ELSEVIER

Analytica Chimica Acta 549 (2005) 166–178

CHIMICA
ACTA

www.elsevier.com/locate/aca

Classification of PDO olive oils on the basis of their sterol composition by multivariate analysis

M. Rui Alves [a,b], Sara C. Cunha [a,c], Joana S. Amaral [a,d], J.A. Pereira [e], M. Beatriz Oliveira [a,*]

[a] *REQUIMTE, Serviço de Bromatologia, Faculdade de Farmácia, Universidade do Porto. Rua Aníbal Cunha, 164, 4099-030 Porto, Portugal*
[b] *ESTG, Instituto Politécnico de Viana do Castelo, Avenida do Atlântico, s/n, 4901-908 Viana do Castelo, Portugal*
[c] *ISEIT, Mirandela, Av. 25 de Abril, 5370 Mirandela, Quinta de Santa Apolónia, 5301-855 Bragança, Portugal*
[d] *ESTiG, Instituto Politécnico de Bragança, Ap. 1134, Quinta de Santa Apolónia, 5301-855 Bragança, Portugal*
[e] *CIMO, Escola Superior Agrária de Bragança, Ap. 1172, Quinta de Santa Apolónia, 5301-855 Bragança, Portugal*

Received 18 March 2005; received in revised form 25 May 2005; accepted 9 June 2005

Abstract

The sterol compositions (GLC/FID/capillary column) of monovarietal olive oils (51 samples) from the most important cultivars of north-eastern Portugal (Cvs. Cobrançosa, Madural and Verdeal Transmontana) and 27 commercial samples of olive oils with protected denomination of origin (PDO) from the same region and cultivars were evaluated.

β-sitosterol, Δ^5-avenasterol and campesterol were the most representative sterols. Cholesterol, stigmasterol, clerosterol and Δ^7-stigmastenol were also found in all samples. All studied samples respected EC Regulation N. 2568, and in all cases total sterols were remarkably higher than the minimum limit set by legislation, ranging from 2003 to 2682 mg/kg.

Results were analysed with the help of several statistical techniques, including reduction of dimensionality by principal component analysis with cross-validation of the number of components, followed by the use of canonical variate predictive biplots for model development and canonical variate interpolative biplots for approximate classification of monovarietal and PDO olive oils. These biplots proved to be a very interesting solution in the present case study, overcoming the problems of interpretation and classification that arise whenever different multivariate analyses are coupled together.

Keywords: Sterols; Olive oils; Multivariate analysis; Biplots

1. Introduction

Clinical studies have demonstrated that dietary intake of phytosterols as a part of the normal diet, or as a supplement, may help reducing blood cholesterol levels through the inhibition of its absorption from the small intestine [1,2]. Additionally, it has been suggested that sterols have anti-inflammatory, antibacterial, antifungal, antiulcerative, antitumural activities [3] and antioxidant activity [4]. As sterols are a major portion of the unsaponifiable components of the olive oils [5], and present a more or less characteristic profile and given benefits, they are widely used to check genuineness [6]. Besides, their determination is of major interest due to their health benefits.

The large amounts of olive oil included in the Mediterranean diet lead to the consumption of high amounts of β-sitosterol, Δ^5-avenasterol and campesterol. Besides these compounds, several other sterols existing in smaller amounts, such as cholesterol, brassicasterol, stigmasterol, clerosterol, sitostanol, Δ^7-stigmastenol and Δ^7-avenasterol can usually be ingested. The knowledge of the sterol composition is very important in the evaluation of the nutritional value as well as in the quality control of olive oils, since they can also be used to determine possible frauds. For example, it is now well established that the presence of large quantities of stigmas-

[*] Corresponding author. Tel.: +351 222078927; fax: +351 222003977.
E-mail address: beatoliv@ff.up.pt (M.B. Oliveira).

0003-2670/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2005.06.033

terol reveal an adulteration with lower priced soybean and/or cottonseed oils [7].

Olive oil is a traditional product from the Northeast of Portugal (named Trás-os-Montes) obeying the European Commission Regulation EC-N.2081/92 [8], and can therefore be labelled as a product with protected designation of origin (PDO). Four olive tree Cvs. are authorised for the production of olive oil, Cv. Cobrançosa, Cv. Madural, Cv. Verdeal Transmontana (which will be referred to as Cv. Verdeal only) and Cv. Cordovil, but the latter has no commercial expression. Being known as high quality products, olive oils from this region attain very high market prices, favouring unfair commercial practices. It is therefore important to use analytical techniques to ensure the assessment of identity and quality of olive oils, and to guarantee the proper product classification.

To evaluate product authenticity it is necessary to build up, test and validate models, against which the characteristics of any new (unseen) products can be compared [9]. Such a work requires appropriate multivariate statistical tools, among which principal component analysis and several types of discriminant analyses [10] occupy a very important position. Powerful statistical software packages are now available that can perform a series of very complicated calculations in a fast and comfortable way for the user, making it very easy to apply sophisticated algorithms to almost any kind of data, with no need for special mathematical background. Therefore, it is now more important to focus studies of statistical analyses on the respective conditions of application, expected type of results and how to carry out interpretations of the statistical outputs. Some good starting points exist for the study of the application of multivariate statistics in the field of chemometrics [11], and in what concerns the particular case of discrimination and classification, important works with reviews and developments are also available [12]. Many recent examples of these concerns in relation to olive oils can be found in the literature, in the search for the best chemical or physical parameters or methodologies to use in authentication and assessment of quality or possible adulterations, recurring to a wide number of statistical techniques [13–19], and the need to couple several statistical techniques in order to attain good discrimination and reliable models for classification is now evident [20].

It is also very important to guarantee that the results of model development are not restricted to investigation, and that can be used on a daily basis, as a part of the laboratory routine [21–23], mainly recurring to the advantages of predictive and interpolative biplots [24].

The work presented in this paper deals with the above mentioned topics, aiming to: (i) build up a model for the characterization of three monovarietal olive oils produced from Cvs. Cobrançosa, Madural and Verdeal, on the basis of their main sterol composition; (ii) use the developed model to classify several commercial PDO olive oils on the basis of their sterol content; (iii) discuss some statistical methodologies that can be used to solve the problem of model building, and simultaneously present and explain some useful statistical operations; (iv) apply predictive biplots for an easier model interpretation, and interpolative biplots to obtain models easier to use on a routine basis.

An effort was made when writing this paper in order to avoid using an excessive statistical language, and a special section, at the end of Section 3, is presented with a set of six statistical remarks, with some statistical details and algorithms. Along the text references are made to these statistical remarks.

2. Materials and methods

2.1. Samples

Monovarietal olive oils (in a total of $N = 51$ samples) were obtained in the laboratory by extraction from olive fruits from Cvs. Cobrançosa ($N_1 = 18$ samples), Madural ($N_2 = 15$ samples) and Verdeal ($N_3 = 18$ samples), all from the N.E. of Portugal, following the method described in Pereira et al. [25]: briefly, olives were collected from identified and carefully marked trees, handpicked and processed in a pilot plant, passing through a mill, a thermo beater and a pulp centrifuge, after which the oil was separated from the pulp by decantation, and kept in dark glass bottles, at 4 °C, in the absence of light.

Samples of commercial PDO olive oils from the same region were randomly purchased in the local market. A total of 27 samples were analysed.

2.2. Sterol composition

The sterol composition was determined by GLC/FID/capillary column following the method described in NP-EN-ISO-12228 (1999) [26]. The oil was previously dehydrated with anhydrous sodium sulphate and subsequently filtered through filter paper. A 250 mg of oil were accurately weighted, mixed with 1.0 mL of internal standard solution (betulin 1.0 mg/mL), and saponified with an ethanolic potassium hydroxide solution; the unsaponifiable fraction was isolated by solid phase extraction on an aluminium oxide column and the steroid fraction was obtained after TLC with *n*-hexane/diethyl ether 1:1 (v/v) as developing solvent and a methanol spray to visualize the band. The trimethylsilylethers were obtained by the addition of 1-methylimidazole and *N*-methyl-*N*-(trimethylsilyl)-hepta-fluorobutyramide (MSHFBA).

The sterol profile was analyzed with a Chrompack CP 9001 chromatograph (Chrompack, Middelburg, The Netherlands) equipped with a split–splitless injector, a FID, and a Chrompack CP-9050 autosampler. Separation was achieved on a fused silica capillary column DB-5MS (30 m × 0.25 mm i.d., 0.25 μm, J & W Scientific, Folsom, CA, USA). The temperature of the injector and the detector were set at 320 °C. The oven temperature was 250 °C and programmed to increase at a rate of 2 °C/min to 300 °C and then held for

12 min. The injected quantity was 1.5 μL, at a split ratio 1:50, using helium as carrier gas at an internal pressure of 100 kPa.

The total sterol content was determined considering all peaks of sterols eluted between cholesterol and Δ^7-avenasterol, and individual sterols were expressed as percentages of the total sterol content. Identification was achieved by comparing the relative retention times from samples with those obtained with standards. Standards used for identification were purchased from Sigma (St. Louis, USA) and included cholestanol, cholesterol, campesterol, stigmasterol, β-sitosterol, β-sitostanol and betulin. Clerosterol, Δ^5-avenasterol and Δ^7-stigmastenol were tentatively identified by comparison with references [7,27,28]. β-sitostanol and Δ^5-avenasterol eluted very close and were therefore quantified as Δ^5-avenasterol. Apparent β-sistosterol, which is an important quality indicator, was calculated as the sum of Δ^5-avenasterol, clerosterol and β-sitosterol.

2.3. Statistical analyses

2.3.1. Initial data matrices

The initial data for monovarietal olive oils consisted of a matrix **X** with $P = 8$ columns representing the following variables (sterols): cholesterol ($\mathbf{x}_1$), campesterol ($\mathbf{x}_2$), stigmasterol ($\mathbf{x}_3$), clerosterol ($\mathbf{x}_4$), β-sitosterol ($\mathbf{x}_5$), Δ^5-avenasterol ($\mathbf{x}_6$), apparent β-sistosterol ($\mathbf{x}_7$) and Δ^7-stigmastenol ($\mathbf{x}_8$). Notation $\mathbf{x}_p$ [$p = 1\ldots P$] is used to refer to any unspecified sterol. Matrix **X** had $N = 51$ rows representing the individual olive oils analysed. The oils belonged to $G = 3$ groups: Cv. Cobrançosa, Cv. Madural and Cv. Verdeal. In the present work all $\mathbf{x}_p$ variables were standardized to mean zero and unit variance, in order to attribute to each sterol the same relative importance. Another matrix was considered, $\mathbf{X}_{\mathrm{PDO}}$, with the same number of columns relative to the same sterols, but with rows relative to $N = 27$ PDO olive oils.

2.3.2. Pre-treatments and univariate analysis

To visualize sterol compositions, the mean and minimum and maximum values for each group were calculated over the standardized variables and plotted, obtaining standardized means and dispersion bars for each monovarietal olive oil. ANOVA and student's *t*-tests relative to a sterol $\mathbf{x}_p$ were calculated by conventional methods, the former leading to the calculation of an observed *F* value (F_{obs}), the latter to the calculation of an observed *t* value (t_{obs}). Observed values were compared to critical *F* and *t* values corresponding to the $\alpha = 0.01$ significance level, referred to $F_{\alpha = 0.01}$ and $t_{\alpha = 0.01}$ respectively.

2.3.3. Multivariate analyses

Principal components analysis (PCA) of initial standardized data was carried out using the NIPALS algorithm to enable the cross-validation of the number of components based on the so-called leave-one-out strategy, as related to the original Wold's method [9,29]. The components with no interest (those showing a decreased prediction ability) were discarded, being left with a reduced set of "*a*" components, referred to as $\mathbf{pc}_1$ to $\mathbf{pc}_a$. MANOVA and Hotelling T^2 tests were carried out on the reduced set of principal components to evaluate the significance of the differences between monovarietal olive oils. Canonical variates analysis (CVA) was also carried out based on the reduced matrix of principal components, obtaining canonical variates generally referred to as **cv**. The resulting combined PCA/CVA model was interpreted in terms of a predictive biplot, and classification of PDO olive oils was carried out by mathematical projection of individual oils, $\mathbf{X}_{\mathrm{PDO}}$, on the plane of the canonical variates, and also recurring to an interpolative CVA biplot.

General algorithms for multivariate analysis [10,30], procedures for PCA with cross-validation [30], for the construction of biplots [24] and ways to solve practical problems [21–23] were carried out according to published hand-books and papers. However, at the end of Section 3 of this paper, main steps and algorithms used to solve the present case study are described with some detail.

2.3.4. Software and algorithms

All analysis were produced in special Programs written by the authors in the Genstat language [31]. All graphs were created in the Statistica for Windows statistical package [32] after conversion of the Genstat ASCII outputs to Statistica files.

3. Results

The monovarietal oils reported in this work were obtained from Cvs. Cobrançosa, Madural and Verdeal, and each of these cultivars is considered a group. Therefore, the terms "monovarietal oils" and "groups" are used with the same meaning. All oils belonging to one cultivar are within-group observations, and differences between these oils are within-group differences. On the other hand, observed differences between oil types, i.e., differences between the mean sterol values of oils from each cultivar, are called the between-group differences.

3.1. Checking conformity with legislation

The sterol composition of 51 monovarietal olive oils from three cultivars (group means, standard deviations and coefficients of variation) and of 27 commercial PDO olive oils (minimum, median and maximum values) is shown in Table 1. It is seen that β-sitosterol, Δ^5-avenasterol and campesterol are the most important sterols, with cholesterol, stigmasterol, clerosterol and Δ^7-stigmastenol being present in all samples but in lower amounts. Regarding the authenticity indices established by the current legislation [33], all samples respect the established limits: cholesterol and campesterol percentages were below the established limits of 0.5% and 4.0%, respectively; the percentages of stigmasterol were lower than those of campesterol and the apparent

Table 1
Mean values ($\bar{x}_p$) and standard deviations (s_p) of the three monovarietal olive oils under study, as well as the minimum (x_{min}), median ($\bar{x}$) and maximum (x_{max}) values for the PDO

Sterols	Cobrançosa			Madural			Verdeal			Overall means[a]	PDO		
	x_{max}	x_{min}	$\bar{x} \pm s$	x_{max}	x_{min}	$\bar{x} \pm s$	x_{max}	x_{min}	$\bar{x} \pm s$		x_{max}	x_{min}	$\bar{x} \pm s$
Cholesterol	0.560	0.162	0.389 ± 0.138	0.519	0.163	0.284 ± 0.132	0.351	0.220	0.265 ± 0.047	0.314	0.578	0.293	0.433 ± 0.076
Campesterol	3.478	1.195	3.277 ± 0.168	2.655	2.487	2.501 ± 0.112	3.363	2.868	3.069 ± 0.193	2.975	3.923	2.245	3.114 ± 0.530
Stigmasterol	1.192	0.600	0.747 ± 0.211	1.493	1.015	1.369 ± 0.218	1.189	0.737	0.958 ± 0.211	1.005	1.101	0.670	0.798 ± 0.131
Clerosterol	0.957	0.838	0.926 ± 0.041	0.965	0.928	0.935 ± 0.070	0.911	0.816	0.844 ± 0.044	0.900	1.091	0.768	0.856 ± 0.075
β-Sitosterol	88.600	85.699	87.112 ± 0.957	86.959	82.812	86.161 ± 3.233	88.739	88.016	88.343 ± 0.348	87.267	85.366	83.094	84.207 ± 0.635
Δ^5-Avenasterol	8.675	6.063	7.483 ± 0.972	11.747	7.940	8.474 ± 3.135	6.827	5.765	6.374 ± 0.407	7.383	11.471	9.387	10.072 ± 0.596
β-Sitosterol (apparent)	97.224	95.114	95.519 ± 0.292	96.040	95.046	95.571 ± 0.367	95.787	95.436	95.559 ± 0.124	95.549	96.275	94.525	95.357 ± 0.594
Δ^7-Stigmastenol	0.206	0.000	0.069 ± 0.075	0.301	0.204	0.277 ± 0.069	0.212	0.000	0.146 ± 0.073	0.157	0.394	0.000	0.267 ± 0.128

All the values obtained are presented in percentages.

[a] Overall means of the three monovarietal cultivars.

β-sitosterol content was higher than the legal minimum of 93% in all olive oils analyzed. Besides, in all cases (data not shown), total sterols were remarkably higher than the minimum limit set by legislation (1000 mg/kg), ranging from 2003 to 2682 mg/kg. This is undoubtedly a good characteristic of olive oils due to the great benefits of these compounds for health, as referred before.

3.2. Univariate analysis

Fig. 1 shows means, extreme values and dispersion bars for each monovarietal oil after standardization of all sterol values to mean zero and unit variance, so that they can all be displayed in the same graph, enabling a helpful visualization of apparent differences and/or similarities between groups. Minimum and maximum initial values for each group are indicated in the graph, to help relating standardized to initial values. According to this figure, it seems that some differences between oil groups may exist, although definite conclusions cannot be drawn due to the existence of apparently high within-groups variations and to the absence of any means of relating individual observations from one sterol to the other.

Since differences in the composition of monovarietal oils seem evident, ANOVAs were carried out for all sterols (statistical Remark 2), and results are presented in Table 2. Each ANOVA carried out for each sterol evaluates the H_0 hypothesis "all oils have the same composition in what concerns this particular sterol", against the alternative H_1 "at least one oil is different from the others in what concerns this particular sterol". With the exception of apparent β-sistosterol, all F_{obs} values were higher than the critical $F_{\alpha=0.01}$ values, indicating that for all the other sterols H_0 is false, and significant differences between oils were found.

Following significant ANOVAs, student's t-tests for the difference between two means were carried out for all possible different pairs of groups, in order to evaluate the H_0 hypothesis "both olive oils have the same composition in what concerns this particular sterol", against the alternative H_1 "the oils have different composition in terms of this sterol". The tests were done for all different pairs and for all sterols. A t_{obs} value higher than the critical $t_{\alpha=0.01}$ indicates that the

Table 2
Condensed ANOVA results for all sterols under analysis

Sterols	SS_{total}	$SS_{between}$	SS_{within}	$S^2_{between}$	S^2_{within}	F_{obs}
Cholesterol	0.761	0.158	0.603	0.079	0.013	6.271
Campesterol	6.446	5.162	1.284	2.581	0.027	96.477
Stigmasterol	5.404	3.226	2.178	1.613	0.045	35.560
Clerosterol	0.216	0.086	0.130	0.043	0.003	15.796
β-Sitosterol	203.563	39.603	163.960	19.802	3.416	5.797
Δ^5-Avenasterol	192.871	36.363	156.509	18.181	3.261	5.576
Ap. β-sitosterol	3.624	0.025	3.599	0.013	0.075	0.168
Δ^7-Stigmastenol	0.612	0.359	0.253	0.180	0.005	34.022
ν	50	2	48			

SS: sums of squares; S^2: mean square; F_{obs}: observed F value; $F_{[\nu_1=2;\, \nu_1=48;\, \alpha=0.01]} \approx 5.08$; ν: degrees of freedom; α: significance level.

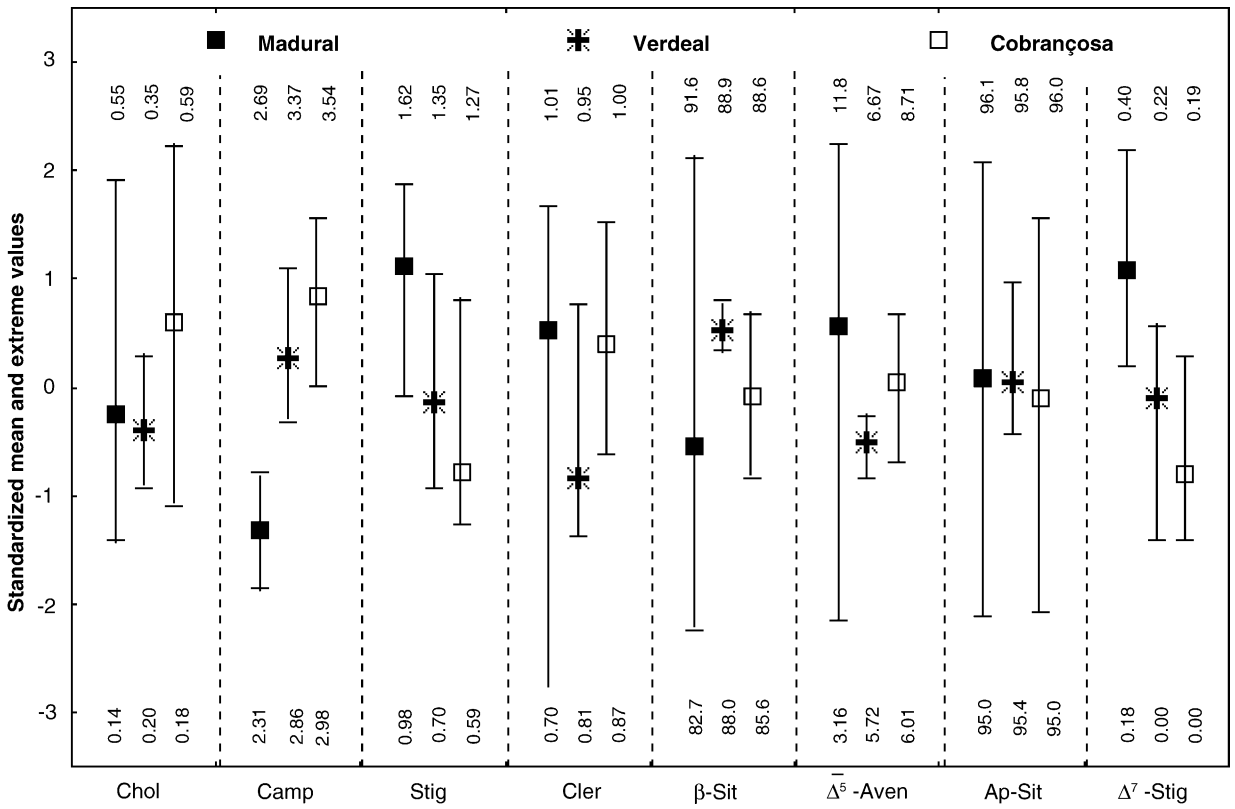

Fig. 1. Standardized sterol composition of the three monovarietal olive oils based on median and extreme values. Minimum and maximum values for each sterol are presented in the bottom and top of the graph, respectively.

alternative hypothesis may be true. Table 3 shows that Cv. Madural is very different from the others, mainly in Campesterol, Stigmasterol and Δ^7-Stigmastenol contents, as judged by the magnitude of t_{obs} values, and less in other sterols, while Cv. Cobrançosa and Cv. Verdeal are different in all but apparent β-sistosterol, although the differences seem less pronounced. These results are in agreement with Fig. 1.

3.3. The need for data simplification

Following conclusions from univariate analysis, it could seem that given the significant differences observed, a multivariate model describing the main characteristics of the monovarietal oils studied could be built up with no difficulty. Just like in a common ANOVA where one compares the between-group differences (b_i) to the within-group differences (w_i) as an $F_{\text{obs}} = b_i/w_i$ (or $F_{\text{obs}} = bw_i{}^{-1}b_i$) for a given sterol $\mathbf{x}_i$ (statistical Remark 2), in a multivariate situation one has to compare the differences between group means (monovarietal oils) taken over all sterols (enclosed in a matrix **B**), with the pooled differences between individual oils and respective group means, also taken over all sterols (enclosed in a matrix **W**). The desired multivariate comparison is achieved by multiplying **B** by the inverse of **W**, i.e., calculating $\mathbf{W}^{-1}\mathbf{B}$, and the analogy with the univariate case is evident. However, the necessary inversion of matrix **W** puts serious problems that must be taken into consideration mainly in two situations:

(1) If the sterols are highly correlated (collinear), if one or more variables are determined as combinations of other

Table 3
t_{obs} and $t_{\alpha=0.01}$ values to evaluate the significance of the observed differences between oils in relation to each sterol

Sterol	Cobrançosa/Madural $t_{[\nu=31;\,\alpha=0.01]} \approx 2.453$	Cobrançosa/Verdeal $t_{[\nu=34;\,\alpha=0.01]} \approx 2.441$	Madural/Verdeal $t_{[\nu=31;\,\alpha=0.01]} \approx 2.453$
Cholesterol	2.22	3.61[a]	0.57
Campesterol	15.27[a]	3.45[a]	10.06[a]
Stigmasterol	8.31[a]	3.00[a]	5.49[a]
Clerosterol	0.44	5.79[a]	4.50[a]
β-Sitosterol	1.19	5.13[a]	2.85[a]
Δ^5-Avenasterol	1.27	4.46[a]	2.82[a]
Ap. β-sitosterol	0.45	0.53	0.13
Δ^7-Stigmastenol	8.26[a]	3.13[a]	5.24[a]

[a] Two oils and one sterol for which differences were found to be significant.

variables (as it is the case with apparent β-sitosterol), or if there are groups with very low variance, **W** becomes ill-conditioned and cannot be inverted (statistical Remark 1, point b). As a consequence, applying some types of multivariate analysis based on the within-groups variations directly to the initial data (e.g., canonical variates analysis for model building, development of discriminant functions for classification, etc.) cannot be done.

(2) In other circumstances matrix inversion is possible, but the inversion of **W** necessary to obtain matrix $\mathbf{W}^{-1}$ always overwhelms the importance of the less informative variables or structures, and the resulting models will then lack good classification properties [12] (statistical Remark 1, point a).

As a consequence of the above-mentioned problems, for classification purposes, a data simplification must be carried out, for which two major routes are available:

(1) If calculation of $\mathbf{W}^{-1}$ is not possible, PCA can be used as the data simplification method. Using PCA, the initial variables ($\mathbf{x}_i$) are not deleted, but are substituted by an (usually) equal number of principal components (**pc**s), ordered by decreasing order of importance. As a component is a set of correlated variables, collinearity problems are overcome by this methodology. It is expected that important information coming from all variables are modelled in the first **pc**s, while errors and spurious information is accommodated in the last **pc**s. Cross-validation of the number of components, is the recommended method to indicate how many components must be retained to guarantee good model predictability [29]. In this way, data is simplified to the most important data structures.

(2) In situations where the calculation of $\mathbf{W}^{-1}$ is possible, some types of discriminant analysis (DA), e.g., standard, forward-selection or backward-elimination stepwise DA, can be a good choice for the selection of the most discriminant variables, removing from the data all the variables that do not contribute to discrimination or proper classification [10,30]. This was used, e.g., to solve a problem similar to the present case study, relative to the classification of vegetable oils based on their fatty acid contents [23]. Thus, in this way, data is simplified to the most discriminant variables.

3.4. Principal components analysis

In the present case study, due to collinearity problems (high correlations between different sterols) increased by the inclusion of a variable (apparent β-sitosterol) calculated as a combination of other sterols, **W** was ill-conditioned, its determinant was null, and $\mathbf{W}^{-1}\mathbf{B}$ could not be calculated. As a consequence, methods based on $\mathbf{W}^{-1}\mathbf{B}$ could not be applied directly, and a PCA was applied to the original data (statistical Remark 3). There are many possible ways to formulate hypothesis testing in PCA [10], but the following simple approach may be useful: since principal components are uncorrelated, each **pc** represents an important aspect of the available information, also called a data structure. Therefore, in simple terms, PCA can be viewed as a test of the hypothesis H_0 "there are no special structures present in the data, all variations observed being random", against the alternative H_1 "there is at least one important data structure". The total number of important structures is best evaluated by cross-validation, which can be formulated in terms of the following question: "how many components must be retained in order to model new, unseen data with the greatest possible accuracy?" This means that one is looking for the set of information that guarantees the best classification ability of the model, and that using less or more components than the cross-validated ones will impoverish the capacity of the model to deal with new samples, which is an important point in classification problems. As a matter of fact, the cross-validation is a balance between those components that can be used to carry out a parsimonious description of the data (using e.g., Mardia et al.s concepts [10]) and that simultaneously display good prediction ability (using the Wold's concepts [9,29]).

Three components (three structures) were found to be important by cross-validation, the main results being shown in Table 4 and Fig. 2a and b. These figures present classical plots of $\mathbf{pc}_1$ versus $\mathbf{pc}_2$ (Fig. 2a) and $\mathbf{pc}_1$ versus $\mathbf{pc}_3$ (Fig. 2b), with $\mathbf{pc}_1$ always represented as a horizontal line and $\mathbf{pc}_2$ and $\mathbf{pc}_3$ represented as vertical lines. In the overall, these two figures show around 80% of the total available information. Olive oils from Cvs. Cobrançosa and Verdeal overlap in relation to $\mathbf{pc}_1$ and $\mathbf{pc}_3$, seeming different in relation to $\mathbf{pc}_2$, although with some degree of overlapping exists. These two oils look very different from Cv. Madural oils. A wider variation is observed for the latter, visible over all three components. These results are in general agreement with the univariate approaches. Observing Table 4, which presents the relationships between sterols and the first three **pc**s, either in the form of PCA eigenvectors or correlations, it is evident that the majority of the information of sterols (around 80%) is condensed in the three first principal components. It is also seen that many sterols are related to more than one **pc**, which makes it difficult to attribute a meaning to each data structure. This problem is sometimes overcome by factor rotation [34], but such a technique was not used in this work since it corresponds to a return to a few of the original variables loosing the benefits of the data modulation effect carried out by PCA.

Nevertheless, through PCA data could be compressed from eight sterols to three components, keeping the majority of the initial information organized in a set of three **pc**s.

3.5. MANOVA and Hotelling T^2

After data compression to principal components, all important information is conveyed to other analysis in an organized way. Now, the between- to within-groups varia-

Table 4
General results from principal components analysis: matrix $\mathbf{L}_a$ of important eigenvectors and $\mathbf{\Lambda}_a$, of the respective eigen values (absolute and relative), and percentage of information explained by each component

Sterols	First three PCA eignevectors and correlations						Information in $\mathbf{pc}_1+\mathbf{pc}_2+\mathbf{pc}_3$
	Eigenvectors			Correlations			
	$\mathbf{l}_1$	$\mathbf{l}_2$	$\mathbf{l}_3$	$\lambda_1^{1/2}\mathbf{l}_1$	$\lambda_2^{1/2}\mathbf{l}_2$	$\lambda_3^{1/2}\mathbf{l}_3$	
Cholesterol	−0.002	−0.472	0.372	−0.002	−0.702	0.463	0.707
Campesterol	0.453	−0.401	0.064	0.736	−0.598	0.081	0.906
Stigmasterol	−0.423	0.328	0.261	−0.685	0.491	0.317	0.811
Clerosterol	−0.307	−0.182	−0.381	−0.503	−0.285	−0.452	0.539
β-Sitosterol	0.450	0.416	0.081	0.728	0.620	0.097	0.924
Δ^5-Avenasterol	−0.445	−0.391	−0.168	−0.721	−0.583	−0.203	0.901
Ap. β-Sitosterol	0.049	0.221	−0.708	0.077	0.323	−0.871	0.869
Δ^7-Stigmastenol	−0.341	0.316	0.326	−0.551	0.482	0.398	0.694
Eigen values (λ)	2.627	2.237	1.496	2.627	2.237	1.496	Sum = 6.351
Eigen values (%)	32.8	27.9	18.7	32.8	27.9	18.7	79.38

tions can be calculated based on these principal components, and all analyses based on a matrix of between- to within-groups distances (any type of $\mathbf{W}^{-1}\mathbf{B}$) are now possible, the difference being the fact that instead of the initial sterols, one is now dealing with **pc**s (data structures). Two such cases, MANOVA and Hotelling T^2 tests [10,29], which are the multivariate equivalents of ANOVA and student's *t*-tests, are presented in this section, with the respective hypothesis testing formulated in terms of the data structures defined by PCA.

MANOVA uses functions of the λ values (eigen values of $\mathbf{W}^{-1}\mathbf{B}$) to test the H_0 hypothesis "there are no significant differences in the composition of the three monovarietal oils", against the alternative H_1 "there is at least one monovarietal oil different from the others in at least one data structure". The common λ functions (see Table 5 for details) are:

(i) The Rao's *R*, based on the Wilks' lambda statistic, which is based on the product of unexplained variances. The lower the amount of unexplained information, the lower the Wilks' lambda is, and the higher and more significant the Rao's *R* tends to be.

(ii) The Pillai's *V*, based on the Pillai–Bartlett trace, which is based on the sum of explained variances. The higher the amount of explained information, the higher the trace is, and the higher and more significant the Pillai's *V* tends to be.

(iii) The η^2_{mult} that is based on the canonical correlations, which are based on the explained variances. The higher the amount of explained information, the higher the canonical correlations are, reflecting a higher degree of association of the data units.

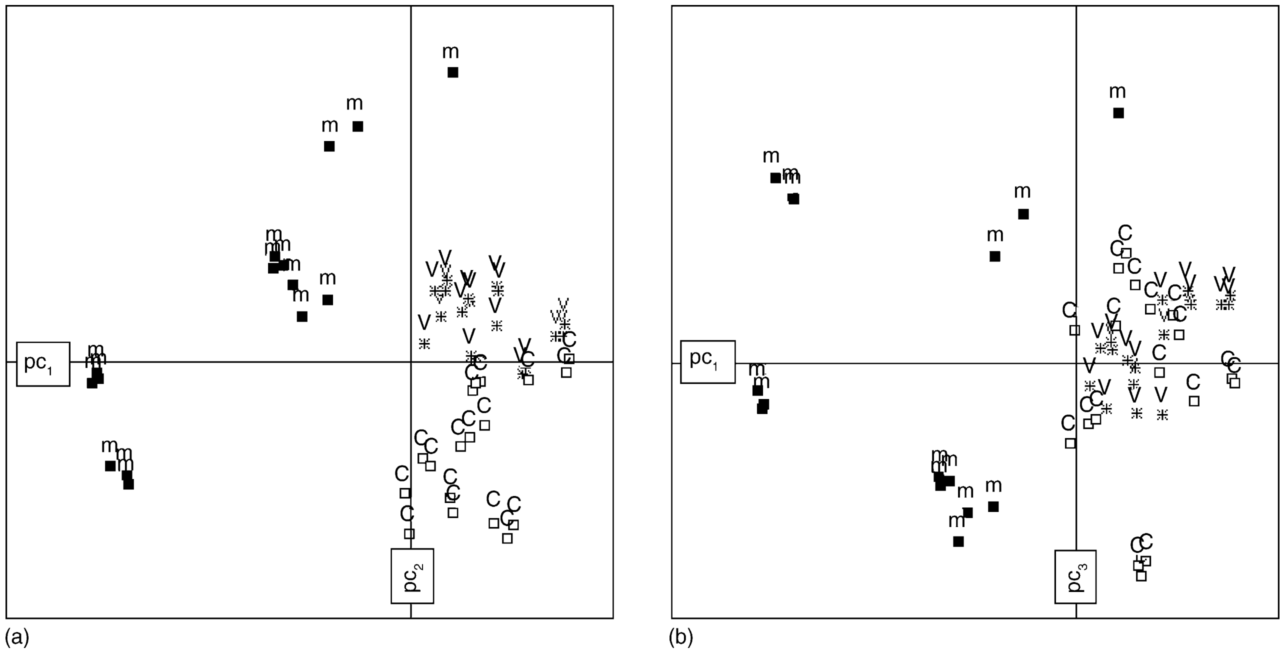

Fig. 2. Plot of principal components. Symbols are: c = Cobrançosa; v = Verdeal; m = Madural (a) Plot of $\mathbf{pc}_1$ vs. $\mathbf{pc}_2$. (b) Plot of $\mathbf{pc}_1$ vs. $\mathbf{pc}_3$.

Table 5
Summary of multivariate measures and test statistics

Canonical variates dimension $\mathbf{v}_q$	Eigen value λ_q	% variance	χ^2_{obs}	$\chi^2_{\alpha=0.001}$	ν	$(1-\lambda_q)^{-1}$	$\lambda_q(1+\lambda_q)^{-1}$	Canonical correlations η_q
1	303.1	0.6564	118.2	22.46	6	0.110	0.890	0.943
2	105.3	0.2282	14.51	13.82	2	0.734	0.266	0.515
Wilks'lambda = $\Pi\ (1-\lambda_q)^{-1}$						0.0809	–	–
Pillai–Bartlett trace = $\Sigma[\lambda_q(1+\lambda_q)^{-1}]$						–	1.156	–
η^2_{mult}						–	–	0.578

Multivariate test	Value	Approximate F_{obs}	Degrees of freedom		α
			ν_1	ν_2	
Rao's R	5.7×10^{-7}	38.59	6	92	<0.001
Pillai's V	2.969	21.44	6	94	<0.001

(iv) The Bartlett test, which is a direct transformation of the eigen values and follows a χ^2 distribution, indicating if the magnitude of the λ value is significantly different from zero.

In Table 5, it is seen that both eigen values are significantly different from zero (significant χ^2_{obs} values) that the unexplained variances tend to zero while the explained variances tend to two, making Rao's R and Pillai's V significant (as seen by the significant F_{obs} values), and that the η^2_{mult} shows that 57.8% of the total variation may be attributed to group membership. But with all these multivariate measures we can only conclude that "at least one oil is different from the others in relation to at least one data structure".

In face of this significant MANOVA, one may wonder which are the different groups. The Hotelling T^2 test is equivalent to a student's t-test in univariate analysis and through complicated algorithms evaluates the H_0 hypothesis "both monovarietal oils are equal", against the alternative H_1 "the two oils are different in at least one data structure". This means that the test must be ran for all different pairs of oils. Table 6 shows the values of the T^2 statistics and the corresponding F_{obs} values, as well as the critical $F_{\alpha=0.001}$ values. It is seen that all oils are different from each other. One should therefore conclude that "the compositions of the three olive oils differ in at least one data structure".

The conclusions that can be drawn from MANOVA and Hotelling T^2 tests, although showing that differences exist, may be faced as poor results, since no justification for the observed differences is provided, and further analysis are therefore necessary.

3.6. PCA/CVA biplots

CVA can be stated in terms of the H_0 hypothesis "groups cannot be separated in the multivariate space", against the alternative H_1 "differences between groups are significant in at least one space dimension" [10]. It can therefore be seen as a test for dimensionality, with the advantage that differences along space dimensions can be easily plotted and "visually" evaluated, and explained in terms of the underlying parameters.

Being a statistical method based on a matrix $\mathbf{W}^{-1}\mathbf{B}$ of between- to within-groups distances, CVA suffers from the mathematical problems discussed above. However, being carried out after data compression, using solely principal components $\mathbf{pc}_1$, $\mathbf{pc}_2$ and $\mathbf{pc}_3$ as the starting point, which as discussed before are conveying the important information (with the best prediction ability) in an organized way, doing a CVA presents no problems (statistical Remark 4), but interpretation of the main outputs is increasingly difficult. It is important to emphasize that interpretation of a CVA is never straightforward, since CVA loadings are not restricted to lie within definite boundaries, as happens e.g. with PCA. This problem is increased in the present case because the previous PCA forces the canonical variates to be interpreted in terms of **pc**s, and if **pc**s are reverted to the original sterols, uncertainty towards interpretation always increases.

The biplots discussed below are a good way to overcome these problems. It is reminded that the way biplot axes are produced [24] and methods for overcoming practical problems have already been discussed [21–23]. The important point here is that PCA and CVA can be coupled together for

Table 6
Hotelling T^2 tests for the difference between oils from two cultivars

	Cobrançosa	Madural	Verdeal
Cobrançosa	0.00	$\delta = 50.10$; $T^2 = 409.9$	$\delta = 11.27$; $T^2 = 101.5$
Madural	$F_{\text{obs}} = 127.83$; $\nu_1 = 3$; $\nu_2 = 29$	0.00	$\delta = 21.26$; $T^2 = 174.0$
Verdeal	$F_{\text{obs}} = 31.83$; $\nu_1 = 3$; $\nu_2 = 32$	$F_{\text{obs}} = 54.25$; $\nu_1 = 3$; $\nu_2 = 29$	0.00

Upper triangle: distances between group means (δ) and corresponding T^2 statistics. Lower triangle: observed F values (F_{obs}) corresponding to the T^2 statistics and respective degrees of freedom (ν_1 and ν_2).

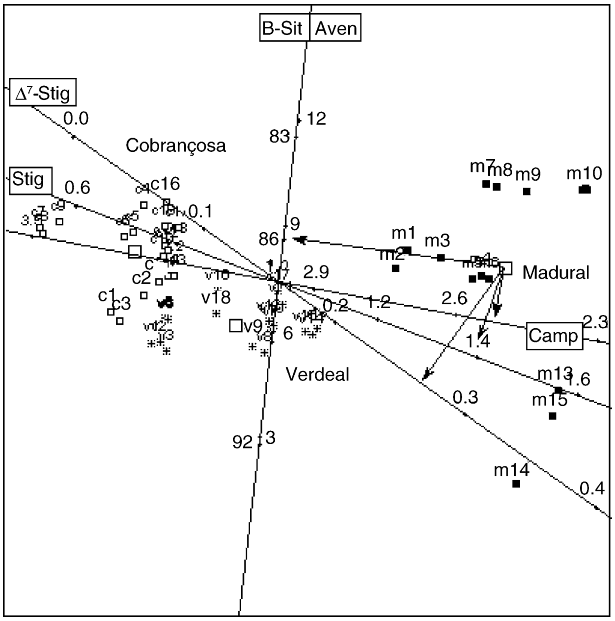

Fig. 3. Predictive biplot constructed on the plane of $\mathbf{cv}_1$ vs. $\mathbf{cv}_2$, after a combined PCA/CVA. Symbols are: c and closed triangles = Cv. Cobrançosa; v and closed circles = Cv. Verdeal; m and closed squares = Cv. Madural. Open squares are group means. Points over axes are the scale markers. Scale values for the markers are shown in italic. Biplot axes are: Aven = Δ^5-avenasterol; Δ^7-Stig = Δ^7-stigmastenol; Stig = stigmasterol; Camp = campesterol; B-sit = β-sitosterol. The axes for Δ^5-avenasterol and β-sitosterol overlapped, the scale for the former is shown on the right, for the latter on the left. Arrows indicate the projection of Madural group average over biplot axes to determine its initial values.

the construction of predictive biplots, which enables the interpretation of final results in terms of the original sterols, while interpolative biplots are a good way to classify new observations comparing favourably with discriminant functions, the appropriate mathematics being presented in statistical Remarks 5 and 6.

3.7. Predictive PCA/CVA biplot

Fig. 3 shows the plot of canonical variates $\mathbf{cv}_1$ versus $\mathbf{cv}_2$. Reminding that in this case there are only two canonical variates (equal to the number of groups minus 1), 100% of the information conveyed by the principal components, which correspond to roughly 80% of the original information, is shown in this figure. It is seen that Cv. Madural oils (on the right side of the graph) are different from the other two oils, and that with this solution Cvs. Cobrançosa and Verdeal olive oils (the former towards the left, the latter in the centre of the graph) are not overlapping, which is a synonym of a good discrimination.

The results from CVA are conventionally shown in graphs very similar to the one presented in Fig. 2 for PCA, with horizontal and vertical axes representing $\mathbf{cv}_1$ and $\mathbf{cv}_2$, respectively. However, as CVA was based on the PCA results, each canonical variate would have to be interpreted in terms of **pc**s, which although possible, could be quite cumbersome. To overcome this problem a predictive biplot was constructed (statistical Remark 5). Axes representing the initial sterols, called biplot axes, equipped with appropriate scales for measurement, were drawn in the figure. Drawing orthogonal projections from any point of interest to an axis representing a sterol determines the initial (percentage) level of the sterol in that point. For example, drawing orthogonal lines from the group mean of Cv. Madural to the sterol axes, as indicated in Fig. 3, the following approximate composition is read from the graph: ~8.7% Δ^5-avenasterol, ~2.52% campesterol, ~0.27% Δ^7-stigmastenol and ~1.39% stigmasterol. Comparing these values with the ones presented in Table 1, it becomes obvious that the approximations are really very good. It is also seen that the model describing the differences between monovarietal olive oils can be based directly in terms of sterols and respective initial values, which is a very simple and straightforward process.

Therefore, it is seen that campesterol, Δ^7-stigmastenol and stigmasterol are the main responsible for the observed differences between oils from different cultivars, in agreement with previous conclusions. It is also seen that these sterols are correlated (running in similar directions). Δ^5-avenasterol and β-sitosterol, assuming a vertical position in the graph, are mainly necessary to describe differences within-groups. The axes for these two sterols are overlapped (although running in opposite directions), reflecting a very high negative correlation, the relationship between correlation and collinearity being here quite evident.

3.8. Interpolative PCA/CVA biplot

Once the model is defined, if one is interested in carrying out an approximate classification of new samples, then discriminant functions are usually employed. The problem with these functions is that they do not apply when new samples are expected to be blends, i.e., mixtures of the previously defined groups, as it happens in this case study with the set of 27 PDO olive oils. An approximate classification can be done with interpolative biplots (statistical Remark 6), as shown in Fig. 4. In relation to the predictive biplot, the interpolative biplot axes will assume different directions, and the magnitude of the scale intervals will also be changed, because interpolation is doing exactly the inverse of prediction. To observe the importance and uses of this type of biplots, the individual monovarietal oils were removed (in order to produce a clearer graph) leaving only the markers representing the group means. The scale values for stigmasterol are not shown in the graph for clarity reasons, but can be deduced from Fig. 3. Also, a set of three theoretical standard mixtures of the three cultivars were calculated and used for validation purposes. These were s1 (30% c16 + 70% m7), s2 (70% c16, 30% m7) and s3 (63% c16, 27% m7 and 10% v8). The samples used for calculation, i.e., samples c1, m2 and v3 are identified in Fig. 5 and in Table 7.

The PDO olive oils were projected mathematically in the plot, and two interpolations are done “by hand” for demon-

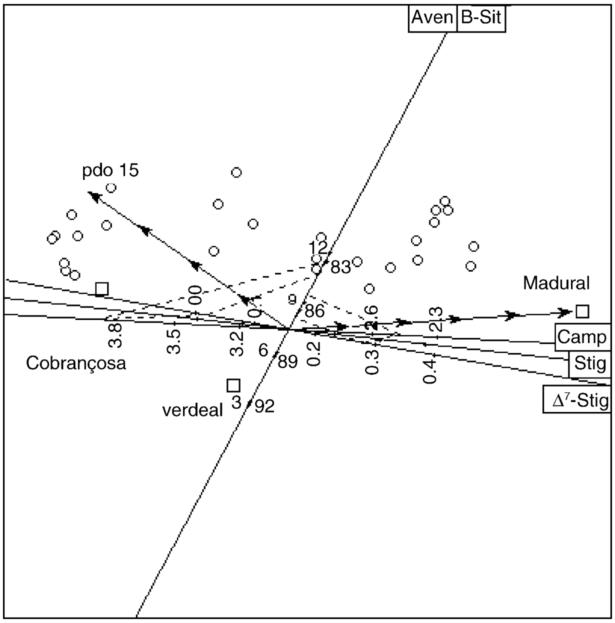

Fig. 4. Interpolative biplot constructed on the plane of $\mathbf{cv}_1$ vs. $\mathbf{cv}_2$, after a combined PCA/CVA. In relation to Fig. 3, individual monovarietal oils were removed and the PDO olive oils were added as open circles, all other features remaining the same. Broken lines and arrows indicate the way vector sums are done for interpolations.

stration purposes. One example is PDO olive oil N. 15 with the following main sterol composition: 3.86% campesterol, 0.67% stigmasterol, 83.09% β-sitosterol, 10.89% Δ^5-avenasterol and 0.21% Δ^7-stigmastenol. Now, marking these values over the respective axes, one forms the vertices of a geometric figure, and these vertices are linked to form the figure (broken lines in Fig. 4). A vector is drawn from the origin (where all axes meet) to the centre of the figure, and is then multiplied by the number of sterols (which in this case is four, because the stigmasterol level falls in the centre of the axes not contributing to interpolation). The apex of the last, resultant vector is the desired sample interpolation, which is seen to be very accurate. Another example applied to the mean of monovarietal Madural oils (whose values are presented in Table 1) is also shown in the figure. Another very accurate interpolation is obtained, and in this case the composition in terms of five sterols was used. An interesting point to recall is that collinearity, which is a problem if no data simplification is applied, is crucial for an accurate interpolation of the existing data, or any new observations.

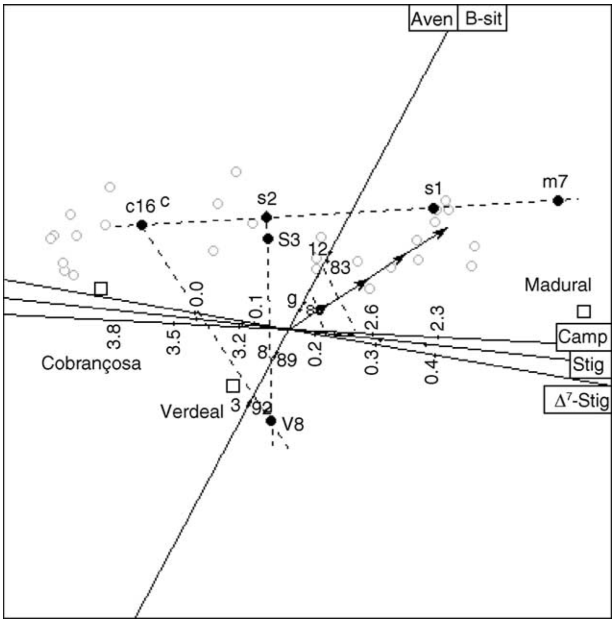

Fig. 5. Interpolative biplot constructed on the plane of $\mathbf{cv}_1$ vs. $\mathbf{cv}_2$, after a combined PCA/CVA, for the evaluation of the of the method's interpolative ability. The PDO olive oils are shown as grey open circles, monovarietal olive oils c16, m7 and v8 are shown as closed circles, as well as theoretical samples s1, s2 and s3. The compositions of these samples are displayed in Table 7. Dotted lines uniting samples are discussed in the text.

Theoretical samples s1, s2 and s3, calculated on the basis of samples c16, m7 and v8, were projected in the $\mathbf{cv}_1$ versus $\mathbf{cv}_2$ plane, in the same way as it was done for PDO olive oils, and the results are shown in Fig. 5 (and not in Fig. 4 for clarity purposes). The positions of samples s1 and s2 show that a mixture of two oils lies in a straight line in between the two original oils (c16 and m7), in a position that reflects the percentage incorporation of individual oils. Adding a third oil in the mixture (as in s3) will displace the sample point towards the third component of the mixture, also reflecting the incorporation percentage. Interpolation of sample s1 by hand, on the basis of four sterols, demonstrates that the interpolative biplots can be faced as very precise.

These interpolative biplots are more flexible than the classification functions, since the position of new samples in relation to group means can be visualized. It is seen that several PDO oils are mainly produced with Cv. Cobrançosa, while others are showing increasing incorporations of Cv. Madural. This conclusion comes from the observation that PDO olive oils are clustered around the Cv. Cobrançosa group mean, and some are displaced towards the Cv. Madural, being interpolated to a position somewhere along a line linking the Cv. Cobrançosa and Cv. Madural group means. Following conclusions from the last paragraph, as no PDO oils approach the Cv. Verdeal group mean, we conclude that this cultivar is used in minor amounts.

3.9. Statistical remarks

To our knowledge there is no software available in the market for the automatic construction of predictive and interpolative biplots (which need interactive graphical facilities), and in many situations, existing software does not provide a satisfactory, automatic answer when several multivariate analysis are coupled together. Consequently, the following statistical remarks are provided as a starting point for those interested in writing their own algorithms. Remarks 1 and 2 are presented to help clarifying some points discussed in this paper. These two remarks, together with Remarks 3 and 4, show the main steps for PCA and CVA, introducing the

Table 7
Sterol percentage composition of samples used for evaluation of the model's interpolative ability

Sterol	Initial sample values of unblended olive oils			30% c16 + 70% m7	70% c16 + 30% m7	63% c16 + 27% m7 + 10% v8
	c16	m7	v8	s1	s2	s3
Cholesterol	0.520	0.550	0.230	0.541	0.529	0.499
Campesterol	3.040	2.630	3.040	2.753	2.917	2.929
Stigmasterol	0.600	1.470	1.090	1.209	0.861	0.884
Clerosterol	1.000	0.920	0.810	0.944	0.976	0.959
β-Sitosterol	86.470	83.010	88.720	84.048	85.432	85.761
Δ^5-Avenasterol	8.250	11.150	5.950	10.280	9.120	8.803
Ap β-Sitosterol	95.730	95.080	95.470	95.275	95.535	95.529
Δ^7-Stigmastenol	0.110	0.270	0.160	0.222	0.158	0.158

Cobrançosa c16, Madural m7 and Verdeal v8 are observed values, and samples s1, s2 and s3 were calculated as mixtures of the observed samples.

notation necessary to follow Remarks 5 and 6 on predictive and interpolative biplots. The complete algorithms, written in the Genstat language, can be supplied to interested readers on request.

Remark 1. The majority of multivariate analyses that can be used to compare groups of observations use matrix $\mathbf{W}^{-1}\mathbf{B}$ which compares the differences between group means (in matrix **B**) with the pooled differences between individual oils and respective group means (enclosed in matrix **W**). For this reason **B** may be called the hypothesis matrix, while **W** is called the error matrix. To calculate $\mathbf{W}^{-1}$, one first calculates the spectral decomposition $\mathbf{W}=\mathbf{V}\boldsymbol{\Delta}\mathbf{V}^{\mathrm{t}}$, where **V** is the matrix of eigenvectors, $\mathbf{v}_q$ $[q=1\ldots Q]$, and $\boldsymbol{\Delta}$ is the diagonal matrix of eigen values ordered by decreasing magnitude $(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_Q)$, with Q equating to the number of variables or the number of groups minus 1, whatever is minimum. **W** can then be expressed as $\mathbf{W}=\mathbf{V}[\lambda_1,\lambda_2,\ldots,\lambda_P]\mathbf{V}^{\mathrm{t}}$, emphasizing the weighting role of the eigen values. Then, for matrix inversion, it suffices to calculate $\mathbf{W}^{-1}=\mathbf{V}\boldsymbol{\Delta}^{-1}\mathbf{V}^{\mathrm{t}}$. If matrix $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$is used instead of $\mathbf{W}^{-1}\mathbf{B}$ due to the advantage of being symmetric, $\mathbf{W}^{-1/2}$ is calculated as $\mathbf{V}\boldsymbol{\Delta}^{-1/2}\mathbf{V}^{\mathrm{t}}$. Two very important facts arise in relation to these matrix inversions, which are emphasized in the following points *a* and *b*:

Point a. It happens that in $\boldsymbol{\Delta}^{-1}$ we are considering the inverse of the eigen values, so that the order of their magnitudes is $\lambda_1^{-1} \leq \lambda_2^{-1} \leq \cdots \leq \lambda_Q^{-1}$. If $\mathbf{W}^{-1}$ is expressed as $\mathbf{W}^{-1}=\mathbf{V}[(1/\lambda_1),(1/\lambda_2),\ldots,(1/\lambda_P)]\mathbf{V}'$, it becomes immediately evident that the smaller an eigen value is, the higher the influence it gets in $\mathbf{W}^{-1}\mathbf{B}$. As a consequence, if the last eigenvectors and values are not deleted, the models incorporate irrelevant information, becoming unstable and lacking good classification properties, as discussed in detail elsewhere [12].

Point b. In situations where collinearity is a problem, **W** becomes ill-conditioned, with a null determinant, the last eigen values are null and some of the products $\mathbf{v}_q(\boldsymbol{\lambda}_q{}^{-1})\mathbf{v}_q{}^{\mathrm{t}}$ cannot be calculated since they correspond to a division by zero. In these situations **W** cannot be inverted, and statistics based on $\mathbf{W}^{-1}\mathbf{B}$ simply cannot be applied.

Remark 2. The elements of the main diagonal of matrix $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$, which equal the number of variables, if weighted by the respective degrees of freedom are the values $F_{\mathrm{obs}} = b_i/w_i = w_i^{-1} \times b_i$ used in ANOVA applied to any variable $\mathbf{x}_i$. This fact also highlights the nature of matrix $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$ (or $\mathbf{W}^{-1}\mathbf{B}$).

Remark 3. PCA uses the total variation in matrix **X**, calculated as $\mathbf{T}=\mathbf{X}^{\mathrm{t}}\mathbf{X}$, and through its spectral decomposition obtains $\mathbf{T}=\mathbf{L}\boldsymbol{\Lambda}\mathbf{L}$, where **L** is a matrix whose columns are the eigenvectors $\mathbf{l}_r$ and $\boldsymbol{\Lambda}$ is a diagonal matrix with eigen values λ_r $[r=1\ldots R]$, with $R=\min(\mathrm{P},\mathrm{N}-1)$, so that in many practical situations R = P. Keeping only the first "a" important dimensions as found by cross-validation, i.e., reducing **L** to $\mathbf{L}_a$, then, $\mathbf{X}\mathbf{L}_a=\mathbf{Y}_a$, with $\mathbf{Y}_a$ representing the matrix whose columns are the most important principal components $(\mathbf{pc}_1,\mathbf{pc}_2,\ldots,\mathbf{pc}_a)$.

Remark 4. CVA based on principal components was carried out starting with matrix $\mathbf{Y}_a$, calculating the between- and within-groups variations as matrices **B** and **W**, followed by calculation of the symmetric matrix $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$ and its spectral decomposition as $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}=\mathbf{U}\boldsymbol{\Delta}\mathbf{U}^{\mathrm{t}}$, and obtaining normalized latent vectors as $\mathbf{V}=(\mathrm{N}-\mathrm{G})^{1/2}\mathbf{W}^{-1}\mathbf{U}$; Then, CVA yields (i) canonical variates $\overline{\mathbf{cv}}_q$ $[q=1\ldots Q]$ as the columns of $\overline{\mathbf{Z}} = \overline{\mathbf{Y}}_a\mathbf{V}$, (ii) projects the individual monovarietal oils in the canonical dimensions, yielding variates $\mathbf{cv}_q$ as the columns of $\mathbf{Z}=\mathbf{Y}_a\mathbf{V}$, and (iii) projects PDO olive oils in the canonical dimensions as $\mathbf{cv}_{\mathrm{PDO}}$ as the columns of $\mathbf{Z}_{\mathrm{PDO}}=\mathbf{X}_{\mathrm{PDO}}\mathbf{L}_a\mathbf{V}$.

Remark 5. The way biplot axes are produced [24] and methods for overcoming practical problems have already been discussed [21–23]. The important point here is the way to couple both analysis and still be relating results to original variables, which can be briefly described as follows.

Each predictive biplot axis is projected in the plot of combined PCA/CVA ρ dimensions as

$$[(\boldsymbol{\mu}_p - 1\bar{x}_p)s_p^{-1}]\,[\mathbf{e}_p^{\mathrm{t}}(\mathbf{L}_a)(\mathbf{V}_\rho^{-1})^{\mathrm{t}}]$$

$$[\mathbf{e}_p^{\mathrm{t}}(\mathbf{L}_a)(\mathbf{V}_\rho^{-1})^{\mathrm{t}}(\mathbf{V}_\rho^{-1})(\mathbf{L}_a)^{\mathrm{t}}\mathbf{e}_p]^{-1}$$

An example of a vector $\boldsymbol{\mu}$ with four markers for Δ^5-avenasterol is $\boldsymbol{\mu} = [3,6,9,12]$, as it can be easily checked in Figs. 3 and 4. Therefore, the unit vector is $\mathbf{1} = [1,1,1,1]$. Vector $\mathbf{e}_p$ indicates the position of the variable in the original matrix: as Δ^5-avenasterol was the sixth out of eight variables in the original data matrix, vector is $\mathbf{e} = [0,0,0,0,0,1,0,0]$, i.e., with zeros everywhere except a 1 in the sixth position. $[(\boldsymbol{\mu}_p - 1\bar{x}_p)s_p^{-1}]$ is the standardization of the scale values by the respective variable's average and standard deviation, respectively x_p and s_p. $[\mathbf{e}_p^{\mathrm{t}}(\mathbf{L}_a)(\mathbf{V}_\rho^{-1})^{\mathrm{t}}]$ is the projection of scale values through PCA to CVA planes, respectively through matrices $\mathbf{L}_a$ and $\mathbf{V}_\rho$. The last part of the equation, $[\mathbf{e}_p^{\mathrm{t}}(\mathbf{L}_a)(\mathbf{V}_r^{-1})^{\mathrm{t}}(\mathbf{V}_\rho^{-1})(\mathbf{L}_a)^{\mathrm{t}}\mathbf{e}_p]^{-1}$, is an adjustment factor necessary for the correct back-projection from any sample point to a variable's axis. Projection of $\boldsymbol{\mu}$ originates points in the biplot graphs (named scale markers), which when joined by a straight line originate the variable's axis.

Remark 6. To construct an interpolative biplot a simpler equation is necessary, since it suffices to delete the right hand side part the equation for prediction seen above, and to substitute $\mathbf{L}_a(\mathbf{V}_\rho^{-1})^{\mathrm{t}}$ by $\mathbf{L}_a\mathbf{V}_\rho$ (and it is seen that prediction is just the multiplication of scale values by the transpose of the inverse of the interpolation matrix). As consequence, the interpolative biplot axes will assume different directions, and the magnitude of the scale intervals will also be changed. The final equation for interpolation is

$$[(\boldsymbol{\mu}_p - 1\bar{x}_p)s_p^{-1}]\ [\mathbf{e}_p^{\mathrm{t}}\mathbf{L}_a\mathbf{V}_\rho]$$

4. Conclusions

In general terms it can be concluded that the aims of the present work were fully achieved, since: (i) it was possible to evaluate the authenticity of PDO olive oils on the basis of a model developed for the characterization of monovarietal olive oils according to their sterol composition; (ii) the model developed can be used in current laboratorial situations with no need for special statistical background, hence its usefulness beyond research; (iii) the approaches used in this work are likely to be applied successfully on the modelling, classification and/or authenticity evaluation of many commercial materials that are blends, and for which it is unrealistic to aim the definition of well defined classes.

In what concerns the statistical questions raised in this work, it was seen that: (i) a model for the characterization of monovarietal olive oils could not be built on the basis of a canonical variates analysis applied directly to the initial data, due to the existence of strong correlations between the classification parameters, and also due to the inclusion of a variable that was calculated as a linear combination of two other variables; (ii) principal components analysis proved to be very useful in carrying out simplification of the initial data to its main structures, enabling the posterior application of multivariate discriminant techniques, like MANOVA, Hotelling T^2 tests and canonical variates analysis; (iii) the application of a PCA followed by a CVA enabled to take into consideration only the sterols that were important for discrimination of monovarietal olive oils, as well as the exclusion of variables that, like apparent β-sitosterol, although important in a legislation point of view, create problems in the statistical analysis of results; (iv) the construction of predictive biplots enabled relating directly the final analyses' outputs to initial variables and respective units of measurement, overcoming problems that arise when different multivariate analysis are coupled together; (v) the interpolative biplots demonstrated a great accuracy, enabling to carry out classifications of PDO olive oils using a model defined for monovarietal oils, thus proving to be more flexible than conventional discriminant functions.

Acknowledgements

M.R. Alves and J.S. Amaral are grateful to the "Programa para o Desenvolvimento Educativo para Portugal (PRODEP III)", and S.C. Cunha is grateful to Subprograma Ciência e Tecnologia do 3° Quadro Comunitário de Apoio (BD 8822/2002) for financial support. The authors are also thankful to the Acção 8.1, of PO Agro N. 482.

References

[1] A. Kamal-Eldin, K. Määttä, J. Toivo, A.-M. Lampi, V. Piironen, Lipids 33 (1998) 1073.
[2] S.L. Abidi, J. Chromatogr. A 935 (2001) 173.
[3] M. Stuchlík, S. Žák, Biomed. Papers 146 (2002) 3.
[4] E. Williamson, PhD Thesis, University of Reading, UK, 1988.
[5] A. Kiritsakis, W.W. Christie, J. Harwood, R. Aparicio (Eds.), Manual del aceite de oliva, AMV ediciones, Madrid, Spain, 2002, p. 135.
[6] A. Ranalli, L. Pollastri, S. Contento, G.D. Loreto, E. Lannucci, L. Lucera, F. Russi, J. Sci. Food Agric. 82 (2002) 854.
[7] W. Kamm, F. Dionisi, C. Hischenhuber, K. Engel, Food Rev. Int. 71 (2001) 249.
[8] Commision Regulation (EC) No. 2081/92 of July 1992 on the protection of traditional food products of the European Community members. Official J. Eur. Commun. L-208 24/07/1992 (1992).
[9] H. Martens, T. Naes (Eds.), Multivariate Calibration, John Wiley and Sons, 1991, p. 73.
[10] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, London, 1979.
[11] CAC, Chemometrics in Analytical Chemistry, Cabral, J.S. (Presiding Chairman), Proceedings of the Ninth International Conference, Instituto Superior Técnico, Lisbon, 2004.
[12] T. Naes, U. Indhal, J. Chemomet. 12 (1998) 205.
[13] E.C.L. Díez, G. Bianchi, R. Goodacre, J. Agric. Food Chem. 51 (2003) 6145.
[14] H.S. Tapp, M. Defernez, E.K. Kemsley, J. Agric. Food Chem. 51 (2003) 6110.
[15] J.E. Spangenberg, N. Ogrinc, J. Agric. Food Chem. 49 (2001) 1534.
[16] D.L. Gonzalez, R. Aparicio, J. Agric. Food Chem. 51 (2003) 3515.
[17] G. Bianchi, L. Giansante, A. Shaw, D.B. Kell, Eur. J. Lipid. Sci. Technol. 103 (2001) 141.

[18] T.G. Diáz, I.D. Merás, C.A. Correa, B. oldán, M.I.R. áceres, J. Agric. Food Chem. 51 (2003) 6934.

[19] D. Ollivier, L. Artaud, C. Pinatel, J.P. Durbec, M. Guérère, J. Agric. Food Chem. 51 (2003) 5723.

[20] R. Bucci, A.D. Magri, A.L. Magri, D. Marini, F. Marini, J. Agric. Food Chem. 50 (2002) 413.

[21] M.B. Oliveira, M.R. Alves, M.A. Ferreira, J. Chemomet. 15 (2001) 71.

[22] M.R. Alves, M.B. Oliveira, J. Chemomet. 17 (2003) 1.

[23] M.R. Alves, M.B. Oliveira, J. Chemomet. 18 (2004) 1.

[24] J.C. Gower, D.J. Hand, Biplots, Chapman & Hall, London, 1996, p. 86.

[25] J.A. Pereira, S. Casal, A. Bento, M.B.P.P. Oliveira, J. Agric. Food Chem. 50 (2002) 6335.

[26] NP-EN-ISO-12228. Animal and vegetal fats and oils – Determination of individual and total sterol contents. Gas cromatographic method, 1999.

[27] J.S. Amaral, S. Casal, J.A. Pereira, R.M. Seabra, B.P.P. Oliveira, J. Agric. Food Chem. 51 (2003) 7698.

[28] D. Firestone, R.J. Reina, P.P. Ashurst, M.J. Dennis (Eds.), Food Authentication, Chapman & Hall, London, 1996, p. 198.

[29] S. Wold, Technometrics 20 (1978) 397.

[30] J.H. Bray, S.E. Maxwell, Multivariate analysis of variance. A Sage University Paper, Series: Quantitative Applications in the Social Sciences, 54, 1985, p. 27.

[31] Lawes Agricultural Trust. Genstat, Release 5.3.1 [Computer program], 1993.

[32] StatSoft Inc. Statistica for Windows, 5.1 Release [Computer program], Tulsa, 1998.

[33] Commision Regulation (EC) No 2568/91, Official J. Eur. Commun. L-248 05/09/1991, 1991.

[34] G. Dunteman, Principal Components Analysis, A Sage University Paper, Series: Quantitative Applications in the Social Sciences, 69, Sage Publications, Newbury Park, 1989, p. 63.