Geographical discrimination of red garlic (Allium sativum L.) using fast and non-invasive Attenuated Total Reflectance-Fourier Transformed Infrared (ATR-FTIR) spectroscopy combined with chemometrics

Alessandra Biancolillo, F. Marini, A. d'Archivio

Geographical discrimination of red garlic (Allium sativum L.) using fast and non-invasive Attenuated Total Reflectance-Fourier Transformed Infrared (ATR-FTIR) spectroscopy combined with chemometrics

Alessandra Biancolillo,[a,b], Federico Marini[a], Angelo Antonio D'Archivio[c]*

[a]Dipartimento di Chimica, Università degli Studi di Roma "La Sapienza" Piazzale Aldo Moro, 00185, Roma, Italy

[b] ITAP, Irstea, Montpellier SupAgro, University of Montpellier, Montpellier, France

[c]Dipartimento di Scienze Fisiche e Chimiche, Università degli Studi dell'Aquila, Via Vetoio, 67010 Coppito, L'Aquila, Italy

*Corresponding author:

dr. Alessandra Biancolillo

Dept. of Chemistry

University of Rome La Sapienza

P.le Aldo Moro 5

I-00185 Rome

Italy

Tel +39 06 49913680

Fax +39 06 4969 3292

e-mail: alessandra.biancolillo@uniroma1.it

Declarations of interest: none

Geographical discrimination of red garlic (Allium sativum L.) using fast and non-invasive Attenuated Total Reflectance-Fourier Transformed Infrared (ATR-FTIR) spectroscopy combined with chemometrics

Declarations of interest: None

Abstract

Four varieties of red garlic (*Allium sativum* L.) cultivated in different Italian territories, Sulmona (Abruzzo), Proceno and Castelliri (Lazio), and Nubia (Sicily), were analysed by Attenuated Total Reflectance-Fourier Transformed Infrared (ATR-FTIR) spectroscopy. ATR-FTIR spectra of bulbils and bulbil tunics were separately acquired and processed by Partial Least Squares Discriminant Analysis (PLS-DA) with the aim of classifying the garlic samples on the basis of their geographical origin. Finally, two multi-block strategies (based on Sequential and Orthogonalized Partial Least Squares and Sequential and Orthogonalized Covariance Selection, coupled with Fisher's Linear Discriminant Analysis) have been applied in order to test whether a joint analysis of data could lead to higher prediction rates. Eventually, the best results were achieved by the multi-block approach based on SO-PLS, which allows obtaining a total classification rate of 95% (corresponding to one misclassified sample over 20) in external validation.

Keywords: *Garlic; ATR-FTIR; geographical classification; PLS-DA; Multi-block; Multi-block; SO-PLS; SO-CovSel.*

1. Introduction

Garlic (*Allium sativum* L.) has been worldwide employed as food condiment and herbal medicine for millennia. Apart from the common culinary use of fresh leaves or cloves, commercial products obtained by various processing methods, including oil maceration, dehydration and lyophilisation, are today marketed for therapeutic purposes (Ramirez et al., 2017). Historical references dating back to 4000 years mention diffuse use of garlic in ancient civilizations, from religious and superstition rituals to prevention and cure of infections and diseases (Corzo-Martínez et al., 2007). Bio-activity of garlic, including anti-inflammatory, antimicrobial, cardioprotective, anticancer and antidiabetic action, has been demonstrated in the last decades by epidemiological and clinical studies (Corzo-Martínez et al., 2007; Martins et al., 2016; Yun et al., 2014; Shukla & Kalra, 2007).

Unique pungent aroma and most of the medical properties attributed to garlic by traditional and modern medicine are related to distinctive organosulfur compounds.

Great attention is paid to the relation between the cultivar and geographical origin of garlic and its metabolomic profile, with specific reference to the aroma precursors and other bio-active constituents (Lu et al., 2011; Beato et al., 2011; Khar et al., 2011; Montaño et al., 2011). Moreover, traditional garlic varieties cultivated in given territories are appreciated by an increasing number of consumers because of their peculiar taste, aroma and functional properties compared to commercial products. In Europe, quality and geographical identity of some traditional garlic varieties of Italy (*Aglio di Voghiera* and *Aglio Bianco Polesano*), France (*Ail violet de Cadours*, *Ail fumé d'Arleux*, *Ail blanche de Lomagne* and *Ail de la Drôme*) and Spain (*Ajo Morado de Las Pedroñeras*) have been officially recognised in recent years through the attribution of PDO (Protected Designation of Origin) or PGI (Protected Geographical Indication) mark (European Commission, Agriculture and Rural Development, 2019). In particular, in the last years, different consortia in Italy have been constituted to valorize and preserve the traditional varieties of specific territories. Beside the protection provided by institutions (through laws and regulations) a wide effort has been put in developing analytical methodologies aimed at authenticating and tracing food specialties awarded of quality marks, for example (Lastra-Mejías e al., 2020; D'Archivio, et al. 2019a; Biancolillo et al., 2018a; Rocha et al., 2019; Giannetti et al., 2019; Mora et al., 2020; Firmani et al. 2019). In this context, analytical/chemometric approaches for the classification of garlic according to the cultivar and/or the geographical origin are essential tools to unveil commercial frauds arising from the intentional substitution of varieties cultivated in specific territories by commercial products.

Various analytical techniques, such as ^{1}H high resolution magic angle spinning-nuclear magnetic resonance spectroscopy (Ritota et al., 2012), infrared spectroscopy (Lu et al., 2011), high performance liquid chromatography (Montaño et al., 2011), high resolution mass spectrometry (Hrbek et al., 2018) and electronic nose (Trirongjitmoah et al., 2015), were applied to characterise garlic for traceability purposes. In these investigations, the organo-sulphur compounds and other

58 metabolome components, such as amino acids, fatty acids, organic acids and sugars were
59 recognised as promising traceability indicators to assess the garlic provenance or variety. In
60 addition, garlic cultivated in different countries (Smith, 2005; Vasi et al., 2016) or close regions
61 (D'Archivio et al., 2019b) were well discriminated using the trace multi-element profile determined
62 by atomic spectroscopy. However, most of the above analytical methods are relatively complex,
63 expensive, time-consuming and require specialized skills. Moreover, a preliminary sample
64 treatment is often necessary, which, apart from further increasing complexity and cost of the
65 characterisation method, may also alter the metabolomic profile of garlic. Attenuated Total
66 Reflectance-Fourier Transformed Infrared (ATR-FTIR) spectroscopy, by contrast, is a relatively
67 simple, fast, cheap and non-invasive technique applicable to both liquids and solids without any
68 complex sample pre-treatment. About garlic, ATR-FTIR spectroscopy was previously used to
69 quantify the total phenol content and antioxidant activity with the aim of differentiating the samples
70 grown in different US states (Lu et al., 2011), but the ATR-FTIR spectra, rather than directly on the
71 garlic samples, were acquired from methanolic extracts.
72 In the light of these considerations, the aim of the present work is to test the potentiality of ATR-
73 FTIR for geographical traceability purposes. In particular, red garlic varieties cultivated in four
74 distinct areas of Italy, namely Sulmona (Abruzzo), Castelliri and Proceno (Lazio), and Nubia
75 (Sicily), were analyzed and the observed spectra were handled by chemometrics. These red garlic
76 ecotypes have been chosen because of their valuable characteristics; additionally, those cultivated in
77 Sulmona and Nubia were also included by Slow Food Foundation for Biodiversity (Slow Food
78 Foundation, 2019) in the list of local plant varieties to safeguard. At first, ATR-FTIR spectra of
79 bulbils and tunics were separately analysed by Partial Least Squares-Discriminant Analysis (PLS-
80 DA) (Sjöström et al., 1986); this approach is widely and satisfactorily applied for authentication of
81 agro-food, in particular handling Infrared (IR) data (Biancolillo & Marini, 2018b). Finally, the two
82 data blocks were jointly analyzed by multi-block classifiers, in order to test whether data fusion
83 strategies would provide more accurate models, allowing a deeper comprehension of the system.

Consequently, IR spectra were analyzed by Sequential and Orthogonalized Partial Least Square (SO-PLS) (Næs et al., 2011) or Sequential and Orthogonalized Covariance Selection (SO-CovSel) (Biancolillo et al., 2019a) coupled with Fisher's Linear Discriminant Analysis (LDA).

2. Materials and methods

2.1. Garlic samples

Bulbs of red garlic varieties cultivated in 2017 in four Italian sites, Sulmona (Abruzzo), Castelliri and Proceno (Lazio), and Nubia (Sicily), were kindly donated by producers working in the respective territories that assured the geographical origin of the samples. In particular, 81 garlic cloves were analyzed with the tunics; of these, 20 were from Castelliri, 20 from Proceno, 20 from Nubia and 21 from Sulmona. Eventually, spectra were collected on 82 skinned samples (21 were from Castelliri, 20 from Proceno, 19 from Nubia and 22 from Sulmona). Only 69 cloves were analyzed on both compartments (i.e., either with and without tunic). Of these, 19 were from Castelliri, 20 from Proceno, 17 from Nubia and 13 from Sulmona. Samples were acquired in July-September 2017, stored under typical domestic conditions (bulbils were located into a box which allowed their perspiration, not exposed under direct light and in a cool room) and analyzed before December to avoid variations in the composition due to aging or sprouting.

2.2 ATR FT-IR measurements

The infrared spectra of garlic cloves and clove tunics were separately recorded on a PerkinElmer Spectrum Two™ (PerkinElmer, Waltham MA, USA) FT-IR spectrometer consisting in a deuterated triglycine sulfate (DTGS) detector and a PerkinElmer Universal Attenuated Total Reflectance (uATR) accessory equipped with a single bounce diamond crystal. Each spectrum was registered from 4000 cm^{-1} to 400 cm^{-1} with 1 cm^{-1} instrumental resolution and ten scans were averaged per spectral replicate. The background was collected with the crystal exposed to the air. Before each measurement, the ATR crystal was cleaned with methanol and air-dried. ATR FT-IR spectra were

collected on intact freshly peeled cloves by contacting a flat part of the clove with the ATR crystal. A consistent force was applied using the pressure monitoring system integrated with the instrument to maximize the spectrum intensity but avoiding crushing the clove. The ATR FT-IR spectra of clove tunics were recorded separately following the same procedure.

----------Insert Figure 1 approx. Here----------

The observed ATR-FTIR spectra (shown in Figure 1) display the typical vibration patterns of the plant constituents (proteins, fats and sugars) (Schulz & Baranska, 2007; Movasaghi et al., 2008) and reflect the composition of garlic clove and tunics. The garlic clove is mainly composed by water (65%), followed by carbohydrates (28%, mainly fructans), sulphur compounds (1-4%), proteins (2%), fibres (1.5%) and free amino acids (1-1.5%) (Rahman, 2003), while polysaccharides (cellulose, hemicellulose and pectin) and lignin are the main constituents of the bulbil skin (Kallel et al., 2015; Reddy & Rhim, 2014). The broad band centred at about 3290 cm^{-1} can be assigned to the N-H stretching of proteins and O-H stretching of carbohydrates and water, while the two sharp signals at 2920 and 2850 cm^{-1} are associated to symmetric and antisymmetric C-H stretching vibrations, respectively. The spectral region between 1200 and 900 cm^{-1}, although showing different intensity and fine structure in the spectra of cloves and skins, takes origin from coupled C–C, C–O stretching and C–O–H, C–O–C deformation modes of oligo- and polysaccharides. The distinctive band at 1025 cm^{-1} in particular can be assigned to the vibrational frequency of CH_2OH groups of carbohydrates. In the same spectral region, the S=O stretching of sulfoxides may contribute to the signal at about 1090 cm^{-1}(Nikolić et al., 2011), well visible in cloves but not in skin samples. The band at about 1160 cm^{-1} arises from the glycosidic linkage (C-O-C) vibrations. The weak bands in the region 880-900 cm^{-1} can be attributed to the C-O-C skeletal modes of carbohydrates and polysaccharides, the signal at 894 cm^{-1} in particular being diagnostic of ??-(1 → 4)-glycosidic bonds. The band at 1735-1740 cm^{-1} observed in both clove and skin samples are typical of the C=O stretching vibration of polysaccharides and cellulose. The signal at 1640 cm^{-1} observed in the spectrum of cloves can be assigned to the bending vibration of water and to the

stretching of carbonyl of proteins (amide I), while the signals at 1552 cm^{-1} and 1252 cm^{-1} can be attributed to the amide II and amide III bands (associated with coupled C–N stretching and N–H bending vibrations of the peptide group). Additionally, the stretching vibrations of aliphatic and aromatic double bonds fall in the region 1640-1500 cm^{-1} as well. The signal at 1225-1230 cm^{-1} of the skin spectra can be attributed to the stretching vibrations of C-O bonds in lignin (Stark et al., 2016). The band observed at about 1600 cm^{-1} in skin samples, partially overlapped to the amide I band of proteins in the clove spectra, can be assigned to the asymmetric stretching of carboxylate groups of amino acids, proteins or polysaccharides. The signals due to the O-H stretching of adsorbed water and asymmetric stretching of lignin aryl rings also fall in this spectral region, while the band at about 1510 cm^{-1} can be assigned to the symmetric stretch of the aromatic groups of lignin (Stark et al., 2016). The two sharp signals at 1460 and 1470 cm^{-1} observed in clove samples and sometimes superimposed, and the bands at 1427 cm^{-1} and 1328 cm^{-1} are due to O-H deformation vibration, and to various vibrational modes of CH_2 groups of lipids, polysaccharides and proteins. The two weak signals at about 720 and 727 cm^{-1} observed in the clove spectra but not in the skin samples can be attributed to C-S stretching vibrations of di-alkyl sulphides and disulphides (Minzhen et al., 2015) and to C-H deformation vibration.

2.3 Chemometric Analysis

The ATR-FTIR signals collected as described in the previous section have been analyzed by means of chemometric tools, in order to classify samples according to their geographical origin. To achieve this goal, three classification methods have been employed: PLS-DA, in order to handle the two data blocks individually, and SO-PLS-LDA or SO-CovSel-LDA, to achieve a simultaneous analysis of both sets of signals.

2.3.1 Partial Least Square-Discriminant Analysis (PLS-DA)

Discriminant classification methods discern samples on the basis of their mutual differences. Applying these approaches, the multi-dimensional samples-space is entirely divided into class-regions, and each object will be assigned to one specific category. One of the first discriminant classifiers proposed is the Linear Discriminant Analysis by Fisher. This approach, despite it performs well and it is still widely used, presents a considerable limitation: it can be used only when the data matrix is invertible. This condition is rarely met, in particular working with instrumental data, where the number of variables is likely higher than the number of samples.

Among the different methods developed in order to enable the application of discriminant analysis on ill-conditioned data matrices, Partial Least Square-Discriminant Analysis (PLS-DA) (Sjöström et al., 1986; Ståhle & Wold, 1987) is probably one of the most widely applied. One of the main reasons for its diffusion is that it is suitable to handle highly correlated variables (e.g., spectroscopic data), which makes it applicable on non-invertible data blocks. PLS-DA exploits the PLS algorithm (Wold et al., 1983) to solve a classification problem as if it were a regression one (Barker, & Rayens, 2003); mathematically, this corresponds to estimate Eq.1:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} \qquad (1)$$

Where $\boldsymbol{X}$ is the data matrix of measures collected on samples, $\boldsymbol{b}$ and $\boldsymbol{e}$ are the regression coefficients and the residuals, respectively, and $\boldsymbol{Y}$ is the so-called *Dummy Matrix*, a binary matrix (of dimensions $N \times G$, where N is the number of the analyzed samples and G is the number of categories present into the system) encoding the class-belonging. Once Eq.1 is solved (i.e., once the calibration model is built), whenever new measures ($\boldsymbol{X}_{new}$) are collected on unknown samples, it is possible to predict their class-belonging (solving $\widehat{\boldsymbol{Y}}_{new}=\boldsymbol{X}_{new}\boldsymbol{b}$). Nevertheless, $\widehat{\boldsymbol{Y}}_{new}$, i.e., the matrix collecting the responses predicted on new samples, is not categorical, but made of continuous values, so that it is necessary to define a criterion for class-attribution. Accordingly, classification may be accomplished, e.g., by assigning the sample to the category corresponding to the highest value of the predicted response, by application of LDA either on scores or on predicted responses or by a bayesian approach based on Gaussian mixture modeling (Perez et al., 2009).

187

188 **2.3.2 Sequential and Orthogonalized Partial Least Square-Linear Discriminant Analysis (SO-**
189 **PLS-LDA)**

190 Sequential and Orthogonalized Partial Least Square-Linear Discriminant Analysis (SO-PLS-LDA)
191 (Biancolillo et al., 2015) is a multi-block discriminant classifier whose algorithm has been
192 developed combining a multi-block regression method, SO-PLS (Næs et al., 2011), and Fisher's
193 Linear Discriminant Analysis, where SO-PLS is used to reduce the dimensionality of the data
194 blocks (solving problems related to ill-conditioned matrices) prior the application of the
195 discriminant approach.

196 In order to create a SO-PLS-LDA model, first of all it is necessary to build the SO-PLS one; given
197 two predictor blocks, $\boldsymbol{X}$ and $\boldsymbol{Z}$, and a dummy $\boldsymbol{Y}$, this can be done applying the procedure described
198 in (Næs et al., 2011). Briefly, $\boldsymbol{Y}$ is fitted to $\boldsymbol{X}$ by PLS (obtaining the $\boldsymbol{X}$-scores $\boldsymbol{T_X}$), $\boldsymbol{Z}$ is
199 orthogonalized with respect to $\boldsymbol{T_X}$ and then the resulting matrix ($\boldsymbol{Z_{Orth}}$) is used to predict (by PLS)
200 the residuals from the previous regression. Finally, the predictive model is calculated summing up
201 the outcomes of the two PLS-models (the reader is addressed to (Næs et al. 2011) for more details).
202 Once the SO-PLS model is built, classification is achieved by applying LDA on the predicted $\boldsymbol{Y}$ (or
203 on the row-augmented scores) (Biancolillo & Næs, 2019b).

204

205 **2.3.3 Sequential and Orthogonalized-Covariance Selection-Linear Discriminant Analysis (SO-**
206 **CovSel-LDA)**

207 As the name suggests, Sequential and Orthogonalized-Covariance Selection-Linear Discriminant
208 Analysis (SO-CovSel-LDA) (Biancolillo et al., 2019a) is a multi-block classification method
209 strictly linked to SO-PLS-LDA. In fact, the two approaches have similar algorithm, but in SO-
210 CovSel-LDA the feature reduction is operated by a variable selection method called Covariance
211 Selection (CovSel) (Roger et al., 2011). CovSel is a feature reduction approach developed to select
212 variables in a regression context; in fact, it points out the predictors which contribute the most to the

213 estimation of a response. Briefly, considering a predictor block $\boldsymbol{X}$, used to estimate a response $\boldsymbol{Y}$,
214 Covariance Selection iteratively selects the $\boldsymbol{X}$-variables presenting the highest covariance with the
215 response. As a consequence, the main divergence between SO-PLS and SO-CovSel, is that, in the
216 latter, calculations are based on the original variables instead on the scores (Biancolillo et al.,
217 2019a). Considering the two predictor blocks case above-mentioned ($\boldsymbol{X}$ and $\boldsymbol{Z}$), in order to create a
218 SO-CovSel model, the first step consists in selecting variables from the $\boldsymbol{X}$-block (by CovSel),
219 obtaining the reduced matrix $\boldsymbol{X_{red}}$, which is used to estimate the $\boldsymbol{Y}$ by ordinary least squares (OLS).
220 Then, $\boldsymbol{Z}$ is orthogonalized with respect to $\boldsymbol{X_{red}}$, obtaining $\boldsymbol{Z_{Orth}}$. Covariance Selection is then used
221 to select the $\boldsymbol{Z_{orth}}$-variables which contribute the most to the prediction of the residuals ($\boldsymbol{E_Y}$) from
222 the previous regression, obtaining $\boldsymbol{Z_{Orth,red}}$, which is used to estimate $\boldsymbol{E_Y}$ by OLS. Finally, the
223 predicted $\boldsymbol{Y}$ is calculated by summing up the individual predictions made by the two regressions.
224 Once the SO-CovSel model is created, LDA can be applied on the predicted $\boldsymbol{Y}$. Also in this case, in
225 order to solve the classification problem, the $\boldsymbol{Y}$ is a binary matrix encoding class-belongings.
226
227 **3. Results and discussion**
228 After the collection of spectra, IR signals were exported in MatLab (The Mathworks, Natick, MA;
229 version 2015b) for the analysis. In order to pursue external validation of the models, spectra are
230 reorganized into a training set, for the optimization of the calibration model, and a test set, for
231 validation, by the Duplex algorithm (Snee, 1977) (more details about the division are reported
232 below).
233 In Figure 1 the average spectra collected on cloves (in red) and on tunics (in blue) are shown. From
234 the figure it is straightforward the IR signals collected on the different compartments of garlic are
235 slightly different, in particular in specific absorption ranges; a wider discussion over the
236 interpretation of the spectra is reported in the related sections.

237

238 Independently on the classifier used, different pretreatments have been tested on training spectra:
239 bare mean centering, 1st or 2nd derivative (following the Savitzsky-Golay approach, using 19 points
240 window and a second or third order for the interpolating polynomial, respectively) (Savitzky &
241 Golay, 1964), Standard Normal Variate (SNV) (Barnes et al., 1989) and combinations of SNV an
242 derivatives; the most suitable preprocessing approach (together with the optimal complexity, i.e.,
243 the number of latent variables) has been defined as the one leading to the lowest classification error
244 in a 7-fold cross-validation procedure. Even if not always explicitly mentioned, data is assumed to
245 be mean-centered prior the creation of any model.

246

247 **3.1 PLS-DA analysis on tunics**

248 As above-mentioned, the IR signals were divided into a calibration and a validation set. The 81
249 signals collected on tunics were divided into a training set of 61 samples (15 belonging to Class
250 Castelliri, 15 appertaining to Class Proceno, 15 for Class Nubia and 16 from Sulmona) and a test set
251 of 20 samples (5 per each category). Then, six different pretreatments (listed in Table 1) have been
252 tested on data, and an equal number of PLS-DA models were calculated (on training samples); the
253 preprocessing approaches, the number of latent variables (LVs) extracted and the average cross-
254 validated classification errors (%) are reported in Table 1.

255 ----------Insert Table 1 approx. here----------

256

257 After inspection of Table 1, the optimal calibration model has been built on data pretreated by SNV,
258 1st derivative and mean centering; this PLS-DA model, applied on the test set (preprocessed
259 accordingly) provided a classification rate of 75% (corresponding to 5 misclassified test samples
260 over 20; of these, 2 belong to class Castelliri, 2 to class Sulmona and 1 appertains to class Nubia).
261 After the creation of any PLS-based model, in order to give a depth insight into the data set under
262 study, it is possible to calculate the Variable Importance in Projection (VIP) indices (Wold et al.,

1993) to understand which variables contribute the most to the model; generally, each spectral variable presenting a VIP index higher than 1 is considered relevant. A graphical representation of VIP analysis is shown in Figure 2; one plot per class is displayed in order to show which variables characterize each category.

----------Insert Figure 2 approx. here----------

In Figure 2, the black solid lines represent the average training signals (offset to avoid overlapping and make them visible) whereas the selected variables are highlighted as bold colored dashed lines: red for Class Castelliri, blue for Class Proceno, green for Class Nubia and cyan for Class Sulmona. From the figure it is possible to see that, as expected, the most relevant instrumental features are more or less the same among the four categories; in fact, independently on the class, variables presenting VIP indices higher than 1 are those around 2849 cm^{-1} and 2920 cm^{-1}, ascribable to the symmetric and antisymmetric stretching of the C-H bond, those in the spectral range between 1640 cm^{-1} and 1629 cm^{-1}, attributable to the stretching of carboxylate groups, variables around 1417cm^{-1} ascribable to CH bending and some from 1181 cm^{-1} to 914 cm^{-1}, linked to the absorptions caused by CH deformations, skeletal stretching of C-O and C-C.

Despite the similarities clearly visible in the plot, it is also possible to spot some variables that are selected for the characterization of some categories, but not for others. For instance, spectral variables from 1192 cm^{-1} to 1292 cm^{-1} present VIP indices higher than 1 for all the categories except for Class Castelliri, probably indicating a different composition in oligo- and polysaccharides among the diverse categories. More details about spectral absorptions can be found in Section 2.2 and in the related literature.

3.2 PLS-DA analysis on cloves

288 The PLS-DA analysis of spectra collected on cloves has been carried out in the same way as
289 described above for tunics. The 82 signals were divided into a training set of 62 objects (containing
290 16 samples per Class Castelliri, 15 Proceno, 14 Nubia, and 17 per Class Sulmona) and a test set of
291 20 samples (5 per category). The same pretreatments have been tested on data, and, also in this
292 case, the most suitable has been defined (together with the optimal complexity) in cross-validation;
293 results are reported in Table 2.
294 ----------Insert Table 2 approx. here----------
295

296 In this case, the model leading to the lowest classification error in cross-validation is the one built
297 on data after 1st derivative (and mean centering). When the optimal model was applied to the test
298 set, only 3 over 20 samples were misclassified (correct classification rate: 85%); among these, 2
299 belong to Class Castelliri and 1 to Class Sulmona.
300 Also in this case the VIP analysis was pursued in order to inspect the variables contributing the
301 most to the observed differentiation among the geographical origin; the agreement regarding the
302 selected variables among the different classes was strong (plot is not shown); the most relevant
303 variables from the classification point of view were those from 2944 cm^{-1} to 2837 cm^{-1}, and some
304 around 1413 cm^{-1}, and 1757 cm^{-1}.
305 Despite the results obtained by the individual analysis of data blocks were quite satisfactory, the
306 above-mentioned multi-block approaches have been used, testing whether it would be possible to
307 improve predictions.
308
309

3.3. Multi-block analysis

Unfortunately, the IR analysis of both tunics and cloves was not available for all the samples discussed for the PLS-DA analysis; consequently, the multi-block data set has been reduced to the 69 samples which have been analyzed on both compartments.

In order to divide samples into a training and a test set taking into account both blocks of measures, two PCA models have been calculated, one per each set of data. Then, the first 5 principal components extracted by each PCA model were row-wise concatenated; finally, samples were divided into training and test set by the Duplex algorithm (Snee, 1977) calculating sample distances in the scores-space defined by the PCs.

Due to the reduction of the available samples, the training set included 49 samples (14 belonging to Class Proceno, 15 from Class Castelliri, 12 appertaining to Class Nubia and 8 of Class Sulmona), and the test set was made of 20 objects (5 objects per class).

As anticipated, two multi-blocks classifiers have been applied at this stage of the work; independently of the approach used, spectra collected on cloves have been used as first input block, while signals collected on tunics have been used as the second one.

Classification models have been built also using an inverted input order (tunics-data modelled as first block and clove-data as second) but results were slightly worse; consequently, these will not be discussed in the following sections.

3.3.1 SO-PLS-LDA analysis

Building the SO-PLS models, all the possible combinations of the above-mentioned preprocessing approaches have been tested on the two data blocks in a cross-validation procedure (7 cancellation groups). The optimal calibration model is the one calculated using the mean-centered spectra collected on cloves as first input block and signals on tunics as the second one (pretreated by 2^{nd} derivative and mean centering). The number of components extracted from the two blocks are 4 and 11 for spectra on cloves and tunics, respectively. This model has been applied on the test set

(pretreated accordingly) and it provided 100% of correct classification for all categories except for class Sulmona, whose correct classification rate was 80% (corresponding to 1 misclassified sample). The results are graphically shown in Figure 3, where samples are projected onto the space of the first two canonical variates.

----------Insert Figure 3 approx. here----------

Looking at the plot, it is possible to recognize a quite clear distinction among the four different classes. In particular, the first canonical variate allows discriminating samples belonging to class Castelliri (red circles, at negative values) from the other three categories; while the second canonical variate allows distinguishing the objects belonging to class Proceno (blue squares) and class Sulmona (green triangle), at negative scores, from those belonging to class Castelliri (red circles) and Nubia (black diamonds) at positive values. From the figure, it is easy to spot the misclassified sample from class Sulmona (green triangle), in fact, this falls closer to the centroid of class Nubia rather than to the one of its own category.

VIP analysis was pursued on the SO-PLS model, following the embedded strategy described in (Biancolillo et al., 2016); nevertheless, the results were not relevantly different from those previously described for the individual PLS-DA analysis, and therefore they are not reported.

3.3.2 SO-CovSel-LDA analysis

Similarly to SO-PLS-LDA analysis, also for SO-CovSel-LDA, several multi-block models have been built (in a cross-validation procedure) in order to test different combinations of pretreatments; simultaneously, also the number of variables to be selected per each block is chosen. The optimal model has been calculated on the clove-block pretreated by 1st derivative whereas the signals collected on tunics were preprocessed by SNV and 1st derivative. The number of selected variables is 1 and 7 on cloves- and tunics-block, respectively. The application of the calibration model to the

test set led to a correct classification rate of 85%, corresponding to 3 misclassified samples in total (2 object from Class Castelliri and 1 belonging to Class Sulmona assigned to Class Proceno).

As described in the Section 2.3.3, SO-CovSel-LDA naturally provides information about the variables contributing the most to the classification. A visual representation of the selected variables is reported in Figure 4, in particular, the mean spectra collected on cloves and tunics are reported in in Figure 4a and Figure 4b, respectively; selected variables are highlighted by red circles.

----------Insert Figure 4 approx. Here----------

As expected, the selection provided by CovSel is sharper than the one achieved by VIP analysis; nevertheless, the two are in agreement. In fact, CovSel selects the variable at 2843 cm^{-1} in the spectra collected on cloves, and those at 2917 cm^{-1} and 3285 cm^{-1} on tunics, probably associated to CH_3, CH_2, O-H and N-H stretching; variables at 1013 cm^{-1}, 1034 cm^{-1}, 1588 cm^{-1}, 400 cm^{-1} and 945 cm^{-1} associable to polysaccharides (for more details the reader is addressed to Section 2.2 and to the related literature).

4. General overview of the results

In general, all the classification models provided acceptable results, indicating ATR FT-IR coupled with discriminant classifiers could represent a suitable approach for assessing the geographical origin of the investigated cultivars of red garlic. The best results, from the prediction point of view, are provided by a multi-block approach; this outcome is somehow expected, because data fusion strategies are supposed to provide comparable or better results than models built on the individual data blocks (Biancolillo, et al. 2019c). In order to ease a comprehensive overview of the classification rates provided by the different models, they are reported all together in Table 3.

----------Inset Table 3 approx. here----------

From the table, it is straightforward the most suitable methodology to solve the classification problem under study is SO-PLS-LDA. Concerning the single-block analysis, the best results are

provided by the model built on data collected on the cloves. This latter achievement suggests this compartments contains more information suitable for distinguishing the different red garlic ecotypes.

5. Conclusions

The aim of the present work was to develop a non-destructive approach suitable for distinguishing different cultivars of red garlic according to their geographical origin. In order to achieve this goal, samples of red garlic harvested in four different Italian towns (Castelliri, Proceno, Nubia and Sulmona) were analyzed by ATR-FTIR spectroscopy and classified. Spectra collection was pursued on both tunics of bulbils and cloves, avoiding any other physical-chemical pretreatment of samples. The data-block obtained were individually analyzed by PLS-DA and involved in multi-block models by the application of SO-PLS-LDA and SO-CovSel-LDA. In general, all the approaches provided good results from the prediction point of view. Concerning the classification pursed on the individual data blocks, the lowest classification error was provided by the PLS-DA model calculated on spectra collected on cloves, which led to the misclassification of three test objects over twenty. Nevertheless, the best results have been provided by a data-fusion strategy, the SO-PLS-LDA approach, which allowed achieving extremely satisfactory results, misclassifying only one sample over the 20 constituting the validation set.

References

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics, 17*, 166–173. https://doi.org/10.1002/cem.785

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy, 43*, 772-777. https://doi.org/10.1366/0003702894202201

Beato, V.M., Orgaz, F., Mansilla, F., & Montaño, A. (2011). Changes in Phenolic Compounds in Garlic (Allium sativum L.) Owing to the Cultivar and Location of Growth, *Plant Foods for Human Nutrition, 66,* 218-223 https://doi.org/10.1007/s11130-011-0236-2.

Biancolillo, A., Måge, I., & Næs, T. (2015). Combining SO-PLS and linear discriminant analysis for multi-block classification. *Chemometrics and Intelligent Laboratory Systems,* 141, 58–67. https://doi.org/10.1016/j.chemolab.2014.12.001

Biancolillo, A., Hovde Liland, K., Måge, I., Næs, T., & Bro, R. (2016). Variable selection in multi-block regression. *Chemometrics and Intelligent Laboratory Systems* 156, 89–101. https://doi.org/10.1016/j.chemolab.2016.05.016

Biancolillo, A., De Luca, S., Bassi, S., Roudier, L., Bucci, R., Magrì, A.D., & Marini. F. (2018a). Authentication of an Italian PDO hazelnut ("Nocciola Romana") by NIR spectroscopy, *Environmental Science and Pollution Research* 25 28780-28786.

Biancolillo, A., & Marini, F. (2018b). Chemometrics Applied to Plant Spectral Analysis. In J. Lopes, & C. Sousa (Eds.), *Vibrational Spectroscopy for Plant Varieties and Cultivars Characterization, Comprehensive Analytical Chemistry, 80* (pp. 69-104). Amsterdam: Elsevier.

Biancolillo, A., Marini, F., & Roger, J-M. (2019a). SO-COVSEL: a novel method for variable selection in a multi-block framework, Journal of Chemometrics, e3120, https://doi.org/10.1002/cem.3120

Biancolillo, A., & Naes, T. (2019b). The sequential and orthogonalised PLS regression (SO-PLS) for multi-block regression: theory, examples and extensions. In M. Cocchi (Ed.), *Data Handling in Science and Technology*, 31 (pp. 157-177). Amsterdam: Elsevier.

Biancolillo, A., Boqué, R., Cocchi, M., & Marini, F. (2019c). Data Fusion strategies in food analysis. In M. Cocchi (Ed.), *Data Handling in Science and Technology,* 31 (pp. 271-310). Amsterdam: Elsevier.

D'Archivio, A.A., Di Vacri, M.L., Ferrante, M., Maggi, M.A., Nisi, S., & Ruggieri, F. (2019a). Geographical discrimination of saffron (Crocus sativus L.) using ICP-MS elemental data and class modeling of PDO Zafferano dell'Aquila produced in Abruzzo (Italy). *Food Analytical Methods*, 1-10.

D'Archivio, A.A., Foschi, M., Aloia, R., Maggi, M.A., Rossi, L. & Ruggieri, F. (2019b). Geographical discrimination of red garlic (Allium sativum L.) produced in Italy by means of multivariate statistical analysis of ICP-OES data. *Food Chemistry, 275*, 333–338. https://10.1016/j.foodchem.2018.09.088.

European Commission, Agriculture and Rural Development. DOOR database, ec.europa.eu/agriculture/quality/door/list.html?locale=en Accessed 4 April 2019.

Firmani, P., Bucci, R., Marini, F., Biancolillo, A. (2019). Authentication of "Avola almonds" by near infrared (NIR) spectroscopy and chemometrics. *Journal of Food Composition and Analysis*, *82* 2019, 103235.

Giannetti, V., Boccacci Mariani, M., Marini, F., Torrelli, P., Biancolillo, A. (2019). Flavour fingerprint for the differentiation of Grappa from other Italian distillates by GC-MS and chemometrics. *Food Control*, 105 123–130.

Hrbek, V., Rektorisova, M., Chmelarova, H., Ovesna, J., & Hajslova, J. (2018) Authenticity assessment of garlic using a metabolomic approach based on high resolution mass spectrometry, *Journal of Food Composition and Analysis, 67*, 19–28. https://doi.org/10.1016/j.jfca.2017.12.020.

Kallel, F., Driss, D., Bouaziz, F., Belghith, L., Zouari-Ellouzi, S., Chaari, F., Haddar, A., Chaabouni, S.E., & Ghorbel, R. (2015). Polysaccharide from garlic straw: Extraction, structural data, biological properties and application to beef meat preservation, *RSC Advances, 5,* 6728–6741. https://10.1039/c4ra11045e.

Khar, A., Banerjee, K. Jadhav, M.R., & Lawande, K.E. (2011). Evaluation of garlic ecotypes for allicin and other allyl thiosulphinates, *Food Chemistry, 128*, 988–996. https://doi.org/10.1016/j.foodchem.2011.04.004.

Lastra-Mejías, M., González-Flores, E., Izquierdo, M., Cancilla, J.C., & Torrecilla, J.S. (2020). Cognitive chaos on spectrofluorometric data to quantitatively unmask adulterations of a PDO vinegar. *Food Control*, *108*, 106860.

Lu, X., Ross, C.F., Powers, J.R. Aston, D.E. & Rasco, B.A. (2011). Determination of total phenolic content and antioxidant activity of garlic (Allium sativum) and elephant garlic (Allium ampeloprasum) by attenuated total reflectance-fourier transformed infrared spectroscopy, *Journal of Agricultural and Food Chemistry, 59*, 5215–5221. https://doi.org/10.1021/jf201254f.

Martins, N., Petropoulos, S., & Ferreira, I.C.F.R. (2016). Chemical composition and bioactive compounds of garlic (Allium sativum L.) as affected by pre- and post-harvest conditions: A review. *Food Chemistry*, *211,* 41–50. https://doi.org/10.1016/j.foodchem.2016.05.029.

Mora, M., Elzo-Aizarna, J., Rozas-Fuertes, S., Velilla-Echeita, L., Vázquez-Araújo, L. (2020). Implicit reaction vs explicit emotional response: Protected designation of origin in apple cider. *Food Quality and Preference*, 79, 103773.

Minzhen, S., Lun, L., Chuanyun, Z., & Deqing, Z. (2015) Identification and Characterization of volatile Organic Compounds of Fresh Plant Using Headspace Combined with Surface-Enhanced Raman Scattering. *Journal of Food Processing & Technology*, *6*, 1-6. https://doi.org/10.4172/2157-7110.1000520

Montaño, A., Beato, V.M., Mansilla, F., & Orgaz, F. (2011). Effect of Genetic Characteristics and Environmental Factors on Organosulfur Compounds in Garlic (Allium sativum L.) Grown in Andalusia, Spain. *Journal of Agricultural and Food Chemistry, 59*, 1301-1307. https://doi.org/10.1021/jf104494j.

Movasaghi, Z., Rehman, S., & Rehman, I. (2008). Transform Infrared (FTIR) Spectroscopy of Biological Tissues. *Applied Spectroscopy Reviews*, *43*, 134-179.

https://doi.org/10.1080/05704920701829043.

Næs, T., Tomic, O., Mevik, B.-H., & Martens, H. (2011). Path modelling by sequential PLS regression. *Journal of Chemometrics, 25*, 28–40. https://doi.org/10.1002/cem.1357

Nikolić, V. D., Ilić, D.P., Nikolić, L.B. Stanković, M.Z., Stanojević, L., P., Savić, I., M., & Savić, I., M. (2012). The synthesis and structure characterization of deoxyalliin and alliin. *Advanced technologies, 1*, 38-46.

Pérez, N.F., Ferré, J., & Boqué, R. (2009). Calculation of the reliability of classification in discriminant partial least-squares binary classification. *Chemometrics and Intelligent Laboratory System, 95*, 122-128. https://doi.org/10.1016/j.chemolab.2008.09.005

Rahman, K. (2003). Garlic and aging: New insights into an old remedy. *Ageing Research Reviews, 2*, 39-56. https://10.1016/S1568-1637(02)00049-1.

Ramirez, D.A., Locatelli, D.A., González, R.E., Cavagnaro, P.F., & Camargo, A.B. (2017). Analytical methods for bioactive sulfur compounds in Allium: An integrated review and future directions. *Journal of Food Composition and Analysis, 61*, 4–19. https://doi.org/10.1016/j.jfca.2016.09.012

Reddy, J.P., & Rhim, J.W. (2014). Isolation and characterization of cellulose nanocrystals from garlic skin, *Materials Letters, 29,* 20-23. https://10.1016/j.matlet.2014.05.019.

Ritota, M., Casciani, L., Han, B.-Z., Cozzolino, S., Leita, L., Sequi, P., & Valentini, M. (2012). Traceability of Italian garlic (Allium sativum L.) by means of HRMAS-NMR spectroscopy and multivariate data analysis. *Food Chemistry, 135*, 684–693. https://doi.org/10.1016/j.foodchem.2012.05.032.

Rocha, S., Pinto, E., Almeida, A., & Fernandes, E. (2019). Multi-elemental analysis as a tool for characterization and differentiation of Portuguese wines according to their Protected Geographical Indication. *Food Control*, *103*, 27-35.

543 Roger, J-M., Palagos, B., Bertrand, D., & Fernandez-Ahumada, E. (2011). CovSel: variable
544 selection for highly multivariate and multi-response calibration, application to IR spectroscopy.
545 *Chemometrics and Intelligent Laboratory Systems, 106*, 216-223.
546 https://doi.org/10.1016/j.chemolab.2010.10.003
547
548 Savitzky, A., & Golay, M.J.E. (1964). Smoothing and differentiation of data by simplified least
549 squares procedures. *Analytical Chemistry, 36*, 1627-1639. https://doi.org/10.1021/ac60214a047
550
551 Schulz, H., & Baranska, M. (2007). Identification and quantification of valuable plant substances by
552 IR and Raman spectroscopy. *Vibrational Spectroscopy*, *43*, 13-25
553 https://doi.org/10.1016/j.vibspec.2006.06.001
554
555 Shukla, Y., Kalra, & N., (2007). Cancer chemoprevention with garlic and its constituents, *Cancer*
556 *Letters, 247*, 167-181. https://doi.org/10.1016/j.canlet.2006.05.009.
557
558 Sjöström, M., Wold, S., & Söderström, B. (1986). PLS discriminant plots. In E.S. Gelsema, &
559 L.N.Kanal (Eds.), *Pattern recognition in practice* (pp. 461-470). Amsterdam: Elsevier.
560
561 Slow Food Foundation. https://www.fondazioneslowfood.com/en/ Accessed 4 April 2019.
562
563 Smith, R.G. (2005). Determination of the country of origin, of garlic (Allium sativum) using trace
564 metal profiling, *Journal of Agricultural and Food Chemistry, 53*, 4041–4045.
565 https://10.1021/jf040166+.
566
567 Snee, R. D. (1977). Validation of regression models: methods and examples. *Technometrics, 19*,
568 415-428.
569
570 Stark, N.M., Yelle, D.J., Agarwal, U.P. (2016). Techniques for Characterizing Lignin. In O. Faruk
571 & M. Sain (Eds.), *Lignin in Polymer Composites* (pp. 49-66) Amsterdam: Elsevier
572 DOI: https://doi.org/10.1016/B978-0-323-35565-0.00004-7
573

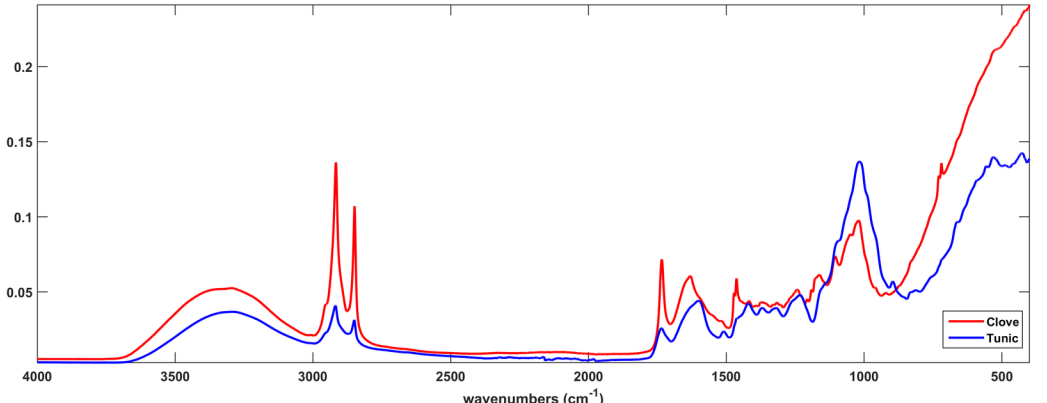
574 Ståhle, L., & Wold, S. (1987). Partial least squares analysis with cross-validation for the two-class
575 problem: a Monte Carlo study. *Journal of Chemometrics, 1,* 185-196.
576 https://doi.org/10.1002/cem.1180010306
577
578 Trirongjitmoah, S., Juengmunkong, Z., Srikulnath, K., & Somboon, P. (2015). Classification of
579 garlic cultivars using an electronic nose, *Computers and Electronics in Agriculture, 113,* 148–153.
580 https://10.1016/j.compag.2015.02.007.
581
582 Vasi, S., Alfa, M., Salvo, A., Bua, G., Giofrè, S., Corsaro, C., Mallamace, D., Cicero, N., Mottese,
583 A., Vadalà, R., & Dugo, G. (2016). Statistical Analysis of Mineral Concentration for the
584 Geographic Identification of Garlic Samples from Sicily (Italy), Tunisia and Spain. *Foods*, *5*, 20.
585 https://10.3390/foods5010020.
586
587 Yun, H.M., Ban, J.O., Park, K.R., Lee, C.K., Jeong, H.S., Han, S.B., & Hong, J.T. (2014). Potential
588 therapeutic effects of functionally active compounds isolated from garlic. *Pharmacology &*
589 *Therapeutics, 142*, 183–195. https://doi.org/10.1016/j.pharmthera.2013.12.005.
590
591 Wold, S., Johansson, E., & Cocchi, M. (1993). PLS: partial least squares projections to latent
592 structures. In H. Kubinyi (Ed.), *3D QSAR in Drug Design: Theory, Methods and Applications* (pp.
593 523-550), Leiden, The Netherlands: KLUWER ESCOM Science Publisher.
594
595 Wold, S., Martens, H., & Wold, H. (1983). The multivariate calibration problem in chemistry
596 solved by the PLS method. In A. Ruhe, B. Kågström (Eds.), *Proceedings of the Conference on*
597 *Matrix Pencils. Lecture Notes in Mathematics* (pp 286-293), Heidelberg, Germany: Springer
598 Verlag.
599
600

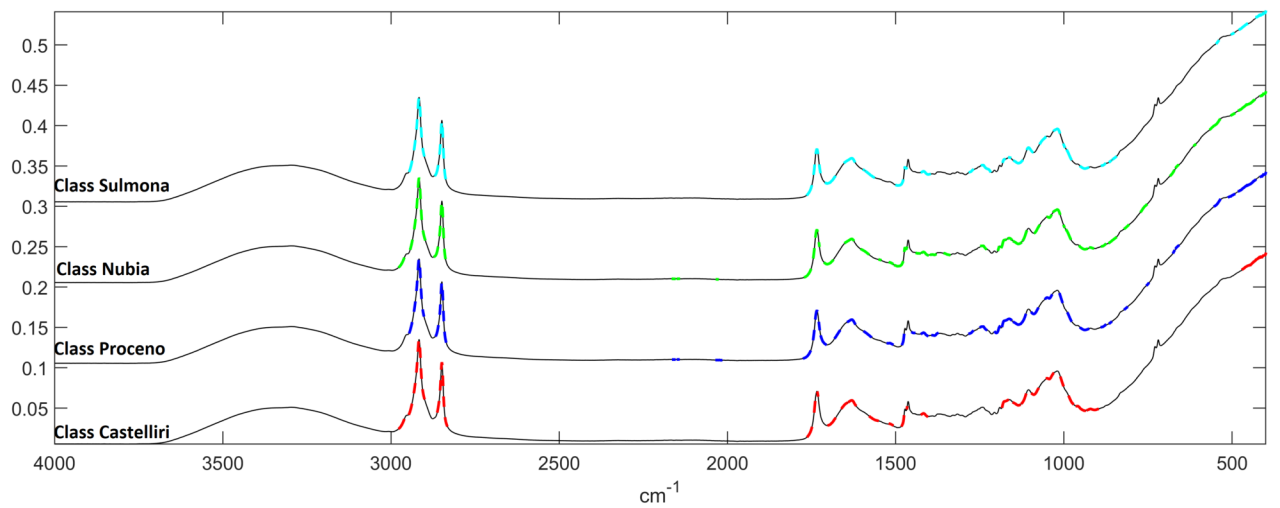
601 **Figure captions**

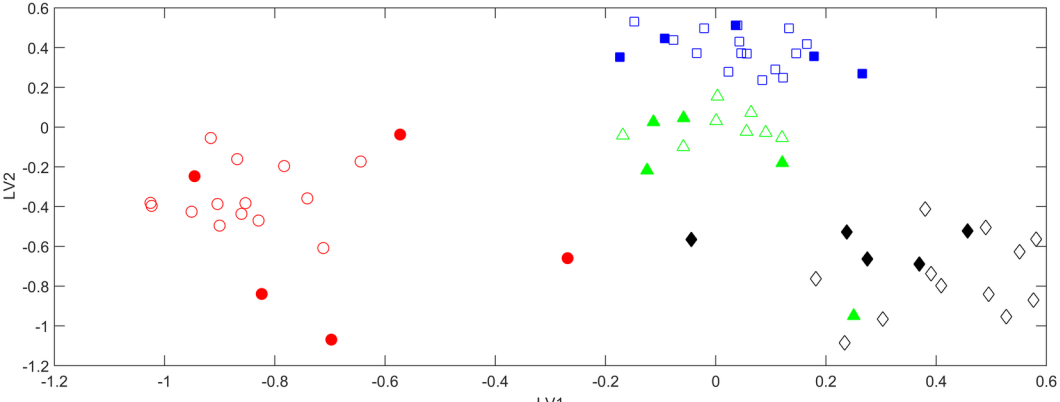
602 Figure 1 Mean spectra collected on cloves (in red) and on tunics (in blue).

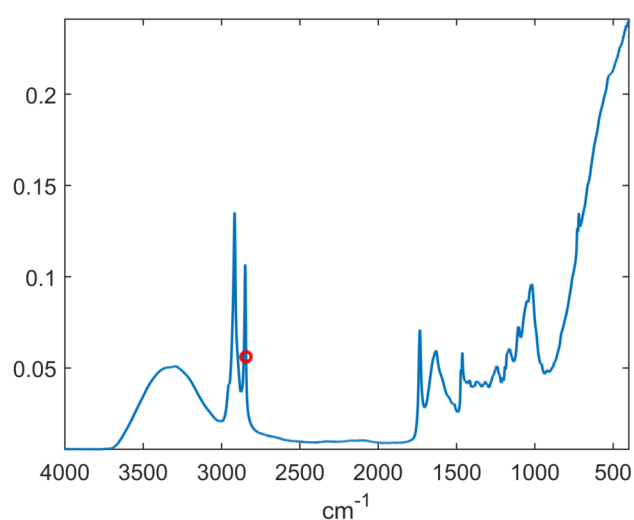
603 Figure 2 VIP analysis. Solid black lines represent mean spectra for the four different categories
604 (offset to make them visible) while bold variables highlight selected features. The upmost plot
605 refers to spectra collected on samples from Sulmona, lines in the middle are mean spectrum for

606 class Nubia and Proceno, respectively; the lowest lines represent the mean signal for samples
607 belonging to class Castelliri.
608
609 Figure 3 SO-PLS-LDA analysis: Samples project onto the first to canonical variates. Legend: Red
610 circles: Class Castelliri; Blue squares: Class Proceno; Black diamonds: Class Nubia; Green
611 triangles: Class Sulmona. Empty and filled symbols represent training and test samples,
612 respectively.
613
614 Figure 4 SO-CovSel Analysis: blue line represents mean spectrum collected on a) cloves b) tunics.
615 Variables selected by SO-CovSel are circled in red.

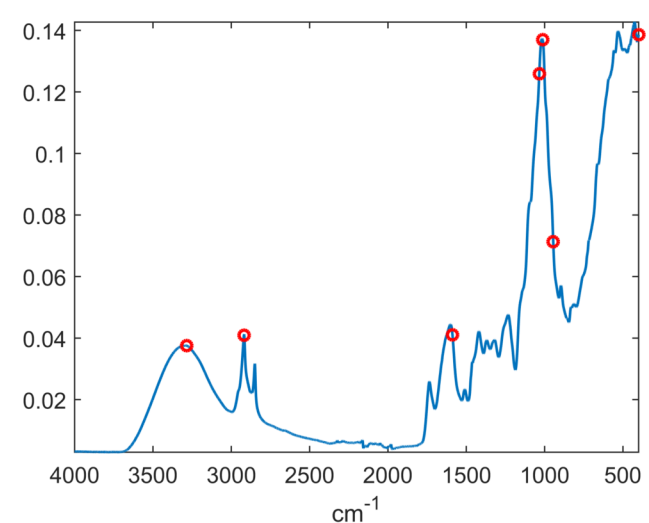

0.2
0.15
0.1
0.05
4000
3500
3000
2500
2000
1500
1000
500
wavenumbers (cm-1)
Clove
Tunic


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
Class Sulmona
Class Nubia
Class Proceno
Class Castelliri
4000
3500
3000
2500
2000
1500
1000
500
cm-1


LV2
LV1
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.2
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6


0.2
0.15
0.1
0.05
4000 3500 3000 2500 2000 1500 1000 500
cm-1

a)


0.14
0.12
0.1
0.08
0.06
0.04
0.02
4000 3500 3000 2500 2000 1500 1000 500
cm-1

b)

Table 1 PLS-DA analysis on tunics: Cross-validated mean classification errors (%) as function of preprocessing and complexity (LVs).

Preprocessing	LVs	Average classification errors (%-CV)
Mean Centering (MC)	10	12.3
1st Derivative + MC	12	10.2
2nd Derivative + MC	7	12.5
SNV + MC	10	11.8
SNV + 1st Derivative + MC	8	8.7
SNV + 2nd Derivative + MC	10	12.0

Table 2 PLS-DA analysis on cloves: Cross-validated mean classification errors (%) as function of preprocessing and complexity (LVs).

Preprocessing	LVs	Average classification errors (%)
Mean Centering (MC)	13	13.4
1st Derivative + MC	15	10.5
2nd Derivative + MC	13	10.9
SNV + MC	16	12.8
SNV + 1st Derivative + MC	14	12.3
SNV + 2nd Derivative + MC	9	16.4

Table 3 Classification rates (on the test set) for the four proposed strategies.

Method	Classification rates on the test set (%)
PLS-DA on tunics	75.0
PLS-DA on cloves	85.0
SO-PLS-LDA	95.0
SO-CovSel-LDA	85.0