JOURNAL OF CHEMOMETRICS, VOL. 4, 413–427 (1990)

STEPWISE CANONICAL DISCRIMINANT ANALYSIS OF CONTINUOUS DIGITALIZED SIGNALS: APPLICATION TO CHROMATOGRAMS OF WHEAT PROTEINS

DOMINIQUE BERTRAND

*Institut National de la Recherche Agronomique, Laboratoire de Technologie Appliquée à la Nutrition, BP 527,
Rue de la Géraudière, F-44026 Nantes Cedex 03, France*

PHILIPPE COURCOUX

*Ecole Nationale d'Ingénieurs des Techniques des Industries Agricoles et Alimentaires, Chaire de Mathématique,
Rue de la Géraudière, F-44072 Nantes, Cedex 03, France*

JEAN-CLAUDE AUTRAN AND REGIS MERITAN

Institut National de la Recherche Agronomique, Place Viala, F-34060, Montpellier Cedex, France

AND

PAUL ROBERT

*Institut National de la Recherche Agronomique, Laboratoire de Technologie Appliquée à la Nutrition, BP 527,
Rue de la Géraudière, F-44026 Nantes Cedex 03, France*

SUMMARY

Continuous digitalized signals such as spectra, electrophoregrams or chromatograms generally have a large number of data points and contain redundant information. It is therefore troublesome performing discriminant analysis without any preliminary selection of variables. A procedure for the application of canonical discriminant analysis (CDA) on this kind of data is studied. CDA can be presented as a succession of two principal component analyses (PCAs). The first is performed directly on the raw data and gives PC scores. The second is applied on the gravity centres of each qualitative group assessed on the normalized PC scores. A stepwise procedure for selection of the relevant PC scores is presented. The method has been tested on an illustrative collection of 165 size-exclusion high-performance (SE-HPLC) chromatograms of proteins of wheat belonging to 55 genotypes and grown in three locations. The discrimination of the growing locations was performed using seven to nine PC scores and gave more than 86% accurate classifications of the samples both in the training sets and the verification sets. The genotypes were also rather well identified, with more than 85% of the samples correctly classified. The studied method gives a way of assessing relevant mathematical distances between digitalized signals according to qualitative knowledge of the samples.

KEY WORDS Discriminant analysis Size-exclusion chromatography Wheat proteins

INTRODUCTION

Discriminant analyses (DAs) are 'supervised learning' methods[1] in which knowledge of the category of samples of a training set makes it possible to develop a classification procedure applicable to unknown samples. The aim of these methods is to predict the qualitative category of samples while knowing the values of a set of predictive variables.[2] Different authors have applied DA on continuous digitalized signals such as spectra, chromatograms or electrophoregrams. Numerous applications have been developed in near-infrared (NIR)

0886–9383/90/060413–15$07.50

Received 17 November 1989
Accepted (revised) 26 June 1990

spectroscopy using the absorbances at various wavelengths as discriminant variables. Mark and Tunnell[3] described a DA procedure applied to spectral data of NIR filter instruments and showed the usefulness of the Malahanobis distance to estimate the similarity between samples and qualitative groups. Bertrand *et al.*[4] attempted to identify wheat cultivars of samples from their NIR spectra. Devaux *et al.*[5] have done similar work in order to grade wheat samples into groups of baking quality. Downey *et al.*[6] used DA to classify skimmed milk powders according to heat treatment. Electrophoregrams have seldom been studied by means of DA. Autran and Abbal[7] applied a computer-aided procedure involving several steps. Three 'similarity indices' between the unknown electrophoregrams and those of known cultivars were estimated from the number of matching bands or nearly matching bands. This procedure had the advantage of closely resembling the human way of identification and of not being too sensitive to the possible shift of the observed bands. However, it did not enable the shape of the peaks to be taken into account. Virion[8] used both DA and computerized identification keys to discriminate wheat cultivars from electrophoregrams.

The application of classical DA on digitalized signals presents certain difficulties. The number of data points is often very large. Spectra or chromatograms may include several hundreds of variables, depending on the measurement intervals. Assessment of the Malahanobis distance involves the inversion of the 'variance–covariance' matrix of the training set. This matrix has dimension equal to $v \times v$ where v is the number of measured variables. Moreover, digitalized signals are often highly redundant: two adjacent data points give almost the same information; the number of digitalized data points can be increased without increasing the independent information which can be extracted from the data collection. If one variable is entirely correlated to any other, the inversion of the variance–covariance matrix cannot be made. Two approaches have been developed to overcome these problems. DA can be performed on a subset or a small number of independent variables. Romeder[9] has developed various algorithms to choose the most discriminant variables. These algorithms are, however, hardly applicable on microcomputers when the initial number of data points is considerable. Data can be put into a more condensed form by using orthogonal transformations such as Fourier transform (FT) before performing DA.[10,11] FT is very efficient but the condensed signal which is obtained cannot be interpreted by the specialist: the values of Fourier coefficients have no immediate meaning. Devaux *et al.*[12] showed the usefulness of presenting DA as a succession of two principal component analyses (PCAs). The whole signal was used without data reduction. Their procedures made it possible to model NIR spectra as a sum of 'discrimant patterns' representative of the discrimination and to have factorial maps showing the qualitative similarity between samples. Because the whole of the spectra were used, the efficiency of the discrimination was in some cases very different between the training set and the evaluation set. Certain principal components artificially presented a discriminant ability on the training set which is not confirmed in routine application of the developed DA. The present work is an attempt to combine the advantages of factorial analyses with a stepwise procedure which introduces only the more relevant pieces of information. Moreover, the case of digitalized signals, where the number of variables is often higher than the number of samples of the training set, is developed. The method has been applied on chromatograms of proteins of wheat differing by their genotypes and their areas of cultivation.

THEORY AND MATHEMATICAL PROCEDURE

It is necessary to briefly recall the procedure of canonical discriminant analysis (CDA) and stepwise discriminant analysis (SDA). The studied procedure is then presented.

General algorithm of canonical discriminant analysis

This procedure, also called 'factorial discriminant analysis', has been used extensively in ecology and other sciences.[13,14] Let us suppose that the training set is represented by a matrix **M**, the dimension of which is $n \times v$ (rows × columns), with n being the number of samples ('observations') and v the number of variables. Each observation can be attributed to a qualitative group G_k which includes n_k samples. Let h be the number of qualitative groups.

The matrix **M** is first centred, i.e. the vector of average variables is subtracted from each row of **M**:

$$\mathbf{x}_i = \mathbf{m}_i - \mathbf{a} \qquad (1)$$

where $\mathbf{x}_i$ is the vector representing the ith row of the centred matrix **X**, $\mathbf{m}_i$ is the corresponding row vector of the matrix **M** and **a** is the $1 \times v$ sample average.

Each of the n centred observations $\mathbf{x}_i$ can be represented as a point in a vector space having v dimensions. CDA creates a new vector space in which the qualitative groups are better separated than in the original one. The observations are characterized by a new set of variables called the 'discriminant scores':

$$\mathrm{S} = \mathbf{X}\mathbf{F}^{\mathrm{T}} \qquad (2)$$

where S is the $n \times f$ matrix of discriminant scores, **F** is the $f \times v$ matrix of discriminant factors and $\mathbf{F}^{\mathrm{T}}$ is the transpose of **F**. The problem is therefore to assess the matrix **F** in order to have the largest separation of the groups.

The 'gravity centre' $\mathbf{g}_k$ of each group k is calculated as

$$\mathbf{g}_k = (1/n_k)\Sigma\{\mathbf{x} \mid \mathbf{x} \in G_k\} \qquad (3)$$

where $\mathbf{g}_k$ represents the $1 \times v$ vector of average values of the n_k observations **x** attributable to group G_k. The h gravity centres can be gathered in an $h \times v$ matrix **G**.

The total variance–covariance matrix **T** is assessed according to

$$\mathbf{T} = \mathbf{X}^{\mathrm{T}}\mathbf{X} \qquad (4)$$

The 'total' matrix **T** can be split into two other matrices:

$$\mathbf{T} = \mathbf{W} + \mathbf{B} \qquad (5)$$

where **W** is the $v \times v$ 'within' matrix which takes into account the variations of the observations within each group and **B** describes the variations between the groups. **B** is estimated according to

$$\mathbf{B} = \mathbf{G}^{\mathrm{T}}\mathbf{H}\mathbf{G} \qquad (6)$$

where **H** is an $h \times h$ diagonal matrix such that $h_{ii} = n_i$, the number of observations of the ith group ($i = 1, \ldots, h$), and $h_{ij} = 0$ (with $i \neq j$). **B** can be seen as a variance–covariance matrix derived from **X** in which each observation $\mathbf{x}_i$ is replaced by its corresponding gravity centre $\mathbf{g}_k$.

It can be shown that the discriminant factors **F** are the eigenvectors of the matrix product $\mathbf{T}^{-1}\mathbf{B}$. The number f of discriminant factors is always less than the number h of qualitative groups. From **F** and **X** the scores **S** can be assessed according to (2). Similarly, the scores of the gravity centres **J** ($h \times f$) are given by

$$\mathbf{J} = \mathbf{G}\mathbf{F}^{\mathrm{T}} \qquad (7)$$

The classification into groups is performed as above from **J** and **S** using the Euclidean distance

$$d_k^2 = (\mathrm{s} - \mathbf{j}_k)(\mathrm{s} - \mathbf{j}_k)^{\mathrm{T}} \tag{8}$$

where s $(1 \times f)$ is the vector of discriminant scores of the observation to be classified and $\mathbf{j}_k$ is the discriminant scores of the group k. The unknown observation is attributed to the group k giving the smallest distance d_k.

Stepwise discriminant analysis

Since some variables may have no discriminant ability, it is worth choosing a relevant subset of the original variables. Romeder[9] showed that variables can be efficiently introduced one after the other. His criterion for introducing a new variable is to maximize the trace of the matrix defined by $\mathbf{T}^{-1}\mathbf{B}$.

Supposing that m variables among v have been introduced at the mth iteration. The procedure consists of assessing any matrix $\mathbf{T}_q^{-1}\mathbf{B}_q$ using the m previously introduced variables and one of the $v - m$ remaining variables. All the values of the $v - m$ traces are then compared. The variable q which gives the largest trace is introduced at iteration $m + 1$ and basic DA is applied on the selected variables. The relevance of the current subset of variables is evaluated by counting the observations of the training set which are rightly reallocated in their actual group. The procedure is iterated with the $v - (m - 1)$ variables if the number of correct classifications is increased.

Stepwise canonical discriminant analysis (SCDA)

It has been shown[15,16] that CDA can be achieved by a succession of two principal component analyses (PCAs). Devaux *et al.*[12] have described the procedure.

A first PCA is performed on the centred matrix $\mathbf{X}$ and gives eigenvectors of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ forming the matrix $\mathbf{U}$, non-null eigenvalues in $\mathbf{I}$ and PC scores $\mathbf{C}$. $\mathbf{C}$ is given by

$$\mathbf{C} = \mathbf{X}\mathbf{U}^{\mathrm{T}} \tag{9}$$

Since the original data are often redundant, the number of components is generally less than v and equal to the number of non-null eigenvalues. Let a be this number. The dimensions of $\mathbf{C}$ are therefore $n \times a$, those of $\mathbf{U}$ are $a \times v$ and $\mathbf{I}$ has dimensions $1 \times a$.

The theory of PCA shows that the matrix $\mathbf{C}^{\mathrm{T}}\mathbf{C}$ is a diagonal matrix $\mathbf{E}$ with the eigenvalues as diagonal elements:

$$\mathbf{C}^{\mathrm{T}}\mathbf{C} = \mathbf{E} \tag{10}$$

with $e_{ii} = l_i$ and $e_{ij} = 0$ (with $i \neq j$). $\mathbf{C}$ is normalized by the corresponding eigenvalues and gives the matrix of normalized PC scores $\mathbf{Y}$:

$$\mathbf{Y} = \mathbf{C}\mathbf{E}^{-1/2} \tag{11}$$

Combining (10) and (11) shows that $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ is a unit matrix with dimensions $a \times a$.

CDA or stepwise DA is easily performed using $\mathbf{Y}$ instead of $\mathbf{X}$ as the data of the training set. Since the new 'total' matrix $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ is unity, the matrix homologous to $\mathbf{T}^{-1}\mathbf{B}$ is reduced to the new 'between' matrix calculated on the gravity centre of $\mathbf{Y}$. The gravity centres are therefore calculated similarly to (3) applied on $\mathbf{Y}$ and give a matrix (homologous to $\mathbf{G}$) called $\mathbf{P}$ $(h \times a)$.

The 'between' matrix is assessed similarly to (6) and is therefore equal to $\mathbf{P}^T\mathbf{HP}$. CDA can be achieved by assessment of the eigenvectors of $\mathbf{P}^T\mathbf{HP}$ with all the components.

The Romeder procedure, performed on **Y** instead of **X**, is simplified because the criterion for introducing a principal component is now to maximize the trace of $\mathbf{P}^T\mathbf{HP}$ which can be assessed without matrix inversion. The elements of the diagonal of $\mathbf{P}^T\mathbf{HP}$ are given by

$$\text{diag}_j = \Sigma g_i p_{ij}^2 \quad \text{for} \quad i = 1, \ldots, h \quad \text{and} \quad j = 1, \ldots, a \tag{12}$$

where p_{ij} is an element of the matrix **P** and diag_j is an element of the vector **diag** ($1 \times a$) containing the diagonal elements of $\mathbf{P}^T\mathbf{HP}$. The order of introduction of components is given at once by classification of the elements of **diag**. Components must be introduced stepwise in decreasing order of the corresponding values of **diag**.

At the qth iteration of the introduction procedure the component having the largest remaining value in **diag** is introduced. Subsets of **Y** and **P** with only the q currently selected components are created. Let $\mathbf{Y}_q$ ($n \times q$) and $\mathbf{P}_q$ ($h \times q$) be the current matrices of selected normalized PC scores and gravity centres respectively.

A second PCA is applied on $\mathbf{P}_q$ by diagonalization of $\mathbf{P}_q^T\mathbf{HP}_q$ and gives the discriminant factors $\mathbf{F}^q$. The discriminant scores are then calculated on normalized PC scores $\mathbf{Y}_q$ and gravity centres $\mathbf{P}_q$ similarly to (2) and (7) by

$$\mathbf{S}_q = \mathbf{Y}_q\mathbf{F}_q^T \tag{13}$$

$$\mathbf{J}_q = \mathbf{P}_q\mathbf{F}_q^T \tag{14}$$

The classification is performed by calculating the Euclidean distances between observations and gravity centres as in (8). The procedure is reiterated until all the components have been introduced or all the observations are correctly classified.

Unknown observations can then be classified in the same way, after assessment of their discriminant scores.

If the studied variables are elements of a continuous digitalized signal, it may be worth examining the 'discriminant patterns', which show bands describing the discrimination. These patterns can be assessed by

$$\mathbf{V} = \mathbf{FE}^{-1/2}\mathbf{U} \tag{15}$$

Implementation of SCDA on computer

The program particularly needs a procedure for diagonalization of symmetric matrices, e.g. the Givens–Householder algorithm, and for multiplication of matrices. The critical steps include the assessment of the variance–covariance matrix $\mathbf{X}^T\mathbf{X}$ and the first diagonalization giving the eigenvectors **U**. In the case when the number v of variables is greater than the number n of observations, it is possible to assess eigenvectors by diagonalization of $\mathbf{XX}^T$, which has dimensions $n \times n$, rather than $\mathbf{X}^T\mathbf{X}$. Let **V** ($a \times n$) be the unit eigenvectors of $\mathbf{XX}^T$. The PC scores are given by

$$\mathbf{C} = \mathbf{V}^T\mathbf{E}^{1/2} \tag{16}$$

and the eigenvectors of $\mathbf{XX}^T$ are

$$\mathbf{U} = \mathbf{VXE}^{-1/2} \tag{17}$$

The number of data points in the studied signal is therefore not critical.

MATERIAL AND METHODS

Sample collection and chromatograms

The procedure was applied on a collection of 165 wheat samples grown in 1987, supplied by INRA (Institut National de la Recherche Agronomique) and 'Club des Cinq', an association of wheat breeders. Each of the 55 genotypes under study was grown in three locations, far apart geographically.

The flour samples were stirred in the presence of a buffer (pH 6·9) and centrifuged in order to extract the proteins. SE-HPLC was achieved on the supernatant. The sample preparation and chromatographic conditions have been described by Dachkevitch and Autran.[17]

The chromatograms were digitalized and stored on an IBM PC. Only the period of time when the peaks appeared were recorded, between 7 and 22 min at intervals of 6 s. Each observation was therefore characterized by 151 data points.

The mathematical procedure was applied three times by changing the samples in the training set and the evaluation set. For each trial the training set and the verification set included 120 and 45 observations respectively.

SCDA was performed twice on the same set of data in order to discriminate locations and genotypes separately.

In the case of the discrimination of genotypes, there were only three observations in each qualitative group. It seemed irrelevant to divide the collection by including two replications in the training set and the third one in the verification set. Forty genotypes of three locations were therefore included in the training set and the 15 × 3 remaining genotypes formed the verification set. The procedure was slightly modified for testing the verification set. SCDA was performed as described on the training set. PC scores of the verification set were assessed and normalized similarly to (9) and (11), giving the matrix $\mathbf{Y}_{ver}$. The gravity centres of the 15 genotypes of the verification set were estimated on $\mathbf{Y}_{ver}$, giving the matrix $\mathbf{P}_{ver}$. The discriminant scores of the observations of the verification set and those of their gravity centres were then assessed similarly to (13) and (14) applied on $\mathbf{Y}_{ver}$ and $\mathbf{P}_{ver}$. The observations of the verification set were then tentatively reclassified in their group of genotype using the Euclidean distance as in (8). The results were compared with those obtained by random allocation of observations of the verification set into 15 groups of three samples. In this way it was possible to test the relevance of the procedure for characterizing new genotypes not present in the training set.

RESULTS AND DISCUSSION

Reference data

Figure 1 shows the chromatograms obtained with various cultivars. According to Dachkevitch and Autran,[17] there are four areas of material absorbing at 214 nm. Upon calibration of the column using five molecular weight standards, the limits between peaks were estimated. Peak F1 elutes at the void volume which corresponds to about 1000 kDa (kilodalton: atomic mass unit) of the column and is likely to correspond to highly aggregated material. Fraction F2, which elutes between 115 and 650 kDa, does not make up a real peak and is likely to consist of smaller aggregates with a continuous range of molecular size. Peaks F3 and F4 correspond to monomeric proteins whose apparent molecular weights agree with the bulk of gliadins and salt-soluble proteins respectively. Direct observation of the

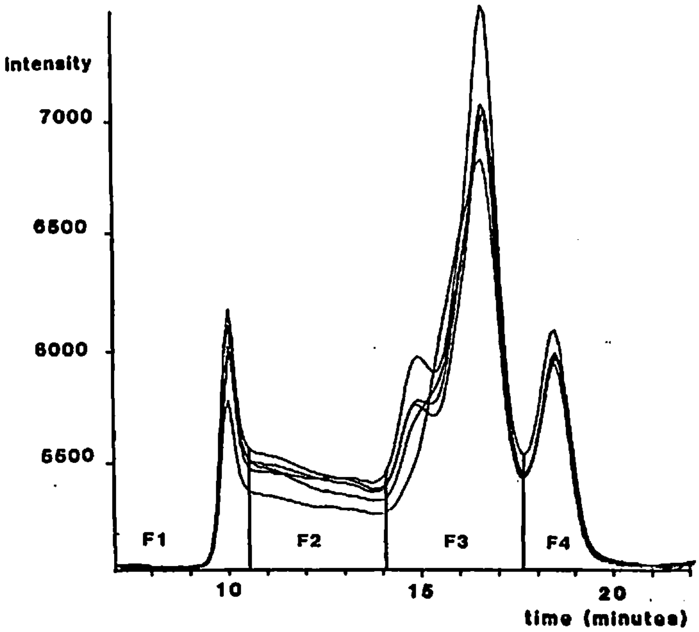

Figure 1. Examples of size-exclusion chromatograms of wheat proteins. F1, F2, F3, F4: areas described by Dachkevitch and Autran[17]

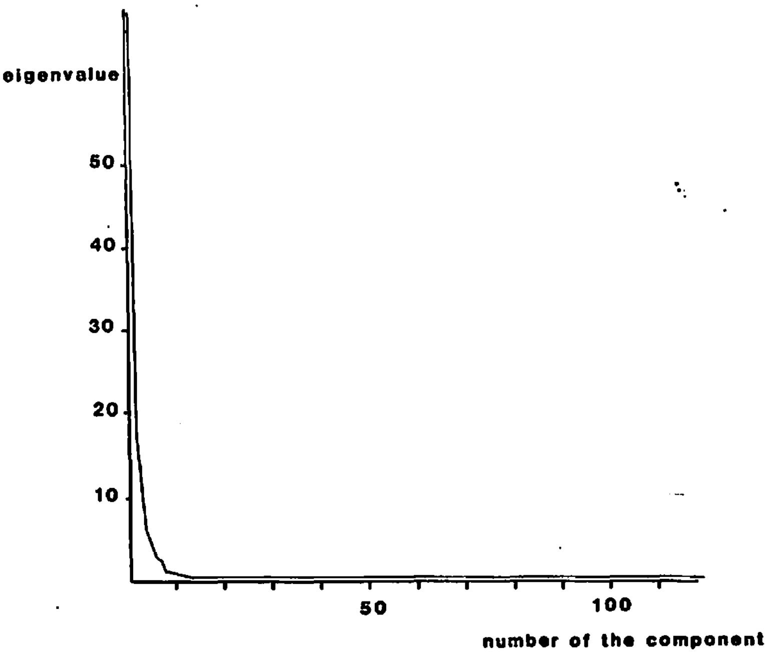

Figure 2. Eigenvalues of the principal component analysis performed on the collection of chromatograms

chromatograms gave little information on the differences between varieties. The only observable differences were a variation of the height of the baseline and weak peaks at 14–15·5 min.

Discrimination between locations

For each of the three trials the eigenvalues of the first PCA decreased very rapidly according to the number of components. The first trial is taken as an example. Less than 20 components were sufficient to include about 100% of the total sum of squares (Figure 2). SCDA was therefore applied on the first 20 components in order to predict the area of cultivation of each kind of wheat. Figure 3 shows the effect of the introduction of each component on the percentage of correctly classified samples. According to the trial, the introduction of seven to

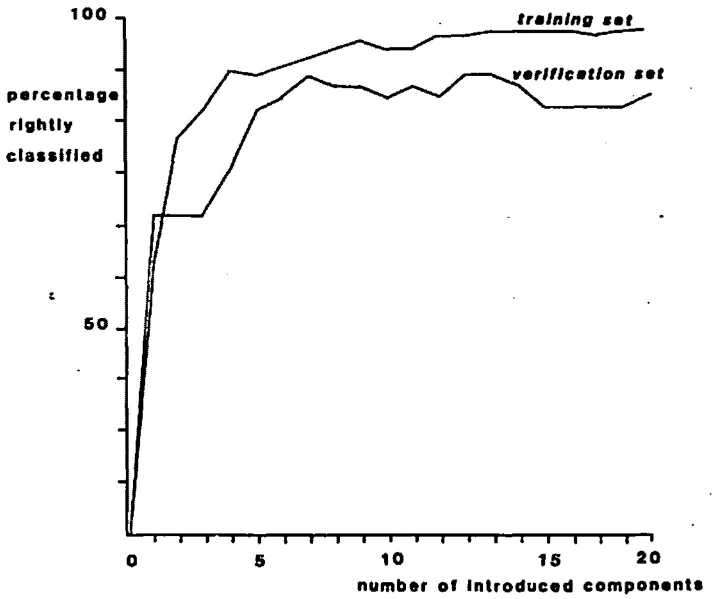

Figure 3. Discrimination of growing locations: influence of the number of principal components introduced on the number of correctly classified observations

Table 1. Discriminations of growing locations and genotypes

	Discrimination of growing locations			Discrimination of genotypes		
Trial	(1)	(2)	(3)	(1)	(2)	(3)
1	7	92·4	88·9	20	89·9	86·7
2	9	94·1	86·7	16	90·8	88·7
3	9	90·6	91·5	14	86·3	85·1

(1) Number of introduced components.
(2) Percentage of samples correctly classified in the training set (120 samples).
(3) Percentage of samples correctly classified in the verification set (45 samples).

nine components among 20 gave the best classification of the verification set. The proportion of samples correctly classified ranged from 90·6% to 94·1% for the training set and from 86·7% to 91·5% for the verification set (Table 1). The introduction of the remaining components gave no improvement.

The order of selection of the components was not related to the size of the corresponding eigenvalues: for instance, in the first trial the first introduced component was the seventh in the order of eigenvalues and represented only 1·8% of the cumulated variance (Table 2). In contrast, the first component, representing 60·7% of the cumulated variance, had no predictive ability. This meant that there was no connection between the intensity of each component and its discriminant ability. These results were not surprising. PCA concentrates

Table 2. Introduction of components in stepwise canonical discriminant analysis example of discrimination of growing locations (trial 1)

Number of introduced components	Percentage of cumulated variance
7	1·8
4	5·1
9	0·6
5	3·5
2	15·6
8	0·7
13	0·1

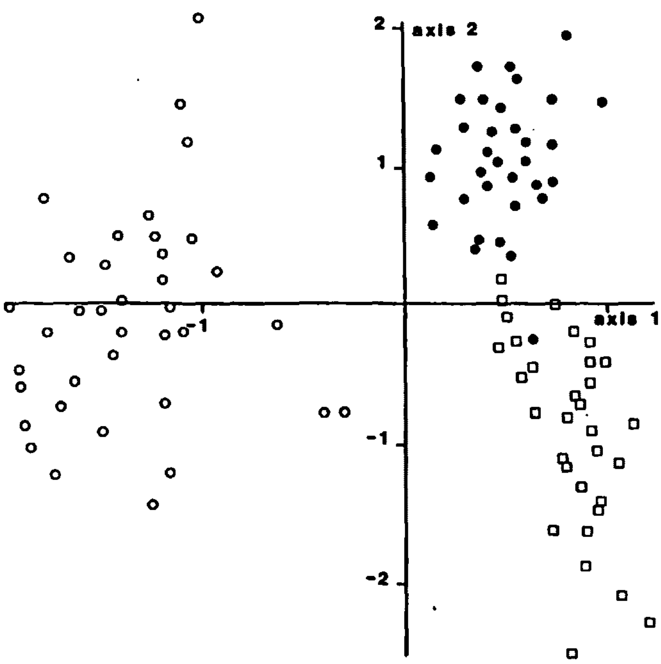

Figure 4. Factorial map of the discrimination of growing locations

the data to the most dominant dimensions. In PCA, irrelevant variations of the chromatograms such as baseline deformations are taken into account as well as significant chromatographic differences. This example shows that the selection of components relevant for discrimination is essential. Since there were only three groups to be discriminated, each chromatogram was characterized by only two discriminant scores. Figure 4 shows the map of the first trial representing the discrimination at the seventh step of the introduction procedure. Each area of cultivation was quite clearly separated. The maps of the second and third trials (not presented here) were very similar.

The discriminant pattern corresponding to the first discriminant score is given in Figure 5. This pattern presented three positive peaks at 9·7, 17·4 and 19·3 min and three negative peaks at 10·0, 16·7 and 18·5 min. It was theoretically representative of the part of the discrimination due to the first discriminant score. This is shown in Figures 6 and 7. Ten chromatograms having negative values of the first discriminant score were averaged and gave the curve labelled A in Figure 6. The same assessment has been made with ten samples having positive values, giving curve B. The averaged values mainly differed in the size of the peaks at 9·7, 17·4 and 19·3 min. The difference between curves A and B was almost identical to the discriminant pattern (Figure 7). Examination of curves A and B and of their difference showed that chromatograms of location A had higher maxima and deeper minima than those of location B. A first explanation could involve a difference in the resolving power of the HPLC columns between the runs of the various samples. This explanation cannot be totally ruled out, although in this study the chromatographic analyses were carried out using the same column for all the samples. One other and more basic hypothesis could be suggested. Chromatograms are representative of the distribution of protein aggregates which include various sub-units such as glutenins having low or high molecular weights or gliadins. The degree of aggregation might vary according to the growing conditions. A high-baking-quality sample of a given

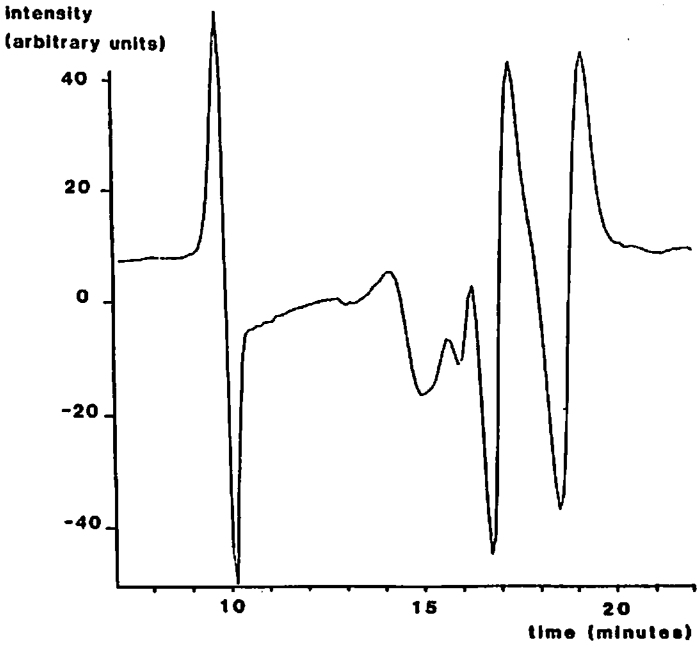

Figure 5. First discriminant pattern of growing locations

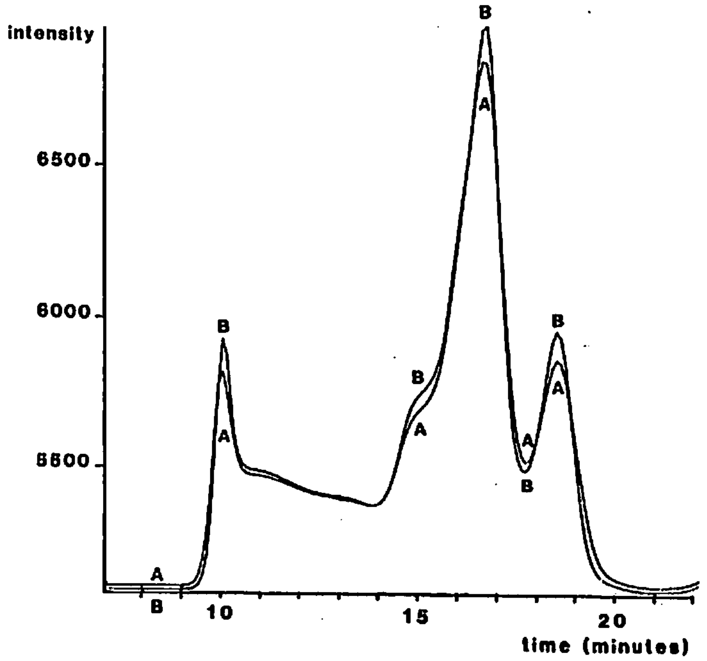

Figure 6. Discrimination of growing locations: average of ten chromatograms having positive and negative values of their first discriminant score

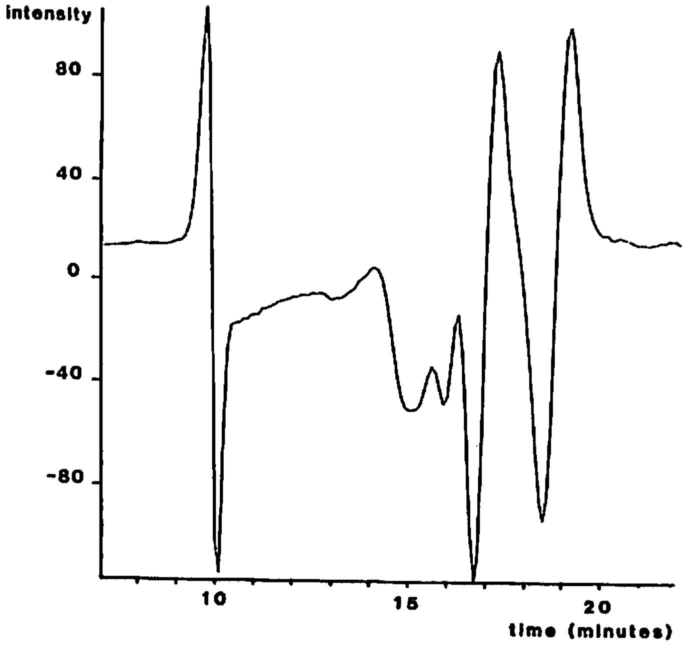

Figure 7. Differences of average chromatograms given in Figure 6

genotype contains more specific associations between sub-units. In this case the molecular weights of the proteins are more clearly separated and the chromatogram shows marked peaks and valleys. In contrast, other growing conditions giving wheats of poor baking quality may result in a reduced degree of protein aggregation and flat chromatograms.

Discrimination of the genotypes

The training and evaluation sets were the same as for the discrimination of the areas of cultivation; the first PCA therefore gave the same results.

Figure 8 shows the evolution of the rightly classified observations according to the number of introduced components in the first trial. The correct results steadily increased with the number of components. The numbers of components giving the best classifications of the verification set ranged from 14 to 20 according to the trial (Table 1). More than 86% of the sample of the training set, including 40 genotypes, were correctly identified in any trial. The samples of the verification set (15 genotypes) were also quite well classified, with more than 85% correct identifications. Random allocation of the verification set into 15 groups of three samples gave only about 4% correct classifications.

The first discriminant pattern (Figure 9) presented a large positive peak at 16 min preceded by a small negative local minimum at about 15 min. As previously, chromatograms of genotypes presenting positive or negative values of their first discriminant scores were averaged (Figure 10). The average values mainly differed in the size of the peaks at 15 and 16 min and were in accordance with the discriminant pattern. The peak at 16 min corresponds to α-, β- and γ-gliadins (molecular weight 30–45 kDa) whereas the peak at 15 min is representative of ω-gliadins (60–65 kDa). It seemed logical that these two types of proteins were antagonistic. The first discriminant pattern showed that their relative proportions were the main criterion

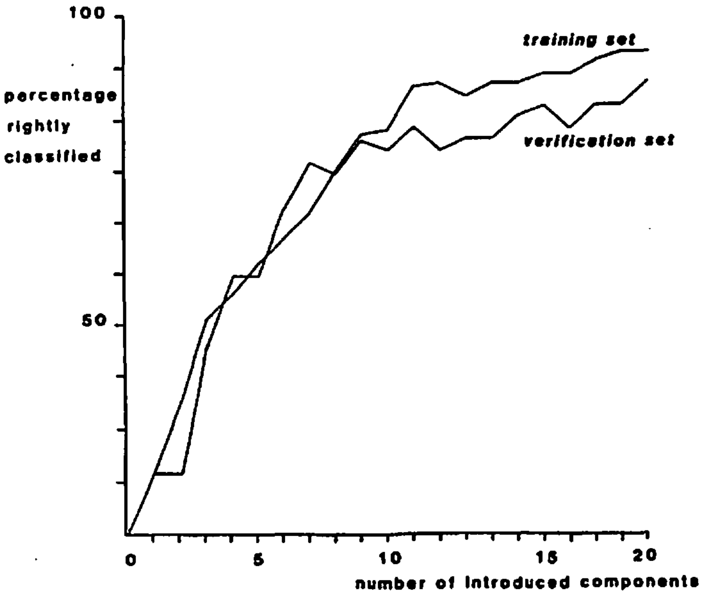

Figure 8. Discrimination of genotypes: influence of the number of introduced principal components on the number of correctly classified observations

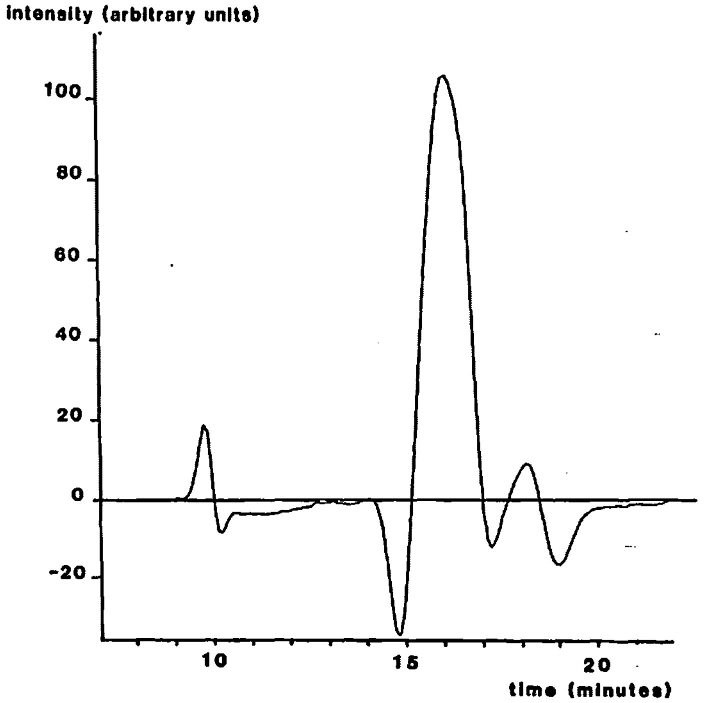

Figure 9. First discriminant pattern of genotypes

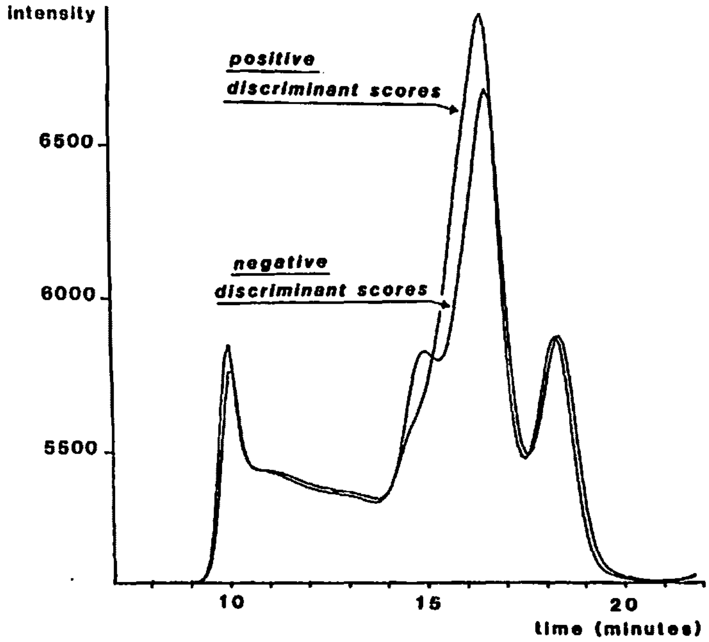

Figure 10.' Discrimination of genotypes: averages of ten chromatograms of cultivars having positive and negative values of their first discriminant score

for identifying genotypes. Rousset and Branlard[18] showed that the proportion of gliadins is related to the baking quality of wheat, but no work was done on genotype identification.

CONCLUSIONS

The procedure described is applicable to signals having a large number of data points. Tests have been done on IBM-compatible microcomputers with collections including up to 120 observations of the training set and 700 data points. In these conditions, SCDA needs about 1 h for completion, and the classification of an unknown sample is performed in a few seconds, depending on the number of groups. When several qualitative classifications are performed on the same set of data, the procedure needs only one time-consuming calculation (achievement of the first PCA) and is therefore very rapid in comparison with other methods. The proposed procedure can be applied when the variance–covariance matrix is singular. When all the components of the first PCA are introduced, SCDA and CDA give identical classifications. In the worst case, SCDA is therefore at least as efficient as basic DA. The Euclidean distance on discriminant scores is equivalent to the Malahanobis distance on the original data. Because the number of discriminant scores for each observation is generally small (in any case, less than than the number of introduced components and the number of qualitative groups), it becomes possible to perform relevant classifications of the observations according to qualitative criteria: the assessment of discriminant distances is very rapid. For example, discriminant scores can be used as variables for automatic clustering and creation of hierarchized databases.[19]

The procedure applied for classification of the genotypes of the verification set shows that the vector basis created from a given collection of chromatograms was relevant for identification of new genotypes which were not present in the training set. The discriminant patterns make it possible to identify the areas of the digitalized signals which are involved in the separation of the qualitative groups. They can take into account not only the intensity of the peaks but also their shapes. This is an improvement in comparison with the usual way of interpreting chromatograms, which consist only of measuring the surface under clearly separated peaks.

In contrast to high-performance liquid chromatography (RP-HPLC), SE-HPLC has rarely been used for identification of genotypes: SE-HPLC chromatograms present only a few large peaks. The present study shows that the efficiency of SE-HPLC seems comparable to that of RP-HPLC. SE-HPLC presents the advantage of being three times faster than RP-HPLC. Other studies are needed before recommending SE-HPLC for varietal identification.

REFERENCES

1. M. A. Sharaf, D. L. Illman and B. R. Kowalski, *Chemometrics*, pp. 228–242, Wiley, New York (1986).
2. L. Lebart and J.-P. Fénelon, *Statistique et Informatique Appliquées*, pp. 280–288, Dunod, Paris (1987).
3. H. L. Mark and D. Tunnell, *Anal. Chem.* 57, 1449 (1985).
4. D. Bertrand, P. Robert and W. Loisel, *J. Sci. Food Agric.* 36, 1120 (1985).
5. M. F. Devaux, D. Bertrand and G. Martin, *Cereal Chem.* 63, 151 (1986).
6. G. Downey, P. Robert, D. Bertrand and P. M. Kelly, *Appl. Spectrosc.* 44, 150 (1990).
7. J.-C. Autran and P. Abbal, *Electrophoresis*, 9, 205 (1988).
8. M.-C. Virioh, 'Methodologies statistiques de la discrimination: application aux électrophorégrammes des farines de blés', *Doctor Thesis*, Université des Sciences et Techniques du Languedoc, Montpellier (1988).

9. J.-M. Romeder, *Méthodes et Programmes d'Analyse Discriminante*, Dunod, Paris (1973).
10. W. F. McClure, A. Hamid, F. G. Giesbrecht and W. W. Weeks, *Appl. Spectrosc.* 38, 322 (1984).
11. M. F. Devaux, D. Bertrand, P. Robert and J.-L. Morat, *J. Chemometrics*, 1, 103 (1987).
12. M. F. Devaux, D. Bertrand, P. Robert and M. Qannari, *Appl. Spectrosc.* 42, 1015 (1988).
13. A. Campbell and G. E. Bradfield, *Can. J. Bot.* 67, 146 (1989).
14. D. H. O'Rourke, B. K. Suarez and J. D. Crouse, *Am. J. Phys. Anthropol.* 67, 241 (1985).
15. T. Foucart, *Analyse Factorielle sur Microordinateurs*, Masson, Paris (1982).
16. J. Lefebvre, *Introduction aux Analyses Statistiques Multidimensionnelles*, Masson, Paris (1983).
17. T. Dachkevitch and J.-C. Autran, *Cereal Chem.* 66, 448 (1989).
18. M. Rousset and G. Branlard, *Ann. Amélior. Plantes*, 30, 133 (1980).
19. J. Zupan, *Clustering of Large Data Sets*, Wiley, Chichester (1982).