applied sciences

MDPI

Review

Chemometric Strategies for Spectroscopy-Based Food Authentication

Alessandra Biancolillo [1], Federico Marini [2,3], Cyril Ruckebusch [4] and Raffaele Vitale [4,*]

1 Dipartimento di Scienze Fisiche e Chimiche, Università degli Studi dell'Aquila, Via Vetoio (Coppito 2, Edificio "Angelo Camillo De Meis"), 67100 Coppito (AQ), Italy; alessandra.biancolillo@univaq.it
2 Dipartimento di Chimica, Università degli Studi di Roma "La Sapienza", Piazzale Aldo Moro 5, 00185 Roma, Italy; federico.marini@uniroma1.it
3 Department of Food Science, Stellenbosch University, Private Bag X1, 7602 Matieland (Stellenbosch), South Africa
4 U. Lille, CNRS, LASIRE, Laboratoire de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, Cité Scientifique, F-59000 Lille, France; cyril.ruckebusch@univ-lille.fr
* Correspondence: raffaele.vitale@univ-lille.fr; Tel.: +33-769-47-66-54

Received: 10 August 2020; Accepted: 15 September 2020; Published: 18 September 2020

Featured Application: This review will offer a global overview of the chemometric approaches most commonly used in the field of spectroscopy-based food analysis and authentication. Three different scenarios will be surveyed: data exploration, calibration and classification. Basic and simple descriptions of the main multivariate techniques exploited in such a domain along with a comprehensive outline of their most recent and interesting applications will be provided.

Abstract: In the last decades, spectroscopic techniques have played an increasingly crucial role in analytical chemistry, due to the numerous advantages they offer. Several of these techniques (e.g., Near-InfraRed—NIR—or Fourier Transform InfraRed—FT-IR—spectroscopy) are considered particularly valuable because, by means of suitable equipment, they enable a fast and non-destructive sample characterization. This aspect, together with the possibility of easily developing devices for on- and in-line applications, has recently favored the diffusion of such approaches especially in the context of foodstuff quality control. Nevertheless, the complex nature of the signal yielded by spectroscopy instrumentation (regardless of the spectral range investigated) inevitably calls for the use of multivariate chemometric strategies for its accurate assessment and interpretation. This review aims at providing a comprehensive overview of some of the chemometric tools most commonly exploited for spectroscopy-based foodstuff analysis and authentication. More in detail, three different scenarios will be surveyed here: data exploration, calibration and classification. The main methodologies suited to addressing each one of these different tasks will be outlined and examples illustrating their use will be provided alongside their description.

Keywords: spectroscopy; food authentication; chemometrics; data exploration; calibration; classification; data fusion; curve resolution; analysis of multivariate designed data

1. Introduction

In recent years, consumers' attention towards the quality of foodstuff has become increasingly lively. The awareness that there exists a close link between health and diet has spread dramatically, leading a growing number of people to develop conscious eating habits. For this reason, and thanks to the so-called *food revolution*, aliments (especially high value-added food products) are commonly subjected to strict quality controls. These tests are of paramount importance, especially for attesting to

some peculiar features (connected, for instance, to their geographical origin, their specific manufacturing process, and/or the know-how of their producers) that might translate into the recognition of distinctive labelling designations like the Protected Designation of Origin (PDO) or the Protected Geographical Indication (PGI). In light of this, a plethora of analytical methodologies aimed at foodstuff authentication and traceability have recently been developed, and a wide variety of applications of such methodologies have been reported in the scientific literature. More specifically, given the notable market value of this type of product, much effort has been put into the design of non-destructive characterization tools for quality control, generally based on the principles of light–matter interaction and spectroscopy. However, unfortunately, the signal profiles yielded by spectroscopy instrumentation (regardless of the spectral range investigated) are often particularly complex and thus, their assessment and interpretation are usually not straightforward. In this context, *multivariate statistical* approaches (also known as *chemometric* approaches) have lately played—and currently still play—a crucial role: they have proven, in fact, to be extremely powerful when large amounts of spectral data (characterized by a considerable degree of intercorrelation among the recorded spectral variables) need to be coped with, and when useful and meaningful information is to be extracted from these data for disparate purposes.

Given, therefore, the sheer relevance of such approaches in the field of spectroscopy-based foodstuff analysis and authentication, this review aims at providing a comprehensive overview of those most commonly exploited in real-world case studies. Specifically, three scenarios will be here surveyed: exploration, calibration (necessary to carry out quantification) and classification (for, e.g., adulteration and/or fraud detection). The first one, intended to pursue direct insights into the data with the possibility of revealing hidden/underlying structures and relationships between samples and/or variables, and making extensive use of plots and graphs to highlight similarities, differences, trends, clusters and/or correlations, relies on so-called *unsupervised* techniques, which do not require any input other than the (spectroscopic) data themselves [1]. On the other hand, calibration and classification problems call for the use of predictive models that exploit the spectroscopic information to predict one or more properties of the objects under study; in order to be reliable, such models have to be constructed by *supervised* approaches, i.e., techniques which actively take advantage not only of the experimental (spectroscopic) data but also of, e.g., reference values of the aforementioned properties or sample labelling [2]. The main chemometric methodologies suited to addressing each one of these different tasks will be outlined, and examples illustrating their use will be provided alongside their description.

2. Data Exploration

Prior to conducting a proper food authentication study, users might be interested in carrying out preliminary investigations to assess, for instance, whether a particular spectroscopic technique is capable of discerning products of distinct geographical origins or whether an instrumental response is sensitive enough to detect the presence of compounds or adulterants of interest in the specific samples at hand. In this regard, chemometric exploratory tools based on the reduction of the dimensionality of the original data collected are of crucial importance, as they dramatically ease the visualization of these data, allowing possible differences among specimens (related to one or multiple of the aforementioned aspects, like the presence of undesired/unexpected substances) to be spotted in a straightforward way. Given, in fact, the complex multidimensional and multivariate nature of spectroscopic measurements, such differences are very often undiscernible when simply plotting raw spectral profiles since they normally lie not in few scattered regions of the concerned wavelength domain, but in the correlation structure intrinsic to the recorded spectral variables or descriptors (this is the case especially for low-selectivity approaches as Near-InfraRed—NIR—spectroscopy). For this reason, one of the most commonly exploited tools in similar scenarios to facilitate and improve the identification of diverse groups of food samples is Principal Component Analysis (PCA) [3,4]. PCA makes it possible to compress, describe and interpret large sets of multidimensional data. Its basic principle can be summarized as follows: let $\mathbf{X}$ be a centered $N \times J$ matrix with J denoting the number of variables

(e.g., J wavelengths of light scanned in a spectroscopy experiment) registered, for instance, at N time instants or for N different objects. As specified, when J is very large, the useful and meaningful information in $\mathbf{X}$ is usually intercorrelated among various of these variables over the whole set of recordings. Then, for a chosen degree of acceptable accuracy, it is possible to reduce the J-dimensional space of the original descriptors to an A_{PCA}-dimensional subspace in which data mostly vary and onto which all the N samples under study can be projected and represented as new points. Mathematically speaking, PCA is based on the bilinear structure model in Equation (1):

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}_{\mathrm{PCA}} \tag{1}$$

where $\mathbf{P}$ ($J \times A_{\mathrm{PCA}}$) is an array of so-called *loadings*, which determine the A_{PCA} basis vectors (*principal components* or *factors*) of the PCA subspace, $\mathbf{T}$ ($N \times A_{\mathrm{PCA}}$) defines the projection coordinates or *scores* of all the N rows of $\mathbf{X}$ on this lower-dimensional space and $\mathbf{E}_{\mathrm{PCA}}$ ($N \times J$) denotes the matrix of unmodelled residuals, i.e., the portion of $\mathbf{X}$ not *explained* at the chosen rank, A_{PCA}—in PCA, the optimal number of principal components to extract can be estimated by a wide range of approaches that, broadly speaking, can be classified into three distinct categories [5]: ad hoc rules (like Kaiser's eigenvalue-greater-than-1 rule [6], Velicer's minimum average partial rule [7], and Cattell's scree test [8]), statistical tests (like Bartlett's Chi-square test [9] and Tracy-Widom's statistics-based test [10]) and computational criteria (like cross-validation [11–13] and permutation testing [14–21].

The PCA solution may be formulated in many equivalent ways and attained by a variety of algorithms, among which the most widespread and popular is certainly Singular Value Decomposition (SVD [22]). SVD decomposes $\mathbf{X}$ as:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}} + \mathbf{E} \tag{2}$$

with the columns of $\mathbf{U}$ ($N \times A_{\mathrm{PCA}}$) and $\mathbf{V}$ ($J \times A_{\mathrm{PCA}}$) being the first A_{PCA} left and right singular vectors of $\mathbf{X}$, respectively, and $\mathbf{D}$ ($A_{\mathrm{PCA}} \times A_{\mathrm{PCA}}$) a square diagonal array whose diagonal elements correspond to its first A_{PCA} non-zero singular values. Therefore, it holds that $\mathbf{T} = \mathbf{U}\mathbf{D}$ and $\mathbf{P} = \mathbf{V}$.

PCA shows the following property:

$$\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I} \tag{3}$$

where $\mathbf{I}$ is an identity matrix of dimensions $A_{\mathrm{PCA}} \times A_{\mathrm{PCA}}$, which translates into the fact that the dimensions of its subspace are orthogonal and, thus, can be inspected assuming that the information they individually capture is mutually unrelated.

The assessment and interpretation of a PCA model is generally performed by examining both scores and loadings profiles (i.e., the columns of $\mathbf{T}$ and $\mathbf{P}$, respectively). The former yield insights into existing relations among analyzed objects, while the latter enable the identification of correlation patterns within ensembles of multiple spectral variables. The basic idea, here, is to simultaneously analyze these profiles not only to possibly pinpoint the presence of distinctive clusters of samples, but also to recognize sets of spectral features shared within each one of these clusters and responsible for their differentiation. Nevertheless, as a consequence of the orthogonality imposed to the principal components, PCA scores and loadings seldom provide physico-chemically meaningful information, since they do not capture single phenomena or events affecting the registered measurements, but rather heterogeneous effects resulting from combinations of such phenomena or events. To overcome this limitation, alternative approaches like Independent Component Analysis (ICA [23,24]) can be resorted to. Unlike PCA, ICA regards the rows of the data matrix $\mathbf{X}$ as a collection of *observed signals* that are mixtures of a certain number of common unknown *source signals* or *independent components* and decomposes it as:

$$\mathbf{X} = \mathbf{A}\mathbf{R}^{\mathrm{T}} + \mathbf{E}_{\mathrm{ICA}} \tag{4}$$

with $\mathbf{A}$ ($N \times A_{\mathrm{ICA}}$) being the so-called *proportion* matrix, carrying the mixing coefficients associated with these independent components contained in every column of $\mathbf{R}$ ($J \times A_{\mathrm{ICA}}$), and $\mathbf{E}_{\mathrm{ICA}}$ ($N \times J$) denoting the ICA residual array.

As for PCA, the interpretation of an ICA model rests on the simultaneous inspection of the retrieved independent components and their respective proportion profiles. Independent components are analogous to PCA loadings, but, differently from them, are not necessarily orthogonal, and therefore, can be easily related to individual phenomena impacting the recorded spectral variables (e.g., Rayleigh or Raman scattering) [25–28]. Subsequently, ICA proportions can give a quantitative idea of how they influence the raw measured spectra.

There exist several algorithms by which the ICA solution can be attained. Examples are Infomax [29], FastICA [30] and JADE (Joint Approximate Diagonalization of Eigenmatrices) [26,31]. Their basic principle, though, is common: source signals are estimated by maximizing their statistical independence. In essence, they can be looked at as the outcome of a rotation of PCA loadings that enhances their aforementioned interpretability properties.

PCA and ICA have been both extensively used in the field of food characterization (see Table 1); by way of illustration, PCA has been coupled to (i) high-resolution ^{1}H Nuclear Magnetic Resonance (NMR) spectroscopy for the discrimination of three apple varieties (Bramley, Russet and Spartan), making it possible to spot significant variations in their malic acid and sucrose content [32]; (ii) synchronous fluorescence spectroscopy for monitoring changes in the flavin composition of beer during storage either under light exposure or in darkness [33]; and (iii) Attenuated Total Reflectance Mid-InfraRed (ATR-MIR) spectroscopy and diffuse reflectance NIR spectroscopy in a feasibility study on the detection of soybean oil adulteration in Camellia oil samples [34]. Conversely, due to its particular methodological basis, ICA has mainly been applied to extract the individual spectral contributions and proportion profiles of certain classes of analytes constituting heterogeneous food matrices, e.g., primary oxidation and polyphenolic compounds, tocopherol, carotenoids and chlorophylls in extra virgin olive oils [35], or fructose, sucrose and glucose in soft drinks [36].

Sometimes, in the context of unsupervised data analysis, one may be interested in having information about the different degrees of similarity/dissimilarity among samples or, more specifically, on whether the distribution of individuals enables the identification of groups or clusters. Clusters correspond to sets of observations which are more similar to one another than they are with respect to all the remaining objects. In this regard, the search for possible clusters of samples is the objective of the so-called clustering techniques, whose basic idea is that similarity among individuals can be inversely related to a distance measure in a multivariate space [37,38]. Based on this concept, clustering techniques can operate non-hierarchically (partitioning methods) or hierarchically. The former [39], encompassing also the popular *k*-means algorithm [40], distribute the samples among a pre-defined number of groups (that can be either known a priori or inferred empirically from the data under study); a sample is basically placed in the cluster with the closest barycenter. They are relatively straightforward, but they suffer from the limitation of needing the number of groups, *k*, to be defined beforehand.

Conversely, as the name suggests, hierarchical approaches proceed iteratively through successive agglomerations of smaller clusters into larger ones [41]—the same approach can be operated top down by progressively dividing larger clusters into smaller ones, but it is more rarely used. Every object is initially assumed to constitute a separate group of observations (*singleton*) and, afterwards, at each iteration, the two most similar groups are joined into a single ensemble. The procedure continues until a certain stopping criterion is met or when all the samples are gathered into a unique cluster. Hierarchical approaches provide more detailed insights into the relations between individuals and/or groups of individuals, but they have the disadvantage that, due to the nature of the agglomeration process, the identification of the most plausible number of clusters may not be easy. The interested reader can find more details on clustering techniques in [37,38,42].

3. Calibration

The analytical characterization of a foodstuff, especially with the aim of its authentication, may imply the quantitative determination of one or more constituents of the samples at hand or of one or more of their global properties, such as the sensory scores for some specific attributes, the iodine

value or the dry matter content. Moreover, one would ideally want such a determination to be rapid (in some cases, almost instantaneous, to allow controls to be conducted on- or in-line), non-invasive or at least non-destructive, not needing the specimens to be subjected to clean-ups or pre-treatments, and, possibly, not encompassing the use of auxiliary reagents or organic solvents. All these requirements point in the direction of spectroscopic techniques, which, depending on the applications concerned, possess most or all of the aforementioned characteristics. However, the possibility of using spectroscopy to quantify the properties of a food product is conditional, on the one hand, to the existence of a relation between the measured signal and the response to be estimated and, on the other hand, to the postulation of a reasonable mathematical formulation to express (or at least approximate) this relation. In other words, quantifying the value of a property, y, based on the measurement of a spectroscopic signal, $\mathbf{x}$, rests on inferring a functional relation, f, that connects them so that:

$$y = f(\mathbf{x}) \tag{5}$$

Unfortunately, the exact form of the function $f(\mathbf{x})$ linking a measured spectrum to the property to be determined is unknown and cannot be retrieved from first principles. A reasonable guess of $f(\mathbf{x})$ should be therefore empirically drawn, through a procedure which is called calibration [43–45].

Calibration makes use of known data to approximate the functional relation in Equation (5). In particular, it requires that, for a sufficient number of samples (which are usually gathered in a so-called *training set*), both the recorded spectra and the values of the property(-ies) of interest are available. For example, if NIR spectroscopy is used for the quantification of the protein content in wheat, in order to proceed with the calibration phase, one should not only collect the NIR spectra of the training samples, but also determine their protein content by, e.g., reference methods such as Kjeldahl [46] or Dumas [47]. Afterwards, a mathematical expression of the function $f(\mathbf{x})$ is explicitly assumed: this function will depend not only on $\mathbf{x}$, but also on the values of some adjustable parameters (*coefficients*), whose estimation is the core issue in calibration. In many cases, $f(\mathbf{x})$ can be assumed to be linear; this means that, if J spectral variables are recorded (i.e., if $\mathbf{x}$ is a J-dimensional vector of light intensity/absorption measurements, with elements $\begin{bmatrix} x_1 & x_2 & , & x_3 & \cdots & x_J \end{bmatrix}$), Equation (5) can be written as:

$$y = \hat{y} + e = b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_J x_J + e \tag{6}$$

where $\hat{y}$ is the approximation of y provided by the linear function defined by the coefficients $b_1, b_2, b_3, \cdots, b_J$, and e is the *residual*, i.e., the difference between the true value of y and $\hat{y}$. In calibration, the available ($\mathbf{x}$, y) pairs constituting the training set of data are used to calculate the coefficients in Equation (6) through regression, according to some user-defined criterion of optimality. Since the aim of calibration is to use the measured signal $\mathbf{x}$ as the basis to obtain the best *prediction* of y, in the majority of cases such a criterion involves the minimization of the residuals (*least squares*). In particular, if the training set contains N samples for which both $\mathbf{x}$ and y are known, for each one of these samples an equation analogous to Equation (6) can be written:

$$y_n = \hat{y}_n + e_n = b_1 x_{n1} + b_2 x_{n2} + b_3 x_{n3} + \cdots + b_J x_{nJ} + e_n \tag{7}$$

where the subscript n indicates that the equality holds for the n-th training object—here, it should be noticed that, since the same functional relation is valid for all the samples, the coefficients $b_1, b_2, b_3, \cdots, b_J$ are identical for all the individuals. By gathering all the training set recordings in the spectral matrix $\mathbf{X}$ (whose rows contain the profiles of all the individuals) and the corresponding true values of the property of interest in the column vector $\mathbf{y}$, Equation (7) can be globally expressed for all the training samples as:

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{8}$$

where the column vector $\mathbf{b}$ carries the regression coefficients ($\mathbf{b} = \begin{bmatrix} b_1 & b_2 & , & b_3 & \cdots & b_J \end{bmatrix}$), the approximated values of the response are the elements of $\hat{\mathbf{y}}$ ($= \mathbf{X}\mathbf{b}$), and the residuals constitute the

vector **e**. As anticipated, the estimation of the optimal value of the coefficients **b** is usually carried out by the so-called least squares approach:

$$\underset{\mathbf{b}}{\operatorname{argmin}} \sum_{n=1}^{N} e_n^2 = \underset{\mathbf{b}}{\operatorname{argmin}}\, \mathbf{e}^{\mathrm{T}}\mathbf{e} = \underset{\mathbf{b}}{\operatorname{argmin}} (\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{b}) \tag{9}$$

where the operator argmin indicates that the optimal set of values is the one minimizing the overall approximation error across all the training objects, i.e., $\sum_{n=1}^{N} e_n^2 = \mathbf{e}^{\mathrm{T}}\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{b})$. Resolving Equation (9) makes it possible to estimate the regression coefficients as:

$$\mathbf{b}_{\mathrm{MLR}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{10}$$

Since the calibration approach, which postulates a linear regression model and uses the least squares criterion for the calculation of these regression coefficients, is also generally referred to as Multiple Linear Regression (MLR [48,49]), the subscript MLR was explicitly added to Equation (10).

Once the values of the coefficients have been estimated using the training samples, it is possible to predict the value of the response for any new (unknown) specimen, $\hat{y}_{\mathrm{new}}$, based on its spectral profile, $\mathbf{x}_{\mathrm{new}}$, as:

$$\hat{y}_{\mathrm{new}} = \mathbf{x}_{\mathrm{new}}\mathbf{b}_{\mathrm{MLR}} \tag{11}$$

Sometimes, the same spectroscopic profile can be used to estimate the values of multiple properties; for instance, one might want to determine, for the same individual, the protein, lipid, starch and moisture content, based on its measured NIR spectrum. Mathematically, this would result in setting up a calibration equation analogous to Equation (8) for each of the L responses to be predicted:

$$\mathbf{y}_l = \hat{\mathbf{y}}_l + \mathbf{e}_l = \mathbf{X}\mathbf{b}_l + \mathbf{e}_l \qquad l = 1, \ldots, L \tag{12}$$

In the case of MLR, it is possible to demonstrate that there is no difference (in terms of estimated values of the regression coefficients $\mathbf{b}_l$ and, subsequently, of the predicted responses $\hat{\mathbf{y}}_l$) between building separate models for all the single properties of interest and constructing a unique global calibration [50] that can be expressed as:

$$\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{E} = \mathbf{X}\mathbf{B} + \mathbf{E} \tag{13}$$

where each of the L columns of the matrices **Y**, $\hat{\mathbf{Y}}$, **B** and **E** contains the values of the corresponding elements in Equation (12) for a particular response:

$$\begin{aligned} \mathbf{Y} &= \begin{bmatrix} \mathbf{y}_1 & \cdots & \mathbf{y}_l & \cdots & \mathbf{y}_L \end{bmatrix} \\ \hat{\mathbf{Y}} &= \begin{bmatrix} \hat{\mathbf{y}}_1 & \cdots & \hat{\mathbf{y}}_l & \cdots & \hat{\mathbf{y}}_L \end{bmatrix} \\ \mathbf{B} &= \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_l & \cdots & \mathbf{b}_L \end{bmatrix} \\ \mathbf{E} &= \begin{bmatrix} \mathbf{e}_1 & \cdots & \mathbf{e}_l & \cdots & \mathbf{e}_L \end{bmatrix} \end{aligned} \tag{14}$$

MLR has the advantages of a clear loss-function, which directly generalizes the least squares criterion to the multivariate case, and of not requiring the tuning of any metaparameter (that is to say, once the **X**-**Y** training pairs are identified, the solution is univocal) [45]. However, due to the mathematical structure of Equation (10), this method cannot be applied in situations where the matrix **X** is ill-conditioned, i.e., when the number of training samples is lower than the number of recorded variables (wavelengths) and/or when such variables are highly correlated [51]. Unfortunately, both these conditions are commonly met by spectroscopic data, and thus, MLR is seldom applied if spectral profiles are concerned, unless some variable selection strategy is adopted [52–55].

On the other hand, a solution to the problem of having to deal with ill-conditioned matrices of descriptors is provided by the use of bilinear models based on the extraction of principal components or *latent variables* [56]. Indeed, as already outlined in Section 2—where the use of principal components to summarize the relevant information encoded in the data was discussed—an effective dimensionality reduction can be achieved by identifying a reduced number of highly informative orthogonal directions in the multivariate space of the original variables and projecting the data onto the subspace spanned by them, obtaining a set of new coordinates or scores, **T**. Using **T** instead of **X** in a calibration framework makes it possible to overcome all the issues related to ill-conditioning, since the number of relevant components is usually lower than the number of samples and because, additionally, they are completely uncorrelated due to their orthogonality. In light of these aspects, various bilinear regression techniques have been proposed in the literature, the most commonly used of which are Principal Component Regression (PCR [48,57,58]) and Partial Least Squares (PLS [59–66]) regression.

PCR, as the name suggests, is a two-step method which is based on the sequential use of PCA and MLR, with the PCA scores, **T**, exploited as predictors:

$$\mathbf{Y} = \hat{\mathbf{Y}}_{\text{PCR}} + \mathbf{E} = \mathbf{T}\boldsymbol{\mathcal{B}} + \mathbf{E} \tag{15}$$

Here, $\boldsymbol{\mathcal{B}}$ is the matrix of regression coefficients relating the scores to the predicted responses through the MLR model. By resorting to the mathematical relationship defining the PCA projection ($\mathbf{T} = \mathbf{XP}$, see also Equation (1)), it is possible to express the linear model in Equation (15) directly as a function of **X** and **Y**, as:

$$\hat{\mathbf{Y}}_{\text{PCR}} = \mathbf{T}\boldsymbol{\mathcal{B}} = \mathbf{XP}\boldsymbol{\mathcal{B}} = \mathbf{XB}_{\text{PCR}} \tag{16}$$

where the matrix of PCR coefficients, $\mathbf{B}_{\text{PCR}}$, is given by:

$$\mathbf{B}_{\text{PCR}} = \mathbf{P}\boldsymbol{\mathcal{B}} \tag{17}$$

With respect to the MLR solution reported in Equation (10), PCR is said to be biased, as not all the information originally present in **X** is used to estimate the parameters of the regression model [67,68]; the calculation of $\mathbf{B}_{\text{PCR}}$, in fact, relies solely on the variability captured by the PCA scores in **T**, and a different model will be returned depending on how many principal components are extracted from the data matrix **X**. Accordingly, the number of principal components to retain in a PCR model is a metaparameter which has to be optimized; normally, models with different numbers of principal components are tested and the one resulting in the lowest prediction error (usually computed by cross-validation or other resampling procedures) is selected [69–71] (see also Section 7.2).

PCR can be considered a highly performant statistical approach as it enables the possibility of calculating regression models for ill-conditioned data matrices. Nevertheless, even in similar contingencies it may not necessarily constitute the best option to choose: indeed, the two steps PCR combines (i.e., data compression by PCA and regression by MLR) have different objectives, and it is not always verified that the directions of maximum explained variance in **X** are also those corresponding to the maximum correlation with the response(s) [72]. Conversely, PLS tries to overcome this issue, by actively using the information in **Y** already at the data compression stage, so that the scores extracted from **X** are relevant for describing at the same time the variance in the descriptors and in the properties of interest.

PLS rests on the extraction of two sets of scores, one from the independent and one from the dependent data block, having maximum covariance with one another:

$$\underset{\mathbf{r}_a, \mathbf{q}_a}{\text{argmax}}\ \mathbf{u}_a^{\text{T}}\mathbf{t}_{a,\text{PLS}} = \underset{\mathbf{r}_a, \mathbf{q}_a}{\text{argmax}}\ \mathbf{q}_a^{\text{T}}\mathbf{Y}^{\text{T}}\mathbf{X}\mathbf{r}_a \tag{18}$$

where $\mathbf{t}_{a,\mathrm{PLS}}$ and $\mathbf{u}_a$ are the **X**- and **Y**-scores along the a-th component, respectively, and $\mathbf{r}_a$ and $\mathbf{q}_a$ represent the vectors of coefficients allowing to retrieve such scores from the corresponding original matrices:

$$\begin{aligned} \mathbf{t}_{a,\mathrm{PLS}} &= \mathbf{X}\mathbf{r}_a \\ \mathbf{u}_a &= \mathbf{Y}\mathbf{q}_a \end{aligned} \tag{19}$$

The functional relation (usually a linear dependence), also known as inner relationship, between the two data blocks is then implemented at the scores level; component by component, the **Y**-scores are approximated by the **X**-scores through a univariate linear regression model:

$$\hat{\mathbf{u}}_a = \mathbf{t}_{a,\mathrm{PLS}} c_a \tag{20}$$

with c_a being a proportionality constant (*inner regression coefficient*). Since more than a single component is usually needed to summarize the relevant information in **X** and **Y**, Equations (19) and (20) can be rewritten as:

$$\begin{aligned} \mathbf{T}_{\mathrm{PLS}} &= \mathbf{X}\mathbf{R} \\ \mathbf{U} = \mathbf{Y}\mathbf{Q} \Longrightarrow \hat{\mathbf{Y}} &= \mathbf{U}\mathbf{Q}^{\mathbf{T}} \\ \hat{\mathbf{U}} &= \mathbf{T}_{\mathbf{PLS}}\mathbf{C} \end{aligned} \tag{21}$$

where the columns of the scores matrices $\mathbf{T}_{\mathrm{PLS}}$ and **U** and those of the coefficient matrices **R** (**X**-weights) and **Q** (**Y**-loadings) are associated with the individual retrieved components (see also Equation (19)), and **C** is a diagonal array whose non-zero elements are the coefficients c_a. The overall regression model can then be expressed as:

$$\hat{\mathbf{Y}} = \mathbf{U}\mathbf{Q}^{\mathbf{T}} = \mathbf{T}_{\mathrm{PLS}}\mathbf{C}\mathbf{Q}^{\mathbf{T}} = \mathbf{X}\mathbf{R}\mathbf{C}\mathbf{Q}^{\mathbf{T}} = \mathbf{X}\mathbf{B}_{\mathrm{PLS}} \tag{22}$$

Equation (22) shows that, despite the regression model be calculated at the level of the scores, due to the linearity of the projection, it is possible to express it directly in terms of the original variables (similarly to PCR, see also Equation (16)). At the same time, since only a part of the information in the original data is used for the regression, and, in particular, for the estimation of the regression coefficients:

$$\mathbf{B}_{\mathrm{PLS}} = \mathbf{R}\mathbf{C}\mathbf{Q}^{\mathbf{T}} \tag{23}$$

PLS constitutes another case of so-called biased regression [67,68], requiring the determination of the optimal number of components to be retained in the model (which, once again, is carried out by selecting the value leading to the minimum prediction error, usually in cross-validation).

A fundamental aspect of PLS regression is that, contrarily to MLR and, therefore, to PCR, when the problem involves the calibration of multiple responses, individual models yield different outcomes compared to a single global model. Since PLS is also based on the extraction of latent variables from the **Y**-block, in fact, the calculation of a single model for calibrating all the responses at the same time entails their sharing part of their systematic variability, i.e., that they will exhibit a certain degree of intercorrelation that is not due to chance or subsampling. Only in this case is it recommended to build a single model for predicting all the responses; if this condition is not met, then it is suggested that each response be calibrated separately [73].

Altogether, MLR, PCR and PLS represent the large majority of regression methods which are used in scenarios demanding calibration based on spectroscopic data. However, all of them return linear models, while circumstances where the complexity of the problem at hand may call for non-linear functional relations might also occur. Given the nature of spectral data, a relatively popular way of implementing a non-linear functional relation—if necessary—is to use locally linear models [74–76]. The concept behind local regression is that a problem which is globally non-linear may be approximated by a combination of locally linear regressions, each one covering a relatively small portion of the variable space. Other options are non-linear extensions of the PLS algorithm [77–82], support vector machines encompassing non-linear kernel transformations [83–86], and artificial neural networks [87–91] (which,

nonetheless, have more stringent requirements regarding the number of training objects needed for constructing a reliable model).

Throughout the years, regression analysis has played and still plays a crucial role in the domain of food characterization. In particular, PLS, together with its non-linear variants, has been widely used for addressing a large number of issues in such a context (see Table 1), from the estimation of lard content in chocolate specimens by ATR-Fourier Transform-IR (ATR-FT-IR) spectroscopy [92] to the quantification of egg content and turmeric adulteration in egg-pasta by NIR spectroscopy [76,93], the estimation of quality parameters of straw wine by FT-MIR spectroscopy (alcohol and sugar content and acidity) [94], and the analytical profiling of sensory attributes of *Trentingrana* cheese by Proton Transfer Reaction–Mass Spectrometry (PTR-MS) [95]. This last study, in particular, was conducted as an attempt to relate specific flavor and odor features to characteristic compounds detected and identified by the PTR-MS instrumental platform (the perception of boiled milk flavor, for instance, was found to be connected to the presence of methanetiol in the samples at hand).

4. Classification

Apart from exploration and regression, another task that users and practitioners commonly perform while conducting food authentication studies is classification. Classifying a sample or object implies predicting one or more of its discrete properties based on the information collected during its characterization [96]. An example would be the determination of the geographical origin of a certain product from its spectral profile recorded within a certain wavelength range (UltraViolet-Visible—UV-Vis—NIR, MIR, etc.). More specifically, classification would aim at assigning such a product to one category or class constituted by objects sharing similar features (in this case, objects sharing the same geographical origin). From a slightly different perspective, considering an individual sample observation as a vector (e.g., a spectrum) corresponding to a point in the multivariate space of the experimental variables (e.g., the wavelength channels), classification approaches can also be regarded as tools for the identification of boundaries in this space separating the various categories at hand: the aforementioned sample is therefore predicted as belonging to a particular class when the respective point falls within its associated boundaries.

Given the extreme relevance of similar strategies in food authentication scenarios and the importance they can have for the resolution of real-world issues, it does not come as a surprise that the applications of existing classification methodologies along with the proposal and development of novel ones have dramatically increased in the last decades [97,98]. Although these methodologies may differ in complexity, requirements and assumptions, two broad groups of techniques can be distinguished: discriminant and modelling techniques [99].

4.1. Discriminant Techniques

Discriminant approaches are probably the most classical solutions for dealing with classification-related problems [97]. As their name suggests, they make it possible to highlight differences between samples belonging to distinct classes. Their basic principle is simple: they divide the multivariate space of the registered variables into several subregions equal to the number of categories of objects considered (say Z) and they assign each one of such objects to a certain class if the point corresponding to its measurement vector falls within the boundaries associated with that class. Subsequently, owing to these characteristics:

1. discriminant classification models need to be calibrated on training sets composed by specimens belonging to all the categories under study;
2. every analyzed sample is always assigned to one and only one of these categories;
3. samples coming from other classes (not considered in the study) will always be (erroneously) recognized as members of one and only one of these categories.

For all these reasons, discriminant techniques can be particularly useful, e.g., when the amount of expected classes is limited and/or if it can be reasonably supposed that all the samples to be assessed are drawn uniquely from them.

The mother of all discriminant classification strategies is undoubtedly Linear Discriminant Analysis (LDA) [100]. In the simplest case encompassing two categories of objects, LDA estimates a direction of maximum separation between classes (i.e., a so-called *canonical variate*) as:

$$\mathbf{w}^{\mathrm{T}} = (\mathbf{m}_1 - \mathbf{m}_2)^{\mathrm{T}}\mathbf{S}^{-1} \tag{24}$$

with $\mathbf{m}_1$ and $\mathbf{m}_2$ (both of dimensions $J \times 1$) being the vectors of variable means (*centroids*) for class #1 and class #2, respectively, and $\mathbf{S}$ ($J \times J$) an estimate of the class covariance matrix (which, here, is assumed to be the same for both categories (this assumption is relaxed in the most direct extension of LDA known as Quadratic Discriminant Analysis (QDA) [101,102])). $\mathbf{m}_1$, $\mathbf{m}_2$ and $\mathbf{S}$ are computed from training samples. A class delimiter, w_0, is then calculated as:

$$w_0 = \mathbf{w}^{\mathrm{T}}\frac{(\mathbf{m}_1 + \mathbf{m}_2)}{2} \tag{25}$$

and the classification rule established such that if the projection coordinate of each observation vector, say $\mathbf{x}_n$ ($J \times 1$), along $\mathbf{w}$ ($t_n = \mathbf{w}^{\mathrm{T}}\mathbf{x}_n$) is found to be lower than w_0, the corresponding object is predicted as belonging to class #1 and *vice versa*. The rationale behind LDA can also be interpreted by assuming that data within each category are Gaussianly distributed so as to calculate the so-called *posterior* probability of $\mathbf{x}_n$ belonging to the z-th class, $p(z|\mathbf{x}_n)$, as:

$$p(z|\mathbf{x}_n) = \frac{p_0(z)}{(2\pi)^{\frac{J}{2}}|\mathbf{S}|}e^{-\frac{1}{2}(\mathbf{x}_n-\mathbf{m}_z)^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}_n-\mathbf{m}_z)} \tag{26}$$

with $p_0(z)$ equal to the probability of observing an individual from the same category before carrying out any measurement (*prior* probability). Accordingly, the final classification rule can be reformulated by stating that a sample should be assigned to the class it has the highest posterior probability of proceeding from.

Even though LDA can be in principle extended to scenarios involving a larger number of categories, it requires the inversion of the covariance matrix $\mathbf{S}$, which can be singular or nearly singular in the presence of highly collinear (inter-correlated) descriptors (as for spectral profiles) [51]. To overcome this limitation, one can perform a proper selection of these descriptors (via, e.g., step-wise algorithms) [103], apply LDA to the scores resulting from a preliminary PCA modelling of the collected data in what is called PCA-LDA [101,104,105] or resort to regularized versions of LDA, like in Regularized Discriminant Analysis (RDA) [106–111]. While, on the one hand, such approaches are capable of stabilizing the LDA solution against collinearity, on the other hand, some of them might suffer from similar drawbacks as seen before for PCR [104,105,112].

Alternatively, one can think of classification as a regression problem in which the class belonging of a sample (i.e., the dependent variable) is to be estimated from the set of (independent) variables returned by a particular analytical platform (e.g., a spectrum). Similarly to what was outlined in Section 3, such an estimation can be carried out by an extension of PLS named Partial Least Squares Discriminant Analysis (PLSDA) [112,113]. PLSDA regresses the entire data matrix $\mathbf{X}$ on a dummy binary-coded response array, say again $\mathbf{Y}$, made up of a set of piled Z-dimensional row vectors, constructed so that, if their corresponding objects/samples are members of the z-th class, they have a 1-value in their z-th entry and 0-values in all the other ones. For instance, in a 2-class scenario, samples belonging to the first category will be described by the dependent vector [1 0], while samples belonging to the second one by the vector [0 1]. Notice that the model structure in Equation (22) applies also in this case and, thus, whenever new objects/samples become available, their projection coordinates onto the PLS latent variables,

as well as their **Y**-predicted values, can be retrieved as shown in Section 2. The class assignation can be finally carried out based on an LDA model constructed either on PLS scores of training objects (this approach is also known as PLS-LDA [112,114,115]) or on their **Y**-predicted values, according to a highest-prediction rule or to a higher-than-a-threshold prediction rule [116]. As for PLS, more complex variants of PLSDA are available for coping with strong non-linear relationships existing between **X** and **Y** (e.g., Locally Weighted PLSDA—LW-PLSDA [117]—Kernel-PLSDA—K-PLSDA [118–120]). Support vector machines and artificial neural networks can also be used for the resolution of classification-related problems [83,91,121,122].

Table 1. Application examples of exploratory, regression and classification analysis of spectroscopic data in the field of food science. PDO, PCA, ICA, PLS, LDA, PLSDA, LW-PLSDA, K-PLSDA and QDA stand for Protected Designation of Origin, Principal Component Analysis, Independent Component Analysis, Partial Least Squares, Linear Discriminant Analysis, Partial Least Squares Discriminant Analysis, Locally Weighted Partial Least Squares Discriminant Analysis, Kernel-Partial Least Squares Discriminant Analysis and Quadratic Discriminant Analysis, respectively.

Aim	Method	Reference
Exploration of apple varieties	PCA	[32]
Beer storage monitoring	PCA	[33]
Adulteration detection of Camellia oils	PCA	[34]
Chemical characterization of Mediterranean olive oils	ICA	[35]
Chemical characterization of honey samples	ICA	[36]
Chemical characterization of soft drinks	ICA	[36]
Determination of lard content in chocolate samples	PLS	[92]
Quantification of turmeric adulteration in egg-pasta	PLS	[93]
Straw wine quality parameter prediction	PLS	[94]
Sensory characterization of *Trentingrana* cheese	PLS	[95]
Egg content quantification in dried egg-pasta	Local PLS	[76]
Characterization of PDO Chianti Classico olive oil	LDA	[123]
Determination of the geographical origin of pistachios	PLSDA	[124]
Classification of rice varieties	LW-PLSDA/K-PLSDA/Artificial neural networks	[117,125]
Classification of honey samples	PLSDA	[126]
Technological classification of egg white powders	PLSDA	[127]
Discrimination of distillates	PLSDA	[128]
Classification of tomato genotypes	LDA/PLSDA/Support vector machines/	[129]
Olive fruit classification	QDA	[130]
Insect infestation detection in stored rice	PLSDA	[131]
Characterization of Italian craft beers	PLSDA	[132]

For the particular type of problems they make it possible to tackle, discriminant methods have had a long history in the field of food quality assessment and control (see Table 1). LDA, QDA, linear and non-linear PLSDA, support vector machines and artificial neural networks have been broadly used to discriminate foodstuff of different origins [123,124], different varieties [117,125,126], different technological properties [127], different purity degree [128], exhibiting different genotypes (i.e., transgenic/non-transgenic) [129] or affected by different diseases [130,131]. PLSDA has been applied, for example, to NIR spectra of pistachios for their geographical authentication making it possible to highlight that seeds from Turkey and USA show a typical absorption behavior between 8000 and 9000 cm^{-1} (second overtone of methylenic stretching vibrations) [124] and to UV-Vis spectra of ale beers brewed by distinct producers pinpointing minimal variations in the tone and intensity of their color which could indirectly influence customer perception [132].

4.2. Modelling Techniques

Contrarily to discriminant techniques, modelling approaches are capable of capturing similarities among samples belonging to the same category [97,133]. They basically define a multivariate boundary

for each considered class, which delimits a specific region of the multidimensional space of the original descriptors where objects proceeding from it are likely to be found. If an analyzed sample falls within this region, it is assigned to the respective class, otherwise it is considered to be a class outlier and rejected as such [134,135].

A fundamental aspect that needs to be taken into account here is that categories are treated and handled separately and that an individual model is constructed for every one of them. Subsequently, (i) modelling methodologies can easily be applied in scenarios in which a unique class of interest exists (a common contingency in food traceability or authentication studies), and (ii) new objects can be predicted as members of one, none or multiple classes also in the light of the fact that the various class spaces do not necessarily have to cover completely the whole original variable space [136].

Given their characteristic nature, class modelling strategies can all be regarded as outlier detection methods. Nonetheless, mainly depending on the *outlyingness* criterion adopted, such strategies can exhibit different advantages and disadvantages and be more or less suitable to be applied in certain situations rather than others. Due to the flexibility that the selection of this outlyingness criterion guarantees, a plethora of class modelling approaches have been proposed throughout the last 45 years. Among those appearing in the chemometric literature (see Table 2), probably the most commonly exploited since their development are UNEQual class spaces (UNEQ [137,138]) and Soft Independent Modelling of Class Analogy (SIMCA [139,140]).

UNEQ was introduced by Derde and Massart in 1986 and constitutes the modelling version of QDA. Briefly, UNEQ defines the class model by the centroid of the concerned category, while the class space is represented by the multidimensional ellipsoid (*hyperellipsoid*) corresponding to a user-defined confidence level, i.e., to a certain probability of finding samples of that category within its boundary. Mathematically speaking, in its initial formulation UNEQ calculates for each investigated object ($\mathbf{x}_n$ – $J \times 1$) its squared Mahalanobis distance from the z-th class centroid ($\mathbf{m}_z$ – $J \times 1$) as:

$$d_n^2 = (\mathbf{x}_n - \mathbf{m}_z)^{\mathrm{T}} \mathbf{S}_z^{-1} (\mathbf{x}_n - \mathbf{m}_z) \tag{27}$$

where $\mathbf{S}_z$ ($J \times J$) denotes an estimate of the z-th class covariance matrix. Once again, $\mathbf{m}_z$ and $\mathbf{S}_z$ are computed from training objects. If d_n^2 is found to be larger than a critical distance value, the respective sample is rejected by the class model as an outlier. Such a critical value can be retrieved as detailed in [137,138,141]. Variants of both the distance statistic and the way its corresponding threshold is determined have been proposed over the years; an alternative UNEQ framework based on the principles of the Hotelling's T^2 statistic was, for instance, developed by Forina et al. in 1995 [142].

Despite its relative simplicity, as for MLR and LDA, UNEQ encompasses the inversion of a covariance matrix, which is seldom attainable in case the registered variables show a high degree of intercorrelation [51]—as a solution, UNEQ can be applied to the scores resulting from a preliminary PCA modelling of the data at hand. In addition, UNEQ requires such variables to follow a multivariate Gaussian distribution, an assumption that is rarely met when distinct sources of variability are present in the data at hand (e.g., in the case of the traceability of a designated food product, these sources of variability may be associated with different cultivars, harvesting years, producers, etc.) [133].

On the other hand, in SIMCA, the data associated with the z-th category of samples ($\mathbf{X}_z$ – $N_z \times J$, with N_z equal to the number of training objects belonging to the z-th class) are modelled by PCA as:

$$\mathbf{X}_z = \mathbf{T}_z \mathbf{P}_z^{\mathrm{T}} + \mathbf{E}_z \tag{28}$$

Every object to be classified ($\mathbf{x}_n$ – $J \times 1$) is afterwards projected onto the class principal component subspace as:

$$\mathbf{t}_{n,z}^{\mathrm{T}} = \mathbf{x}_n^{\mathrm{T}} \mathbf{P}_z \tag{29}$$

and its distance from such a subspace (of dimensionality A_z) is calculated as:

$$s_n^2 = \frac{\mathbf{e}_{n,z}^{\mathrm{T}} \mathbf{e}_{n,z}}{(J - A_z)} \tag{30}$$

with $\mathbf{e}_{n,z}^{\mathrm{T}} = \mathbf{x}_n^{\mathrm{T}} - \mathbf{t}_{n,z}^{\mathrm{T}} \mathbf{P}_z^{\mathrm{T}}$.

This distance value is compared with the average distance to the model subspace of the N_z training samples (computed as $s_z^2 = \sum_{n_z=1}^{N_z} \sum_{j=1}^{J} \frac{e_{n_z,j,z}^2}{(J-A_z)(N_z-A_z-1)}$, where $e_{n_z,j,z}^2$ results from the projection of the n_z-th training object onto the PCA class model subspace) by means of an F test with $(J - A_z)$ and $(N_z - A_z - 1)$ degrees of freedom. If the null-hypothesis of the test is rejected, the corresponding sample is labelled as an outlier and rejected by the class model.

This first implementation of SIMCA was proposed by Wold et al. in 1976 and was almost immediately amended in order to additionally account for the sample distance within the principal component subspace and not only from the principal component subspace [140]. In addition, together with this alternative formulation, many others have been reported in literature generally based on diverse ways of defining the class space and/or the classification rule (see, e.g., [143]).

Finally, it is important to note that the performance of class modelling approaches mainly depends on the settings of several metaparameters such as the confidence level at which the class space is constructed and/or the number of latent variables extracted from the original data. Different approaches for the optimization of these metaparameters exist, whose efficiency and robustness can vary according to the specific scenario faced (e.g., high class overlap, presence of outliers in the training set, etc.) [144–149].

For their particular nature and for the main implication resulting from their use (i.e., each category of objects is handled independently and separately), modelling classification strategies are nowadays regarded as ad hoc solutions for coping with food authentication problems (see Table 2). In particular SIMCA has lately been widely resorted to for the traceability of high value-added foodstuffs like wine, beer, extra virgin olive oil, olive seeds, walnuts, coffee and rice [132,150–163], and has been found to guarantee extremely satisfactory performances when PDO and/or PGI products are to be discerned from lower-quality ones. In this regard, several application studies involving the combination of NIR spectroscopy and SIMCA for the characterization of Avola almonds [164], Gragnano pasta [165], Italian hazelnuts [166] and Vallerano chestnuts [167] have been recently reported in literature.

Table 2. Main class modelling methods and application examples in the field of spectroscopy-based food analysis. PGI and PDO stand for Protected Geographical Indication and Protected Designation of Origin, respectively.

Methods		
Name	**Aim**	**Reference**
UNEQual class spaces (UNEQ)	Exploratory	[137,138]
Soft Independent Modelling of Class Analogy (SIMCA)	Regression	[139,140]
Non-parametric class modelling	Regression	[151]
Neural networks-based class modelling	Regression	[152,153]
Partial Least Squares Density Modeling (PLSDM)		[154]
Potential function (kernel density) method (POTFUN)	Regression	[168–170]
Pattern Recognition by Independent Multicategory Analysis (PRIMA)	Regression	[171]
Multivariate Range Modeling (MRM)	Regression	[172]
Support Vector Domain Description (SVDD)		[173]

Table 2. *Cont.*

Applications		
Aim	**Method**	**Reference**
Traceability of rice varieties	UNEQ/SIMCA/Neural networks-based class modelling	[152]
Authentication of wine samples	UNEQ/SIMCA/MRM/Neural networks-based class modelling	[153,155,156]
Authentication of beer samples	UNEQ/SIMCA/POTFUN	[132,157,158]
Traceability of extra virgin olive oils	SIMCA/Non-parametric class modelling	[151,159,160]
Authentication of olive seeds	UNEQ/SIMCA/PLSDM	[154,161]
Traceability of coffee	SIMCA/POTFUN	[150,162]
Traceability of walnuts	SIMCA	[163]
Authentication of Avola almonds	SIMCA	[164]
Authentication of PGI Gragnano pasta	SIMCA	[165]
Authentication of Italian PDO hazelnut	SIMCA	[166]
Authentication of Vallerano chestnut	SIMCA	[167]
Plant ripening monitoring	PRIMA	[174]
Wheat straw fermentation monitoring	SVDD	[175]

5. Data Fusion

Although all the illustrated methods certainly represent suitable tools to achieve the purposes they are conceived for, sometimes, the complexity of an analytical problem requires to be regarded from several different perspectives to be comprehensively embraced. As an example, many present-day food chemistry issues cannot be easily addressed unless distinct instrumental techniques (e.g., Gas Chromatography–Mass Spectrometry—GC-MS—and MIR spectroscopy) are combined for a more extensive characterization of the samples at hand. Similar strategies are commonly defined *multi-block* and the chemometric approach used for the concerted analysis of multi-platform data is known as *Data Fusion* (DF).

Broadly speaking, DF enables the simultaneous extraction of meaningful and useful information from diverse analytical sources. In recent years, DF has been found to provide more exhaustive descriptions of studied systems compared to when single datasets are assessed separately. For this reason, in the last decades, numerous DF methodologies have been proposed in the literature with the three tasks surveyed in the previous sections of this article in mind: data exploration, regression and classification. Nonetheless, despite the ultimate purposes behind their development, such methodologies are usually classified according to the *level* at which the data fusion is implemented. In the so-called *low-level* DF, for instance, raw datasets are simply concatenated and analyzed at the same time as a unique global ensemble. *Mid-level* DF extracts specific features (e.g., principal components or latent variables) from each considered dataset which are subsequently gathered and investigated concurrently. Finally, in *high level* DF, the outcomes of individual models constructed for every concerned dataset are pooled in a unique solution. Another common distinction among DF strategies is based on the nature of their underlying algorithms (*sequential*—if joint information is retrieved iteratively—or *parallel*—if joint information is retrieved collectively) [176].

Due to the large number of existing DF approaches, the following subsections will offer only a basic and brief description of those most commonly used in the field of food authentication. In Table 3, a detailed list of applications of DF in this domain is also reported.

5.1. Multi-Block Data Exploration

One of the first exploratory multi-block methods dates back to 1987. It was proposed by Wold and Hellberg and it is called Consensus PCA (CPCA) [177]. CPCA is the natural extension of PCA for the analysis of multiple datasets. Basically, the algorithm searches for directions of maximum variance common to all the data blocks at hand. CPCA encompasses four different steps: (i) one measured variable is selected as super-score; (ii) block loadings and scores are estimated from such a super-score; (iii) block scores are gathered in an updated super-score block which is used to calculate

a set of super-weights; and (iv) super-weights are normalized and a new super-score is computed. The procedure is iterated until convergence.

A few years later, in 1996, Wold himself, together with Kettaneh and Tjessem, developed a novel version of multi-block PCA, called Hierarchical PCA (HPCA), sharing similar algorithmic features with CPCA [178]. Their main difference, though, relates to the fact that in HPCA the super-scores are normalized instead of the super-weights. It was observed that both CPCA and HPCA exhibit severe convergence issues, eventually overcome through the modifications proposed by Westerhuis and Coenegracht [179].

Later on, in 2001, a new sequential exploratory approach (Generalized PCA—GPCA [180]) appeared in literature. GPCA constructs a PCA model on the matrix resulting from the concatenation of the datasets under study and extracts normalized super-scores. These super-scores are resorted to for the retrieval of block scores and loadings. The outcomes of a comprehensive comparison of different variants of CPCA, HPCA and GPCA are reported in [181].

Almost concomitantly, in 2000, Qannari et al. implemented Common Components and Specific Weights Analysis (CCSWA) [182,183]. CCSWA seeks the shared information underlying all the investigated datasets. Specifically, A_{com} orthogonal common components are extracted from a combination of the cross-product matrices corresponding to the I data blocks coped with ($\mathbf{X}_i\mathbf{X}_i^{\text{T}}$ $\forall i \in [1, I]$), which is weighted according to their so-called *salience*, i.e., the portion of variance they explain in each one of them. Over the years, different formulations of CCSWA have been designed. CCSWA has nowadays become very well-known with the name ComDim.

More recently, many alternative multi-block techniques (thoroughly reviewed in [184,185]), based on the extraction of common factors, but at the same time also aimed at the recognition of the distinctive information carried by each of the individual considered datasets, have been proposed (e.g., DIStinctive and COmmon Simultaneous Component Analysis—DISCO-SCA [186]—or 2-block Orthogonal Projections to Latent Structures—O2PLS [187]).

5.2. Multi-Block Regression and Classification

The chemometric literature is rich in multi-block methods for regression and classification. One of the most commonly used is undoubtedly Multi-Block-PLS (MB-PLS). MB-PLS constitutes the direct extension of PLS to the data fusion field. Different variants of its algorithm have been proposed all over the years [179], the most recent being developed by Qin et al. in 2001 [188]. The basic idea behind this latest formulation is that multiple predictor blocks can be concatenated and modelled simultaneously by PLS. Although extremely intuitive and yielding accurate predictions, this strategy does not provide information about the single datasets. Conversely, the solution implemented by Westerhuis and Coenegracht can be exploited for the interpretation of block scores and loadings [179]. In this regard, it is important to stress that MB-PLS requires the datasets at hand to be scaled separately (by, e.g., their individual sum-of-squares or their Frobenius' norm), to prevent structures with greater variance from preferentially driving the model and blurring the information carried by the others [179].

Due to its remarkable flexibility, MB-PLS is nowadays probably the most widely applied multi-block regression approach. It is also often used in combination with LDA for addressing classification problems—MB-PLS is here known as MB-PLS-LDA or MB-PLSDA. So far, many applications of MB-PLS have been reported in the field of food analysis (see Table 3); it has been resorted to (i) for the investigation of sensory parameters of different nature and of their relationships with technological properties of cheese and bread samples [189,190], (ii) for the prediction of meat spoilage time, wine ageing time and crude protein and moisture content in soybean flour by MIR and NIR spectroscopy [191–193], (iii) for the discrimination of botanical varieties of extra virgin olive oil, lemon essential oils and wines of different geographical origin by MIR, NIR and Raman spectroscopy [194–196], and (iv) for distinguishing added-value from low-quality products [132].

Despite all the aforementioned advantages MB-PLS offers, usually redundant information and/or the presence of categorical regressors may jeopardize the predictive capability of MB-PLS. In these

situations, techniques like the recent Sequential and Orthogonalized-PLS (SO-PLS [197,198]) could represent feasible options to overcome such limitations. SO-PLS enables the sequential modelling of each regressor set. The keystone of this methodology is a preliminary orthogonalization step performed to filter out redundancies among blocks. Briefly, taking into consideration the case where a generic response $\mathbf{Y}$ is to be estimated from two predictor matrices, say $\mathbf{X}_1$ and $\mathbf{X}_2$, the SO-PLS algorithmic scheme encompasses the following four steps (readers are addressed to [140,141] for a more detailed description of the SO-PLS algorithm):

1. $\mathbf{Y}$ is regressed onto $\mathbf{X}_1$ by PLS;
2. $\mathbf{X}_2$ is orthogonalized with respect to the $\mathbf{X}_1$-scores calculated in step 1, yielding the array $\mathbf{X}_{2,\text{orth}}$. This ensures the common information shared by $\mathbf{X}_1$ and $\mathbf{X}_2$ is removed from the latter;
3. The PLS residuals resulting from step 1 are regressed onto $\mathbf{X}_{2,\text{orth}}$ by PLS;
4. The SO-PLS predictive model is expressed by combining the outcomes of steps 1 and 3 as:

$$\mathbf{Y} = \mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_2\mathbf{B}_2 + \mathbf{E}_{\text{SO-PLS}} \tag{31}$$

where $\mathbf{B}_1$ and $\mathbf{B}_2$ contain the regression coefficients for $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively, and $\mathbf{E}_{\text{SO-PLS}}$ carry the final $\mathbf{Y}$-residuals.

As a consequence of the deflation of the common information shared by the different datasets under study, SO-PLS provides useful insights into the unique and distinctive sources of variation within each one of them. Analogously to MB-PLS, SO-PLS has already been broadly employed in food science (see Table 3), also coupled to LDA for tackling classification tasks [199]; it has been used (i) to correlate sensory perception and chemical composition descriptors [200], (ii) for the quantification of particular compounds in dietary supplements [201], (iii) for the authentication of spirits (i.e., Italian grappa—Italian grape marc spirit) by MIR and NIR spectroscopy [202], and (iv) for the determination of the geographical origin of foodstuff (e.g., red garlic, semolina and saffron) by multiple spectroscopic modalities (UV-Vis, NIR, MIR and ATR-FT-IR) [203–205]. Furthermore, the robustness and versatility of SO-PLS has recently led to the design of hybrid techniques conceived for very disparate purposes: Sequential Multi-Block PLS (SMB-PLS [206]), merging features of both MB-PLS and SO-PLS; Sequential and Orthogonalized-N-PLS (SO-N-PLS [207]), for the multi-block analysis of multi-way structures; Sequential Preprocessing through ORThogonalization (SPORT [208]), for the pretreatment of the SO-PLS predictor blocks; Sequential and Orthogonalized Covariance Selection (SO-CovSel [209]), for the low-level DF of preselected variables; and all their discriminant extensions. Yet, one of the main drawbacks associated with SO-PLS regards the fact that a deep and complete inspection of the factors shared by the regressor matrices cannot be easily carried out.

In similar scenarios, P-ComDim, the predictive version of CCSWA, constitutes the designated approach [210–212]. The algorithmic procedures behind P-ComDim and CCSWA are strictly resemblant with the difference that, in the former, for each regressor block, the cross-product matrices $\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}$ are replaced by the core matrices $\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$. Common components (together with their salience values) are iteratively retrieved as in ComDim, but, here, also $\mathbf{Y}$-scores are computed. Additionally, P-ComDim has been rather often applied for tackling complex food authentication problems, not only due to its generally high prediction accuracy, but also because it may open wide horizons for the identification of commonalities underlying regressors proceeding from very diverse sources.

Table 3. Additional data fusion methods and application examples in the field of spectroscopy-based food analysis. MB-PLS, SO-PLS, SO-CovSel and OnPLS stand for Multi-Block-Partial Least Squares, Sequential and Orthogonalized-Partial Least Squares, Sequential and Orthogonalized Covariance Selection and n-block Orthogonal Projections to Latent Structures, respectively.

Methods		
Name of the Method	**Aim**	**Reference**
Hierarchical PLS (H-PLS)	Regression	[178]
Joint and Individual Variation Explained (JIVE)	Exploratory	[213]
Multiblock PLS serial extension	Regression	[214]
Network-Induced Supervised Learning (NI-SL)	Regression	[215]
Parallel Orthogonalized Partial Least Squares (PO-PLS)	Regression	[216]
Multiblock Redundancy Analysis	Regression	[217]
OnPLS	Regression	[218]
Applications		
Aim	**Multi-Block method**	**Reference**
Prediction of bread sensory properties	MB-PLS	[190]
Prediction of wine ageing time	MB-PLS/H-PLS/NI-SL/SO-PLS	[192]
Quantification of protein and moisture in soybean flour	MB-PLS	[193]
Discrimination of lemon essential oils	MB-PLS	[194]
Determination of the geographical origin of wine	MB-PLS	[196]
Authentication of spirits	SO-PLS/SO-CovSel	[128,202]
Determination of the geographical origin of saffron	SO-PLS/SO-CovSel	[205]
Path modelling	SO-PLS	[219]
Path modelling	SO-PLS	[220]

6. Other Approaches

Data exploration and predictive modelling represent the large majority of chemometric applications in the context of spectroscopy-based assessment of the quality of foodstuff. Nevertheless, there also exist other strategies which may provide relevant insights into the products or raw materials under investigation when the aim is their characterization and/or authentication. In this section, these techniques will be briefly illustrated.

6.1. Curve Resolution

In the previous sections, it was shown how a spectral data matrix **X** can be decomposed according to a bilinear model and how this decomposition can be regarded as its projection onto a subspace of latent (*abstract*) directions. However, under normal experimental conditions, it is well known that spectroscopic data can also be approximated by a bilinear model whose elements are directly interpretable in chemical terms. Indeed, by assuming that **X** contains the spectral intensities or absorbance values for the various analyzed samples at different wavelengths, and, without any loss of generality, that the optical path is unitary, Beer-Lambert-Bouger law can be expressed as:

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathrm{T}} \tag{32}$$

where **C** and **S** are the arrays collecting the concentrations of the chromophores constituting the spectral mixture across all the specimens and their individual spectroscopic signatures (molar absorbivities), respectively [221]. Based on the relation described by Equation (31), the aim of curve resolution methods is to perform a data-driven unmixing of the matrix **X**, i.e., to estimate the number of constituents in the aforementioned mixture and obtain their pure concentration and spectral profiles from the information encoded in the recorded data. Among the possible approaches for accomplishing such a task, the most popular (due to its flexibility and to the possibility of being applied also to multi-set or multi-way data) is Multivariate Curve Resolution–Alternating Least Squares (MCR-ALS [222–227]). As the name suggests, MCR-ALS is an iterative algorithm based on the alternating least squares concept, which involves the following steps:

1. The number of components (mixture constituents) is estimated (e.g., according to a priori knowledge of the systems under study or by SVD/PCA).
2. A first guess of either $\mathbf{C}$ or $\mathbf{S}^T$ is calculated by methodologies like Evolving Factor Analysis (EFA [228]) or SIMPLe-to-use Interactive Self-Modeling Analysis (SIMPLISMA [229]).
3. $\mathbf{C}$ and $\mathbf{S}^T$ are iteratively updated using alternating least squares under appropriate constraints (e.g., non-negativity of the values in $\mathbf{C}$ and/or $\mathbf{S}^T$) as:

$$\begin{aligned} \mathbf{C} &= \mathbf{X}\mathbf{S}\left(\mathbf{S}^T\mathbf{S}\right)^{-1} \\ \mathbf{S}^T &= \left(\mathbf{C}^T\mathbf{C}\right)^{-1}\mathbf{C}^T\mathbf{X} \end{aligned} \tag{33}$$

4. Step 3 is repeated until a certain convergence criterion is met.

It is evident for the algorithmic procedure outlined before that one of the main features of MCR-ALS is its "self-modeling" nature, i.e., in principle, it does not require any specific preliminary information about the data at hand: it is only needed that the bilinear model in Equation (31) is satisfied and that some very generic characteristics of the pure concentration or spectral profiles (e.g., their non-negativity) are known. If further information is also available, though, it can be actively exploited in the form of additional mathematical restrictions. Basically, it is the possibility of implementing in a rather straightforward way a large variety of these restrictions what has made and still makes MCR-ALS very popular compared to other curve resolution strategies. Indeed, they not only reduce the ambiguity associated with the final model (that is to say the range of possible $\mathbf{C}/\mathbf{S}^T$ sets returning the same fit when approximating $\mathbf{X}$) but, at the same time, they render the resolved profiles more physico-chemically meaningful [230,231]. Examples of possible restrictions are unimodality (if the constrained profiles are expected to exhibit a unique global maximum), closure (if a mass or concentration balance among all or part of the mixture constituents holds), and selectivity (if some of the resolved species are absent in some experiments or do not absorb at specific wavelengths) [222,226,232–234].

Moreover, as already anticipated, MCR-ALS can easily be applied to multiple data matrices in a multi-set configuration. This provides a different way of analyzing multi-block data with respect to what discussed in Section 5, as here the information shared by the different investigated arrays is resorted to for improving the unmixing of the pure constituents. Concurrently, since some of these constituents can be present only in one or a few matrices and some others may be present in all, multi-set MCR-ALS is one of those multi-block techniques by which information about both common and distinctive components can be obtained [235].

Similarly to ICA, MCR-ALS has mainly been employed in order to retrieve the spectral contributions of particular compounds constituting complex food samples and to determine how their abundance/concentration varies during, e.g., a renneting process (see Table 4). In [236–238], for instance, the dynamic evolution of various forms of milk (liquid, sol-gel, coagulated) over the progression of a lactic acid fermentation reaction was monitored and their individual FT-NIR fingerprints disentangled. In [239], a similar study was conducted on beer fermentation: here, maltose, maltotriose, fructose, sucrose, dextrins and ethanol spectral and concentration profiles were retrieved for enabling a better understanding of the various phenomena behind their biotransformation.

6.2. Analysis of Multivariate Designed Data

The rational design of the experiments to be conducted is a fundamental step in any scientific discipline and the only way in which it can be guaranteed that the information sought be present in the collected data. The design of experiments involves four steps: (i) identifying some critical variables (*factors*) which could affect one or more properties of interest (*responses*), (ii) defining their range of variability that could be worth inspecting, (iii) planning the trials to be performed (so as to be at the same time as informative and as low in number as possible), and (iv) interpreting the final results [240–243]. In this subsection, attention will be essentially focused on this last step, as it

is the one where traditional statistical approaches commonly fail when instead of measuring a few (uncorrelated) responses for each experimental condition setting, a full spectroscopic signature is registered to characterize the concerned samples [244].

As summarized before, based on a rational design, experiments are conducted under different conditions, characterized by the fact that the factors whose effect is investigated are fixed (or controlled) at specific values (*levels*). In other words, each combination of factor levels is a so-called *design point*, at which the measurement of the responses is carried out. For instance, one could be interested in studying how the spectroscopic profile of a specific product varies as a function of temperature and pH, and then decide to run experimental trials only at three levels of temperature (25 °C, 40 °C and 55 °C) and at two levels of pH (5 and 9). In case all possible combinations of factor levels are explored, the design is called *full-factorial*. In the example sketched before, a full-factorial design would encompass six distinct experiments: 25 °C and pH = 5; 40 °C and pH = 5; 55 °C and pH = 5; 25 °C and pH = 9; 40 °C and pH = 9; 55 °C and pH = 9. At each condition setting, the desired responses should be measured; if the aim is a spectroscopic characterization, the response is highly multivariate and it is represented by the spectroscopic signature of the respective sample. Moreover, very often, the whole design is replicated at least twice to have an estimate of the variability not to be ascribed to the controlled factors.

Given a particular design and having collected the data corresponding to every one of its points, the successive step is to analyze such data to first evaluate whether any of the controlled factors exhibits a statistically significant effect on the responses and (if this is the case) to interpret and describe in greater detail how this effect occurs (for instance, which spectral regions change with the increase of the temperature and how). The traditional statistical tool used in this type of contexts is the analysis of variance—univariate (ANOVA [245–248]) or multivariate (MANOVA [58,241,244,249–252]), depending on the nature of the response(s) measured. However, when dealing with a high number of possibly very correlated descriptors (like spectral variables) and, therefore, with ill-conditioned data matrices, MANOVA suffers from the same drawbacks as MLR or LDA. To overcome these limitations, in the last 15 years, various approaches for the assessment of multivariate or megavariate data resulting from designed experiments have been proposed, all of them having in common the same initial partitioning of the experimental data array, which follows the classical ANOVA scheme based on the linear additivity of the effects. In particular, by assuming that the experimental design at hand involves two factors (generically named A and B), the data matrix $\mathbf{X}$ made up of all the spectra collected at the different design points after mean centering (i.e., after subtraction of the mean spectrum calculated across the entire dataset) is partitioned as follows:

$$\mathbf{X} = \mathbf{X}_{\mathrm{A}} + \mathbf{X}_{\mathrm{B}} + \mathbf{X}_{\mathrm{AB}} + \mathbf{X}_{\mathrm{res}} \tag{34}$$

where $\mathbf{X}_{\mathrm{A}}$, $\mathbf{X}_{\mathrm{B}}$, and $\mathbf{X}_{\mathrm{AB}}$ account for the effect of the two factors and their interaction, respectively, while $\mathbf{X}_{\mathrm{res}}$ contains the residual variance in $\mathbf{X}$ not explained by the model. More specifically, $\mathbf{X}_{\mathrm{A}}$, $\mathbf{X}_{\mathrm{B}}$, and $\mathbf{X}_{\mathrm{AB}}$ carry identical replicates of the mean spectra corresponding to the different factor or interaction levels. Starting from the common ground defined by Equation (33), the various methods developed for the analysis of multivariate designed data proceed along different directions. Among these techniques, the most popular is surely ANOVA–Simultaneous Component Analysis (ASCA [253,254]), due to its relative simplicity and remarkable interpretability properties. In ASCA, the effect of each design term (factor or interaction) is estimated through the sum of squares of the elements of the corresponding matrix in Equation (29). For example, the effect of factor A is quantified as:

$$SSQ_{\mathrm{A}} = \sum_n \sum_j x_{n,j,\mathrm{A}}^2 = {\mathbf{X}_{\mathrm{A}}}^2 \tag{35}$$

and its statistical significance is evaluated non-parametrically by comparing SSQ_{A} with its null-distribution retrieved by permutations [255,256]. The interpretation of the significant effects is finally carried out by applying PCA to the respective matrices.

Alternative techniques described in the literature are ANOVA-PCA (APCA [257,258]), ANOVA-Target Projection (ANOVA-TP [259–261]), ANOVA–Common Dimensions (AComDim [262]) and regularized MANOVA (rMANOVA [263]).

Considering their underlying model structure, the methodologies outlined in this subsection have been widely exploited to obtain insights into how certain technological factors affect food-related processes like ageing (see Table 4); ASCA, for instance, has often been coupled to MIR and/or NIR spectroscopy to determine and examine the influence of (i) drying temperature and time on dried egg-pasta manufacturing [76], (ii) varietal origin and roasting time on coffee bean spectral properties [150], and (iii) storage temperature and conditions on Cheddar cheese ripening [264]. In all these case studies, the application of ASCA has generally permitted to highlight in a rapid, non-destructive, relatively cheap and green fashion the occurrence of specific phenomena (e.g., proteo- and lypolisis) ongoing along with the evolution of the monitored process itself.

Table 4. Examples of applications of curve resolution and analysis of variance-based techniques in the fields of spectroscopy-based food analysis. MCR-ALS and ASCA stand for Multivariate Curve Resolution-Alternating Least Squares and ANOVA-Simultaneous Component Analysis, respectively. The abbreviation ANOVA denotes the analysis of variance.

Aim	Method	Reference
Chemical characterization of milk lactic acid fermentation	MCR-ALS	[236]
Milk renneting characterization and monitoring	MCR-ALS	[237,238]
Chemical characterization of beer fermentation	MCR-ALS	[239]
Assessment of coconut oil purity/adulteration degree	MCR-ALS	[265]
Chemical characterization of chocolate samples	MCR-ALS	[266]
Egg-pasta characterization	ASCA	[76]
Coffee bean roasting monitoring	ASCA	[150]
Cheddar cheese ripening monitoring	ASCA	[264]
Eggplant chilling injury characterization	ASCA	[267]

7. Additional Fundamental Aspects of Chemometric Modelling: Data Preprocessing and Validation

In the previous sections, the main chemometric approaches for tackling data exploration, calibration and classification were critically discussed and examples of their application were illustrated both for single- and multi-block cases. Nevertheless, data analysis is not limited to the model building phase only; rather, there exist at least two further steps (before and after the construction of a chemometric model itself) that play a fundamental role in determining the quality of the obtained outcomes and their reliability: preprocessing and validation.

7.1. Data Pre-Processing

Raw data are almost always not suited to be analyzed as such, i.e., directly in the form in which they are yielded by the instrument, and spectroscopic profiles in particular result from multiple contributions, and only a few of these contributions might be relevant for the problem under study. The combination of multiple sources of unwanted variability can have a significant impact both on the model performance and on its interpretation; for instance, bilinear approaches extract components according to some variance/covariance-based criteria which may lead to nuisance artifacts in the presence of a high amount of spurious/uninteresting information. Accordingly, data are preliminarily subjected to one or more transformations (preprocessing), so as to remove or reduce their so-called detrimental fraction [268]. When considering the possible ways of pretreating experimental data, it is necessary to point out that there are some strategies which are valid independently from the nature of the registered signal, while others are conceived for more specific types of instrumental responses. Within the first family of techniques, the two most commonly used operations are centering and scaling [269]. Centering aims at removing variable offsets; normally, mean-centering is adopted, i.e.,

the mean value computed across each column of a data matrix is subtracted from all the elements of that column. This makes it possible to discard what is shared by the entire set of samples and magnify the differences among individuals. On the other hand, scaling is frequently applied to guarantee that the scales of the different recorded variables are comparable as well as their a priori contribution to the chemometric model. Mathematically, it consists of dividing all the elements of each column of the data matrix by a constant term (usually its standard deviation), and it is highly recommended when the descriptors have been collected using distinct analytical platforms and, therefore, have been reported in different measurement units. It goes without saying that scaling is rarely performed on spectral profiles.

As mentioned before, there also exist several other preprocessing strategies which directly focus on the nature of the instrumental response itself and are designed to correct/remove unwanted sources of variability that are related to its specific characteristics, such as stochastic noise, non-linearities, baseline or wavelength shifts, or undesired variations due to particular chemical and physical phenomena (such as light scattering). The impact of stochastic noise can be reduced by the use of methods which filter high-frequency contributions out of the investigated signal. Fourier [270] or wavelet [271] transforms or Savitzy-Golay smoothing [272] are possible solutions to cope with such an issue. The Savitzky-Golay approach, in particular, is worth mentioning in greater detail, as it can also be exploited for signal or spectral differentiation: it fits successive subsets of adjacent variables with a low-order (usually second or third) polynomial and approximates the variable around which each subset is centered with the value resulting from the respective estimated function.

Differentiation methodologies make it possible to remove additive (first derivative) and multiplicative (second derivative) effects from the registered profiles and to deconvolve, at least partially, overlapping peaks. However, they exhibit the drawback of significantly decreasing the signal-to-noise ratio. This is why they often require a preliminary smoothing operation—Savitzky-Golay differentiation, for example, couples the smoothing procedure described before to the point-to-point calculation of the derivative of the interpolating polynomial.

The impact of light scattering on spectroscopic data, especially in the NIR domain, can also be regarded as a multiplicative effect and corrected by the use of second-order derivation. Alternatively, scatter correction can be achieved by techniques like Standard Normal Variate (SNV) [273], Multiplicative Scatter Correction (MSC) [274] and Extended Multiplicative Signal Correction (EMSC) [275], which enables the simultaneous removal of the effects of scattering, baseline and potential interferents (if their spectroscopic signatures are available).

Finally, for the correction of a non-constant baseline, detrending [273] and penalized asymmetric least squares regression [276] represent feasible options. The former corresponds to fitting a polynomial function to all the variables constituting the spectral signal, but in those cases in which the baseline does not account for a large portion of the overall data variance, severe artifacts may be produced. The latter is underlain by a more effective weighted fitting procedure that automatically identifies the variables most likely to be affected by the baseline contribution and weights them more for the calculation of the approximating curve.

7.2. Validation

Having built a chemometric model does not guarantee per se its quality and/or validity, as these two aspects are strictly influenced by user-dependent factors like the chosen analysis technique or the values of certain metaparameters [2]. To assess whether it can lead to solid conclusions, the outcomes/predictions resulting from it are reliable and generalizable and its interpretation is sound and meaningful, such a model needs, therefore, to be *validated* [277]. Operationally, validation consists in evaluating the model's performance on a dataset (called a *test set*) different from the one used in the training or calibration phase (referred to as a *training* or *calibration set*, as also specified before), but reflecting as closely as possible the distribution of future individuals and accounting for all potential sources of expected variability (e.g., in terms of geographical origin, composition, manufacturing

process, etc.)—it is fundamental not to use for validation the same data exploited for model building as this could lead to overoptimistic considerations (being data-driven, chemometric models are conceived so as to return the best performance on those data). Unfortunately, in many real-world scenarios, it is not possible to design additional sampling campaigns and assemble completely new test sets. In these circumstances, it is common practice to split all the recorded measurement observations into two different blocks, one for model building and the other for model validation. However, for the validation procedure to be effective, the sample splitting scheme should ensure that both training and test sets span a representative amount of the overall data variability: in this sense, one should avoid random partitioning and, instead, resort to targeted subset selection algorithms like Kennard-Stone [278], Duplex [279,280] or D-optimal design-based approaches [281].

An alternative is constituted by internal resampling strategies [282], like cross-validation [70], in which the original data are repeatedly divided into a training and a test set and the procedure is iterated either until each object has been left out at least once from the former or until a maximum number of computational cycles has been reached. Cross-validation is often exploited when small sample sizes are concerned but, due to its underlying principle, usually yields optimistically biased outputs, as training and test sets are never completely independent from one another. Nevertheless, cross-validation can be fruitfully adopted to estimate optimal values for model metaparameters like the number of latent variables in PLS regression or PLSDA classification.

8. Software

Nowadays, multiple software tools for chemometric modelling, both commercial and open-source, are available and easily accessible. A detailed list can be found in [1]. In particular, a collection of freely downloadable Matlab (The Mathworks Inc., Natick, MA, USA) functions for the implementation of most of the methods described in this review can be found at https://www.chem.uniroma1.it/romechemometrics/research/algorithms/.

9. Conclusions

This article offers a global overview of the chemometric approaches most commonly used in the field of spectroscopy-based food analysis and authentication. Three different scenarios were surveyed: data exploration, regression and classification. Basic and simple descriptions of the main multivariate techniques employed in such a domain along with a comprehensive outline of their most recent and (as far as the authors are concerned) interesting applications were provided. Data preprocessing- and model validation-related issues were also thoroughly covered.

Author Contributions: Conceptualization, A.B., F.M. and R.V.; methodology, A.B., F.M., C.R. and R.V.; resources, A.B., F.M., C.R. and R.V.; writing—original draft preparation, A.B., F.M. and R.V.; writing—review and editing, A.B., F.M., C.R. and R.V.; supervision, R.V.; project administration, R.V.; funding acquisition, F.M. and C.R. 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 conflict of interest.

References

1. Brereton, R.; Jansen, J.; Lopes, J.; Marini, F.; Pomerantsev, A.; Rodionova, O.; Roger, J.; Walczak, B.; Tauler, R. Chemometrics in analytical chemistry—Part I: History, experimental design and data analysis tools. *Anal. Bioanal. Chem.* **2017**, *409*, 5891–5899. [CrossRef] [PubMed]
2. Brereton, R.; Jansen, J.; Lopes, J.; Marini, F.; Pomerantsev, A.; Rodionova, O.; Roger, J.; Walczak, B.; Tauler, R. Chemometrics in analytical chemistry—Part II: Modeling, validations and applications. *Anal. Bioanal. Chem.* **2018**, *410*, 6691–6704. [CrossRef] [PubMed]
3. Pearson, K. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **1901**, *2*, 559–572. [CrossRef]

4. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, *24*, 417–441. [CrossRef]
5. Saccenti, E.; Camacho, J. Determining the number of components in Principal Components Analysis: A comparison of statistical, crossvalidation and approximated methods. *Chemometr. Intell. Lab.* **2015**, *149*, 99–116. [CrossRef]
6. Kaiser, H. The application of electronic computers to Factor Analysis. *Educ. Psychol. Meas.* **1960**, *20*, 141–151. [CrossRef]
7. Velicer, W. Determining the number of components from the matrix of partial correlations. *Psychometrika* **1976**, *41*, 321–327. [CrossRef]
8. Cattell, R. The scree test for the number of factors. *Multivar. Behav. Res.* **1966**, *1*, 245–276. [CrossRef]
9. Bartlett, M. A note on the multiplying factors for various χ^2 approximations. *J. R. Stat. Soc. B Met.* **1954**, *16*, 296–298. [CrossRef]
10. Saccenti, E.; Smilde, A.; Westerhuis, J.; Hendriks, M. Tracy-Widom statistic for the largest eigenvalue of autoscaled real matrices. *J. Chemometr.* **2011**, *25*, 644–652. [CrossRef]
11. Bro, R.; Kjeldahl, K.; Smilde, A.; Kiers, H. Cross-validation of component models: A critical look at current methods. *Anal. Bioanal. Chem.* **2008**, *390*, 1241–1251. [CrossRef] [PubMed]
12. Camacho, J.; Ferrer, A. Cross-validation in PCA models with the element-wise *k*-fold (*ekf*) algorithm: Theoretical aspects. *J. Chemometr.* **2012**, *26*, 361–373. [CrossRef]
13. Camacho, J.; Ferrer, A. Cross-validation in PCA models with the element-wise *k*-fold (*ekf*) algorithm: Practical aspects. *Chemometr. Intell. Lab.* **2014**, *131*, 37–50. [CrossRef]
14. Horn, J. A rationale and test for the number of factors in factor analysis. *Psychometrika* **1965**, *30*, 179–185. [CrossRef] [PubMed]
15. Dray, S. On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. *Comput. Stat. Data Anal.* **2008**, *52*, 2228–2237. [CrossRef]
16. Vieira, V. Permutation tests to estimate significances on Principal Components Analysis. *Comput. Ecol. Softw.* **2012**, *2*, 103–123.
17. Peres-Neto, P.; Jackson, D.; Somers, K. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Comput. Stat. Data Anal.* **2005**, *49*, 974–997. [CrossRef]
18. Endrizzi, I.; Gasperi, F.; Rødbotten, M.; Næs, T. Interpretation, validation and segmentation of preference mapping models. *Food Qual. Prefer.* **2014**, *32*, 198–209. [CrossRef]
19. Saccenti, E.; Timmerman, M. Considering Horn's parallel analysis from a random matrix theory point of view. *Psychometrika* **2017**, *82*, 186–209. [CrossRef]
20. Vitale, R.; Westerhuis, J.; Næs, T.; Smilde, A.; de Noord, O.; Ferrer, A. Selecting the number of factors in Principal Component Analysis by permutation testing—Numerical and practical aspects. *J. Chemometr.* **2017**, *31*, e2937. [CrossRef]
21. Vitale, R.; Saccenti, E. Comparison of dimensionality assessment methods in Principal Component Analysis based on permutation tests. *Chemometr. Intell. Lab.* **2018**, *181*, 79–94. [CrossRef]
22. Eckart, C.; Young, G. The approximation of one matrix by another of lower rank. *Psychometrika* **1936**, *1*, 211–218. [CrossRef]
23. Comon, P. Independent component analysis, a new concept? *Signal Process.* **1994**, *36*, 287–314. [CrossRef]
24. Hyvärinen, A.; Karhunen, J.; Oja, E. *Independent Component Analysis*, 1st ed.; John Wiley & Sons Inc.: New York, NY, USA, 2001.
25. Jouan-Rimbaud Bouveresse, D.; Benabid, H.; Rutledge, D. Independent component analysis as a pretreatment method for parallel factor analysis to eliminate artefacts from multiway data. *Anal. Chim. Acta* **2007**, *589*, 216–224. [CrossRef] [PubMed]
26. Rutledge, D.; Jouan-Rimbaud Bouveresse, D. Independent Component Analysis with the JADE algorithm. *TRAC-Trends Anal. Chem.* **2013**, *50*, 22–32. [CrossRef]
27. Rutledge, D. Comparison of Principal Components Analysis, Independent Components Analysis and Common Components Analysis. *J. Anal. Test.* **2018**, *2*, 235–248. [CrossRef]
28. Monakhova, Y.; Rutledge, D. Independent components analysis (ICA) at the "cocktail-party" in analytical chemistry. *Talanta* **2020**, *208*, 120451. [CrossRef]
29. Bell, A.; Sejnowski, T. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* **1995**, *7*, 1129–1159. [CrossRef]

30. Hyvärinen, A.; Oja, E. A fast fixed-point algorithm for Independent Component Analysis. *Neural Comput.* **1997**, *9*, 1483–1492. [CrossRef]
31. Cardoso, J.; Souloumiac, A. Blind beamforming for non-Gaussian signals. *IEE Proc. F* **1993**, *140*, 362–370. [CrossRef]
32. Belton, P.; Colquhon, I.; Kemsley, E.; Delgadillo, I.; Roma, P.; Dennis, M.; Sharman, M.; Holmes, E.; Nicholson, J.; Spraul, M. Application of chemometrics to the ^{1}H NMR spectra of apple juices: Discrimination between apple varieties. *Food Chem.* **1998**, *61*, 207–213. [CrossRef]
33. Sikorska, E.; Gorecki, T.; Khmelinskii, I.; Sikorski, M.; De Keukeleire, D. Monitoring beer during storage by fluorescence spectroscopy. *Food Chem.* **2006**, *96*, 632–639. [CrossRef]
34. Wang, L.; Lee, F.; Wang, X.; He, Y. Feasibility study of quantifying and discriminating soybean oil adulteration in Camellia oils by attenuated total reflectance MIR and fiber optic diffuse reflectance NIR. *Food Chem.* **2006**, *95*, 529–536. [CrossRef]
35. Alves, F.; Coqueiro, A.; Março, P.; Valderrama, P. Evaluation of olive oils from the Mediterranean region by UV-Vis spectroscopy and independent component analysis. *Food Chem.* **2019**, *273*, 124–129. [CrossRef]
36. Monakhova, Y.; Tsikin, A.; Kuballa, T.; Lachenmeier, D.; Mushtakova, S. Independent component analysis (ICA) algorithms for improved spectral deconvolution of overlapped signals in ^{1}H NMR analysis: Application to foods and related products. *Magn. Reson. Chem.* **2014**, *52*, 231–240. [CrossRef]
37. Massart, D.; Kaufmann, L. *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*, 1st ed.; John Wiley & Sons Inc.: New York, NY, USA, 1983.
38. Kaufmann, L.; Rousseeuw, P. *Finding Groups in Data: An Introduction to Cluster Analysis*, 1st ed.; John Wiley & Sons Inc.: New York, NY, USA, 1990.
39. Sammut, C.; Webb, G. *Encyclopedia of Machine Learning*, 1st ed.; Springer: Boston, MA, USA, 2010.
40. MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1st ed.; Le Cam, L., Neyman, J., Eds.; University of California Press: Berkeley, CA, USA, 1967; Volume 1, pp. 281–297.
41. Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: An overview. *WIREs Data Min. Knowl.* **2012**, *2*, 86–97. [CrossRef]
42. Marini, F.; Amigo, J. Unsupervised exploration of hyperspectral and multispectral images. In *Hyperspectral Imaging*, 1st ed.; Amigo, J., Ed.; Elsevier B.V.: Amsterdam, The Netherlands, 2020; Volume 2, pp. 93–114.
43. Brown, P. *Measurement, Regression, and Calibration*, 1st ed.; Clarendon Press/Oxford University Press: New York, NY, USA, 1993.
44. Martens, H.; Næs, T. *Multivariate Calibration*, 1st ed.; John Wiley & Sons Inc.: New York, NY, USA, 1989.
45. Oliveri, A. *Introduction to Multivariate Calibration*, 1st ed.; Springer Nature: Cham, Switzerland, 2018.
46. Kjeldahl, J. Neue Methode zur Bstimmung des Stickstoffs in organischen Körpern. *Z. Anal. Chem.* **1883**, *22*, 366–383. [CrossRef]
47. Dumas, J. Lettre de M. Dumas à M. Gay-Lussac sur les procedes de l'analyse organique. *Ann. Chim. Phys.* **1831**, *2*, 198–215.
48. Draper, N.; Smith, H. *Applied Regression Analysis*, 1st ed.; John Wiley & Sons Inc.: New York, NY, USA, 1966.
49. Krzanowski, W. *Principles of Multivariate Analysis*, 1st ed.; Clarendon Press/Oxford University Press: New York, NY, USA, 1988.
50. Johnson, R.; Wichern, D. *Applied Multivariate Statistical Analysis*, 6th ed.; Pearson Education Inc.: Upper Saddle River, NJ, USA, 2007.
51. Dodge, Y. *The Oxford Dictionary of Statistical Terms*, 6th ed.; Oxford University Press: Oxford, UK, 2006.
52. Jolliffe, I. A note on the use of principal components in regression. *J. R. Stat. Soc. C Appl.* **1982**, *31*, 300–303. [CrossRef]
53. Halinski, R.; Feldt, L. The selection of variables in multiple regression analysis. *J. Educ. Meas.* **1970**, *7*, 151–157. [CrossRef]
54. Thompson, M. Selection of variables in multiple regression: Part I. A review and evaluation. *Int. Stat. Rev.* **1978**, *46*, 1–19. [CrossRef]
55. Thompson, M. Selection of variables in multiple regression: Part II. Chosen procedures, computations and examples. *Int. Stat. Rev.* **1978**, *46*, 129–146. [CrossRef]
56. Suárez, E.; Pérez, C.; Rivera, R.; Martínez, M. *Applications of Regression Models in Epidemiology*, 1st ed.; John Wiley & Sons: New York, NY, USA, 2017.

57. Jackson, J. *A User's Guide to Principal Components*, 1st ed.; John Wiley & Sons Inc.: New York, NY, USA, 1991.
58. Mardia, K.; Kent, J.; Bibby, J. *Multivariate Analysis*, 1st ed.; Academic Press: London, UK, 1980.
59. Wold, H. Soft modelling. The basic design and some extensions. In *Systems under Indirect Observation*, 1st ed.; Jöreskog, K., Wold, H., Eds.; North-Holland Publishing Co.: Amsterdam, The Netherlands, 1982; Volume 2, pp. 1–54.
60. Kowalski, B.; Gerlach, R.; Wold, H. Chemical Systems under Indirect Observation. In *Systems under Indirect Observation*, 1st ed.; Jöreskog, K., Wold, H., Eds.; North-Holland Publishing Co.: Amsterdam, The Netherlands, 1982; Volume 2, pp. 191–209.
61. Wold, S.; Martens, H.; Wold, H. The multivariate calibration problem in chemistry solved by the PLS method. In *Matrix Pencils. Lecture Notes in Mathematics*, 1st ed.; Kågström, B., Ruhe, A., Eds.; Springer: Berlin/Heidelberg, Germany, 1983; Volume 973, pp. 286–293.
62. Wold, S.; Ruhe, A.; Wold, H.; Dunn, W., III. The collinearity problem in linear regression. The partial least squares approach to generalized inverses. *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 735–743. [CrossRef]
63. Höskuldsson, A. PLS regression methods. *J. Chemometr.* **1988**, *2*, 211–228. [CrossRef]
64. Höskuldsson, A. *Prediction Methods in Science and Technology*, 1st ed.; Thor Publishing Co.: Copenhagen, Denmark, 1996.
65. Wold, S.; Johansson, E.; Cocchi, M. PLS—Partial Least Squares projections to latent structures. In *3D QSAR in Drug Design, Theory, Methods, and Applications*, 1st ed.; Kubinyi, H., Ed.; ESCOM Science Publishers B.V.: Leiden, The Netherlands, 1993; pp. 523–550.
66. Tenenhaus, M. *La Regression PLS: Theorie et Pratique*, 1st ed.; Editions Technip: Paris, France, 1998.
67. Myers, R.H. *Classical and Modern Regression with Applications*, 1st ed.; Duxbury Press: Boston, MA, USA, 1986.
68. Burnham, A.; MacGregor, J.; Viveros, R. Latent variable multivariate regression modeling. *Chemometr. Intell. Lab.* **1999**, *48*, 167–180. [CrossRef]
69. Allen, D. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **1974**, *16*, 125–127. [CrossRef]
70. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. B Met.* **1974**, *36*, 111–133. [CrossRef]
71. Geisser, S. A predictive approach to the random effect model. *Biometrika* **1974**, *61*, 101–107. [CrossRef]
72. Geladi, P.; Kowalski, B. Partial Least Squares regression: A tutorial. *Anal. Chim. Acta* **1986**, *185*, 1–17. [CrossRef]
73. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemometr. Intell. Lab.* **2001**, *58*, 109–130. [CrossRef]
74. Martens, H.; Næs, T. Multivariate calibration. I. Concepts and distinctions. *TRAC-Trends Anal. Chem.* **1984**, *3*, 204–210. [CrossRef]
75. Centner, V.; Massart, D. Optimization in locally weighted regression. *Anal. Chem.* **1998**, *70*, 4206–4211. [CrossRef] [PubMed]
76. Bevilacqua, M.; Bucci, R.; Materazzi, S.; Marini, F. Application of near infrared (NIR) spectroscopy coupled to chemometrics for dried egg-pasta characterization and egg content quantification. *Food Chem.* **2013**, *140*, 726–734. [CrossRef] [PubMed]
77. Wold, S.; Kettaneh-Wold, N.; Skagerberg, B. Nonlinear PLS modelling. *Chemometr. Intell. Lab.* **1989**, *7*, 53–65. [CrossRef]
78. Wold, S. Nonlinear partial least squares modelling II. Spline inner relation. *Chemometr. Intell. Lab.* **1992**, *14*, 71–84. [CrossRef]
79. Jaeckle, C.; MacGregor, J. Product design through multivariate statistical analysis of process data. *AICHE J.* **1998**, *44*, 1105–1118. [CrossRef]
80. Walczak, B.; Massart, D. The Radial Basis Functions-Partial Least Squares approach as a flexible non-linear regression technique. *Anal. Chim. Acta* **1996**, *331*, 177–185. [CrossRef]
81. Walczak, B.; Massart, D. Application of Radial Basis Functions-Partial Least Squares to non-linear pattern recognition problems: Diagnosis of process faults. *Anal. Chim. Acta* **1996**, *331*, 187–193. [CrossRef]
82. Vitale, R.; Palací-López, D.; Kerkenaar, H.; Postma, G.; Buydens, L.; Ferrer, A. Kernel-Partial Least Squares regression coupled to pseudo-sample trajectories for the analysis of mixture designs of experiments. *Chemometr. Intell. Lab.* **2018**, *175*, 37–46. [CrossRef]
83. Vapnik, V. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, NY, USA, 2000.

84. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, 1st ed.; Cambridge University Press: Cambridge, UK, 2000.
85. Schölkopf, B.; Smola, A. *Leaning with Kernels*, 1st ed.; MIT Press: Cambridge, MA, USA, 2002.
86. Li, H.; Liang, Y.; Xu, Q. Support vector machines and its applications in chemistry. *Chemometr. Intell. Lab.* **2009**, *95*, 188–198. [CrossRef]
87. Gasteiger, J.; Zupan, J. Neural networks in chemistry. *Angew. Chem. Int. Ed. Engl.* **1993**, *32*, 503–527. [CrossRef]
88. Vandeginste, B.; Massart, D.; Buydens, L.; De Jong, S.; Lewi, P.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part B*, 1st ed.; Elsevier B.V.: Amsterdam, The Netherlands, 1998.
89. Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH Verlag: Weinheim, Germany, 1999.
90. Marini, F.; Bucci, R.; Magrì, A.L.; Magrì, A.D. Artificial neural networks in chemometrics: History, examples and perspectives. *Microchem. J.* **2008**, *88*, 178–185. [CrossRef]
91. Marini, F. Non-linear Modeling: Neural Networks. In *Comprehensive Chemometrics*, 2nd ed.; Brown, S., Tauler, R., Walczak, B., Eds.; Elsevier B.V.: Amsterdam, The Netherlands, 2020; Volume 3, pp. 519–541.
92. Che Man, Y.; Syahariza, Z.; Mirghani, M.; Jinap, S.; Bakar, J. Analysis of potential lard adulteration in chocolate and chocolate products using Fourier transform infrared spectroscopy. *Food Chem.* **2005**, *90*, 815–819. [CrossRef]
93. Biancolillo, A.; Santoro, A.; Firmani, P.; Marini, F. Identification and quantification of turmeric adulteration in egg-pasta by near infrared spectroscopy and chemometrics. *Appl. Sci.* **2020**, *10*, 2647. [CrossRef]
94. Croce, R.; Malegori, C.; Oliveri, P.; Medici, I.; Cavaglioni, A.; Rossi, C. Prediction of quality parameters in straw wine by means of FT-IR spectroscopy combined with multivariate data processing. *Food Chem.* **2020**, *305*, 125512. [CrossRef]
95. Biasoli, F.; Gasperi, F.; Aprea, E.; Endrizzi, I.; Framondino, V.; Marini, F.; Mott, D.; Märk, T. Correlation of PTR-MS spectral fingerprints with sensory characterization of flavour and odour profile of "Trentingrana" cheese. *Food Qual. Prefer.* **2006**, *17*, 63–75. [CrossRef]
96. Bevilacqua, M.; Nescatelli, R.; Bucci, R.; Magrì, A.D.; Magrì, A.L.; Marini, F. Chemometrics classification techniques as a tool for solving problems in analytical chemistry. *J. AOAC Int.* **2014**, *97*, 19–28. [CrossRef]
97. Bevilacqua, M.; Bucci, R.; Magrì, A.D.; Magrì, A.L.; Nescatelli, R.; Marini, F. Classification and class-modelling. In *Chemometrics in Food Chemistry*, 1st ed.; Marini, F., Ed.; Elsevier B.V.: Amsterdam, The Netherlands, 2013; Volume 28, pp. 171–233.
98. Brereton, R. *Chemometrics for Pattern Recognition*, 1st ed.; John Wiley & Sons Inc.: New York, NY, USA, 2009.
99. Albano, C.; Dunn III, W.; Edlund, U.; Johansson, E.; Nordén, B.; Sjöström, M.; Wold, S. Four levels of pattern recognition. *Anal. Chim. Acta* **1978**, *103*, 429–443. [CrossRef]
100. Fisher, R. The use of multiple measurements in taxonomic problems. *Ann. Eugenic.* **1936**, *7*, 179–188. [CrossRef]
101. McLachlan, G. *Discriminant Analysis and Statistical Pattern Recognition*, 1st ed.; John Wiley & Sons Inc.: New York, NY, USA, 1992.
102. Tharwat, A. Linear vs. quadratic discriminant analysis classifier: A tutorial. *Int. J. Appl. Pattern Recogn.* **2016**, *3*, 145–180. [CrossRef]
103. Lavine, B.; Davidson, C.; Rayens, W. Machine learning based pattern recognition applied to microarray data. *Comb. Chem. High Throughput Screen.* **2004**, *7*, 115–131. [CrossRef] [PubMed]
104. Liu, Y.; Rayens, W. PLS and dimension reduction for classification. *Comput. Stat.* **2007**, *22*, 189–208. [CrossRef]
105. Liu, Y.; Rayens, W.; Andersen, A.; Smith, C. Partial least squares discrimination with heterogeneous covariance structures. *J. Chemometr.* **2011**, *25*, 109–115. [CrossRef]
106. Friedman, J. Regularized discriminant analysis. *J. Am. Stat. Assoc.* **1989**, *84*, 165–175. [CrossRef]
107. Greene, T.; Rayens, W. Partially pooled covariance matrix estimation in discriminant analysis. *Commun. Stat.* **1989**, *18*, 3679–3702. [CrossRef]
108. Rayens, W. A role for covariance stabilization in the construction of the classical mixture surface. *J. Chemometr.* **1990**, *4*, 159–170. [CrossRef]
109. Rayens, W.; Greene, T. Covariance pooling and stabilization for classification. *Comput. Stat. Data Anal.* **1991**, *11*, 17–42. [CrossRef]
110. Hastie, T.; Buja, A.; Tibshirani, R. Penalized discriminant analysis. *Ann. Stat.* **1995**, *23*, 73–102. [CrossRef]

111. Ripley, B. *Pattern Recognition and Neural Networks*, 1st ed.; Cambridge University Press: Cambridge, UK, 2008.
112. Barker, M.; Rayens, W. Partial least squares for discrimination. *J. Chemometr.* **2003**, *17*, 166–173. [CrossRef]
113. Wold, S.; Albano, C.; Dunn, W.; Esbensen, K.; Hellberg, S.; Johansson, E.; Sjöström, M. Pattern recognition: Finding and using regularities in multivariate data. In *Food Research and Data Analysis*, 1st ed.; Martens, H., Russwurm, H., Jr., Eds.; Applied Science Publishers Ltd.: London, UK, 1983; Volume 3, pp. 147–188.
114. Nocairi, H.; Qannari, E.; 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]
115. Indahl, U.; Martens, H.; Næs, T. From dummy regression to prior probabilities in PLS-DA. *J. Chemometr.* **2007**, *21*, 529–536. [CrossRef]
116. Pérez, N.; 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]
117. Bevilacqua, M.; Marini, F. Local classification: Locally-Weighted-Partial Least Squares-Discriminant Analysis (LW-PLS-DA). *Anal. Chim. Acta* **2014**, *838*, 20–30. [CrossRef] [PubMed]
118. Postma, G.; Krooshof, P.; Buydens, L. Opening the kernel of kernel partial least squares and support vector machines. *Anal. Chim. Acta* **2011**, *705*, 123–134. [CrossRef] [PubMed]
119. Smolinska, A.; Blanchet, L.; Coulier, L.; Ampt, K.; Luider, T.; Hintzen, R.; Wijmega, S.; Buydens, L. Interpretation and visualization of non-linear data fusion in kernel space: Study on metabolomic characterization of multiple sclerosis. *PLoS ONE* **2012**, *7*, e38163. [CrossRef] [PubMed]
120. Vitale, R.; de Noord, O.; Ferrer, A. A kernel-based approach for fault diagnosis in batch processes. *J. Chemometr.* **2014**, *28*, 697–707. [CrossRef]
121. Lu, B.; Ito, M. Task decomposition and module combination based on class relations: A modular neural network for pattern classification. *IEEE Trans. Neural Netw.* **1999**, *10*, 1244–1256.
122. Cheng, H.; Tang, P.; Jin, R. Efficient algorithm for localized support vector machine. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 537–549. [CrossRef]
123. Forina, M.; Oliveri, P.; Bagnasco, L.; Simonetti, R.; Casolino, M.; Nizzi Grifi, F.; Casale, M. Artificial nose, NIR and UV-visible spectroscopy for the characterisation of the PDO Chianti Classico olive oil. *Talanta* **2015**, *144*, 1070–1078. [CrossRef]
124. Vitale, R.; Bevilacqua, M.; Bucci, R.; Magrì, A.D.; Magrì, A.L.; Marini, F. A rapid and non-invasive method for authenticating the origin of pistachio samples by NIR spectroscopy and chemometrics. *Chemometr. Intell. Lab.* **2013**, *121*, 90–99. [CrossRef]
125. Marini, F.; Zupan, J.; Magrì, A.L. On the use of counterpropagation artificial neural networks to characterize Italian rice varieties. *Anal. Chim. Acta* **2004**, *510*, 231–240. [CrossRef]
126. Nasab, S.; Yazd, M.; Marini, F.; Nescatelli, R.; Biancolillo, A. Classification of honey applying high performance liquid chromatography, near-infrared spectroscopy and chemometrics. *Chemometr. Intell. Lab.* **2020**, *202*, 104037. [CrossRef]
127. Grassi, S.; Vitale, R.; Alamprese, C. An exploratory study for the technological classification of egg white powders based on infrared spectroscopy. *LWT-Food Sci. Technol.* **2018**, *96*, 469–475. [CrossRef]
128. 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]
129. Xie, L.; Ying, Y.; Ying, T. Classification of tomatoes with different genotypes by visible and short-wave near-infrared spectroscopy with least-squares support vector machines and other chemometrics. *J. Food Eng.* **2009**, *94*, 34–39. [CrossRef]
130. Ayora-Cañada, M.; Muik, B. Fourier-transform near-infrared spectroscopy as a tool for olive fruit classification and quantitative analysis. *Spectrosc. Lett.* **2005**, *38*, 769–785. [CrossRef]
131. Biancolillo, A.; Firmani, P.; Bucci, R.; Magrì, A.D.; Marini, F. Determination of insect infestation on stored rice by near infrared (NIR) spectroscopy. *Microchem. J.* **2019**, *145*, 252–258. [CrossRef]
132. Biancolillo, A.; Bucci, R.; Magrì, A.L.; Magrì, A.D.; Marini, F. Data-fusion for multiplatform characterization of an Italian craft beer aimed at its authentication. *Anal. Chim. Acta* **2014**, *820*, 23–31. [CrossRef]
133. De Luca, S.; Bucci, R.; Magrì, A.D.; Marini, F. Class modeling techniques in chemometrics: Theory and applications. In *Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation*, 1st ed.; Meyers, R., Ed.; John Wiley & Sons Inc.: New York, NY, USA, 2018; pp. 1–24.

134. Forina, M.; Oliveri, P.; Lanteri, S.; Casale, M. Class-modeling techniques, classic and new, for old and new problems. *Chemometr. Intell. Lab.* **2008**, *93*, 132–148. [CrossRef]
135. Oliveri, P.; Downey, G. Multivariate class modeling for the verification of food-authenticity claims. *TRAC-Trends Anal. Chem.* **2012**, *35*, 74–86. [CrossRef]
136. Marini, F. Classification methods in chemometrics. *Curr. Anal. Chem.* **2010**, *6*, 72–79. [CrossRef]
137. Derde, M.; Massart, D. UNEQ: A disjoint modelling technique for pattern recognition based on normal distribution. *Anal. Chim. Acta* **1986**, *184*, 33–51. [CrossRef]
138. Derde, M.; Massart, D. UNEQ: A class modelling supervised pattern recognition technique. *Microchim. Acta* **1986**, *89*, 139–152. [CrossRef]
139. Wold, S. Pattern recognition by means of disjoint principal component models. *Pattern Recogn.* **1976**, *8*, 127–139. [CrossRef]
140. Wold, S.; Sjöström, M. SIMCA: A method for analyzing chemical data in terms of similarity and analogy. In *Chemometrics: Theory and Application*, 1st ed.; Kowalski, B., Ed.; American Chemical Society: Washington, DC, USA, 1977; Volume 52, pp. 243–282.
141. Defrise-Gussenhoven, E. Ellipses equiprobables et taux d'éloignement en biometric. *Bull. Inst. R. Sci. Nat. Belg.* **1955**, *31*, 1–31.
142. Forina, M.; Lanteri, S.; Sarabia, L. Distance and class space in the UNEQ class-modeling technique. *J. Chemometr.* **1995**, *9*, 69–89. [CrossRef]
143. Pomerantsev, A. Acceptance areas for multivariate classification derived by projection methods. *J. Chemometr.* **2008**, *22*, 601–609. [CrossRef]
144. Rodionova, O.; Oliveri, P.; Pomerantsev, A. Rigorous and compliant approaches to one-class classification. *Chemometr. Intell. Lab.* **2016**, *159*, 89–96. [CrossRef]
145. Pirro, V.; Oliveri, P.; Sciutteri, B.; Salvo, R.; Salomone, A.; Lanteri, S.; Vincenti, M. Multivariate strategies for screening evaluation of harmful drinking. *Bioanalysis* **2013**, *5*, 687–699. [CrossRef]
146. Rodionova, O.; Balyklova, K.; Titova, A.; Pomerantsev, A. Quantitative risk assessment in classification of drugs with identical API content. *J. Pharm. Biomed.* **2014**, *98*, 186–192. [CrossRef]
147. Oliveri, P. Class-modelling in food analytical chemistry: Development, sampling, optimization and validation issues—A tutorial. *Anal. Chim. Acta* **2017**, *982*, 9–19. [CrossRef] [PubMed]
148. Vitale, R.; Marini, F.; Ruckebusch, C. SIMCA modeling for overlapping classes: Fixed or optimized decision threshold? *Anal. Chem.* **2018**, *90*, 10738–10747. [CrossRef] [PubMed]
149. Małyjurek, Z.; Vitale, R.; Walczak, B. Different strategies for class model optimization. A comparative study. *Talanta* **2020**, *215*, 120912. [CrossRef] [PubMed]
150. De Luca, S.; De Filippis, M.; Bucci, R.; Magrì, A.D.; Magrì, A.L.; Marini, F. Characterization of the effects of different roasting conditions on coffee samples of different geographical origins by HPLC-DAD, NIR and chemometrics. *Microchem. J.* **2016**, *129*, 348–361. [CrossRef]
151. Derde, M.; Kaufman, L.; Massart, D. A non-parametric class-modelling technique. *J. Chemometr.* **1989**, *3*, 375–395. [CrossRef]
152. Marini, F.; Zupan, J.; Magrì, A.L. Class-modeling using Kohonen artificial neural networks. *Anal. Chim. Acta* **2005**, *544*, 306–314. [CrossRef]
153. Marini, F.; Magrì, A.L.; Bucci, R. Multilayer feed-forward artificial neural networks for class modeling. *Chemometr. Intell. Lab.* **2007**, *88*, 118–124. [CrossRef]
154. Oliveri, P.; López, M.; Casolino, M.; Ruisánchez, I.; Callao, M.; Medini, L.; Lanteri, S. Partial least squares density modeling (PLS-DM)—A new class-modeling strategy applied to the authentication of olives in brine by near-infrared spectroscopy. *Anal. Chim. Acta* **2014**, *851*, 30–36. [CrossRef]
155. Marini, F.; Bucci, R.; Magrì, A.L.; Magrì, A.D. Authentication of Italian CDO wines by class-modeling techniques. *Chemometr. Intell. Lab.* **2006**, *84*, 164–171. [CrossRef]
156. Forina, M.; Oliveri, P.; Jäger, H.; Römisch, U.; Smeyers-Verbeke, J. Class modeling techniques in the control of the geographical origin of wines. *Chemometr. Intell. Lab.* **2009**, *99*, 127–137. [CrossRef]
157. Di Egidio, V.; Oliveri, P.; Woodcock, T.; Downey, G. Confirmation of brand identity in foods by near infrared transflectance spectroscopy using classification and class-modelling chemometric techniques—The example of a Belgian beer. *Food Res. Int.* **2011**, *44*, 544–549. [CrossRef]

158. Mannina, L.; Marini, F.; Antiochia, R.; Cesa, S.; Magrì, A.L.; Capitani, D.; Sobolev, A. Tracing the origin of beer samples by NMR and chemometrics: Trappist beers as a case study. *Electrophoresis* **2016**, *37*, 2710–2719. [CrossRef] [PubMed]
159. Karunathilaka, S.; Fardin Kia, A.; Srigley, C.; Kyu Chung, J.; Mossoba, M. Nontargeted, rapid screening of extra virgin olive oil products for authenticity using near-infrared spectroscopy in combination with conformity index and multivariate statistical analyses. *J. Food Sci.* **2016**, *81*, C2390–C2397. [CrossRef] [PubMed]
160. Bevilacqua, M.; Bucci, R.; Magrì, A.D.; Magrì, A.L.; Marini, F. Tracing the origin of extra virgin olive oils by infrared spectroscopy and chemometrics: A case study. *Anal. Chim. Acta* **2012**, *717*, 39–51. [CrossRef] [PubMed]
161. Casale, M.; Zunin, P.; Cosulich, M.; Pistarino, E.; Perego, P.; Lanteri, S. Characterisation of table olive cultivar by NIR spectroscopy. *Food Chem.* **2010**, *122*, 1261–1265. [CrossRef]
162. Esteban-Díez, I.; González-Sáiz, J.; Pizarro, C. An evaluation of orthogonal signal correction methods for the characterization of *arabica* and *robusta* coffee varieties by NIRS. *Anal. Chim. Acta* **2004**, *514*, 57–67. [CrossRef]
163. Li, B.; Wang, H.; Zhao, Q.; Ouyang, J.; Wu, Y. Rapid detection of authenticity and adulteration of walnut oil by FTIR and fluorescence spectroscopy: A comparative study. *Food Chem.* **2015**, *181*, 25–30. [CrossRef]
164. 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]
165. Firmani, P.; La Piscopia, G.; Bucci, R.; Marini, F.; Biancolillo, A. Authentication of P.G.I. Gragnano pasta by near infrared (NIR) spectroscopy and chemometrics. *Microchem. J.* **2020**, *152*, 104339. [CrossRef]
166. 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]
167. Nardecchia, A.; Presutto, R.; Bucci, R.; Marini, F.; Biancolillo, A. Authentication of the geographical origin of "Vallerano" chestnut by near infrared spectroscopy coupled with chemometrics. *Food Anal. Method* **2020**, *13*, 1782–1790. [CrossRef]
168. Loftsgaarden, D.; Queesenberry, C. A nonparametric estimate of a multivariate density function. *Ann. Math. Stat.* **1965**, *36*, 1049–1051. [CrossRef]
169. Coomans, D.; Massart, D.; Broeckaert, I.; Tassin, A. Potential methods in pattern recognition: Part 1. Classification aspects of the supervised method ALLOC. *Anal. Chim. Acta* **1981**, *133*, 215–224. [CrossRef]
170. Forina, M.; Armanino, C.; Leardi, R.; Drava, G. A class-modelling technique based on potential functions. *J. Chemometr.* **1991**, *5*, 435–453. [CrossRef]
171. Juricskay, I.; Veress, G. PRIMA: A new pattern recognition method. *Anal. Chim. Acta* **1985**, *171*, 61–76. [CrossRef]
172. Forina, M.; Oliveri, P.; Casale, M.; Lanteri, S. Multivariate range modeling, a new technique for multivariate class modeling: The uncertainty of the estimates of sensitivity and specificity. *Anal. Chim. Acta.* **2008**, *622*, 85–93. [CrossRef]
173. Tax, D.; Duin, R. Support vector domain description. *Pattern Recogn. Lett.* **1999**, *20*, 1191–1199. [CrossRef]
174. Eröss-Kiss, K.; Kiss, Z.; Wiener, E.; Szakálas, G. New data on the evaluation of the infrared (IR) spectra of substances of complicated structure and their application for identification with PRIMA pattern recognition method. Part I. *Period. Polytech. Chem.* **1991**, *35*, 3–22.
175. Jiang, H.; Liu, G.; Xiao, X.; Mei, C.; Ding, Y.; Yu, S. Monitoring of solid-state fermentation of wheat straw in a pilot scale using FT-NIR spectroscopy and support vector data description. *Microchem. J.* **2012**, *102*, 68–74. [CrossRef]
176. Biancolillo, A.; Boqué, R.; Cocchi, M.; Marini, F. Data fusion strategies in food analysis. In *Data Fusion Methodology and Applications*, 1st ed.; Cocchi, M., Ed.; Elsevier B.V.: Amsterdam, The Netherlands, 2019; Volume 31, pp. 271–310.
177. Wold, S.; Hellberg, S.; Lundstedt, T.; Sjostrom, M.; Wold, H. PLS modeling with latent variables in two or more dimensions. In Proceedings of the Symposium on PLS Model Building: Theory and Application, Frankfurt am Main, Germany, 23–25 September 1987.
178. Wold, S.; Kettaneh, N.; Tjessem, K. Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *J. Chemometr.* **1996**, *10*, 463–482. [CrossRef]

179. Westerhuis, J.; Coenegracht, P. Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *J. Chemometr.* **1997**, *11*, 379–392. [CrossRef]
180. Casin, P. A generalization of principal component analysis to K sets of variables. *Comput. Stat. Data Anal.* **2001**, *35*, 417–428. [CrossRef]
181. Smilde, A.; Westerhuis, J.; De Jong, S. A framework for sequential multiblock component methods. *J. Chemometr.* **2003**, *17*, 323–337. [CrossRef]
182. Qannari, E.; Wakeling, I.; Courcoux, P.; MacFie, H. Defining the underlying sensory dimensions. *Food Qual. Prefer.* **2000**, *11*, 151–154. [CrossRef]
183. Mazerolles, G.; Hanafi, M.; Dufour, E.; Bertrand, D.; Qannari, E. Common components and specific weights analysis: A chemometric method for dealing with complexity of food products. *Chemometr. Intell. Lab.* **2006**, *81*, 41–49. [CrossRef]
184. Van Deun, K.; Smilde, A.; Thorrez, L.; Kiers, H.; Van Mechelen, I. Identifying common and distinctive processes underlying multiset data. *Chemometr. Intell. Lab.* **2013**, *129*, 40–51. [CrossRef]
185. Vitale, R.; de Noord, O.; Westerhuis, J.; Smilde, A.; Ferrer, A. *Divide et impera*: How disentangling common and distinctive variability in multiset data analysis can aid industrial process troubleshooting and understanding. *J. Chemometr.* **2020**, in press. [CrossRef]
186. Schouteden, M.; Van Deun, K.; Pattyn, S.; Van Mechelen, I. SCA with rotation to distinguish common and distinctive information in linked data. *Behav. Res. Methods* **2013**, *45*, 822–833. [CrossRef]
187. Trygg, J. O2-PLS for qualitative and quantitative analysis in multivariate calibration. *J. Chemometr.* **2002**, *16*, 283–293. [CrossRef]
188. Qin, S.; Valle, S.; Piovoso, M. On unifying multiblock analysis with application to decentralized process monitoring. *J. Chemometr.* **2001**, *15*, 715–742. [CrossRef]
189. Xu, Y.; Correa, E.; Goodacre, R. Integrating multiple analytical platforms and chemometrics for comprehensive metabolic profiling: Application to meat spoilage detection. *Anal. Bioanal. Chem.* **2013**, *405*, 5063–5074. [CrossRef] [PubMed]
190. Jourdren, S.; Saint-Eve, A.; Panouillé, M.; Lejeune, P.; Déléris, I.; Souchon, I. Respective impact of bread structure and oral processing on dynamic texture perceptions through statistical multiblock analysis. *Food Res. Int.* **2016**, *87*, 142–151. [CrossRef] [PubMed]
191. Guichard, E.; Repoux, M.; Qannari, E.; Laboure, H.; Feron, G. Model cheese aroma perception is explained not only by in vivo aroma release but also by salivary composition and oral processing parameters. *Food Funct.* **2017**, *8*, 615–628. [CrossRef] [PubMed]
192. Campos, M.; Sousa, S.; Pereira, A.; Reis, M. Advanced predictive methods for wine age prediction: Part II—A comparison study of multiblock regression approaches. *Talanta* **2017**, *171*, 132–142. [CrossRef] [PubMed]
193. Brás, L.; Bernardino, S.; Lopes, J.; Menezes, J. Multiblock PLS as an approach to compare and combine NIR and MIR spectra in calibration of soybean flour. *Chemometr. Intell. Lab.* **2005**, *75*, 91–99. [CrossRef]
194. Maléchaux, A.; Laroussi-Mezghan, S.; Le Drêau, Y.; Artaud, J.; Dupuy, N. Multiblock chemometrics for the discrimination of three extra virgin olive oil varieties. *Food Chem.* **2020**, *309*, 125588. [CrossRef]
195. Mehl, F.; Marti, G.; Merle, P.; Delort, E.; Baroux, L.; Sommer, H.; Wolfender, J.; Rudaz, S.; Boccard, J. Integrating metabolomic data from multiple analytical platforms for a comprehensive characterization of lemon essential oils. *Flavour Frag. J.* **2015**, *30*, 131–138. [CrossRef]
196. Teixeira dos Santos, C.; Páscoa, R.; Sarraguca, M.; Porto, P.; Cerdeira, A.; González-Sáiz, J.; Pizarro, C.; Lopes, J. Merging vibrational spectroscopic data for wine classification according to the geographic origin. *Food Res. Int.* **2017**, *102*, 504–510. [CrossRef]
197. Næs, T.; Tomic, O.; Mevik, B.; Martens, H. Path modelling by sequential PLS regression. *J. Chemometr.* **2011**, *25*, 28–40. [CrossRef]
198. Biancolillo, A.; Næs, T. The sequential and orthogonalized PLS regression for multiblock regression: Theory, examples, and extensions. In *Data Fusion Methodology and Applications*, 1st ed.; Cocchi, M., Ed.; Elsevier B.V.: Amsterdam, The Netherlands, 2019; Volume 31, pp. 157–177.
199. Biancolillo, A.; Måge, I.; Næs, T. Combining SO-PLS and linear discriminant analysis for multi-block classification. *Chemometr. Intell. Lab.* **2015**, *141*, 58–67. [CrossRef]
200. Niimi, J.; Tomic, O.; Næs, T.; Jeffery, D.; Bastian, S.; Boss, P. Application of sequential and orthogonalised-partial least squares (SO-PLS) regression to predict sensory properties of Cabernet Sauvignon wines from grape chemical composition. *Food Chem.* **2018**, *256*, 195–202. [CrossRef] [PubMed]

201. Tao, L.; Via, B.; Wu, Y.; Xiao, W.; Liu, X. NIR and MIR spectral data fusion for rapid detection of Lonicera japonica and Artemisia annua by liquid extraction process. *Vib. Spectrosc.* **2019**, *102*, 31–38. [CrossRef]
202. Giannetti, V.; Mariani, M.; Marini, F.; Torrelli, P.; Biancolillo, A. Grappa and Italian spirits: Multiplatform investigation based on GC-MS, MIR and NIR spectroscopies for the authentication of the geographical indication. *Microchem. J.* **2020**, *157*, 104896. [CrossRef]
203. Biancolillo, A.; Marini, F.; D'Archivio, 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]
204. 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]
205. Biancolillo, A.; Foschi, M.; D'Archivio, A. Geographical classification of Italian saffron (*Crocus sativus* L.) by multi-block treatments of UV-Vis and IR spectroscopic data. *Molecules* **2020**, *25*, 2332. [CrossRef]
206. Lauzon-Gauthier, J.; Manolescu, P.; Duchesne, C. The sequential multi-block PLS algorithm (SMB-PLS): Comparison of performance and interpretability. *Chemometr. Intell. Lab.* **2018**, *180*, 72–83. [CrossRef]
207. Biancolillo, A.; Næs, T.; Bro, R.; Måge, I. Extension of SO-PLS to multi-way arrays: SO-N-PLS. *Chemometr. Intell. Lab.* **2017**, *164*, 113–126. [CrossRef]
208. Roger, J.; Biancolillo, A.; Marini, F. Sequential preprocessing through orthogonalization (SPORT) and its application to near infrared spectroscopy. *Chemometr. Intell. Lab.* **2020**, *199*, 103975. [CrossRef]
209. Biancolillo, A.; Marini, F.; Roger, J. SO-CovSel: A novel method for variable selection in a multiblock framework. *J. Chemometr.* **2020**, *34*, e3120. [CrossRef]
210. El Ghaziri, A.; Cariou, V.; Rutledge, D.; Qannari, E. Analysis of multiblock datasets using ComDim: Overview and extension to the analysis of (K + 1) datasets. *J. Chemometr.* **2016**, *30*, 420–429. [CrossRef]
211. Cariou, V.; Qannari, E.; Rutledge, D.; Vigneau, E. ComDim: From multiblock data analysis to path modeling. *Food Qual. Prefer.* **2018**, *67*, 27–34. [CrossRef]
212. Ríos-Reina, R.; Callejón, R.; Savorani, F.; Amigo, J.; Cocchi, M. Data fusion approaches in spectroscopic characterization and classification of PDO wine vinegards. *Talanta* **2019**, *198*, 560–572. [CrossRef]
213. Lock, E.; Hoadley, K.; Marron, J.; Nobel, A. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.* **2013**, *7*, 523–542. [CrossRef] [PubMed]
214. Berglund, A.; Wold, S. A serial extension of multiblock PLS. *J. Chemometr.* **1999**, *13*, 461–471. [CrossRef]
215. Reis, M. Network-induced supervised learning: Network-induced classification (NI-C) and network-induced regression (NI-R). *AIChE J.* **2013**, *59*, 1570–1587. [CrossRef]
216. Måge, I.; Menichelli, E.; Næs, T. Preference mapping by PO-PLS: Separating common and unique information in several data blocks. *Food Qual. Prefer.* **2012**, *24*, 8–16. [CrossRef]
217. Bougeard, S.; Qannari, E.; Rose, N. Multiblock redundancy analysis: Interpretation tools and application in epidemiology. *J. Chemometr.* **2011**, *25*, 467–475. [CrossRef]
218. Löfstedt, T.; Trygg, J. OnPLS—A novel multiblock method for the modelling of predictive and orthogonal variation. *J. Chemometr.* **2011**, *25*, 441–455. [CrossRef]
219. Nguyen, Q.; Liland, K.; Tomic, O.; Tarrega, A.; Varela, P.; Næs, T. SO-PLS as an alternative approach for handling multi-dimensionality in modelling different aspects of consumer expectations. *Food Res. Int.* **2020**, *133*, 109189. [CrossRef]
220. Næs, T.; Romano, R.; Tomic, O.; Måge, I.; Smilde, A.; Liland, K. Sequential and orthogonalized PLS (SO-PLS) regression for path analysis: Order of blocks and relations between effects. *J. Chemometr.* **2020**, in press.
221. Ruckebusch, C. *Resolving Spectral Mixtures with Applications from Ultrafast Time-Resolved Spectroscopy to Super-Resolution Imaging*, 1st ed.; Elsevier B.V.: Amsterdam, The Netherlands, 2016.
222. Tauler, R.; Smilde, A.; Kowalski, B. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *J. Chemometr.* **1995**, *9*, 31–58. [CrossRef]
223. Tauler, R. Multivariate curve resolution applied to second order data. *Chemometr. Intell. Lab.* **1995**, *30*, 133–146. [CrossRef]
224. Ruckebusch, C.; Blanchet, L. Multivariate curve resolution: A review of advanced and tailored applications and challenges. *Anal. Chim. Acta* **2013**, *765*, 28–36. [CrossRef] [PubMed]
225. de Juan, A.; Jaumot, J.; Tauler, R. Multivariate curve resolution (MCR). Solving the mixture analysis problems. *Anal. Methods* **2014**, *6*, 4964–4976. [CrossRef]

226. Jaumot, J.; Gargallo, R.; de Juan, A.; Tauler, R. A graphical user-friendly interface for MCR-ALS: A new tool for multivariate curve resolution in MATLAB. *Chemometr. Intell. Lab.* **2005**, *76*, 101–110. [CrossRef]
227. Jaumot, J.; de Juan, A.; Tauler, R. MCR-ALS GUI 2.0: New features and applications. *Chemometr. Intell. Lab.* **2015**, *140*, 1–12. [CrossRef]
228. Maeder, M. Evolving factor analysis for the resolution of overlapping chromatographic peaks. *Anal. Chem.* **1987**, *59*, 527–530. [CrossRef]
229. Windig, W.; Guilment, J. Interactive self-modeling mixture analysis. *Anal. Chem.* **1991**, *63*, 1425–1432. [CrossRef]
230. Abdollahi, H.; Tauler, R. Uniqueness and rotation ambiguity in multivariate curve resolution methods. *Chemometr. Intell. Lab.* **2011**, *108*, 100–111. [CrossRef]
231. Golshan, A.; Abdollahi, H.; Beyramysoltan, S.; Maeder, M.; Neymeyr, K.; Rajkó, R.; Sawall, M.; Tauler, R. A review of recent methods for the determination of ranges of feasible solutions resulting from soft modelling analyses of multivariate data. *Anal. Chim. Acta* **2016**, *911*, 1–13. [CrossRef]
232. Tauler, R.; Izquierdo-Ridorsa, A.; Casassas, E. Simultaneous analysis of several spectroscopic titrations with self-modelling curve resolution. *Chemometr. Intell. Lab.* **1993**, *18*, 293–300. [CrossRef]
233. Bro, R.; De Jong, S. A fast non-negativity-constrained least squares algorithm. *J. Chemometr.* **1997**, *11*, 393–401. [CrossRef]
234. Esteban, M.; Ariño, J.; Díaz-Cruz, J.; Díaz-Cruz, M.; Tauler, R. Multivariate curve resolution with alternating least squares optimisation: A soft-modelling approach to metal complexation studies by voltammetric techniques. *TRAC-Trends Anal. Chem.* **2000**, *19*, 49–61. [CrossRef]
235. Tauler, R.; Maeder, M.; de Juan, A. Multiset Data Analysis: Extended Multivariate Curve Resolution. In *Comprehensive Chemometrics*, 2nd ed.; Brown, S., Tauler, R., Walczak, B., Eds.; Elsevier B.V.: Amsterdam, The Netherlands, 2020; Volume 2, pp. 305–336.
236. Grassi, S.; Alamprese, C.; Bono, V.; Casiraghi, E.; Amigo, J. Modelling milk lactic acid fermentation using multivariate curve resolution-alternating least squares (MCR-ALS). *Food Bioprocess Technol.* **2014**, *7*, 1819–1829. [CrossRef]
237. Grassi, S.; Strani, L.; Casiraghi, E.; Alamprese, C. Control and monitoring of milk renneting using FT-NIR spectroscopy as a process analytical technology tool. *Foods* **2019**, *8*, 405. [CrossRef]
238. Strani, L.; Grassi, S.; Alamprese, C.; Casiraghi, E.; Ghiglietti, R.; Locci, F.; Pricca, N.; de Juan, A. Effect of physicochemical factors and use of milk powder on milk rennet-coagulation: Process understanding by near infrared spectroscopy and chemometrics. *Food Control.* **2020**, in press.
239. Grassi, S.; Amigo, J.; Lyndgaard, C.; Foschino, R.; Casiraghi, E. Assessment of the sugars and ethanol development in beer fermentation with FT-IR and multivariate curve resolution models. *Food Res. Int.* **2014**, *62*, 602–608. [CrossRef]
240. Fisher, R. *The Design of Experiments*, 5th ed.; Oliver & Boyd: Edinburgh, UK, 1951.
241. Box, G.; Hunter, J.; Hunter, W. *Statistics for Experimenters: Design, Innovation and Discovery*, 2nd ed.; John Wiley & Sons Inc.: New York, NY, USA, 2005.
242. Montgomery, D. *Design and Analysis of Experiments*, 8th ed.; John Wiley & Sons Inc.: New York, NY, USA, 2012.
243. Leardi, R. Experimental design in chemistry: A tutorial. *Anal. Chim. Acta.* **2009**, *652*, 161–172. [CrossRef]
244. Morrison, D. *Multivariate Statistical Methods*, 4th ed; Duxbury Press: Boston, MA, USA, 2003.
245. Fisher, R. The correlation between relatives on the supposition of mendelian inheritance. *Philos. Trans. R. Soc. Edinb.* **1918**, *52*, 399–433. [CrossRef]
246. Fisher, R. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* **1921**, *1*, 3–32.
247. Fisher, R. *Statistical Methods for Research Workers*, 1st ed.; Oliver & Boyd: Edinburgh, UK, 1925.
248. Ståhle, L.; Wold, S. Analysis of variance (ANOVA). *Chemometr. Intell. Lab.* **1989**, *6*, 259–272. [CrossRef]
249. Cooley, W.; Lohnes, P. *Multivariate Data Analysis*, 1st ed.; John Wiley & Sons Inc.: New York, NY, USA, 1971.
250. Scheffé, H. *The Analysis of Variance*, 1st ed.; John Wiley & Sons Inc.: New York, NY, USA, 1959.
251. Bray, J.; Maxwell, S. *Multivariate Analysis of Variance*, 1st ed.; SAGE Publications Inc.: Beverly Hills, CA, USA, 1986.
252. Ståhle, L.; Wold, S. Multivariate analysis of variance (MANOVA). *Chemometr. Intell. Lab.* **1990**, *9*, 127–141. [CrossRef]

253. Jansen, J.; Hoefsloot, H.; van der Greef, J.; Timmerman, M.; Westerhuis, J.; Smilde, A. ASCA: Analysis of multivariate data obtained from an experimental design. *J. Chemometr.* **2005**, *19*, 469–481. [CrossRef]
254. Smilde, A.; Jansen, J.; Hoefsloot, H.; Lamers, R.; van der Greef, J.; Timmerman, M. ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data. *Bioinformatics* **2005**, *21*, 3043–3048. [CrossRef]
255. Anderson, M.; ter Braak, C. Permutation tests for multi-factorial analysis of variance. *J. Stat. Comput. Sim.* **2003**, *73*, 85–113. [CrossRef]
256. Vis, D.; Westerhuis, J.; Smilde, A.; van der Greef, J. Statistical validation of megavariate effects in ASCA. *BMC Bioinform.* **2007**, *8*, 322. [CrossRef]
257. de Boves Harrington, P.; Vieira, N.; Chen, P.; Espinoza, J.; Nien, J.; Romero, R.; Yergey, A. Analysis of variance-principal component analysis: A soft tool for proteomic discovery. *Anal. Chim. Acta* **2005**, *544*, 118–127. [CrossRef]
258. de Boves Harrington, P.; Vieira, N.; Espinoza, J.; Nien, J.; Romero, R.; Yergey, A. Proteomic analysis of amniotic fluids using analysis of variance-principal component analysis and fuzzy rule-building expert systems applied to matrix-assisted laser desorption/ionization mass spectrometry. *Chemometr. Intell. Lab.* **2006**, *82*, 283–293. [CrossRef]
259. Marini, F.; de Beer, D.; Joubert, E.; Walczak, B. Analysis of variance of designed chromatographic data sets: The analysis of variance-target projection approach. *J. Chromatogr. A* **2015**, *1405*, 94–102. [CrossRef]
260. Marini, F.; de Beer, D.; Walters, N.; de Villiers, A.; Joubert, E.; Walczak, B. Multivariate analysis of variance of designed chromatographic data. A case study involving fermentation of rooibos tea. *J. Chromatogr. A* **2017**, *1489*, 115–125. [CrossRef] [PubMed]
261. Marini, F.; Walczak, B. ANOVA-Target Projection (ANOVA-TP). In *Comprehensive Chemometrics*, 2nd ed.; Brown, S., Tauler, R., Walczak, B., Eds.; Elsevier B.V.: Amsterdam, The Netherlands, 2020; Volume 1, pp. 495–520.
262. Bouveresse, D.; Pinto, R.; Schmidtke, L.; Locquet, N.; Rutledge, D. Identification of significant factors by an extension of ANOVA-PCA based on multi-block analysis. *Chemometr. Intell. Lab.* **2011**, *106*, 173–182. [CrossRef]
263. Engel, J.; Blanchet, L.; Bloemen, B.; van den Heuvel, L.; Engelke, U.; Wevers, R.; Buydens, L. Regularized MANOVA (rMANOVA) in untargeted metabolomics. *Anal. Chim. Acta* **2015**, *899*, 1–12. [CrossRef] [PubMed]
264. Firmani, P.; Vitale, R.; Ruckebusch, C.; Marini, F. ANOVA-simultaneous component analysis modelling of low-level-fused spectroscopic data: A food chemistry case-study. *Anal. Chim. Acta.* **2020**, *1125*, 308–314. [CrossRef]
265. Zhang, X.; de Juan, A.; Tauler, R. Multivariate curve resolution applied to hyperspectral imaging analysis of chocolate samples. *Appl. Spectrosc.* **2011**, *69*, 993–1003. [CrossRef]
266. Neves, M.; Poppi, R. Monitoring of adulteration and purity in coconut oil using Raman spectroscopy and multivariate curve resolution. *Food Anal. Method* **2018**, *11*, 1897–1905. [CrossRef]
267. Babellahi, F.; Amodio, M.; Marini, F.; Chaudry, M.; de Chiara, M.; Mastrandrea, L.; Colelli, G. Using chemometrics to characterise and unravel the near infra-red spectral changes induced in aubergine fruit by chilling injury as influenced by storage time and temperature. *Biosyst. Eng.* **2020**, *198*, 137–146. [CrossRef]
268. Roger, J.; Boulet, J.; Zeaiter, M.; Rutledge, D. Pre-processing Methods. In *Comprehensive Chemometrics*, 2nd ed.; Brown, S., Tauler, R., Walczak, B., Eds.; Elsevier B.V.: Amsterdam, The Netherlands, 2020; Volume 3, pp. 1–75.
269. Bro, R.; Smilde, A. Centering and scaling in component analysis. *J. Chemometr.* **2003**, *17*, 16–33. [CrossRef]
270. Kimball, B. Smoothing data with Fourier transformations. *Agron. J.* **1974**, *66*, 259–262. [CrossRef]
271. Walczak, B.; Massart, D. Noise suppression and signal compression using the wavelet packet transform. *Chemometr. Intell. Lab.* **1997**, *36*, 81–94. [CrossRef]
272. Savitzky, A.; Golay, M. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **1964**, *36*, 1627–1639. [CrossRef]
273. Barnes, R.; Dhanoa, M.; Lister, S. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.* **1989**, *43*, 772–777. [CrossRef]
274. Geladi, P.; MacDougall, D.; Martens, H. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Appl. Spectrosc.* **1985**, *39*, 491–500. [CrossRef]

275. Martens, H.; Stark, E. Extendend multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy. *J. Pharm. Biomed.* **1991**, *9*, 625–635. [CrossRef]
276. Eilers, P. Parametric time warping. *Anal. Chem.* **2004**, *76*, 404–411. [CrossRef] [PubMed]
277. Westad, F.; Marini, F. Validation of chemometric models—A tutorial. *Anal. Chim. Acta* **2015**, *893*, 14–24. [CrossRef] [PubMed]
278. Kennard, R.; Stone, L. Computer aided design of experiments. *Technometrics* **1969**, *11*, 137–148. [CrossRef]
279. Snee, R. Validation of regression models: Methods and examples. *Technometrics* **1977**, *19*, 415–428. [CrossRef]
280. Daszykowski, M.; Walczak, B.; Massart, D. Representative subset selection. *Anal. Chim. Acta* **2002**, *468*, 91–103. [CrossRef]
281. Wu, W.; Walczak, B.; Massart, D.; Heuerding, S.; Erni, F.; Last, I.; Prebble, K. Artificial neural networks in classification of NIR spectral data: Design of the training set. *Chemometr. Intell. Lab.* **1996**, *33*, 35–46. [CrossRef]
282. Efron, B. *The Jacknife, the Bootstrap and Other Resampling Plans*, 1st. ed.; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 1982.