLS step only ensures that the retained latent variables have higher covariance with the membership degrees than the cancelled ones. In [11], it is shown that the regression based discrimination is equivalent to a true discrimination only if the a priori probability densities are the same for all classes. The separability criterion used by the FEF-DA is based on much less strong assumptions. Then, FEF-DA should permit to realize more correct models than PLS-DA.
- The FEF-DA parameters are continuous values; then they can be adjusted with the desired resolution level. For PLS-DA, the parameter (n_{LV}) is discrete. Every refinement consists in adding a dimension to the model.
- The discriminant vectors are very little dependent, even orthogonal, as shown by Table 4, which reports cos($\mathbf{u}_1$,

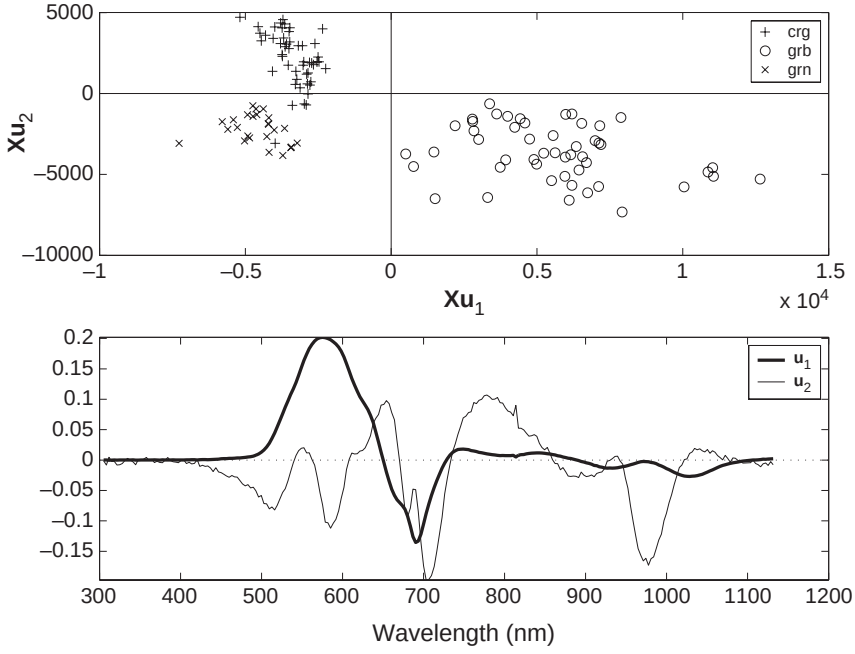

Fig. 7. Top: Factorial map of the test set projected by the orthogonal scanning model. Bottom: Discriminant vectors of model.

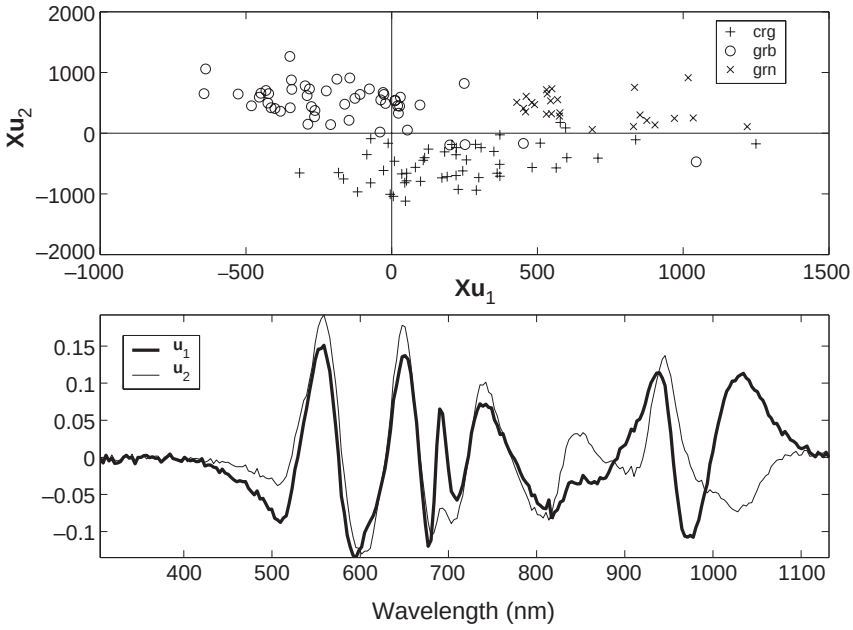

Fig. 8. Top: Factorial map of the test set projected by the PLS-DA model (n_{LV}=10). Bottom: Discriminant vectors of model.

$\mathbf{u}_2$) for the four models calibrated above. The first advantage of this property is algebraic: the more independent the discriminant vectors, the more reliable the Z space definition. The second one addresses the interpretation of the vectors (especially for spectrometry): if they are independent their interpretation is easier. As a comparative illustration, the discriminant vectors of PLS-DA are reported in Fig. 8, to be compared to those of orthogonal FEF-DA, Fig. 7.

5. Conclusion

In this paper, a new method of discriminant space identification has been presented. It is particularly adapted to ill-dimensioned and ill-conditioned problems, typical of spectrometry, while staying compatible with the classical discriminant analysis on well-dimensioned problems. By scanning a restricted number of scalar functions (Focal Eigen Functions), this method provides a sequence of spaces with an increasing discriminant power, which can then be selected with regard to a over-learning criterion. Some interesting properties of these Focal Eigen Functions, which improve the efficiency of the scanning algorithms, are proven. The potential of this new method, in comparison to the well known PLS-DA, is illustrated through a real case study of visible/NIR spectrometric data. A spectrometry oriented analysis of the discriminant vectors also shows the relevance of the calculated dicriminant space. The computing time required by this method nevertheless remains a disadvantage. Further work will be undertaken to highlight other properties of these functions, especially concerning their shape, and thus to improve the performances of calculation. The Focal Eigen Functions appear as very straightforward tools to investigate high dimensional spaces for discrimination problem.

Table 4
Cosine between the two discriminant vectors, for the four models

Method	PLS-DA 10 LV	FPF-AD vertical	FPF-AD asynchronous	FPF-AD orthogonal
cos($\mathbf{u}_1$, $\mathbf{u}_2$)	−0.6153	0.0084	0.0046	0.0000

Appendix A

Proof of Property 1. Each F_i function is derivable and strictly decreasing; its derivative is given by: $F_i'(z)=\mathbf{u}_i^T(z)\mathbf{T}\mathbf{u}_i(z)$. □

Proof. Let $z \in [0, 1]$; Let δz be a scalar so that $z+\delta z \in [0, 1]$; Let $\mathbf{u}=\mathbf{u}_i(z)$ and $\delta\mathbf{u}=\mathbf{u}_i(z+\delta z)-\mathbf{u}_i(z)$.

Computing F_i at $z+\delta z$ gives:

$$(\mathbf{B}-(z+\delta z)\mathbf{T})(\mathbf{u}+\delta\mathbf{u}) = F_i(z+\delta z)(\mathbf{u}+\delta\mathbf{u})$$

$$(\mathbf{B}-z\mathbf{T})\mathbf{u}+(\mathbf{B}-z\mathbf{T})\delta\mathbf{u}-\delta z\mathbf{T}\mathbf{u}-\delta z\mathbf{T}\delta\mathbf{u} = F_i(z+\delta z)\mathbf{u}+F_i(z+\delta z)\delta\mathbf{u}$$

Left-multiplying by $\mathbf{u}^T$, substituting $(\mathbf{B}-z\mathbf{T})\mathbf{u}$ by $F_i(z)\mathbf{u}$ and $\mathbf{u}^T\mathbf{u}$ by 1 yields:

$$F_i(z)+F_i(z)\mathbf{u}^T\delta\mathbf{u}-\delta z\mathbf{u}^T\mathbf{T}\mathbf{u}-\delta z\mathbf{u}^T\mathbf{T}\delta\mathbf{u} = F_i(z+\delta z)+F_i(z+\delta z)\mathbf{u}^T\delta\mathbf{u}$$

$$(F_i(z+\delta z)-F_i(z))\left(1+\mathbf{u}^T\delta\mathbf{u}\right) = -\delta z\left(\mathbf{u}^T\mathbf{T}\mathbf{u}+\mathbf{u}^T\mathbf{T}\delta\mathbf{u}\right)$$

$$\frac{F_i(z+\delta z)-F_i(z)}{\delta z} = \frac{\mathbf{u}^T\mathbf{T}\mathbf{u}+\mathbf{u}^T\mathbf{T}\delta\mathbf{u}}{1+\mathbf{u}^T\delta\mathbf{u}}$$

However, when $\delta z\to 0$, $\|\delta\mathbf{u}\|\to 0$, then $\mathbf{u}^T\mathbf{T}\delta\mathbf{u}\to 0$ and $\mathbf{u}^T\delta\mathbf{u}\to 0$. Finally:

$$\frac{F_i(z+\delta z)-F_i(z)}{\delta z}\to -\mathbf{u}^T\mathbf{T}\mathbf{u}$$

$$F_i'(z) = -\mathbf{u}_i^T(z)\mathbf{T}\mathbf{u}_i(z) < 0$$

Proof of Property 2. Only the functions $F_1,\ldots, F_{c-1}$ can equate 0 on [0, 1]. □

Proof. At $z=0$, $c-1$ eigenvalues $F_1(0),\ldots, F_{c-1}(0)$ are positive and $r-c+1$ eigenvalues $F_c(0),\ldots, F_r(0)$ are null, because rank($\mathbf{B}$)$=c-1$. Since F_i are strictly decreasing, $0>F_c(z)\geq F_{c+1}(z)\geq,\ldots,\geq F_r(z)$, $\forall z\in[0,\ 1]$. Consequently, only $F_1,\ldots, F_{c-1}$ can equate 0 on [0, 1].

Proof of Property 3. In well-dimensioned cases ($n>p$), the zero of a focal eigen function belongs to [0, 1]. □

Proof. For well-dimensioned problems, the zeros of $F_1,\ldots, F_{c-1}$ correspond to the $c-1$ solutions of the classical FDA. These values cannot be null, since we assumed that rank($\mathbf{B}$)$=c-1$, i.e. that $F_1(0),\ldots, F_{c-1}(0)\neq 0$.

Proof of Property 4. In ill-dimensioned cases ($n-p$), the focal eigen functions are strictly positive on [0, 1] and null at 1. □

Proof. For ill-dimensioned problems, $r=n-1$;

At $z=0$, $\mathbf{B}-z\mathbf{T}=\mathbf{B}$. Then $c-1$ eigenvalues are positive and $n-c$ eigenvalues are null. Then, $F_1(0)\geq F_2(0)\geq\ldots\geq F_{c-1}(0)>0$ and $F_c(0)=\cdots=F_{n-1}(0)=0$.

At $z=1$, $\mathbf{B}-z\mathbf{T}=-\mathbf{W}$ and rank($\mathbf{W}$)$=n-c$. Then $n-c$ eigenvalues are non-null and $c-1$ eigenvalues are null.

Since all the functions F_i are strictly decreasing, $F_1(1),\ldots, F_{c-1}(1)=0$ and $F_1(z)>0,\ldots, F_{c-1}(Z)>0$; $\forall z\in[0,\ 1]$.

Proof of Property 5. The curvature of the focal eigen functions is positive. □

Proof. Let $z\in[0,\ 1]$. Let $u=\mathbf{u}_i(z)$. Replacing $\mathbf{u}^T\mathbf{T}\mathbf{u}$ by $-F_i'(z)$ in 3 yields:

$$F_i(z) = u^TBu + zF_i'(z)$$

Deriving with respect to z yields:

$$F_i'(z) = \frac{\mathrm{d}}{\mathrm{d}z}\left(\mathbf{u}^T\mathbf{B}\mathbf{u}\right) + zF_i''(z) + F_i'(z)$$

$$F_i''(z) = -\frac{1}{z}\frac{\mathrm{d}}{\mathrm{d}z}\left(\mathbf{u}^T\mathbf{B}\mathbf{u}\right)$$

However, $\mathbf{u}$ is an eigenvector of $\mathbf{B}-z\mathbf{T}$. It is then clear that, the larger z, the smaller the part of $\mathbf{B}$ captured by $\mathbf{u}$, i.e smaller $\mathbf{u}^T\mathbf{B}\mathbf{u}$, then $(\mathrm{d}/\mathrm{d}z)(\mathbf{u}^T\mathbf{B}\mathbf{u})<0$. Then $F_i''(z)>0$.

Proof of Property 6. The discriminant power of the focal eigen vectors increases with z, i.e. that the function $L_i(z)=\Lambda(\mathbf{u}_i(z))$ is increasing. □

Proof. Let $z\in[0,\ z_i^*]$; From the Eqs. (3) and (2), we have:

$$\frac{F_i(z)}{F_i'(z)} = -\Lambda(u_i(z)) + z$$

$$L_i(z) = z - \frac{F_i(z)}{F_i'(z)}$$

Deriving with respect to z gives:

$$L_i'(z) = 1 - \frac{F_i'^2(z) - F_i(z)F_i''(z)}{F_i'^2(z)}$$

$$L_i'(z) = \frac{F_i(z)F_i''(z)}{F_i'^2(z)}$$

As $F_i(z)>0$ and $F_i''(z)>0$, $L_i'(z)>0$.

References

[1] C.W. Huck, R. Maurer, G.K. Bonn, Quality control of liquid plant extracts in the phytopharmaceutical industry in near infrared spectroscopy, in: A.M.C. Davis, R. Giangiacomo (Eds.), Near Infrared Spectroscopy: Proceedings of the 9th International Conference, NIR Publications, 2000, pp. 487–491.

[2] G. Lachenal, Structural investigations and monitoring of polymerisation by nir spectroscopy, Journal of Near Infrared Spectroscopy 1–4 (1998) 299–306.

[3] H. Martens, T. Naes, Multivariate Calibration, Wiley, New York, 1989.

[4] E. Kemsley, Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods, Chemometrics and Intelligent Laboratory Systems 33 (1996) 47–61.

[5] R. Fisher, The use of multiple measurements in taxinomic problems, Ann. Eugenics 7 (1936) 179–188.

[6] U. Indahl, N. Sahni, B. Kirkhus, T. Ns, Multivariate strategies for classification based on nir-spectra—with application to mayonnaise, Chemometrics and Intelligent Laboratory Systems 49 (1999) 19–31.

[7] D. Foley, J. Sammon, An optimal set of discriminant vectors, IEEE Transactions on Computers 24 (3) (1975) 281–289.

[8] K. Liu, Y. Cheng, J. Yang, An generalized optimal set of discriminant vectors, Pattern Recognition 25 (7) (1992) 731–739.

[9] W. Xiao-Jun, J. Kittler, Y. Jing-Yu, W. Shi-Tong, An analytical algorithm for determining the generalized optimal set of discriminant vectors, Pattern Recognition 37 (9) (2004) 1949–1952.

[10] L. Lebart, A. Morineau, K. Warwick, Multivariate Descriptive Statistical Analysis, Wiley and Sons, New York, 1984.

[11] B. Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, Cambridge, 1996.