applied sciences

MDPI

Article

Authentication of Sorrento Walnuts by NIR Spectroscopy Coupled with Different Chemometric Classification Strategies

Luigi Amendola [1], Patrizia Firmani [1], Remo Bucci [1], Federico Marini [1,*] and Alessandra Biancolillo [2,*]

[1] Department of Chemistry, University of Rome "La Sapienza", P.le Aldo Moro, 500185 Rome, Italy; luigiamendola92@gmail.com (L.A.); patrizia.firmani@uniroma1.it (P.F.); remo.bucci@uniroma1.it (R.B.)

[2] Department of Physical and Chemical Sciences, University of L'Aquila, Via Vetoio, Coppito, 67100 L'Aquila, Italy

* Correspondence: federico.marini@uniroma1.it (F.M.); alessandra.biancolillo@univaq.it (A.B.)

Received: 18 May 2020; Accepted: 8 June 2020; Published: 9 June 2020

Abstract: Walnuts have been widely investigated because of their chemical composition, which is particularly rich in unsaturated fatty acids, responsible for different benefits in the human body. Some of these fruits, depending on the harvesting area, are considered a high value-added food, thus resulting in a higher selling price. In Italy, walnuts are harvested throughout the national territory, but the fruits produced in the Sorrento area (South Italy) are commercially valuable for their peculiar organoleptic characteristics. The aim of the present study is to develop a non-destructive and shelf-life compatible method, capable of discriminating common walnuts from those harvested in Sorrento (a town in Southern Italy), considered a high quality product. Two-hundred-and-twenty-seven walnuts (105 from Sorrento and 132 grown in other areas) were analyzed by near-infrared spectroscopy (both whole or shelled), and classified by Partial Least Squares-Discriminant Analysis (PLS-DA). Eventually, two multi-block approaches have been exploited in order to combine the spectral information collected on the shell and on the kernel. One of these latter strategies provided the best results (98.3% of correct classification rate in external validation, corresponding to 1 misclassified object over 60). The present study suggests the proposed strategy is a suitable solution for the discrimination of Sorrento walnuts.

Keywords: Walnuts; Classification; Traceability; Near Infrared Spectroscopy; Partial Least Squares-Discriminant Analysis; PLS-DA; Multi-Block; Data Fusion; Sequential and Orthogonalized Partial Least Squares Linear Discriminant Analysis (SO-PLS-LDA); Sequential and Orthogonalized Covariance Selection Linear Discriminant Analysis (SO-CovSel-LDA)

1. Introduction

Walnut is the fruit of the *Juglans regia* L. tree. It is an economically interesting arboreal species, appreciated for its wood and edible fruits, which grows in temperate climate areas. Its seed is an important source of phospholipids, tocopherol, proteins, and mono- and polyunsaturated fatty acids. Overall, it is a noticeable source for some microelements, such as iron, copper, selenium, and zinc. In addition, as it has been observed that their consumption is capable to reduce the incidence of coronary diseases, walnuts have been deeply studied in the last few years [1]. In Italy, walnuts are harvested throughout the national territory; nevertheless, the fruits produced in the Sorrento area (South Italy) are commercially valuable for their peculiar organoleptic characteristics, confirmed by genetic criteria established by Foroni et al. [2]. From this reason arises the necessity of characterizing Sorrento walnuts, in order to discriminate it from common fruits (having lower market value), preventing possible

commercial frauds (e.g., counterfeits). In general, many analytical techniques have been exploited in walnut analysis: chromatographic approaches have been involved in studies on polyphenolic fraction [3,4], to quantify bioactive compounds [5], or lipids [6]; moreover, thermogravimetric analysis coupled with gas chromatography/mass spectroscopy (GC/MS) has been used to study the thermal behavior of walnut shell [7], while Inductively Coupled Plasma Optical Emission Spectroscopy (ICP-OES) was exploited to study the mineral composition of walnuts and walnut oils [8]. In some works, also Simple Sequence Repeats (SSRs) is used to study different walnuts [1,9], while Ercisli et al. used image processing to distinguish different cultivars on the basis of dimensional, gravimetric, and morphological features of the fruits [10]. Eventually, Sinesio and Moneta discriminated walnuts varieties on the basis of morphological and organoleptic characteristics [11].

Near-infrared spectroscopy (NIRS) presents several advantages: if coupled with an integrating sphere, it allows carrying out analysis on raw samples, without any physical pre-treatment, resulting in a fast, green, relatively cheap, noninvasive, and automatable technique. Moreover, it is possible to analyze samples before selling them, in a time range compatible with their shelf-life and avoiding any loss of product. For this reason, NIRS was elected as the tool of choice to address the Sorrento walnut authentication problem. There are many examples in literature about the use of NIRS coupled with chemometric classification tools in food-related authentication issues; for instance, on dried foodstuff (cereal, fruits, and nuts) such as rice [12,13], tea [14,15], macadamia [16], hazelnuts [17], almonds [18,19], and several others [20] can be found. NIR (coupled with linear discriminant analysis) has been used for varietal discrimination on Portuguese walnuts, providing indication of the feasibility of this approach [21]. Under this perspective, NIRS has been paired with two different classification strategies, with the aim of distinguishing Sorrento from non-Sorrento walnuts. In particular, it has been coupled with Partial Least Squares-Discriminant Analysis (PLS-DA), which has been applied on data collected on the shell and on spectra of the kernels. This classifier is one of the most used in different fields [22–25], and it has been chosen because it had performed well in similar situations [26–29], and it is therefore a common choice in this context. Eventually, in order to investigate whether a simultaneous analysis of both sets of spectra could improve predictions, two multi-block strategies have been tested. In particular, Sequential and Orthogonalized Partial Least Squares Linear Discriminant Analysis (SO-PLS-LDA) [30] and Sequential and Orthogonalized Covariance Selection Linear Discriminant Analysis (SO-CovSel-LDA) [31] have been exploited, handling together both data blocks. These methods have been pursued because it is acclaimed that, when possible, data fusion provides more accurate results than individual models, and because the same strategies have been already applied with the same aim providing successfully results [32–35].

2. Materials and Methods

2.1. Samples and Dataset

Two-hundred-and-thirty-seven (237) walnut samples, belonging to different varieties, were gathered from different areas, as shown in Table 1.

Table 1. Provenance of the examined walnuts samples.

Type	Number of Samples	Class
California	32	Non-Sorrento
Italy (no Sorrento)	87	Non-Sorrento
Moldavia	13	Non-Sorrento
Sorrento	105	Sorrento

For each sample, NIR spectra were collected on the nutshell (two replicates, one per side) and on the kernel (two replicates, one per side), for a total of 4 spectra collected on each walnut. This procedure led to a total of 948 (237 × 4) NIR spectra. These measurements were organized into two

data matrices, the first one ($\mathbf{X_1}$ -of dimensions 237 × 3112-), containing all the spectra collected on the shell (averaged over the two replicates) and the second one ($\mathbf{X_2}$ -of dimensions 237 × 3112-), made of the spectra collected on the kernel (averaged over the two replicates).

NIR spectra were collected by the OMNIC software (Thermo Scientific Inc., Madison, WI), and imported in MATLAB 2015b (The Mathworks, Natick, MA) for calculations. Prior to the creation of classification models, spectra were divided into training and test sets by using the Duplex algorithm [36] to pursue external validation of the models. In order to divide samples into subsets taking into account the variability of both $\mathbf{X_1}$ and $\mathbf{X_2}$, the procedure described in [37] was applied. Eventually, the training set included 177 samples (77 samples belonging to the "Sorrento class" and 100 to "non-Sorrento class"), while the test set comprehended 60 objects (28 Sorrento walnuts and 32 non-Sorrento); obviously, the same division was used for both $\mathbf{X_1}$ and $\mathbf{X_2}$.

2.2. Chemometric Tools

2.2.1. Partial Least Squares Discriminant Analysis (PLS-DA)

Partial Least Squares Discriminant Analysis (PLS-DA) [38,39] is a widely applied tool in the context of discriminant classification. One of its major benefits is that it allows handling ill-conditioned data matrices (a condition often encountered working with spectral data) [40]. This approach, despite being a classifier, starts from the resolution of a regression problem defined between a predictor data matrix $\mathbf{X}$ and a dummy response $\boldsymbol{y}$ [41]. The dummy $\boldsymbol{y}$ has a key role in the application of the PLS-DA algorithm; in fact, it encodes class information through binary ciphering, and it allows solving the classification problem by estimating the regression equation represented by Equation (1):

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{b} \tag{1}$$

The solution is achieved by Partial Least Squares (PLS) [42,43]. In Equation (1), $\boldsymbol{b}$ represents the regression coefficients.

When the investigated problem involves only two classes (as in the present work), $\boldsymbol{y}$ is a binary vector whose elements represent whether the corresponding sample belongs to one class ($y = 1$) or to the other ($y = 0$). For example, in a two-category case, for six samples equally distributed between the two classes (the first three objects belonging to the first class and the remaining ones to the second category), the $\boldsymbol{y}$ dummy would be $\boldsymbol{y} = [\,1\ 1\ 1\ 0\ 0\ 0\,]^{\mathrm{T}}$. Once the calibration model is built on the training samples (i.e., a set of objects whose class-membership is known), it is possible to classify unknown samples ($\mathbf{X_{new}}$) and estimate the predicted $\hat{\boldsymbol{y}}_{new}$. Nevertheless, $\hat{\boldsymbol{y}}_{new}$ will be made of real numbers, and, consequently, the class-membership cannot be directly deduced. Different classification rules have been proposed to face this issue (see, e.g., in [44–47]; in the present work, the solution suggested by Indahl and collaborators, i.e., to apply linear discriminant analysis on the predicted response, has been applied [46].

In the present work, PLS-DA has been (separately) applied on spectra collected on the shell ($\mathbf{X_1}$) and on the kernel ($\mathbf{X_2}$).

2.2.2. Sequential and Orthogonalized Partial Least Squares Linear Discriminant Analysis (SO-PLS-LDA)

Sequential and Orthogonalized-PLS is a multi-block regression approach developed to handle data matrices removing redundant information possibly present [48]. For two predictor blocks ($\mathbf{X_1}$ and $\mathbf{X_2}$) and a response matrix y, the algorithm can be ensemble in four steps:

1. $\mathbf{X_1}$ is used to estimate $\boldsymbol{y}$ by PLS. Scores $\mathbf{T_{x1}}$ and $\boldsymbol{y}$-residuals $\mathbf{e}$ are calculated.
2. $\mathbf{X_2}$ is orthogonalized with respect to $\mathbf{T_{x1}}$, obtaining $\mathbf{X_{2orth}}$
3. $\mathbf{X_{2,orth}}$ is used to estimate $\mathbf{e}$ by PLS.

4. The equation $\boldsymbol{y = X_1 b + X_2 c + f}$ is solved ($\boldsymbol{b}$ and $\boldsymbol{c}$ being the regression coefficients and $\boldsymbol{f}$ the residuals).

Sequential and Orthogonalized Partial Least Squares Linear Discriminant Analysis leans on SO-PLS, exploiting it for features reduction. In fact, in order to apply SO-PLS-LDA analysis [30] it is sufficient to create the SO-PLS model, and then applying LDA on the predicted $\boldsymbol{y}$ or on the concatenated scores. For more details on SO-PLS-LDA the reader is referred to the works in [30,49]. Calculations were made using in-house written functions running under Matlab, which are freely downloadable at [50].

In the present work, SO-PLS-LDA has been used to distinguish Sorrento and Non-Sorrento walnuts, simultaneously handling data collected on the shell ($\mathbf{X_1}$) and on the kernel ($\mathbf{X_2}$).

2.2.3. Sequential and Orthogonalized Covariance Selection Linear Discriminant Analysis (SO-CovSel-LDA)

Sequential and Orthogonalized Covariance Selection Linear Discriminant Analysis [31] is a multi-block classifier based on the combination of the regression approach called SO-CovSel and LDA. The algorithm of SO-CovSel is similar to the one for SO-PLS; the main difference being the fact the feature reduction operated by PLS (in SO-PLS) is replaced by the variable selection achieved by CovSel [51]. Briefly, considering the two predictor blocks $\mathbf{X_1}$ and $\mathbf{X_2}$, for the prediction of the response matrix $\boldsymbol{y}$, the SO-CovSel algorithm can be summarized as follows.

1. Variables in $\mathbf{X_1}$ are selected by CovSel, obtaining the reduced matrix $\mathbf{X_{1sel}}$
2. $\mathbf{X_{1sel}}$ is used to predict $\boldsymbol{y}$ by ordinary least squares
3. $\mathbf{X_2}$ is orthogonalized with respect to $\mathbf{X_{1sel}}$, obtaining $\mathbf{X_{2orth}}$
4. Variables in $\mathbf{X_{2orth}}$ are selected by CovSel, obtaining the reduced matrix $\mathbf{X_{2orth,sel}}$
5. $\mathbf{X_{2orth,sel}}$ is used to estimate the residuals from step 2
6. The equation $\boldsymbol{y = X_1 b + X_2 c + f}$ is solved ($\boldsymbol{b}$ and $\boldsymbol{c}$ being the regression coefficients and $\boldsymbol{f}$ the residuals).

Eventually, if the aim is to create a classification model, LDA can be calculated on the y predicted at step 6. For more details on SO-CovSel-LDA, the reader is addressed to the work in [31]. Calculations were made using in-house written functions running under Matlab, which are freely downloadable at [52].

3. Results

After the division into training a test set described in Section 2.1, data collected on the shell ($\mathbf{X_1}$) and on the kernel ($\mathbf{X_2}$) were analyzed by PLS-DA. The outcomes of these analyses are reported in Sections 3.1 and 3.2, respectively. In both cases, different spectral pretreatments were tested on the spectra, in order to remove spurious information possibly present. The tested preprocessing approaches are 1st and 2nd derivatives calculated according to the Savitzky–Golay approach (19 points window, and second- or third-order interpolating polynomial, respectively) [53], Standard Normal Variate (SNV) [54], and their combinations. These pretreatments were chosen because derivatives are expected to remove both addictive and multiplicative effect from spectra, whereas SNV has been conceived to attenuate the artifacts given by the scattering. Moreover, the width of the interpolation window was selected on the basis of our previous experience with similar NIR data as the one providing the best compromise between noise reduction and excessive (artifact) smoothing.

Eventually, a multi-block strategy has been exploited for the joint analysis of both sets of spectra. The results of this latter analysis are reported in Section 3.3. In all the classification models described in these three sections, the optimal data pretreatment model parameters were selected as the ones leading to the lowest classification error in a 7-fold cross-validation procedure on the training samples. Regardless the pretreatment used, blocks were always mean-centered prior to the creation of any model.

3.1. PLS-DA Analysis of NIR Spectra Collected on the Shell

As mentioned, NIR spectra were collected on the whole nuts (i.e., on the shellnut); the average spectra for samples belonging to Class Sorrento (red line) and Class Non-Sorrento (blue line) are reported in Figure 1. From the plot is clear that the two spectra are very similar.

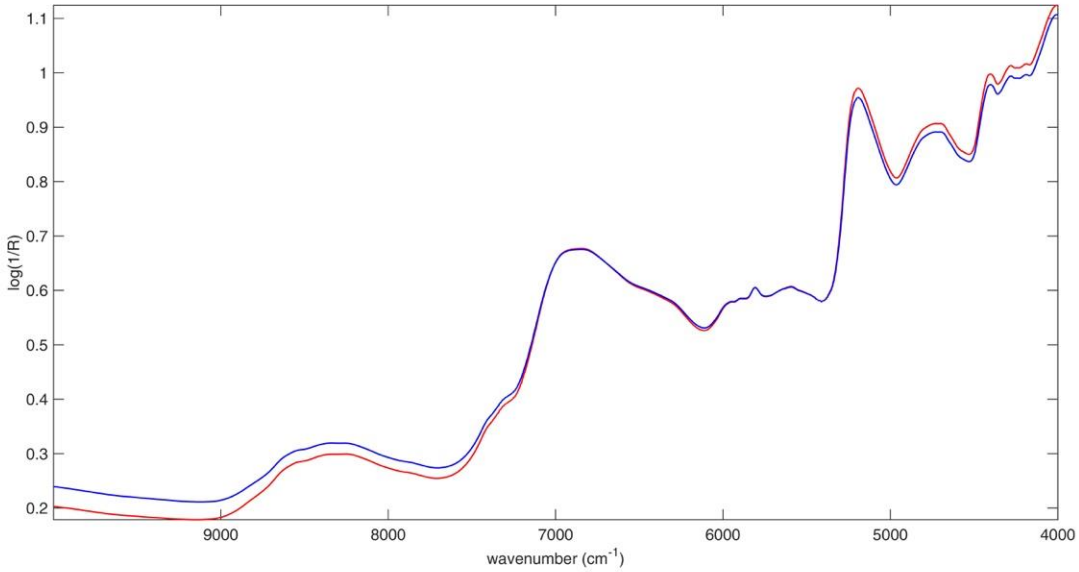

Figure 1. Average raw spectra collected on Sorrento (blue line) and Non-Sorrento (red line) walnuts (nutshell).

Different pretreatments were tested on the data. The average classification error in cross-validation (reported in Table 2) was used to select the optimal preprocessing.

Table 2. Partial Least Squares-Discriminant Analysis (PLS-DA) modeling of the spectra collected on the nutshell–results of cross-validation (LVs: Latent variables).

Pre-Treatment	LVs	Average Classification Error (%CV)
Mean Centering (MC)	9	2.8
1st derivative (+ MC)	11	1.6
2nd derivative (+ MC)	9	2.9
SNV (+ MC)	12	3.4
SNV+ 1st derivative (+ MC)	11	2.8
SNV+ 2nd derivative (+ MC)	9	2.9

The model providing the lowest classification error was the one built on data preprocessed by 1st derivative. The application of this model to the test set led to 92.9% sensitivity and 96.9% specificity for Class Sorrento; naturally, due to the symmetry of the classification results for a two class-problem, these values are reversed in the case of Class Non-Sorrento, for which sensitivity was 93.6% and specificity 92.9%. Altogether, two Sorrento and one Non-Sorrento test objects were misclassified. A graphical representation of this outcome is also reported in Figure 2, where the predicted **y** is displayed as a function of the sample index. In the figure, training objects are represented by empty symbols, while the test ones are displayed as filled items. The black dashed line in the plot is the threshold: samples falling above it are assigned by the model to Class Sorrento, whereas those below the line are predicted as belonging to Class Non-Sorrento. From the representation, it is easy to spot the three misclassified test samples: one object belonging to Class Non Sorrento (Blue square) and two samples appertaining to Class Sorrento (red diamonds).

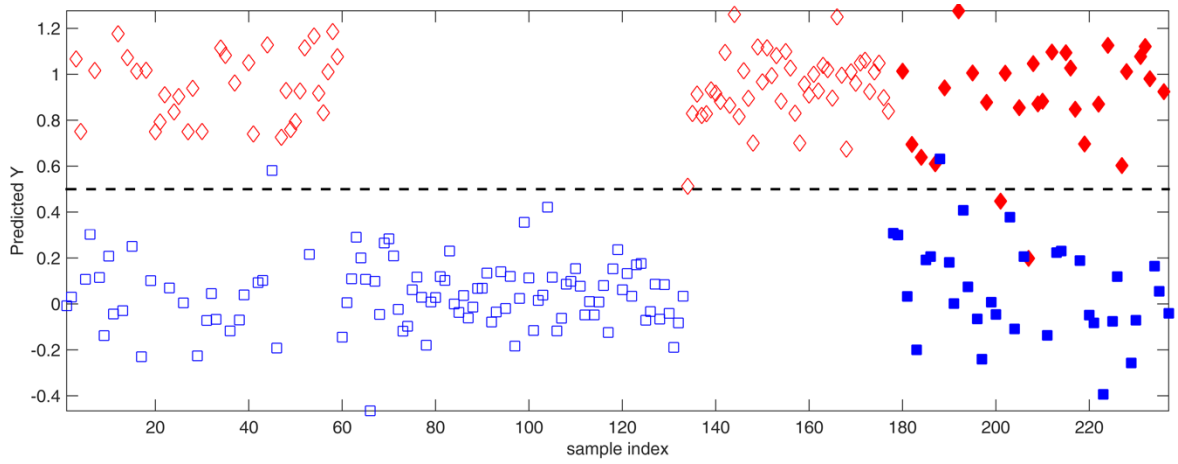

Figure 2. PLS-DA analysis: Predicted Y vs. sample index. Legend: Class Sorrento: Red diamonds; Class Non-Sorrento: Blue squares. The black dashed line represents the classification threshold between the two classes (Sorrento for y > threshold; Non-Sorrento for y < threshold). Empty and filled symbols represent training and test samples, respectively.

Eventually, in order to understand which spectral variables contribute the most to the discrimination between the two categories, Variable Importance in Projection (VIP) [55] analysis was performed. Applying this approach, it is possible to obtain a VIP index for each predictor (i.e., spectral variable), reflecting its contribution to the discriminant model. Customarily, a variable presenting a VIP index higher than 1 is counted as relevant. Handling spectral data, the outcomes of VIP analysis can be straightforwardly inspected through a graphical representation such as the one reported in Figure 3.

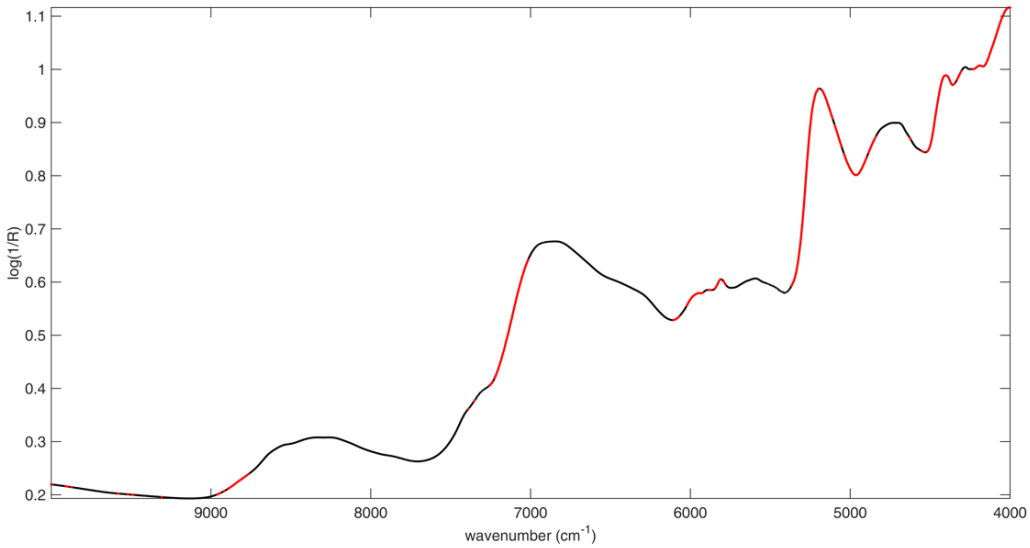

Figure 3. Variable Importance in Projection (VIP) Analysis. Mean spectrum (black line). Variables presenting a VIP index > 1 are highlighted in red.

In the plot, variables presenting a VIP index higher than 1 are highlighted in red, over the mean spectrum of the samples. From the figure, it can be noticed that the area between 4000 and 4200 cm^{-1} is selected. These variables interest the combination bands of C–H bonds and are probably due to the presence of fatty acids in walnuts. The spectral features constituting the peak at 5199 cm^{-1} (approximately from 4840 cm^{-1} to 5363 cm^{-1}) also present VIP index > 1. These variables are linked to the CC and CH combination modes of unsaturated fatty acids [54]. A VIP index higher than 1 are is

shown by variables in the spectral range 7000 to 7200 cm^{-1}; this area is associable to the presence of carbohydrates, or to the absorption of non-bonded O-H groups in fatty acids [56]. Eventually, some variables between 8800 cm^{-1} and 8900 cm^{-1} are selected by VIP analysis. In this range, the absorptions of the second overtone of the C–H bonds and the combinations bands of the O–H bond take place [57].

3.2. PLS-DA Analysis of NIR Spectra Collected on the Kernel

As discussed before, after measuring the shellnut, each walnut was opened and NIR spectra were collected on the kernels. Mean spectra for samples belonging to Class Sorrento (red line) and Class Non-Sorrento (blue line) are reported in Figure 4. Moreover, in this case, it is not possible to appreciate a significant difference between the spectra collected on samples belonging to the two categories.

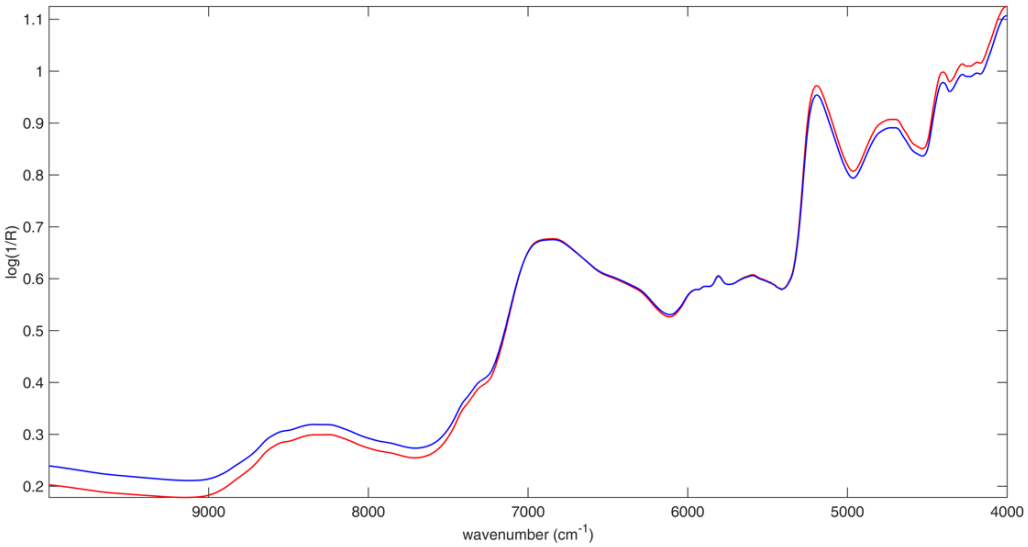

Figure 4. Average raw spectra collected on Sorrento (blue line) and Non-Sorrento (red line) walnuts (kernel).

PLS-DA analysis was carried out following the same procedure used for spectra collected on the nutshell. Consequently, different pretreatments were tested, and the optimal pretreatment and complexity to build the final calibration model were selected on the basis of smallest average classification error in cross validation. The results from this part of the analysis are summarized in Table 3.

Table 3. PLS-DA modeling of the spectra collected on the kernel: results of cross-validation (LVs: Latent variables).

Pre-Treatment	LVs	Average Classification Error (%CV)
Mean Centering (MC)	10	2.8
1st derivative (+ MC)	12	1.6
2nd derivative (+ MC)	10	4.1
SNV (+ MC)	10	7.1
SNV+ 1st derivative (+ MC)	12	2.8
SNV+ 2nd derivative (+ MC)	10	2.8

The PLS-DA model providing the lowest average classification error is the one built on data preprocessed by 1st derivative. Consequently, this pretreatment was considered the most suitable for the investigated data. When the calibration model was used to predict test samples, it correctly classified all Class Sorrento objects (i.e., 100% of sensitivity) with a specificity of 93.8%, and it misclassified two

(out of 32) Non-Sorrento samples (corresponding to a sensitivity of 93.8%), the specificity being 100%, due to the symmetry of the classification results for a two-class problem.

The predicted **y** is displayed as a function of the sample index in Figure 5. The plot is quite self-explanatory. As before, samples associated to a **y** higher than the threshold are predicted as belonging to Class Sorrento (otherwise, they are predicted as Class Non-Sorrento). From the figure, it is clear only two samples are misclassified: two Non-Sorreto samples (blue squares) predicted as belonging to Class Sorrento (red diamonds).

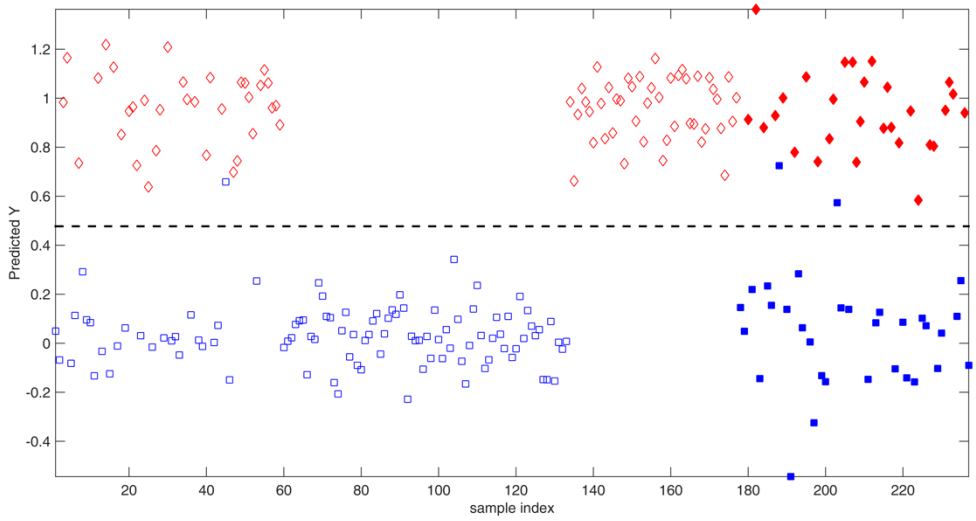

Figure 5. PLS-DA analysis: Predicted Y vs. sample index. Legend: Class Sorrento: Red diamonds; Class Non-Sorrento: Blue squares. The black dashed line represents the threshold between the two classes (Sorrento for y > threshold; Non-Sorrento for y < threshold). Empty and filled symbols represent training and test samples, respectively.

VIP analysis was run also on this model. The variables identified as important were in agreement with the ones discussed in Section 3.1; consequently, they will not be discussed again here, but they are shown in Figure A1 in Appendix A.

Comparing the predictions provided by the PLS-DA models built on spectra collected on the nutshell and on the kernel, it can be observed how one of the misclassified samples (belonging to Class Non-Sorrento) was wrongly predicted by both models. This is not completely surprising because, as detailed in Table 1, some Non-Sorrento walnuts are Italian, so they could have been harvested in an area nearby Sorrento or in a town presenting similar pedoclimatic conditions.

3.3. Multi-Block Analysis

3.3.1. SO-PLS-LDA Analysis

The sequential data fusion model was built using data preprocessed by the optimal pretreatment selected in individual analysis: first derivative. Building SO-PLS-LDA models, the optimal number of latent variables, six for X_1 and seven for X_2, was selected based on a cross-validation procedure. The corresponding optimal classification model provided a sensitivity of 98.7% and a specificity of 98.0% for Class Sorrento and vice versa for Class Non-Sorrento (i.e., 98.0% sensitivity and 98.7% specificity). When this SO-PLS-LDA model was used to predict validation samples, the classification rates were extremely satisfying. In fact, it correctly classified all test samples except one belonging to Class Non-Sorrento. In Figure 6 the histograms representing the scores of the training and test samples along the canonical variate are displayed both as scatterplot (panel a) and as histograms (panels b and c). Taking a look at this graphical representation of the results, it appears that samples belonging to Class Sorrento (red bars) present negative value of the canonical variate score; on the contrary, Non-Sorrento

samples present positive or slightly negative values of CV1. The misclassified test sample is the Non-Sorrento object whose score on the canonical variate is at around −0.1. VIP analysis has been carried out also on this model, following the procedure described in [58]. The selected variables are approximately the same highlighted before, so the discussion is not reported here; the graphical representation is displayed in Figure A2 in Appendix A.

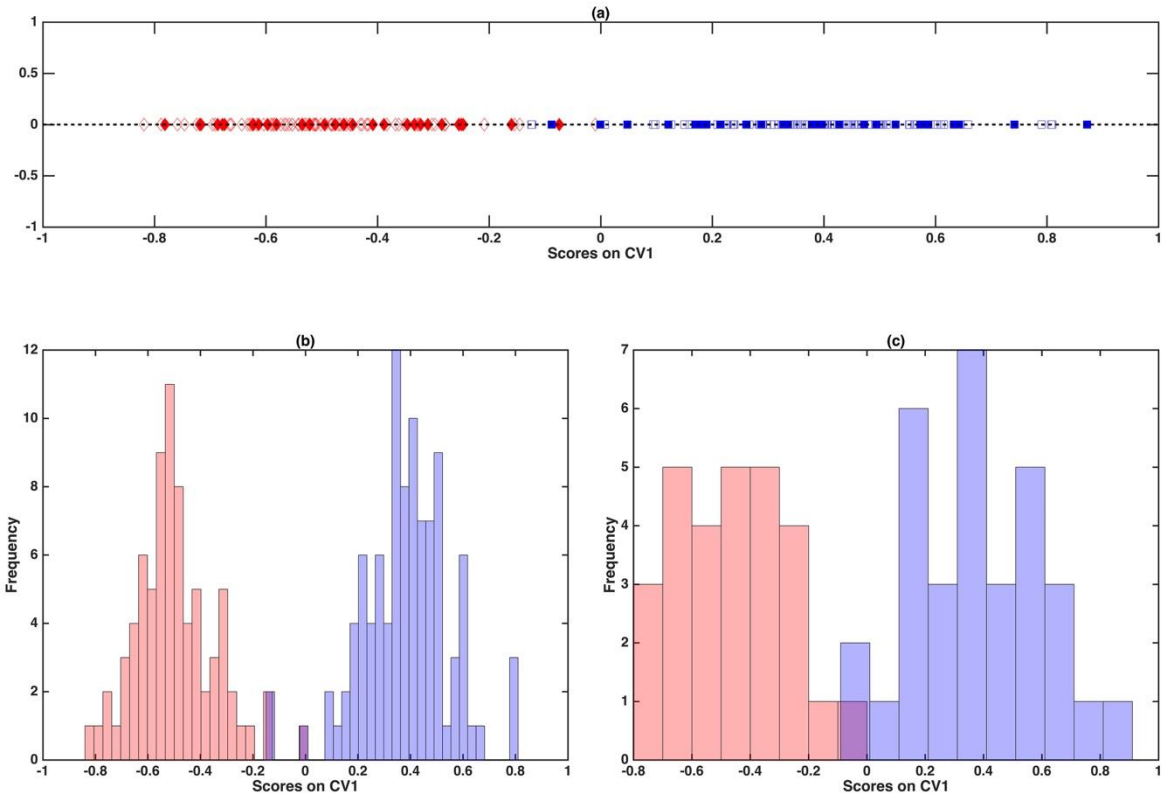

Figure 6. Sequential and Orthogonalized Partial Least Squares Linear Discriminant Analysis (SO-PLS-LDA) analysis: Investigation of the canonical variate. (**a**) Plot of the samples' scores onto the only canonical variate for the training (empty symbols) and test (filled symbols) sets. Legend: Sorrento—red diamonds, Non-Sorrento—blue squares; (**b**) distribution of scores for the calibration model; (**c**) distribution of scores for the validation model.

3.3.2. SO-CovSel-LDA Analysis

Similarly to the procedure described in Section 3.3.1, the SO-CovSel-LDA model was built using both predictors blocks preprocessed by 1st derivative. The optimal number of selected variables (defined on the basis of cross-validation) was 1 for X_1 and 20 for X_2. The cross-validated calibration model provided sensitivities of 96.1% and 96.0% for Class Sorrento and Class Non-Sorrento, respectively. The application of this model to the test samples led to the correct classification of 58 over 60 validation objects. This outcome is good, comparable to those obtained by PLS-DA, but less satisfying than SO-PLS-LDA analysis.

4. Discussion

All the discussed models served as suitable tools for the discrimination of Sorrento walnuts from all the other inspected samples. This outcome was not that obvious, because, among the investigated walnuts, there are fruits produced on the Italian territory, not necessarily far from Sorrento and/or grown in particularly different soils and climatic conditions. Ideally, the best solution for the problem under consideration would be to allow discrimination by using the spectra collected on the nutshell, because this means avoiding any loss of product, with the consequence of having a lower economic

impact. The results described in Section 3.1 demonstrate that this is actually possible, with a relatively low total classification error (5%, corresponding to 3 over 60 misclassified test samples). It has to be noticed that, among the misclassified samples, two belong to Class Sorrento and only one to Class Non-Sorrento, indicating the possibility of false positive (i.e., Non-Sorrento walnuts predicted as Sorrento) is definitely reasonable (1 sample out of 32, corresponding to ~3% of error).

Despite the results obtained on the individual analysis were satisfying, the application of the multi-block strategies, and, in particular, of SO-PLS-LDA, provided an improvement from the prediction point of view. Consequently, whether the aim is to maximize the efficiency of the analysis, even considering the possibility of losing part of the product (which could anyhow be sold as shelled walnuts) SO-PLS-LDA represents a definitely suitable solution, with a rather low total error rate of ~1% in prediction.

5. Conclusions

Two-hundred-and-thirty-seven walnuts have been investigated by NIR spectroscopy coupled with chemometrics in order to understand whether it is possible to discriminate fruits harvested in the Sorrento area from other walnuts. NIR spectra were collected on the whole fruit (i.e., with shell) and on the kernels, and then classified by PLS-DA. As auspicated, the PLS-DA model built on data collected on the shells provided satisfying results (3% of total classification error rate in external validation), indicating the proposed strategy is a suitable solution to discriminate Sorrento samples avoiding any loss of product (walnuts can be sold as they are after NIR analysis on the shell). Nevertheless, whether a more accurate solution is required, the multi-block strategy represents the ideal approach. In fact, SO-PLS-LDA led to a total classification rate of 1% in external validation.

Author Contributions: Conceptualization, A.B. and F.M.; methodology, A.B. and F.M.; software, A.B. and F.M.; validation, A.B. and F.M.; formal analysis, L.A. and P.F.; investigation, A.B. and L.A.; resources, R.B. and F.M.; data curation, R.B. and L.A.; writing—original draft preparation, L.A. and A.B.; writing—review and editing, A.B. and F.M.; visualization, A.B. and F.M.; supervision, A.B. and F.M.; project administration, F.M. and R.B.; funding acquisition, R.B. and F.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

In Figure A1, the graphical representation of VIP analysis on spectra collected on the kernels is displayed.

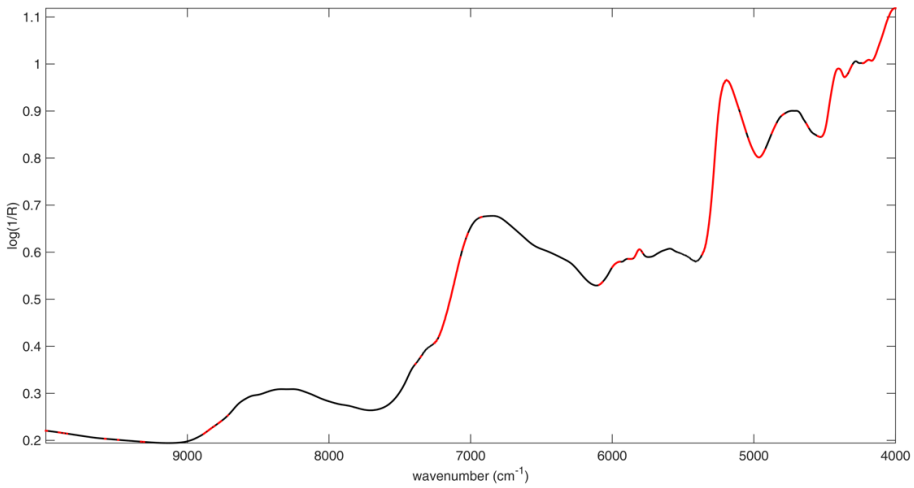

Figure A1. VIP Analysis: Mean spectrum (black line). Variables presenting a VIP index > 1 are highlighted in red.

In Figure A2, the graphical representation of VIP analysis on the SO-PLS-LDA model is displayed.

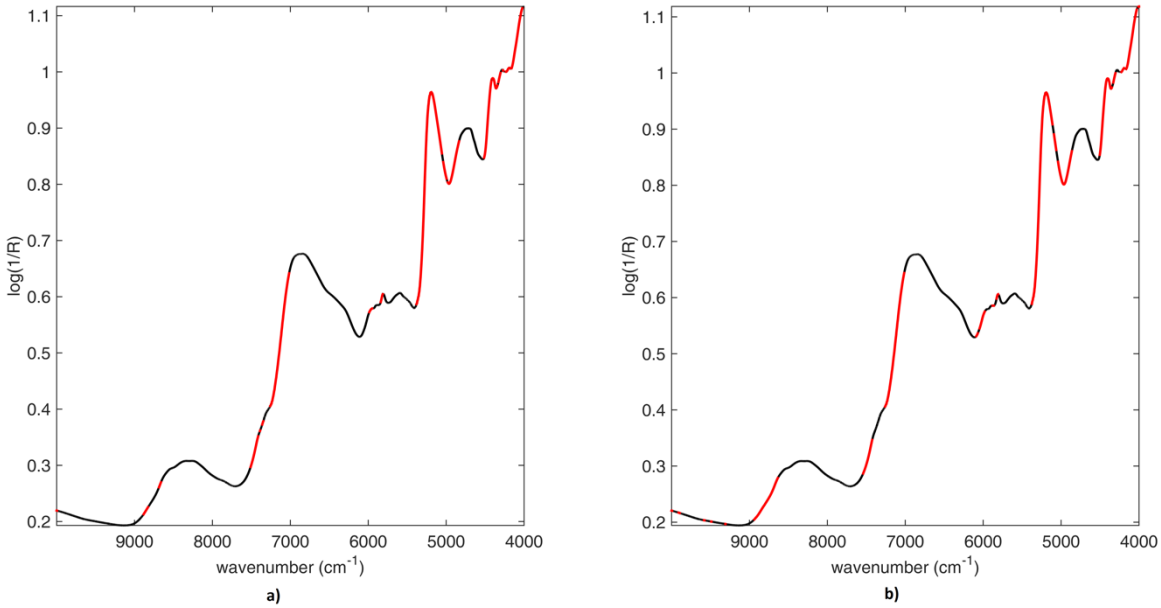

Figure A2. VIP Analysis: Mean spectrum (black line). Variables presenting a VIP index > 1 are highlighted in red. (**a**) Variables selected on the spectra collected on the nutshells. (**b**) Variables selected on the spectra collected on the kernels.

References

1. Hayes, D.; Angove, M.J.; Tucci, J.; Dennis, C. Walnuts (*Juglans regia*) Chemical Composition and Research in Human Health. *Crit. Rev. Food Sci. Nutr.* **2016**, *56*, 1231–1241. [CrossRef] [PubMed]
2. Foroni, I.; Woeste, K.; Monti, L.M.; Rao, R. Identification of *Sorrento* walnut using simple sequence repeats (SSRs). *Genet. Resour. Crop Evol.* **2007**, *54*, 1081–1094. [CrossRef]
3. Regueiro, J.; Sánchez-González, C.; Vallverdú-Queralt, A.; Simal-Gándara, J.; Lamuela-Raventós, R.; Izquierdo-Pulido, M. Comprehensive identification of walnut polyphenols by liquid chromatography coupled to linear ion trap–Orbitrap mass spectrometry. *Food Chem.* **2014**, *152*, 340–348. [CrossRef] [PubMed]
4. Figueroa, F.; Marhuenda, J.; Zafrilla, P.; Villaño, D.; Martínez-Cachá, A.; Tejada, L.; Mulero, J. High-performance liquid chromatography-diode array detector determination and availability of phenolic compounds in 10 genotypes of walnuts. *Int. J. Food Prop.* **2017**, *20*, 1074–1084. [CrossRef]
5. Grace, M.H.; Warlick, C.W.; Neff, S.A.; Lila, M.A. Efficient preparative isolation and identification of walnut bioactive components using high-speed counter-current chromatography and LC-ESI-IT-TOF-MS. *Food Chem.* **2014**, *158*, 229–238. [CrossRef] [PubMed]
6. Verardo, V.; Bendini, A.; Cerretani, L.; Malaguti, D.; Cozzolino, E.; Caboni, M.F. Capillary gas chromatography analysis of lipid composition and evaluation of phenolic compounds by micellar electrokinetic chromathography in Italian walnut (*Juglans regia* L.): Irrigation and fertilization influence. *J. Food Qual.* **2009**, *32*, 262–281. [CrossRef]
7. Fan, F.; Li, H.; Xu, Y.; Liu, Y.; Zheng, Z.; Kan, H. Thermal behaviour of walnut shells by thermogravimetry with gas chromatography–mass spectrometry analysis. *R. Soc. Open Sci.* **2018**, *5*, 180331. [CrossRef]
8. Juranović Cindrić, I.; Zeiner, M.; Hlebec, D. Mineral Composition of Elements in Walnuts and Walnut Oils. *Int. J. Environ. Res. Public Health* **2018**, *15*, 2674. [CrossRef]
9. Foroni, I.; Rao, R.; Woeste, K.; Gallitelli, M. Characterisation of *Juglans regia* L. with SSR markers and evaluation of genetic relationships among cultivars and the *Sorrento* landrace. *J. Hortic. Sci. Biotech.* **2005**, *80*, 49–53. [CrossRef]
10. Ercisli, S.; Sayinci, B.; Kara, M.; Yildiz, C.; Ozturk, I. Determination of size and shape features of walnut (*Juglans regia* L.) cultivars using image processing. *Sci. Hortic.* **2012**, *133*, 47–55. [CrossRef]
11. Sinesio, F.; Moneta, E. Sensory evaluation of walnut fruit. *Food Qual. Prefer.* **1997**, *8*, 35–43. [CrossRef]

12. Sampaio, P.S.; Castanho, A.; Almeida, A.S.; Oliveira, J.; Brites, C. Identification of rice flour types with near-infrared spectroscopy associated with PLS-DA and SVM methods. *Eur. Food Res. Technol.* **2020**, *246*, 527–537. [CrossRef]
13. Biancolillo, A.; Firmani, P.; Bucci, R.; Magrì, A.; Marini, F. Determination of insect infestation on stored rice by near infrared (NIR) spectroscopy. *Microchem. J.* **2019**, *145*, 252–258. [CrossRef]
14. Huang, J.; Ren, G.; Sun, Y.; Jin, S.; Li, L.; Wang, Y.; Ning, J.; Zhang, Z. Qualitative discrimination of Chinese dianhong black tea grades based on a handheld spectroscopy system coupled with chemometrics. *Food Sci. Nutr.* **2020**, *8*, 1–10. [CrossRef] [PubMed]
15. Firmani, P.; De Luca, S.; Bucci, R.; Marini, F.; Biancolillo, A. Near infrared (NIR) spectroscopy-based classification for the authentication of Darjeeling black tea. *Food Control* **2019**, *100*, 292–299. [CrossRef]
16. Carvalho, L.C.; Morais, C.L.M.; Lima, K.M.G.; Leite, G.W.P.; Oliveira, G.S.; Casagrande, I.P.; Santos Neto, J.P.; Teixeira, G.H.A. Using Intact Nuts and Near Infrared Spectroscopy to Classify Macadamia Cultivars. *Food Anal. Meth.* **2018**, *11*, 1857–1866. [CrossRef]
17. Biancolillo, A.; De Luca, S.; Bassi, S.; Roudier, L.; Bucci, R.; Magrì, A.D.; Marini, F. Authentication of an Italian PDO hazelnut ("Nocciola Romana") by NIR spectroscopy. *Environ. Sci. Pollut. Res.* **2018**, *25*, 28780–28786. [CrossRef]
18. Borraz-Martínez, S.; Boqué, R.; Simó, J.; Mestre, M.; Gras, A. Development of a methodology to analyze leaves from Prunus dulcis varieties using near infrared spectroscopy. *Talanta* **2019**, *204*, 320–328. [CrossRef]
19. Firmani, P.; Bucci, R.; Marini, F.; Biancolillo, A. Authentication of *Avola almonds* by near infrared (NIR) spectroscopy and chemometrics. *J. Food Compos. Anal.* **2019**, *82*, 103235. [CrossRef]
20. Biancolillo, A.; Marini, F. Chemometrics Applied to Plant Spectral Analysis. In *Vibrational Spectroscopy for Plant Varieties and Cultivars Characterization, Comprehensive Analytical Chemistry*; Lopes, J., Sousa, C., Eds.; Elsevier: Amsterdam, The Netherlands, 2018; pp. 69–104.
21. Nogales-Bueno, J.; Feliz, L.; Baca-Bocanegra, B.; Hernández-Hierro, J.M.; Heredia, F.J.; Barroso, J.M.; Rato, A.E. Comparative study on the use of three different near infrared spectroscopy recording methodologies for varietal discrimination of walnuts. *Talanta* **2020**, *206*, 120189. [CrossRef]
22. Lee, K.-M.; Jeon, J.-Y.; Lee, B.-J.; Lee, H.; Choi, H.-K. Application of metabolomics to quality control of natural product derived medicines. *Biomol. Ther.* **2017**, *25*, 559–568. [CrossRef]
23. Christensen, K.; Liland, K.H.; Kvaal, K.; Risvik, E.; Biancolillo, A.; Scholderer, J.; Nørskov, S.; Næs, T. Mining online community data: The nature of ideas in online communities. *Food Qual. Prefer.* **2017**, *62*, 246–256. [CrossRef]
24. dos Santos, A.; Oumar, Z.; Arnhold, A.; da Silva, N.; Oliveira Silva, C.; Zanetti, R. Multispectral characterization, prediction and mapping of *Thaumastocoris peregrinus* (Hemiptera: Thamascoridae) attack in Eucalyptus plantations using remote sensing. *J. Spat. Sci.* **2017**, *62*, 127–137.
25. Marzetti, E.; Picca, A.; Marini, F.; Biancolillo, A.; Coelho-Junior, H.J.; Gervasoni, J.; Bossola, M.; Cesari, M.; Onder, G.; Landi, F.; et al. Inflammatory signatures in older persons with physical frailty and sarcopenia: The frailty "cytokinome" at its core. *Exp. Gerontol.* **2019**, *122*, 129–138. [CrossRef]
26. Di Donato, F.; Di Cecco, V.; Torricelli, R.; D'Archivio, A.A.; Di Santo, M.; Albertini, E.; Veronesi, F.; Garramone, R.; Aversano, R.; Marcantonio, G.; et al. Discrimination of Potato (*Solanumtuberosum* L.) Accessions Collected in Majella National Park (Abruzzo, Italy) Using Mid-Infrared Spectroscopy and Chemometrics Combined with Morphological and Molecular Analysis. *Appl. Sci.* **2020**, *10*, 1630. [CrossRef]
27. De Luca, S.; Ciotoli, E.; Biancolillo, A.; Bucci, R.; Magrì, A.D.; Marini, F. Simultaneous quantification of caffeine and chlorogenic acid in coffee green beans and varietal classification of the samples by HPLC-DAD coupled with chemometrics. *Environ. Sci. Pollut. Res.* **2018**, *25*, 28748–28759. [CrossRef]
28. Campmajó, G.; Saez-Vigo, R.; Saurina, J.; Núñez, O. High-performance liquid chromatography with fluorescence detection fingerprinting combined with chemometrics for nut classification and the detection and quantitation of almond-based product adulterations. *Food Control* **2020**, *114*, 107265. [CrossRef]
29. Medina, S.; Perestrelo, R.; Silva, P.; Pereira, J.A.M.; Câmara, J.S. Current trends and recent advances on food authenticity technologies and chemometric approaches. *Trends Food Sci. Technol.* **2019**, *85*, 163–176. [CrossRef]
30. Biancolillo, A.; Måge, I.; Næs, T. Combining SO-PLS and linear discriminant analysis for multi-block classification. *Chemom. Intell. Lab.* **2015**, *141*, 58–67. [CrossRef]

31. Biancolillo, A.; Marini, F.; Roger, J.-M. SO-CovSel: A novel method for variable selection in a multiblock framework. *J. Chemom.* **2020**, *34*, e3120. [CrossRef]
32. Picca, A.; Ponziani, F.R.; Calvani, R.; Marini, F.; Biancolillo, A.; Coelho-Junior, H.J.; Gervasoni, J.; Primiano, A.; Putignani, L.; Del Chierico, F.; et al. Gut microbial, inflammatory and metabolic signatures in older people with physical frailty and sarcopenia: Results from the BIOSPHERE study. *Nutrients* **2020**, *12*, 65. [CrossRef] [PubMed]
33. Biancolillo, A.; Marini, F.; D'Archivio, A.A. 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. *J. Food Compos. Anal.* **2020**, *86*, 103351. [CrossRef]
34. Schiavone, S.; Marchionni, B.; Bucci, R.; Marini, F.; Biancolillo, A. Authentication of Grappa (Italian grape marc spirit) by Mid and Near Infrared spectroscopies coupled with chemometrics. *Vib. Spectrosc.* **2020**, *107*, 103040. [CrossRef]
35. Biancolillo, A.; Boqué, R.; Cocchi, M.; Marini, F. Data Fusion strategies in food analysis. In *Data Handling in Science and Technology*; Cocchi, M., Ed.; Elsevier: Amsterdam, The Netherlands, 2019; Volume 31, pp. 271–310.
36. Snee, R.D. Validation of regression models: Methods and examples. *Technometrics* **1977**, *19*, 415–428. [CrossRef]
37. Firmani, P.; Nardecchia, A.; Nocente, F.; Gazza, L.; Marini, F.; Biancolillo, A. Multi-block classification of Italian semolina based on Near Infrared Spectroscopy (NIR) analysis and alveographic indices. *Food Chem.* **2020**, *309*, 125677. [CrossRef]
38. Ståhle, L.; Wold, S. Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. *J. Chemom.* **1987**, *1*, 185–196. [CrossRef]
39. Sjöström, M.; Wold, S.; Söderström, B. PLS discriminate plots. In *Pattern Recognition in Practice*; Kanal, E.S.H.N., Ed.; Elsevier: Amsterdam, The Netherlands, 1986; pp. 461–470.
40. Biancolillo, A.; Marini, F. Chemometric methods for spectroscopy-based pharmaceutical analysis. *Front. Chem.* **2018**, *6*, 576. [CrossRef]
41. Barker, M.; Rayens, W. Partial least squares for discrimination. *J. Chemom.* **2003**, *17*, 166–173. [CrossRef]
42. Martens, H.; Naes, T. *Multivariate Calibration*; John Wiley & Sons: New York, NY, USA, 1991.
43. Geladi, P.; Kowalski, P.B. Partial least squares regression: A tutorial. *Anal. Chim. Acta* **1986**, *185*, 1–17. [CrossRef]
44. Kemsley, E.K. Discriminant analysis of high-dimensional data: A comparison of principal components analysis and partial least squares data reduction methods. *Chemometr. Intell. Lab.* **1996**, *33*, 47–61. [CrossRef]
45. Nocairi, H.; Qannari, E.M.; Vigneau, E.; Bertrand, D. Discrimination on latent components with respect to patterns. Application to multicollinear data. *Comput. Stat. Data Anal.* **2005**, *48*, 139–147. [CrossRef]
46. Indahl, U.G.; Martens, H.; Næs, T. From dummy regression to prior probabilities in PLS-DA. *J. Chemometr.* **2007**, *21*, 529–536. [CrossRef]
47. Pérez, N.F.; Ferré, J.; Boqué, R. Calculation of the reliability of classification in discriminant partial least-squares binary classification. *Chemometr. Intell. Lab.* **2009**, *95*, 122–128. [CrossRef]
48. Næs, T.; Tomic, O.; Mevik, B.H.; Martens, H. Path modelling by sequential PLS regression. *J. Chemometr.* **2011**, *25*, 28–40. [CrossRef]
49. Biancolillo, A.; Næs, T. The sequential and orthogonalised PLS regression (SO-PLS) for multi-block regression: Theory, examples and extensions. In *Data Handling in Science and Technology*; Cocchi, M., Ed.; Elsevier: Amsterdam, The Netherlands, 2019; Volume 31, pp. 157–177.
50. Matlab Functions for SO-PLS and SO-PLS-LDA. Available online: https://www.chem.uniroma1.it/romechemometrics/research/algorithms/so-pls/ (accessed on 9 June 2020).
51. Roger, J.M.; Palagos, B.; Bertrand, D.; Fernandez-Ahumada, E. CovSel: Variable selection for highly multivariate and multi-response calibration application to IR spectroscopy. *Chemom Intel. Lab. Syst.* **2011**, *106*, 216–223. [CrossRef]
52. Matlab Functions for SO-CovSel and SO-CovSel-LDA. Available online: https://www.chem.uniroma1.it/romechemometrics/research/algorithms/so-covsel/ (accessed on 9 June 2020).
53. Savitzky, A.; Golay, M.J.E. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **1964**, *36*, 1627–1639. [CrossRef]
54. Barnes, R.J.; Dhanoa, M.S.; Lister, S.J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectr.* **1989**, *43*, 772–777. [CrossRef]

55. Wold, S.; Johansson, E.; Cocchi, M. PLS: Partial least squares projections to latent structures. In *3D QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Ed.; Kluwer Escom Science Publisher: Leiden, The Netherlands, 1993; pp. 523–550.
56. Grabska, J.; Bec, K.B.; Ishigaki, M.; Huck, C.W.; Ozak, Y. NIR Spectra Simulations by Anharmonic DFT-Saturated and Unsaturated Long-Chain Fatty Acids. *J. Phys. Chem. B* **2018**, *122*, 6931–6944. [CrossRef]
57. Sun, D.-W. (Ed.) *Infrared Spectroscopy for Food Quality Analysis and Control*, 1st ed.; Elsevier: Amsterdam, The Netherlands; Academic Press: Cambridge, MA, USA, 2009.
58. Biancolillo, A.; Liland, K.H.; Måge, I.; Næs, T.; Bro, R. Variable selection in multi-block regression. *Chemometr. Intell. Lab.* **2016**, *156*, 89–101. [CrossRef]