CHAPTER TEN

Chemometric Methods for Classification and Feature Selection

Marina Cocchi*, Alessandra Biancolillo†, Federico Marini†,1
*Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Modena, Italy
†Department of Chemistry, University of Rome La Sapienza, Rome, Italy
[1]Corresponding author: e-mail address: federico.marini@uniroma1.it

Contents

1. MULTIVARIATE CLASSIFICATION

In chapter "Exploratory analysis of metabolomic data", by Stanimirova and Daszykowski, attention was focused on the chemometric approaches for the exploratory analysis [1] of multivariate data sets in the framework of the -omic disciplines, i.e., on those techniques (mostly based on bilinear projections) which allow capturing the essential characteristics of the samples under investigation, providing a sort of "snapshot" of the system by means of highly informative plots. On the other hand, quite often experimental data are collected with the aim of predicting the value of one or more properties of the system (responses), which can be of quantitative or qualitative nature [2]. In particular, the case where the response to be predicted

Comprehensive Analytical Chemistry, Volume 82
ISSN 0166-526X
https://doi.org/10.1016/bs.coac.2018.08.006

is qualitative, i.e., when it may assume only a discrete set of values (and there is not necessarily an ordered relationship among them), is of particular relevance in the -omic field, as many problems of interest can be formulated in such terms, and it is the domain of application of classification methods [3–6]. Indeed, a qualitative variable induces a categorization, according to which each of the values it can take is said to be a class (or category): for instance, if one is interested in investigating the metabolic effects of different dosages of a drug onto a set of individuals, and the response (drug dose) is studied at three levels (low, medium and high dose), this can be formulated as classification problems involving three classes (each one identically corresponding to a particular level of the qualitative response variable). Another example could be the possibility of differentiating between healthy and ill patients based on the collected experimental data: in such case, the qualitative response to be predicted would be the health status of the individuals and its two discrete values (classes) "ill" and "healthy", respectively.

Classification methods are chemometric tools building models aiming at predicting which class (qualitative attribute) more accurately describes the individuals under investigation based on the experimental data collected. In mathematical terms, the statement above can be formulated as:

$$c_i = f(\boldsymbol{x}_i) \qquad c_i \in \{1, 2, \ldots, C\} \tag{1}$$

where $\boldsymbol{x}_i$ is the vector of measurements collected on the ith sample, c_i is the predicted class for the same individual (which can be one of the C possible values of the qualitative response) and f is a generic function relating the experimental data to the predicted outcome. In Eq. (1), a generic numerical coding for the values of the qualitative response was adopted, so that $c_i = 1$ indicates one possible value of the response (class 1, e.g. "low dose"), $c_i = 2$ another possible value of the response (class 2, e.g. "medium dose") and so on. Here it is important to stress that, in order to be able to define as accurately as possible the functional relationship generally described by Eq. (1), the possibility of building a predictive (classification) model relies on the availability of a (representative) set of samples for which not only the experimental data ($\boldsymbol{x}_i$) have been measured but also the corresponding "true" class labels c_i are known (training set).

1.1 Discriminant vs. Modelling Tools

Looking for classification methods in the literature, despite each of them has its own characteristics, which make it peculiar and suitable for specific

purposes, one could find herself/himself caught up into the several proposed approaches in this field; a little taxonomy, providing similarities/differences among the different methods, could address to the most suitable approach for a given problem.

One first distinction, which could help to circumstantiate the number of interesting approaches, is the one between *discriminant* and *class-modelling* methods [5,7].

One typical feature which differentiates these two families of methods is the fact that discriminant classifiers focus more on dissimilarities among objects belonging to diverse categories, while class-modelling approaches exploit interclass similarities in order to solve the classification problem. Applying a discriminant method, the multivariate space of the samples is divided into as many regions as the number of classes present in the system under study. As a consequence of this, the application of this kind of classifiers provides unique class assignments, i.e., each sample is predicted as belonging to one and only category. Considering a three-class problem (for instance, categories A, B and C) the application of discriminant classifiers leads to the case displayed in Fig. 1A. A discriminant method allows defining three class boundaries which divide the multivariate sample space in as many class regions: a sample will be classified according to the class region it falls into. In the example represented in Fig. 1A, the space is divided into three class regions: a sample falling into the region of category A (red squares) will be considered as belonging to this class, one present into the space of class B (green diamonds) will be assigned to this category while one falling into class region C (blue circles) will be predicted as belonging to the C class.

Conversely, as mentioned above, using class-modelling classifiers, the definition of class boundaries is based on similarities among intercategory samples, and each class region is defined independently from the others [8]. Briefly, applying a class-modelling method means defining, per each class in the system (or per a specific class of interest), the region in the multivariate sample space, where an object belonging to that specific category is the most likely expected to fall. As a consequence of this, different from the discriminant analysis, the multivariate space of samples is not completely divided into class regions, but they only partially cover it. Moreover, since each class boundary is individually defined, they can overlap, leading to regions where sample belonging to different categories could fall into. More specifically, taking once again into account a three-class case, the application of class-modelling methods leads to three different scenarios (displayed in Fig. 1B). In fact, a sample could fall into a class region, and it is therefore

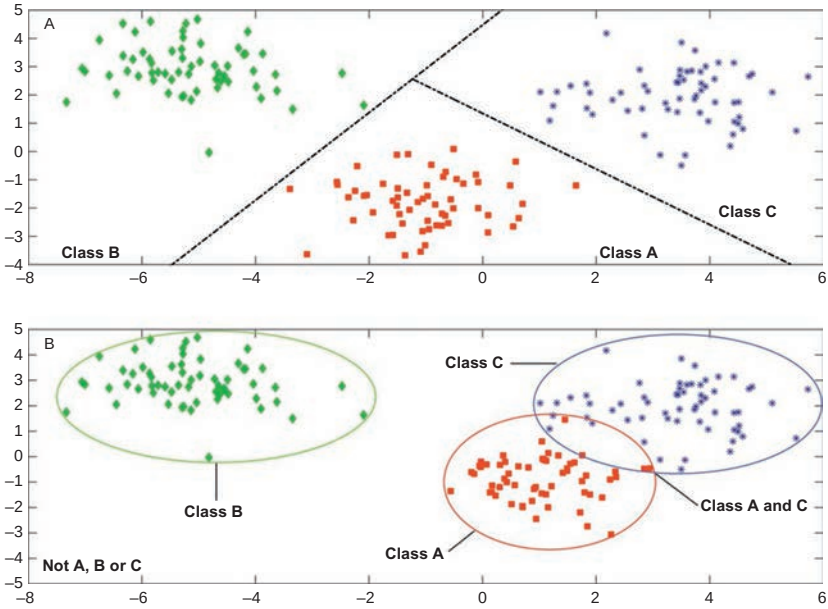

Fig. 1 Illustration of discriminant (A) and modelling (B) classification approaches.

assigned to that category (for instance, objects represented as green diamonds in Fig. 1B, assigned to class B); otherwise, it could be accepted by more than one class model (falling in an overlapping regions between two categories, like some objects represented as red squares or blue dots in Fig. 1B, which are accepted by both class A or class C models) meaning it has the same probability of belonging to all of them (and it is called *confused sample*). Finally, an object could be rejected by all the models (falling outside any class region), and it is therefore not assigned to any category.

1.2 Discriminant Methods

As explained above, discriminant methods operate a "strict" classification by associating with each of the samples under investigation one and only one of the possible values of the qualitative response variable, i.e., assigning each to one and only one of the available classes. Considering the geometric interpretation discussed in Section 1.1, they operate by identifying multivariate surfaces (decision boundaries) which partition the variable hyperspace into as many regions as the number of categories, so that there is a one-to-one correspondence between the coordinates of the samples ($\boldsymbol{x}_i$) and their predicted class. These decision surfaces are calculated in a way as to minimize some sort

of error criterion for the training samples, the most common of which is the overall classification error E_{TOT} defined as:

$$E_{TOT} = \frac{\sum_{i=1}^{N_{TOT}} e_i}{N_{TOT}} \tag{2}$$

where N_{TOT} is the total number of training samples, and e_i represents the error on the prediction of the ith sample

$$e_i = \begin{cases} 0 & \text{if } \widehat{c}_i = c_i \\ 1 & \text{if } \widehat{c}_i \neq c_i \end{cases} \tag{3}$$

$\widehat{c}$ and c_i being the predicted and the true value of the class label for that individual, respectively. The criterion in Eq. (2) assumes that all the classification errors are equally costly, i.e., misclassifying an individual from class1 by predicting it as class2 is as bad as wrongly assigning a sample from class2 to class1. However, this assumption is not always true, especially in biomedical applications, where, for instance, the consequences of wrongly predicting that a person is ill when instead he/she is healthy (false positive) are less severe than those of wrongly predicting that a person is healthy when, in reality, he/she is ill (false negative). In such cases, the decision boundaries can be calculated in a way as to minimize a loss function which, together with the classification errors, also includes the misclassification costs and, in the case where only two classes are present, takes the form

$$E^* = l_{21}E_{21} + l_{12}E_{12} = l_{21}\frac{\sum_{j=1}^{N_1} e_j}{N_1} + l_{12}\frac{\sum_{k=1}^{N_2} e_k}{N_2} \tag{4}$$

where l_{21} and l_{12} are the cost of wrongly predicting as class 2 a sample from class 1 and vice versa, respectively, while E_{21} and E_{12} are the percentage of class 1 samples and class 2 samples wrongly predicted, respectively [9]. The impact of the introduction of a classification cost into the error criterion is graphically shown in Fig. 2 for the two-class case. In particular, let us assume that the aim is to build a classification model which, on the basis of the measured variables (here, only two, for the sake of graphical simplicity), allows predicting whether an individual is healthy or ill. When assuming that all the classification errors are equally costly (Fig. 2A), the decision boundary is placed so that there is an equal probability of false positives and false negatives. On the other hand, when introducing a higher misclassification cost

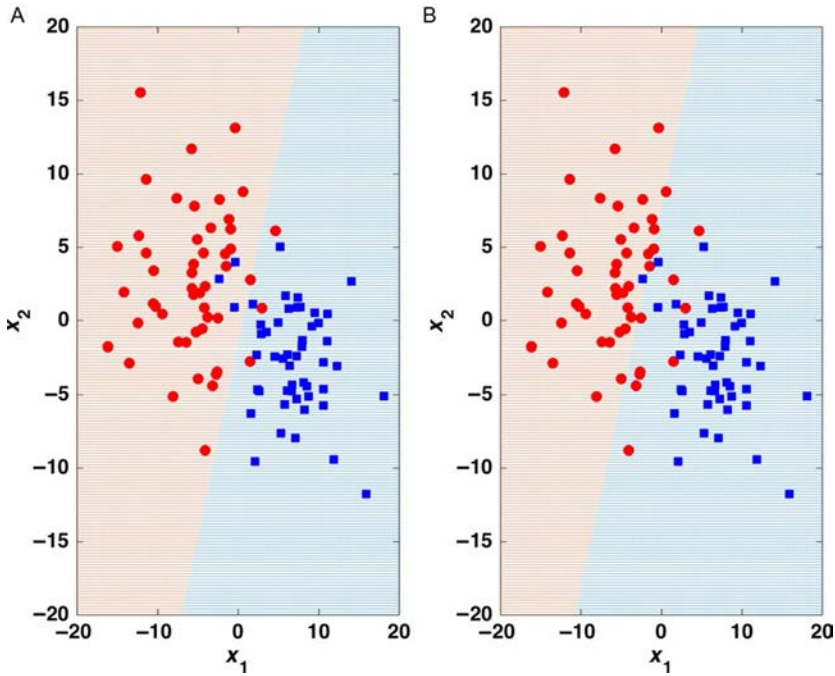

Fig. 2 Illustration of the effect of the cost of misclassification on the decision threshold for a general linear classifier in a two-class problem. As an example, the case of discrimination between healthy (*red circles*) and ill (*blue squares*) individuals is considered. (A) Decision threshold is based only on minimum classification error: number of false positives (4) is almost equal to the number of false negatives (3). (B) Decision threshold takes into account the cost of misclassification: there are no false negatives, but the number of false positives increases significantly (11).

for false negatives (Fig. 2B), the decision boundary is moved closer to the healthy samples and allows a higher rate of false positives.

In general, building a classification model involves the definition of the boundaries which separate the regions associated with the different categories in the multivariate space or, in other words, the formulation of a classification rule which is nothing else than the functional relationship described in Eq. (1), relating the experimental data to the predicted class label. In particular, when the error criterion to be minimized is the one in Eq. (2), assuming equal misclassification costs for all the categories, the classification rule underlying all discriminant approaches is the so-called Bayes' rule, which states that a sample should be assigned to the class it has the highest probability of belonging to [10]. Accordingly, Bayes' rule represents the probabilistic core of all the discriminant classification

approaches, even if not all the available methods rely on it for the actual calculation of the optimal model parameters. The probability that a sample belongs to the gth category estimated based on the measurements (*posterior*) $p(c_i = g|\boldsymbol{x}_i)$ is defined as:

$$p(c_i = g|\boldsymbol{x}_i) \propto p(\boldsymbol{x}_i|c_i = g)\omega_g \tag{5}$$

where $p(\boldsymbol{x}_i|c_i = g)$ is the probability of observing the vector $\boldsymbol{x}_i$ from samples truly coming from class g (likelihood) and ω_g is the a priori (i.e. before taking any measurement) probability of observing an individual from that class; the proportionality sign indicates that a normalization constant (necessary for the probabilities to sum up to one) is not considered in the equation. Building classification models according to the Bayes' rule, then, involves calculating the probability that a sample belongs to each of the C investigated categories:

$$p(c_i = 1|\boldsymbol{x}_i), p(c_i = 2|\boldsymbol{x}_i), \ldots, p(c_i = C|\boldsymbol{x}_i) \tag{6}$$

and assigning the individual to the class corresponding to the highest value of the posterior probability. On the other hand, not all the discriminant approaches make explicit use of Bayes' rule (even if it still remains underlying) to calculate the decision boundaries, as the classification rule can be more straightforwardly expressed in other terms (e.g. based on distances). In the remainder of this section, the discriminant methods most commonly used in the context of -omic applications will be illustrated and discussed.

1.2.1 Linear Discriminant Analysis

Linear discriminant analysis (LDA) initially proposed by Fisher in 1936 was the first and it is still one of the most commonly used classification method in the literature [11]. As the name suggests, it is a linear technique, meaning that these decision boundaries separating the regions of the variable hyperspace associated to the different categories are linear surfaces (i.e. hyperplanes), whose calculation directly follows from Bayes' rule, as LDA is a probabilistic (parametric) method. In particular, LDA assumes that, for each category, the likelihood in Eq. (5) follows a multivariate normal distribution, whose variance–covariance matrix $\boldsymbol{S}$ is the same for all the classes:

$$p(\boldsymbol{x}_i|c_i = g) = \frac{1}{(2\pi)^{\frac{v}{2}}|\boldsymbol{S}|} e^{-\frac{1}{2}\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_g\right)^T \boldsymbol{S}^{-1}\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_g\right)} \tag{7}$$

where $\overline{\boldsymbol{x}}_g$ is the barycentre (centroid) of the gth category and v is the number of measured variable (i.e. the dimensionality of the hyperspace).

By substituting Eq. (7) into Eq. (5), one obtains that the posterior probability that a sample belongs to the gth category, under the assumptions of LDA, becomes:

$$p(c_i = g|\boldsymbol{x}_i) = \frac{\omega_g}{K_{norm}\ (2\pi)^{\frac{p}{2}}|\boldsymbol{S}|} e^{-\frac{1}{2}\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_g\right)^T \boldsymbol{S}^{-1}\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_g\right)} \tag{8}$$

where the normalization constant K_{norm} is the same for all the classes.

Since, according to Bayes' rule, a sample should be assigned to the class it has the highest posterior probability of belonging to, in a probabilistic technique such as LDA, the decision boundaries, separating the portions of the multivariate hyperspace, associated to the different categories, are characterized by being the surfaces where:

$$p(c_i = j|\boldsymbol{x}_i) = p(c_i = k|\boldsymbol{x}_i) \ \ \forall j, k \tag{9}$$

i.e., there is an equal probability for an individual to belong to each of the classes on the two sides of the delimiter. When inserting Eq. (8) into (9) and taking the logarithm of both sides of the equality, after cancelling out all the terms which are the same for both categories one obtains:

$$\overline{\boldsymbol{x}}_j{}^T \boldsymbol{S}^{-1}\boldsymbol{x} - \frac{1}{2}\overline{\boldsymbol{x}}_j{}^T \boldsymbol{S}^{-1}\overline{\boldsymbol{x}}_j + \ln(\omega_j) = \overline{\boldsymbol{x}}_k{}^T \boldsymbol{S}^{-1}\boldsymbol{x} - \frac{1}{2}\overline{\boldsymbol{x}}_k{}^T \boldsymbol{S}^{-1}\overline{\boldsymbol{x}}_k + \ln(\omega_k) \tag{10}$$

which is the equation of a linear surface (i.e. a hyperplane):

$$\boldsymbol{w}^T\boldsymbol{x} - w_0 = 0 \tag{11}$$

where

$$\boldsymbol{w}^T = \left(\overline{\boldsymbol{x}}_j - \overline{\boldsymbol{x}}_k\right)^T \boldsymbol{S}^{-1} \tag{12}$$

and

$$w_0 = \frac{1}{2}\overline{\boldsymbol{x}}_j{}^T \boldsymbol{S}^{-1}\overline{\boldsymbol{x}}_j - \frac{1}{2}\overline{\boldsymbol{x}}_k{}^T \boldsymbol{S}^{-1}\overline{\boldsymbol{x}}_k - \ln\left(\frac{\omega_j}{\omega_k}\right). \tag{13}$$

An example of a linear decision boundary (hyperplane) calculated by LDA for a problem involving two classes in three dimensions is graphically shown in Fig. 3.

Since it explicitly assumes a specific probability distribution function (in particular, the multivariate Gaussian), LDA is said to be a *parametric* method: calculation of the model parameters, i.e., the coefficients $\boldsymbol{w}$ and w_0 identifying the decision boundaries, just requires the estimation of the

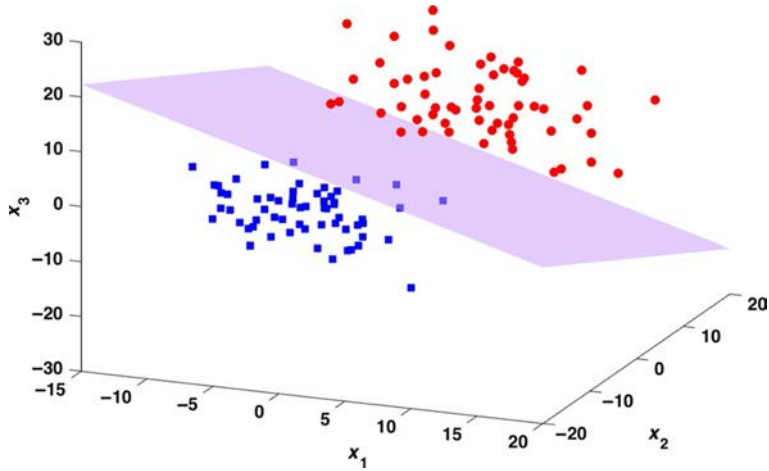

Fig. 3 Illustration of a linear decision boundary (hyperplane in *violet*) for a problem involving two classes (*red circles* and *blue squares*) in three dimensions.

parameters characterizing the probability density functions (the centroids $\overline{\boldsymbol{x}}_j$ and $\overline{\boldsymbol{x}}_k$, and the variance–covariance matrix $\boldsymbol{S}$) plus the priors (ω_j and ω_k). This may represent an advantage in terms of model building as, once the training data are collected, the calculation of the model parameters is straightforward, and no model selection phase is needed. On the other hand, due to the need of inverting the variance/covariance matrix $\boldsymbol{S}$, the training set should contain a higher number of samples than variables, and the latter should be as uncorrelated as possible. Unfortunately, these conditions are rarely met in the data sets normally resulting from omic studies, and therefore LDA can only seldom be directly applied to those kinds of problems. Accordingly, one could think of improving the samples to variables ratio by operating a preliminary variable (or feature) selection stage (see Section 3) or use alternative classification techniques which are insensitive to ill-conditioned matrices (e.g. partial least squares-discriminant analysis (PLS-DA), see Section 1.2.2).

1.2.2 Partial Least Squares-Discriminant Analysis

As already highlighted at the end of the previous section, most of the experiments in the omic field (especially in the case of untargeted studies) involve the use of high-throughput instrumental fingerprinting platforms on a (usually) limited number of samples. From a data analytical standpoint, this translates to the fact that the corresponding data matrices, containing very correlated descriptors whose number is often much higher than that of

the analysed specimens, are ill-conditioned and cannot be processed with traditional statistical methods such as LDA. In contrast, the possibility of extending the concept of bilinear projection, which is one of the pillars of chemometrics, to multivariate classification, by overcoming the limitations discussed above (since the samples are described by a low-dimensional set of orthogonal components) resulted in algorithms, such as PLS-DA, which are becoming more and more utilized in the omic disciplines.

PLS-DA [12–14] was originally proposed to extend the advantages of partial least squares regression (PLS-R) [15,16], a calibration technique based on the partial least squares algorithm [17], to discriminant classification problems. Therefore, in order to understand how PLS-DA works, it is first necessary to illustrate the original regression algorithm.

PLS-R is a multivariate linear regression technique, whose aim is, therefore, to find the best (linear) relationship between a set of predictors (the vector $\boldsymbol{x}$) and one or more responses (the scalar y or the vector $\boldsymbol{y}$, respectively):

$$\begin{cases} y = \boldsymbol{x}\boldsymbol{b} & \text{single response} \\ \boldsymbol{y} = \boldsymbol{x}\boldsymbol{B} & \text{multiple responses} \end{cases} \tag{14}$$

the vector $\boldsymbol{b}$ or the matrix $\boldsymbol{B}$ (depending on the case) collecting the proportionality (regression) coefficients. Regression coefficients are the model parameters and are calculated from a set of samples (the training set) for which the values of both the predictors and the response(s) are known (and organized in the matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively). For those samples, the regression model can be formulated in matrix form as:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}_Y = \widehat{\boldsymbol{Y}} + \boldsymbol{E}_Y \tag{15}$$

where

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{B} \tag{16}$$

is the matrix collecting the responses predicted by the model, while $\boldsymbol{E}_Y$ contains the residuals (i.e. the unmodelled variation in $\boldsymbol{Y}$). In this framework, since classical regression approaches, such as multiple linear regression, which calculates the values of the coefficients $\boldsymbol{B}$ according to the ordinary least squares approach (i.e. by minimizing the Frobenius' norm of $\boldsymbol{E}_Y$), suffer from the same limitations as LDA, when the predictor matrix contains more variables than samples and/or the variables are correlated, PLS-R was introduced as a technique able to deal with the cases where $\boldsymbol{X}$ is ill-conditioned.

Indeed, PLS-R operates by projecting both the $\boldsymbol{X}$ and the $\boldsymbol{Y}$ matrix onto a reduced (low-dimensional) common set of latent variables (components)

$$\boldsymbol{T} = \boldsymbol{XR} \tag{17}$$

$$\boldsymbol{U} = \boldsymbol{YQ} \tag{18}$$

which are oriented in a way as to provide the maximum covariance between the corresponding scores for the two blocks:

$$\text{argmax}_{\boldsymbol{r}_i, \boldsymbol{q}_i}(\text{cov}(\boldsymbol{t}_i, \boldsymbol{u}_i)). \tag{19}$$

In Eqs. (17)–(19), $\boldsymbol{T}$ and $\boldsymbol{U}$ are the score matrices for the $\boldsymbol{X}$ and the $\boldsymbol{Y}$ blocks, respectively, $\boldsymbol{R}$ are the $\boldsymbol{X}$ weights and $\boldsymbol{Q}$ the $\boldsymbol{Y}$ loadings; $\boldsymbol{t}_i$, $\boldsymbol{u}_i$, $\boldsymbol{r}_i$ and $\boldsymbol{q}_i$ are the ith columns of the corresponding matrices, i.e., the $\boldsymbol{X}$ and $\boldsymbol{Y}$ scores, the $\boldsymbol{X}$ weights and the $\boldsymbol{Y}$ loadings for the ith latent variable, respectively. The linear dependence between the $\boldsymbol{X}$ and the $\boldsymbol{Y}$ blocks is achieved by imposing $\boldsymbol{U}$ to be linearly proportional to $\boldsymbol{T}$ component-wise (*inner relation*):

$$\boldsymbol{u}_i = \boldsymbol{t}_i c_i \Longrightarrow \boldsymbol{U} = \boldsymbol{TC} \tag{20}$$

c_i being the inner regression coefficients, collected in the diagonal matrix $\boldsymbol{C}$. By combining Eqs. (18) and (20), it is possible to explicitly express the predicted responses in terms of the scores:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{TCQ}^T \tag{21}$$

so that, by further incorporating Eq. (17), one obtains:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{XRCQ}^T = \boldsymbol{XB}_{PLS} \tag{22}$$

where the PLS regression coefficients $\boldsymbol{B}_{PLS}$ are defined as:

$$\boldsymbol{B}_{PLS} = \boldsymbol{RCQ}^T. \tag{23}$$

The use of few orthogonal components (the scores $\boldsymbol{T}$) as regressors for the responses allows overcoming all the limitations related to the matrix $\boldsymbol{X}$ being ill-conditioned and allows to calculate reliable estimates of the model parameters. Moreover, Eq. (22) shows that the model can still be expressed in terms of the original (measured) variables through the set of regression coefficients $\boldsymbol{B}_{PLS}$, not only making interpretation more straightforward but also allowing to more easily formulate predictions on new

samples. Indeed, whenever a new sample is analysed, obtaining the experimental vector $\boldsymbol{x}_{new}$, the predicted responses $\widehat{\boldsymbol{y}}_{new}$ can be obtained directly as:

$$\widehat{\boldsymbol{y}}_{new} = \boldsymbol{x}_{new}\boldsymbol{B}_{PLS}. \tag{24}$$

In order to be able to use the PLS-R method for calculating classification models, it is necessary to code class information (qualitative response) into a numerical $\boldsymbol{Y}$ matrix. This is accomplished by building the $\boldsymbol{Y}$ for the training samples so to have as many columns as the number of classes and introducing the following binary coding: for each sample, the corresponding row of the $\boldsymbol{Y}$ matrix contains all zeros, except for the column corresponding to its category where there is a one. For example, in a problem involving three classes, all the rows of the $\boldsymbol{Y}$ matrix corresponding to samples from the first class will be identically equal to:

$$\boldsymbol{y}_{class1} = [1\,0\,0], \tag{25}$$

those corresponding to individuals from the second class to:

$$\boldsymbol{y}_{class2} = [0\,1\,0], \tag{26}$$

and the ones associated with the third category, to:

$$\boldsymbol{y}_{class3} = [0\,0\,1]. \tag{27}$$

The dummy coding of class information into the binary matrix $\boldsymbol{Y}$ allows the PLS-R algorithm to be used as the basis for calculating discriminant classification models, giving rise to the technique called partial least squares-discriminant analysis. Indeed, the starting point of PLS-DA is the possibility of calculating a PLS regression model between the experimental data $\boldsymbol{X}$ and the binary-coded matrix $\boldsymbol{Y}$ and, in particular, of obtaining, for each sample, a vector of predicted responses $\widehat{\boldsymbol{y}}$, as described in Eq. (24). However, since PLS regression is a calibration method and, therefore, it produces quantitative outcomes, the predicted responses $\widehat{\boldsymbol{y}}$ will not be binary coded as well but, instead, real-valued. So, the second, and fundamental, step which characterizes PLS-DA is the definition of a classification rule which allows formulating predictions about the category the samples belong to, based on the values of $\widehat{\boldsymbol{y}}$. In this respect, during the years, different variants of PLS-DA have been proposed in the literature, differing from one another on the way this last step was accomplished. In particular, the most naive approach operates classification by assigning each sample to the category corresponding to the highest value of the predicted

response. For example, in a problem involving four classes, if the response predicted for a particular sample were:

$$\hat{\boldsymbol{y}} = [0.03 \quad -0.12 \quad 0.85 \quad 0.19] \tag{28}$$

that individual would be assigned to the third class. On the other hand, alternative approaches involve the calculation of an LDA model on $\hat{\boldsymbol{Y}}$.

Here it is worth mentioning that, when the problem involves only two categories, the dummy matrix $\boldsymbol{Y}$ can be substituted, without loss of information, by a single-column dummy vector $\boldsymbol{y}$, assuming the value 1 for all the individuals from class 1 and 0 for all the samples from the second class. In such cases, the classification process corresponds to setting a threshold of ϑ_y to the value of the predicted response, so that if $\hat{y} > \vartheta_y$ the sample will be assigned to the first class, whereas if $\hat{y} < \vartheta_y$, it will be assigned to the second category. Accordingly, in the most naive approach, the threshold is usually fixed to 0.5, but different values can be obtained if alternative classification schemes are adopted. This situation is graphically illustrated in Fig. 4.

Since PLS-DA requires a projection onto a low-dimensional space of components, different than in the case of LDA, a model selection stage,

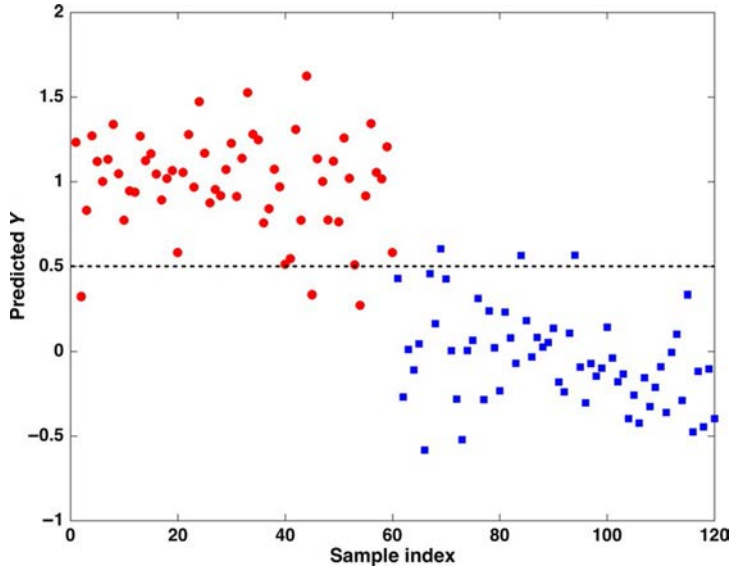

Fig. 4 Illustration of naive PLS-DA classification based on predicted **Y** for a problem involving two categories (*red circles* and *blue squares*). The *black dashed line* represents the classification threshold ($\hat{y} = 0.5$): samples falling above the line are assigned to the *red class*, while samples falling below are assigned to the *blue one*.

where the optimal number of latent variables is estimated, is needed. This is generally accomplished by choosing the model dimensionality which corresponds to the minimum classification error (defined as in Eq. 2), usually in cross-validation (see Section 4).

1.2.3 Other PLS-DA-Based Methods

Since PLS-DA (as its analog PLS-R) has the advantage of being applicable to all sorts of problems, even those involving matrices with a very high number of variables measured onto a very limited number of samples (highly ill-conditioned problems), at the same time providing a straightforward interpretation of the results, various modifications have been proposed in the literature in order to exploit the capabilities of the technique for specific purposes.

In this context, a technique that is largely used in the -omic field is orthogonal partial least squares-discriminant analysis (OPLS-DA) [18], which actively uses the information in the $\boldsymbol{Y}$ block to partition the systematic variation in $\boldsymbol{X}$ in two contributions which are related and orthogonal to the responses, respectively. In particular, the categorical information contained in $\boldsymbol{Y}$ is used to partition $\boldsymbol{X}$ into three contributions:

$$\boldsymbol{X} = \boldsymbol{T}_P \boldsymbol{P}_P^T + \boldsymbol{T}_O \boldsymbol{P}_O^T + \boldsymbol{E}_X \tag{29}$$

where $\boldsymbol{T}_P$ and $\boldsymbol{P}_P$ are the scores and loadings accounting for the predictive (i.e. correlated to $\boldsymbol{Y}$) part of the variation in $\boldsymbol{X}$, $\boldsymbol{T}_O$ and $\boldsymbol{P}_O$ are the scores and loadings for the orthogonal (to $\boldsymbol{Y}$) systematic variation in $\boldsymbol{X}$ and $\boldsymbol{E}_X$ collects the residuals (i.e. the portion of the variability in $\boldsymbol{X}$ not accounted for by the bilinear model). In OPLS-DA, the predicted responses $\widehat{\boldsymbol{Y}}$ are calculated only using the predictive scores $\boldsymbol{T}_P$, which can also be plotted to have a graphical idea of the separation between the classes. However, in order to further exploit the partition of the variability in $\boldsymbol{X}$ summarized by Eq. (29), it was proposed to include also a criterion based on the orthogonal variation to fine-tune the predictions of OPLS-DA (in this way, resembling class-modelling techniques, see Section 3).

On the other hand, kernel PLS-DA [19,20] was proposed as a valid approach to deal with problems where the decision boundaries could not be assumed to be linear anymore. This is because it is possible to express various algorithms, including PLS, in terms of inner products in specific function spaces (kernels). In particular, it is possible to calculate a PLS

(and, consequently, PLS-DA) model in terms of the inner product matrix K_X (linear kernel)

$$K_X = \boldsymbol{X}\boldsymbol{X}^T \tag{30}$$

instead of $\boldsymbol{X}$.

When a problem is not linearly separable, one possibility would be to extend the dimensionality of the variable space by some nonlinear transformation ϕ of the original data:

$$\boldsymbol{\xi} = \phi(\boldsymbol{x}) \tag{31}$$

$\boldsymbol{\xi}$ being the coordinates of a generic sample in the transformed space. If the transformation ϕ is chosen properly, the data should be linearly separable in this new space; however, it is generally not very easy to know a priori which transformation should be applied to the data for a specific purpose. Luckily, due to what is called the "kernel trick", in general, the knowledge of the explicit mapping ϕ is not needed in order to operate nonlinear modelling. Indeed, as explained above, many algorithms can be expressed directly in terms of kernel matrices so, by testing a limited number of general purpose nonlinear kernels (the most commonly used being the polynomial or the radial basis functions), it is possible to implement various degrees of nonlinearity in different linear algorithms as, in particular, PLS-DA.

On the other hand, since the model is calculated using square matrices which carry only the information about similarities or dissimilarities among samples, information about the experimental variable is lost, and interpretation may be less straightforward than with the standard PLS-DA algorithm. Indeed, in order to obtain information about the contribution of the different experimental variables to the kernel PLS-DA model, an approach based on pseudo-samples has been proposed in the literature [21].

1.3 SIMCA and Class Modelling

Although there are only a few examples of the use of class-modelling techniques in the context of -omic disciplines, still it is worth to describe their main characteristics and, in particular, the technique which is most frequently used in this context, which is soft independent modelling of class analogies (SIMCA). As already explained in Section 1.1, class-modelling techniques aim at defining a (usually) bound region in the multivariate space where it is likely to find individuals from a particular category. Since the model of a particular class is built from a training set made only

of individuals from that category, the definition of the class boundary relies on the identification of the salient features of the samples from that group and resembles outlier detection techniques.

SIMCA was the first, and it is still the most widely used class-modelling technique in chemometrics [22,23]. It assumes that the salient features which characterize the samples for a particular class (systematic variation) can be captured by a principal component model of appropriate dimensionality:

$$\boldsymbol{X}_g = \boldsymbol{T}_g \boldsymbol{P}_g^T + \boldsymbol{E}_g \tag{32}$$

where $\boldsymbol{T}_g$, $\boldsymbol{P}_g$ and $\boldsymbol{E}_g$ are the scores, loadings and residuals, while the suffix g indicates that the model is built only using samples from the gth category. Once the model has been built, a distance to the model is calculated as the combination of a contribution which takes into account the distance of a sample to the centre of the scores space (scores distance SD_g) and the residuals (orthogonal distance OD_g):

$$d_g = \sqrt{SD_g^2 + OD_g^2}. \tag{33}$$

The scores distance is often defined through the use of the T^2 statistics [24], which represents the Mahalanobis distance of the sample to the centre of the PC space:

$$T_{i,g}^2 = \sum\nolimits_{j=1}^{F} \frac{t_{ij,g}^2}{\lambda_{j,g}} \tag{34}$$

where $t_{ij,g}$ is the score of the ith samples on the jth component of the PCA model of class g, $\lambda_{j,g}$ is the eigenvalue (variance) of the jth component of the PCA model of class g and F is the dimensionality of the PCA subspace. On the other hand, the orthogonal distance is usually calculated, for each individual, as the sum of its squared residuals (Q statistics) [25]:

$$Q_{i,g} = \sum\nolimits_{j=1}^{v} e_{ij,g}^2 \tag{35}$$

where $e_{ij,g}$ is the jth component of the residual vector of the ith sample for the PCA model of class g, while v is the total number of measured variables. Accordingly, the distance of a sample to the model of a particular category (here, generically indicated as g) is usually defined as:

$$d_g = \sqrt{\left(T_{red,g}^2\right)^2 + Q_{red,g}^2} = \sqrt{\left(\frac{T_g^2}{T_{g,0.95}^2}\right)^2 + \left(\frac{Q_g}{Q_{g,0.95}}\right)^2} \tag{36}$$

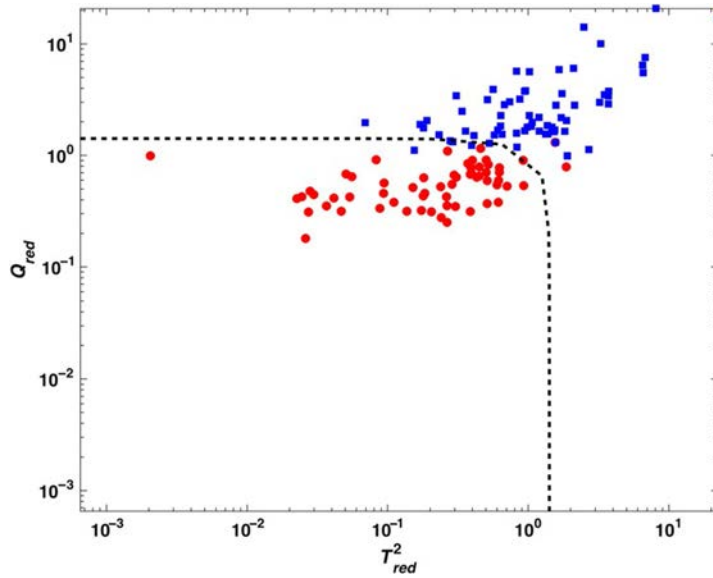

Fig. 5 Illustration of SIMCA model space onto a T^2_{red} vs. Q_{red} plot. The *black dashed line* indicates the acceptance threshold $d = \sqrt{2}$: samples falling below the line are accepted by the class model while those falling above are rejected.

where the suffix *red* indicates that, in order to make the two contributions, which are defined in different subspaces, comparable, both statistics should be normalized by their critical values (95th percentiles of their distributions under the null hypothesis: $T^2_{g,0.95}$ and $Q_{g,0.95}$, respectively).

Accordingly, the class boundary is identified by setting a threshold to the acceptable values of the distance to the model defined in Eq. (36), so that samples having a distance lower than the threshold are accepted, and all the others are rejected. Due to the normalization adopted in Eq. (36), usually the threshold value for acceptance is set to $\sqrt{2}$. This acceptance criterion is graphically illustrated in Fig. 5.

2. DATA FUSION

In omic sciences, it is quite common to have different sets of measurements collected on the same samples; for instance, in genomics, it could be very likely to have amplified fragment length polymorphism data together with spectroscopic measurements and/or characterization of the phenotype. When information coming from different platforms is available for the entire set of samples, the data set is a so-called *multiblock data set*.

Independently from the nature of the multitab data, it is quite established that it is preferable to handle multiblock data sets with methods conceived for it, i.e., by *data fusion* techniques. In fact, it has been demonstrated that it can be more efficient to jointly extract information from the different blocks rather than creating individual models on each of them [26,27].

The multiblock field has been quite widely investigated in the last years, and several approaches have been proposed in the literature. One main distinction among all the different methods could be given taking into account at what *level* data are fused together. With respect to this, they can be divided into three categories: *low*-, *mid*- and *high*-level data fusion techniques [28]. Applying a low-level data fusion approach, the original data are concatenated together, and the resulting matrix is modelled (Fig. 6A). For instance, taking into account a three-block data set ($\boldsymbol{X}_{1(N\times V_1)}$, $\boldsymbol{X}_{2(N\times V_2)}$ and $\boldsymbol{X}_{3(N\times V_3)}$), in order to apply a low-level approach, these will be row-wise concatenated, obtaining a final data matrix $\boldsymbol{X}_{Conc}=[\boldsymbol{X}_1\boldsymbol{X}_2\boldsymbol{X}_3]$ of dimensions $N\times(V_1+V_2+V_3)$.

The main advantage of this class of methods is that, once the final matrix is obtained, all the classical chemometric approaches, unless for other constraints, can be applied. On the other hand, one of the main drawbacks of low-level data fusion strategies is that they lead to quite large data matrices, difficult to handle, in particular by methods which require invertibility (for instance, LDA).

Mid-level data fusion is a good solution when a multiblock method approach is required, but it is also necessary to reduce the number of features. In fact, applying a method from this family of techniques, the concatenation is not among original variables but latent ones (Fig. 6B). Latent variables (such as scores or principal component) can be extracted by applying whatever method conceived for this purpose (e.g. PLS or PCA). Taking once again into account the three-block data set case, and assuming that a set

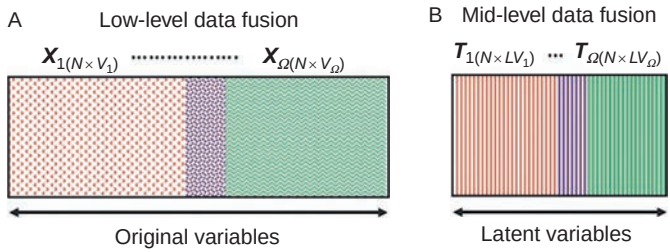

Fig. 6 Schematic illustration of low-level (A) and mid-level (B) data fusion.

of latent variables has been extracted from each of these blocks ($\boldsymbol{T}_{1(N\times LV_1)}$, $\boldsymbol{T}_{2(N\times LV_2)}$ and $\boldsymbol{T}_{3(N\times LV_3)}$), the resulting matrix is obtained by row-wise concatenation of features from each data block $\boldsymbol{T}_{Conc}=[\boldsymbol{T}_1\boldsymbol{T}_2\boldsymbol{T}_3]$ and its dimensions will be $N\times(LV_1+LV_2+LV_3)$.

Obviously, a mid-level data fusion approach is slightly more computationally challenging than a low-level data fusion one, since, in order to obtain the final matrix $\boldsymbol{T}_{Conc}$, a number of models equal to the number of blocks available are required. Anyhow, in general, it is worth to apply these approaches because, due to the features reduction step, part of the non-informative variance is removed from the blocks, and, as a consequence, when these methods are used for regression or classification, often they provide better results (in terms of predictions) than low-level data fusion strategies.

Finally, high-level data fusion techniques combine predictions from other models in order to obtain further and, hopefully, more accurate estimations; for instance, Bayesian approaches belong to this family of techniques. The application of these methods in the omic field is not very common; consequently, their relevance in this context is quite limited.

Independently from the level the concatenation among blocks takes place, it is possible to distinguish the different data fusion approaches also considering their final aim. In fact, they can be applied in diverse fields, such as the exploratory analysis, regression or classification. In the first case, they are focused on solving multiblock component problems: the most widely applied methods in this context are CPCA [29], DISCO [30], JIVE [31] and ComDim [32]. For more details, the reader is addressed to the literature.

For what concerns regression or classification methods, several multiblock methods have been developed to solve these kinds of problems, such as multiblock-PLS (MB-PLS) [26,27,33–35], hierarchical PLS [36], sequential and orthogonalized partial least squares (SO-PLS) [37,38] and On-PLS [39] just to mention few of them.

3. VARIABLE/FEATURE SELECTION

The need of assessing which are the most relevant variables in model building can be motivated by different purposes: (i) increase the understanding of the underlying process that generated the data (enhance interpretation); (ii) improve the performance of the model (increase prediction/classification rate); (iii) reduce the noise generated by irrelevant features

and (iv) to select variables in order to reduce measurement effort or to design tailored instrumental devices.

This explains why there is an ever-growing proposal of variable/feature selection methods in the literature [40–42]. Providing a comprehensive review of them is beyond the scope of this chapter; our aim, instead, is to frame the different approaches, describing the few per each category which are more diffuse in omic research fields.

In the omic context, feature selection is closely linked to putative biomarker assessment, i.e., finding the molecular markers (e.g. genes, proteins, metabolites) that are most relevant in discriminating among control/disease or treated/untreated samples in terms of their relative expression/concentration. In this respect, there is a quite high risk of spurious finding. In fact, it is often underestimated that the excessive false discovery rate due to multiple hypothesis testing (when univariate pruning of features is applied) or the inherent overfitting arises in the absence of adequate validation. Moreover, given the high dimensionality and high degree of correlation present in omic data sets both issues of missing part of the relevant features and of retaining part of the not relevant one can be present.

In the following section, we will only discuss multivariate feature selection methods in the framework of the chemometrics methods illustrated in the previous sections. Table 1 organizes feature selection methods with respect to the category they belong to, and classifiers to which are associated.

An established taxonomy for variable selection allocates the methods in three main groups [40]:

(1) filter methods: these evaluate the variable performance on the basis of some inherent property after model building and generally are implemented by defining a ranking criterion and applying a threshold. Among these most used in omic applications are variable importance in projection (VIP) and selectivity ratio (SR) for PLS-DA based classification and the modelling power in SIMCA class modelling;

(2) wrapper methods: in this case feature selection is accomplished iteratively. A search algorithm is used to span the features space, and distinct models are built for each subset of variables. A measure of model performance is then used to assess the best subset of variables. A further distinction can be done on the basis of the search algorithm being deterministic, e.g., based on successive elimination loops such as uninformative variable elimination in PLS (UVE-PLS), interval PLS (iPLS) or random, e.g., genetic algorithm (GA);

Table 1 Conceptual Organization of Variable Selection Methods

	Method	Classifier
Filter	VIP	PLS-DA
	SR	PLS-DA
	Discriminant power	SIMCA
	Regression coefficients, weights	PLS-DA
	Loadings, canonical vectors	PCA-LDA, LDA
Wrapper	UVE	PLS-DA
	GA	PLS-DA, LDA, SVM, ANN, etc.
	Interval based (iPLS, iECVA)	PLS-DA, ECVA
Embedded	Sparsity in regression coefficients, weights	PLS-DA, SVM
	Sparsity in loadings	PCA-LDA
	Sparsity in Mahalanobis distance	LDA, SHM

(3) embedded methods, where the variable selection is integrated in model derivation. This is a one-step procedure, i.e., at variance with wrapper methods model refitting is not required, and the best subset of variables is assessed simultaneously to model derivation. Typical of this category are methods based on the introduction of sparsity [43–45], e.g., in weight vectors, regression coefficients (PLS), loadings (PCA) or canonical vector (LDA).

Each category has different implications from the point of view of validation, computational costs and purposes pursued by feature selection. Filter methods, being independent of the modelling step, do not require additional validation nor increase the adjustable model parameters, e.g., in PLS (PLS-DA) the estimation of the variables ranking according to VIP or SR will be disjoint from the assessment of the number of components; this will also mean that no additional computational cost is introduced. These methods may be more advantageous when the objective is model interpretation as they also provide a variable ranking and more criteria can be used, as it will be discussed further. Wrapped methods are the most computationally intensive because a search in the space of candidate variables subsets has to be

accomplished, and also a more intensive validation scheme is required, e.g., double cross-validation (DCV); they may be very efficient to obtain a parsimonious model with improved performance, but on the other hand, their application becomes critical with a limited number of samples. Wrapper methods can also be suboptimal when there is an interest in maintaining redundancy. As an example, in proteomic studies, often the aim is identifying all proteins whose expression rate change because of a disease status even if for prediction purposes only a subset may be sufficient. Embedded methods may achieve a good trade-off between computational costs and performance since they do not require model refitting. However, to control the degree of sparsity an additional parameter has to be tuned; mostly this is done by cross-validation. As for model interpretation, it has to be taken into account that depending on the way sparsity is implemented correlated (redundant) variables may be altogether selected or only some of them, arbitrarily.

A further approach to feature selection is to apply one of the approaches mentioned above after data transformation, e.g., in the wavelet domain [46–48].

Recently [49], in connection with classifiers such as random forest or SVM, it is emerging the use of an ensemble of feature selection criteria, mainly for the binary classification context, which are generally based on information criteria, e.g., mutual information, Bayesian information criterion (BIC) or Akaike's information criterion (AIC). The majority vote is then used as a criterion to fuse decision from the different feature selection methods. In conjunction with the evaluation of classification performance, by using resampling methods, the stability and similarity of the selected feature subsets can also be taken into account.

In the following section, the most diffuse methods for each category will be detailed.

3.1 Filter Method

3.1.1 VIP, SR (PLS-DA Classifier)

VIP scores are defined for each $\mathbf{X}$ variable, j, as the sum, over latent variables (LV), of its PLS-weight value (w_j) weighted by the percentage of explained Y variance (SSY) by each specific LV, according to the formula [50,51]:

$$\mathrm{VIP}_j^2 = \sum\nolimits_f w_{jf}^2 \mathrm{SSY}_f J / \left(\mathrm{SSY}_{tot.expl.} \mathrm{F}\right) \tag{37}$$

where F is the number of latent variables of the PLS model and J the number of $\mathbf{X}$ variables.

In the case of multivariate $\mathbf{Y}$ a single VIP value, referring to the overall model, can be calculated if SSY refers, as above, to the percentage of explained variance for the whole $\mathbf{Y}$ block; otherwise a VIP value for each $\mathbf{Y}$ variable, k, can be calculated, as it is usually done in PLS-DA application, by using SSy_{kf} and $SSy_{k,\ tot.expl}$ in (37), where k refers to a specific $\mathbf{Y}$ variable.

In the case of a single response, y ($I \times 1$), the following equalities also hold [52]:

$$SSY_f = b_{ff}^2 \mathbf{t}_f^{\mathrm{T}} \mathbf{t}_f \quad SSY_{tot.expl} = \mathbf{b}^2 \mathbf{T}^{\mathrm{T}} \mathbf{T} \tag{38}$$

where $\mathbf{T}$ is the $\mathbf{X}$ scores matrix and $\mathbf{b}$ the PLS inner relation coefficients.

The VIP formulation as originally proposed [50] is a parameter varying in a fixed range since the sum of squared VIP for all variables sum to the number of variables. This explains why it is common to assume as a threshold a VIP value larger than 1 (i.e. larger than the average of squared VIP values), which means that a selected variable will have an above average influence on the model explaining Y. This criterion is very reasonable to discard irrelevant variables, while it may have drawbacks if used for assessing the significance of features; alternative threshold values include increasing the threshold to 2/3 or considering the average of VIP values as a threshold.

Other criteria have been proposed to verify that the variables have a significantly higher than 1 VIP value, i.e., using jack-knifed confidence intervals or resampling methods such as bootstrap [53,54]. A modification of the bootstrap, which is coupling it with a random permutation of each variable in turn, as to obtain simultaneously a threshold and a significance criterion, has been proposed in Ref. [53]. Obviously, this comes at a computational price.

Selectivity ratio (SR) [55–58] has been introduced to assess the relevance of variable in a PLS (PLS-DA) model considering only $\mathbf{Y}$-related variation; it is defined, for each variable, j, as the ratio of the explained variance for that variable on the target component (TP) and its residual variance:

$$\mathrm{SR}_j = \mathrm{var}\left(\mathbf{X}_{\mathrm{TP},\,j}\right) / \mathrm{var}\left(\mathbf{X}_{\mathrm{O},\,j} + \mathbf{e}_j\right) \tag{39}$$

where $\mathbf{X}_{\mathrm{TP}}$ is referred to the systematic variation in $\mathbf{X}$ correlated with $\mathbf{Y}$, $\mathbf{X}$o to the systematic variation but orthogonal to $\mathbf{Y}$ and $\mathbf{e}$ to the $\mathbf{X}$ residual variation.

$\mathbf{X}_{TP}$ is estimated by the target component (i.e. the **y**-related or predictive component of the PLS model), which is obtained by:

$$\mathbf{t}_{TP} = \mathbf{X}\mathbf{w}_{TP} \quad \mathbf{w}_{TP} = \mathbf{b}/\|\mathbf{b}\| \tag{40}$$

$$\mathbf{p}_{TP} = \mathbf{X}^{T}\mathbf{t}_{TP}/\left(\mathbf{t}_{TP}^{T}\mathbf{t}_{TP}\right) \tag{41}$$

$$\mathbf{X}_{TP} = \mathbf{t}_{TP}\mathbf{p}_{TP}^{T} \tag{42}$$

where **b** is the vector of PLS regression coefficients for the given **y** variable.

As for VIP different criteria have been proposed to define a significance threshold for SR, such as using an *F*-test (since it is a variance ratio) [56], using the mean over all variables [59] and using bootstrap [60].

According to the proposing author, SR should overcome the problem of missing explanatory variables with small variance but relevant to predict Y and ease interpretation because of focus only on Y-related information.

SR and VIP have been compared in several studies [59,61–63] and the obtained results depend indeed on the type of data set. Common conclusions to Refs [59,61], where a wide systematic simulation study by DoE on the level of relevant/irrelevant systematic variation and the noise was also conducted, are that both are effective in skipping the variables showing systematic variation but irrelevant to Y-prediction and avoiding selection of noise variables.

VIP may tend to select more "false positive" (variables not causally related to Y) and SR is affected by more false negative (it may miss some relevant variables; in particular with spectral data when these are non-selective for the calibrated analyte). However, it has to be remembered that the total number of selected features depends on how the threshold value is set. In particular, using the *F*-test as the criterion for SR tends to be too conservative (few variables selected) while other criteria are not; on the other hand, for VIP the threshold value of 1 (which tends to select quite many variables) may be increased taking into account, e.g., that in a spectrum several variables define a band [51].

Further considerations concern interpretation, as neither VIP nor SR alone are sufficient to interpret the contribution of ranked/selected features; to this aim, the sign of regression coefficients and target loadings, respectively, has, for instance, to be taken into account [57].

3.1.2 Modelling Power (SIMCA Classifier)

In SIMCA, the most salient features for discriminating the samples with respect to the modelled classes can be identified by calculating the variable

discriminatory power, DPow [23]. Dpow is calculated, for each variable, as the ratio of the square residual standard deviations, calculated for the objects when projected on the model of the class(es) they do not belong, with respect to that of the class they belong. The motivation is that we expect the residuals of samples on a given variable being higher when fitted to a different class model with respect to their own, if this variable is important in discrimination (thus giving a high Dpow value).

Variables can then be ranked according to Dpow; again analogously to VIP and SR a threshold criterion has to be established to accomplish a selection or to state which features are significantly contributing to class discrimination and which are not.

The proposing authors did not indicate a privileged criterion, generically indicated much above 1, and in literature, most used thresholds were 1–3, average, median. In ref. [64], a criterion based on the use of probability plots, preceded by the normalization of the DPow distribution through the Box–Cox transformation, has been proposed.

3.2 Wrapper Method

3.2.1 Genetic Algorithms, GA (Any Classifier)

GA belong to randomized search algorithms category; they are inspired by biological evolution theory and natural selection. In general, they allow an efficient search of the "best solutions" in a given domain that, because of dimension/complexity, cannot be explored systematically, according to a fitness function (optimized), which takes the role of "fitness to the environment" in natural evolution.

In the specific context of feature selection, the domain explored is constituted by all the possible subsets of variables, and it is assumed that variables that yield fitted models showing high performance (maximum fitness) have a higher probability to "survive" and thus to be included in the selected set. In practice, closely following the analogy with natural evolution: (1) an initial generation (whose individual chromosomes are codifying presence/absence of variables, i.e., are formed by a gene of 1 bit) is randomly generated; the number of individuals in this generation is a tunable parameter; (2) the fitness function (e.g. model classification performance in cross-validation) is calculated for each; (3) a number (to be set) of individuals with the highest fitness are selected to survive until the next generation; (4) the "survivors" undergo crossover (new individuals, number to be set, are created by combining part of the bit of each "parent" chromosome) and mutation (analogous of reproduction a small probability of random switch between 0 and 1 in the

chromosome is programmed) and (5) new individuals and "survivors" enter the next generation and fitness is reestimated (back to step 2).

Thus, summarizing four parameters: population size, the probability of initial selection (number of bits set equal to 1), the maximum number of variables in a chromosome and the probability of mutation have to be defined. They influence the balance between exploration (capability to converge to the min/max in a region) and exploitation (capability of moving from local min/max and assaying different regions of the domain). The overall number of generations (i.e. how many iterations of steps 2–5, in other words stopping of the algorithm) also has to be defined. Indeed, if set too high, this could increase the risk of overfitting of the model, as also using GA on data set with a small sample/variables ratio could do.

A detailed discussion of these and other aspects can be found in Leardi [65], together with guidelines for parameters setting, and a specific implementation of GA which includes: a preliminary evaluation of the feasibility of applying a GA to the data at hand; implementation of a randomization test to assess when GA should stop; a quite high repeated number of run (restarting from scratch) on the basis of which frequency of selection of each variable can be evaluated; and a hybrid criterion for final feature selection.

GA may be linked to the different types of classifiers as far as a suitable fitness function is defined, e.g., classification error in cross-validation.

Another drawback of GA, when used for spectral data, is that the number of selected variables is not contiguous; a modification to take into account the signal structure is to divide it in contiguous intervals and to consider a bit as defining an interval instead of a single variable.

3.2.2 UVE-PLS and Interval-Based Selection (PLS-DA, ECVA Classifiers)

These methods belong to the ones operating a deterministic search, meaning that a guided search of possible subsets of features is operated. Deterministic wrapper methods require less computation than randomized ones, e.g., GA; the number of parameters to tune is smaller and is less prone to overfitting. On the other hand, they have a higher risk of finding local optima.

Obviously are computationally more intense than filter methods.

Uninformative variable elimination (UVE) [66] is based on the introduction in the original data set of noise variables in a number similar to an explanatory variable with similar internal variability but multiplied by a very small constant as to prevent that the random added variables will affect the value of the regression coefficients of the model. A regression model

(in the case of PLS-DA with the number of components equal to the one selected for the full model with no random variables added) is then estimated on the augmented data. By jackknife, the stability (reliability) of the regression coefficients for each variable is estimated as the ratio between its average value and its standard deviation (standard deviation is calculated distinctly for the original variables and the noise variables). In this way, a reference stability value is chosen as the max of the ones estimated for the noise variables set is used as a cutoff to eliminate uninformative variables. A new model is then refitted, and its performance in the classification of a test set is evaluated; the procedure is iterated reducing the model dimensionality by one unit at each step, until no further improvement is observed.

A variant of the method uses resampling approaches as an alternative to jackknife, while another implementation, MC-UVE, uses a Monte Carlo approach to generate an ensemble of train/test splitting.

Interval-based methods, iPLS [67] and iECVA [68], are based on splitting the whole set of variables in a set of small intervals (not necessarily of the same size and may be allowed to overlap); a classification model (PLS-DA or extended canonical variate analysis, i.e. ECVA) is then fitted for each interval. The method can work in "forward" or "backward" mode. In "forward" mode the intervals having the smallest cross-validated classification error are selected in turn until including a number of specified intervals. In "backward" mode the intervals having the largest error are removed iteratively until a number of specified intervals are left. The selected intervals can also be optimized by including or eliminating single variables. The interval-based selection has mainly been proposed in the context of spectral data.

Several other wrapper methods have been proposed [40], and filter methods may be turned to wrapper methods if the ranking/selection procedure is iterated more than once.

In general, the main limitations in the application of these methods are due to the lack of proper validation and unawareness of the applicability condition depending on the ratio between the number of samples and features.

Moreover, when the focus is on interpretation and recovering, e.g., in proteomic or metabolomics all proteins or metabolites changing their expression as a consequence of a disease or a treatment, wrapper methods (especially random search algorithm) may tend to have quite high "false negative" rate, since they tend to select a small number of features.

3.3 Embedded Method (PLS-DA, LDA, PCA-DA Classifiers)

Distinctive of these methods is the fact that feature selection is nested in the classification algorithm; the most used approaches are based on implementation of sparsity constraints, i.e., adding a penalty term in the least square fitting, similarly to what has been implemented in ordinary least squares regression by the LASSO [44,69] where it is minimized:

$$\sum_{i=1}^{n}\left(y_i - b_0 - \sum_{j=1}^{m} b_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} \left|b_j\right| \tag{42}$$

the first term in (42) is the usual least square errors minimization while the second term is a penalty (l_1 norm) term, which forces some regression coefficients to become zero, hence operates feature selection. Other used penalties are: the l_2 norm ($\lambda \sum_{j=1}^{p} b_j{}^2$), i.e., Ridge regularization, however in this case regression coefficients become small but not zero; combination of l_1 and l_2 norm (elastic net); or l_0 norm, where the penalty term is on the number of nonzero coefficients.

Sparse PLS [44,70] and PLS-DA [43,71] instead of using the penalty on the regression coefficients vector impose it on the weights vector, sparse PCA [72,73] on the loadings vectors and sparse LDA [45,71] on the canonical variate vector.

In metabolomics, the most used method has been sparse PLS-DA [43,44,71,74–76].

Moreover, recently an approach called statistical health monitoring (SHM) [45] where a sparsity constraint is implemented on Mahalanobis distance has been proposed, with the aim of selecting characteristic feature of not-healthy status, i.e., the author formulates a one-class problem (class modelling) where the modelled class is the healthy one and distance is calculated in the original domain (PCA compression is skipped).

The degree of sparseness and hence the number of selected features depend on the tuning of the λ parameter; there are various suggestions and often a "manual" tuning is adopted (passing by visual inspection or consistency of interpretation of the selected features). In general, good results are reported with sparse methods; however, there is no guarantee that all salient biomarkers will be automatically identified.

3.4 Feature Selection in Wavelet Domain

Wavelet coefficients obtained from WT decomposition (discrete or continuous) of signals (spectra, chromatographic profiles) can be seen as a set of

alternative variables and can be linked to any multivariate classifier. In particular, features selection can be done in the WT domain, i.e., selecting wavelet coefficients instead of raw variables. The main advantage is that in WT the different frequency contents of a signal are split into disjoint decomposition blocks (approximations and details at a given scale); hence, wavelet coefficients will reflect only part of the information present in the original data and selection of salient feature may be enhanced [46,77].

In general, almost all the feature selection methods described in the previous sections can be applied to WT coefficients; most used are filter and wrapper methods, e.g., GA [46,78]. Specific of feature selection in WT domain is that it can be applied scale/level-wise [79] or coefficients-wise [46,80,81], as well as the fact that WT decomposition parameters have to be set, such as a wavelet filter, decomposition level; some algorithms try to automatically iterate on these settings [46].

Even if, not always operated, since classification models are obtained based on using wavelet coefficients, reconstruction of selected coefficients in the original domain by the inverse WT is particularly useful for interpretation of the models.

4. VALIDATION

Validation is a key stage of chemometric modelling as it is the ensemble of operations which allows evaluating the quality of a model and how reliable all its aspects (predictions, interpretation, subspace estimation, just to cite a few) are [82,83]. Although validation should follow each kind of chemometric approach, its role and impact are absolutely fundamental whenever predictive modelling is involved. Indeed, since most of the practical applications of predictive modelling (especially in the omic field) involve measuring a very high number of variables on a limited number of individuals, issues such as undersampling (not enough specimens to accurately describe the class distribution) or chance correlations may result in overfitting, i.e., on model parameters adjusted so well to the training data that are not able to produce reliable outcomes for new unknown samples [84].

This is because chemometric models are almost always "soft", i.e., data-driven models, built as to approximate as better as possible the system under investigation (to capture a high share of the experimental variance) and, when needed, to provide predictions of the responses which are as close as possible to their true values.

In this context, validation allows evaluating how appropriate and reliable the model is for the specific purpose, by the use of specific strategies and diagnostics which can capture in one or few figures of merit the model quality. In particular, especially when predictive modelling is concerned, all the most frequently used diagnostics are based on the calculation of some sort of residuals (or classification error, such as the one in Eq. 2). Accordingly, based on the considerations reported above, to avoid overoptimistic estimates, it is essential that the validation diagnostics be not calculated based on the fitted residuals (i.e. those resulting from the application of the model to the samples used for model building), as they will never be distributed as those which one expects to occur on new data (i.e. from individuals which never took part in model building or model selection). It is then evident that, in order to properly validate a (classification) model, one should apply the model to a completely independent set of samples (test set) which should be as representative as possible of the practical use of the model on future individuals and for which the true values of the responses to be predicted (e.g. the class labels of the samples) should be known as well. The last condition is necessary in order to calculate the validation diagnostics, which are in general defined in terms classification error (see Eq. 2), even if, especially in the cases where only two classes are present, other figures of merit such as the area under the ROC curve [85], or the discriminant Q^2 [85,86] are also used.

When the number of samples is not large, and there is no possibility of keeping aside a set of samples as an independent validation set, as this would involve either one or both sets (training and test) not to be representative enough to formulate reliable conclusions, one could make use of resampling strategies, the most commonly used of which is cross-validation. In cross-validation, the available set of data is split into a predetermined number M of so-called cancellation groups, so that, in turn, one of the cancellation groups is left aside as a test set, and the model is built using the remaining $M-1$ as the training set. Sometimes the procedure can be repeated more than once, with different data splittings, in order to obtain results which are less insensitive to the specific sample partitioning scheme. The advantage of the use of cross-validation is that it can provide reasonable estimates also in cases where there are not many samples in the data set; on the other hand, due to the resampling strategy, the validation samples may never be considered as truly independent from the training set, so that there is still the possibility of obtaining overoptimistic results.

In general, whenever possible, it is preferable to use cross-validation only for the model selection stage, when there is the need to optimize model

parameters and/or metaparameters and, once the optimal model has been calculated, to use a truly independent set for its validation. Here it is fundamental to stress that validation should comprise all stages of model selection, including variable selection, which should be subjected to complete validation through what is called a cross-model validation procedure [87].

When the number of samples is limited, but one still wants to obtain reliable estimates of the prediction error (or of other figures of merit), a possibility is to use DCV [84], which consists in two nested cross-validation loops, in which the inner one is used for model selection and model building and the outer one for model validation. Permutation tests may then be used to further safeguard against overoptimism and model overinterpretation, since they allow to verify whether the observed discrimination between the classes may be considered significant or just due to chance (e.g. as a consequence of undersampling). In permutation tests, the class label of the samples is randomly permuted and classification models are built on the permuted data: the results obtained on permuted data are used to build the nonparametric empirical distribution of the classification figures of merit under the null hypothesis, which, in turn, are used to test whether the observed discrimination could be considered as statistically significant or not.

REFERENCES

[1] J.W. Tukey, Exploratory Data Analysis, Addison Wesley, Boston, MA, 1977.
[2] S. Geysser, Predictive Inference: An Introduction, Chapman & Hall, New York, NY, 1993.
[3] R. Madsen, T. Lundstedt, J. Trygg, Chemometrics in metabolomics—a review in human disease diagnosis, Anal. Chim. Acta 659 (2010) 23–33.
[4] R.G. Brereton, Chemometrics for Pattern Recognition, John Wiley and Sons, New York, NY, 2009.
[5] M. Bevilacqua, R. Bucci, A.D. Magrì, A.L. Magrì, R. Nescatelli, Classification and class-modeling, in: F. Marini (Ed.), Chemometrics in Food Chemistry, Elsevier, Oxford, UK, 2013, pp. 171–233.
[6] L. Coulier, S. Wopereis, C. Rubingh, H. Hendriks, M. Radonjic, R.H. Jellema, Systems biology, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, vol. 4, Elsevier, Amsterdam, The Netherlands, 2009, pp. 279–312.
[7] C. Albano, W. Dunn III, U. Edlund, E. Johansson, B. Nordén, M. Sjöström, S. Wold, Four levels of pattern recognition, Anal. Chim. Acta 103 (1978) 429–443.
[8] S. De Luca, R. Bucci, A.D. Magrì, F. Marini, Class modeling techniques in chemometrics: theory and applications, in: R. Meyers (Ed.), Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation, John Wiley and Sons, New York, NY, 2018. https://doi.org/10.1002/9780470027318.a9578.
[9] U. Grouven, F. Bergel, A. Schulz, Implementation of linear and quadratic disciminant analysis incorporating costs of misclassification, Comput. Methods Programs Biomed. 49 (1995) 55–60.
[10] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., Wiley Interscience, New York, NY, 2000.

[11] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (1936) 179–188.
[12] M. Sjöström, S. Wold, S. Söderström, PLS discriminant plots, in: E.S. Gelsema, L.N. Kanal (Eds.), Pattern Recognition in Practice II, Elsevier, Amsterdam, The Netherlands, 1986, pp. 461–470.
[13] L. Ståhle, S. Wold, Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study, J. Chemometr. 1 (1987) 185–196.
[14] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemometr. 17 (2003) 166–173.
[15] S. Wold, H. Martens, H. Wold, The multivariate calibration problem in chemistry solved by the PLS methods, in: A. Ruhe, B. Kågstrøm (Eds.), Matrix Pencils: Proceedings of a Conference Held at Pite Havsbad, Sweden, March 22–24, 1982, Springer Verlag, Heidelberg, Germany, 1983, pp. 286–293.
[16] P. Geladi, B.R. Kowalski, Partial least squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.
[17] H. Wold, Estimation of principal components and related models by iterative least squares, in: P.R. Krishnaiaah (Ed.), Multivariate Analysis, Academic Press, New York, NY, 1966, pp. 391–420.
[18] M. Bylesjo, M. Rantalainen, O. Cloarec, J.K. Nicholson, E. Holmes, J. Trygg, OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification, J. Chemometr. 20 (2006) 341–351.
[19] B. Walczak, D.L. Massart, The radial basis function—partial least squares approach as a flexible non-linear regression technique, Anal. Chim. Acta 331 (1996) 177–185.
[20] R. Rosipal, L.J. Trejo, Kernel partial least squares regression in reproducing kernel Hilbert space, J. Mach. Learn. Res. 2 (2001) 97–123.
[21] G.J. Postma, P.W. Krooshof, L.M.C. Buydens, Opening the kernel of kernel partial least squares and support vector machines, Anal. Chim. Acta 705 (2011) 123–134.
[22] S. Wold, Pattern recognition by means of disjoint principal components models, Pattern Recogn. 8 (1976) 127–139.
[23] S. Wold, M. Sjöström, SIMCA: a method for analysing chemical data in terms of similarity and analogy, in: B.R. Kowalski (Ed.), Chemometrics, Theory and Application, American Chemical Society Symposium Series, vol. 52, American Chemical Society, Washington, DC, 1977, pp. 243–282.
[24] H. Hotelling, The generalization of Student's ratio, Ann. Math. Statist. 2 (1931) 360–378.
[25] J.E. Jackson, G.S. Muldholkar, Control procedures for residuals associated with principal component analysis, Dent. Tech. 21 (1979) 341–349.
[26] I.E. Frank, B.R. Kowalski, Prediction of wine quality and geographic origin from chemical measurements by partial least-squares regression modeling, Anal. Chim. Acta 162 (1984) 241–251.
[27] T. Skov, A.H. Honoré, H.M. Jensen, T. Næs, S.B. Engelsen, Chemometrics in foodomics: handling data structures from multiple analytical platforms, Trends Anal. Chem. 60 (2014) 71–79.
[28] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment—a review, Anal. Chim. Acta 891 (2015) 1–14.
[29] S. Wold, S. Hellberg, T. Lundstedt, M. Sjostrom, H. Wold, Proceedings of Symposium on PLS Model Building: Theory and Application, Frankfurt am Main, 1987; also Technical Report, Department of Organic Chemistry, Umeå University (1987).
[30] M. Schouteden, K. Van Deun, S. Pattyn, I. Van Mechelen, SCA with rotation to distinguish common and distinctive information in linked data, Behav. Res. Methods 45 (2013) 822–833.

[31] E.F. Lock, K.A. Hoadley, J.S. Marron, A.B. Nobel, Joint and individual variation explained (jive) for integrated analysis of multiple data types, Ann. Appl. Stat. 7 (2013) 523–542.
[32] A. El Ghaziri, V. Cariou, D.N. Rutledge, E.M. Qannari, Analysis of multiblock datasets using ComDim: overview and extension to the analysis of (K + 1) datasets, J. Chemometr. 30 (2016) 420–429.
[33] L.E. Wangen, B.R. Kowalski, A multiblock partial least squares algorithm for investigating complex chemical systems, J. Chemometr. 3 (1988) 3–20.
[34] J.A. Westerhuis, A.K. Smilde, Deflation in multiblock PLS, J. Chemometr. 15 (2001) 485–493.
[35] S.J. Qin, S. Valle, M.J. Piovoso, On unifying multiblock analysis with application to decentralized process monitoring, J. Chemometr. 15 (2001) 715–742.
[36] S. Wold, N. Kettaneh, K. Tjessem, Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection, J. Chemometr. 10 (1996) 463–482.
[37] I. Måge, B.H. Mevik, T. Næs, Regression models with process variables and parallel blocks of raw material measurements, J. Chemometr. 22 (2008) 443–456.
[38] A. Biancolillo, I. Måge, T. Næs, Combining SO-PLS and linear discriminant analysis for multiblock classification, Chemom. Intel. Lab. Syst. 141 (2015) 58–67.
[39] T. Löfstedt, J. Trygg, OnPLS—a novel multiblock method for the modelling of predictive and orthogonal variation, J. Chemometr. 25 (2011) 441–455.
[40] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in partial least squares regression, Chemom. Intel. Lab. Syst. 118 (2012) 62–69.
[41] R.K.H. Galvao, M.C.U. Araujo, Variable selection, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, vol. 3, Elsevier, Amsterdam, The Netherlands, 2009, pp. 233–283.
[42] P.S. Gromski, Y. Xu, E. Correa, D.I. Ellis, M.L. Turner, R. Goodacre, A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data, Anal. Chim. Acta 829 (2014) 1–8.
[43] G. Fu, B. Zhang, H. Kou, L. Yi, Stable biomarker screening and classification by subsampling-based sparse regularization coupled with support vector machines in metabolomics, Chemom. Intel. Lab. Syst. 160 (2017) 22–31.
[44] P. Filzmoser, M. Gschwandtner, V. Todorov, Review of sparse methods in regression and classification with application to chemometrics, J. Chemometr. 26 (2012) 42–51.
[45] J. Engel, L. Blanchet, U.F.H. Engelke, R.A. Wevers, L.M.C. Buydens, Sparse statistical health monitoring: a novel variable selection approach to diagnosis and follow-up of individual patients, Chemom. Intel. Lab. Syst. 164 (2017) 83–93.
[46] M. Li Vigni, M. Cocchi, Multiresolution analysis and and chemometrics for pattern enhancement and resolution in spectral signals and images, in: C. Ruckebusch (Ed.), Resolving Spectral Mixtures—With Applications From Ultrafast Time-Resolved Spectroscopy to Super-Resolution Imaging, Elsevier, Amsterdam, The Netherlands, 2016, pp. 409–451.
[47] D.A. Donald, Y.L. Everingham, L.W. McKinna, D. Coomans, Feature selection in the wavelet domain: adaptive wavelets. Spectral matrix, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, vol. 3, Elsevier, Amsterdam, The Netherlands, 2009, pp. 647–680.
[48] J.B. Ghasemi, Z. Heidari, A. Jabbari, Toward a continuous wavelet transform-based search method for feature selection for classification of spectroscopic data, Chemom. Intel. Lab. Syst. 127 (2013) 185–194.
[49] B. Pes, N. Dessì, M. Angioni, Exploiting the ensemble paradigm for stable feature selection: a case study on high-dimensional genomic data, Inform. Fusion 35 (2017) 132–147.

[50] S. Wold, E. Johansson, M. Cocchi, PLS: partial least squares projections to latent structures, in: H. Kubinyi (Ed.), 3D QSAR in Drug Design, ESCOM Science Publishers, Leiden, The Netherlands, 1993, pp. 523–550.
[51] S. Favilla, C. Durante, M. Li Vigni, M. Cocchi, Assessing feature relevance in NPLS models by VIP, Chemom. Intel. Lab. Syst. 129 (2013) 76–86.
[52] I. Chong, C. Jun, Performance of some variable selection methods when multicollinearity is present, Chemom. Intel. Lab. Syst. 78 (2005) 103–112.
[53] N.L. Afanador, T.N. Tran, L.M.C. Buydens, Use of the bootstrap and permutation methods for a more robust variable importance in the projection metric for partial least squares regression, Anal. Chim. Acta 768 (2013) 49–56.
[54] R. Gosselin, D. Rodrigue, C. Duchesne, A bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications, Chemom. Intel. Lab. Syst. 100 (2010) 12–21.
[55] T. Rajalahti, R. Arneberg, F.S. Berven, K. Myhr, R.J. Ulvik, O.M. Kvalheim, Biomarker discovery in mass spectral profiles by means of selectivity ratio plot, Chemom. Intel. Lab. Syst. 95 (2009) 35–48.
[56] T. Rajalahti, R. Arneberg, A.C. Kroksveen, M. Berle, K. Myhr, O.M. Kvalheim, Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles, Anal. Chem. 81 (2009) 2581–2590.
[57] O.M. Kvalheim, R. Arneberg, O. Bleie, T. Rajalahti, A.K. Smilde, J.A. Westerhuis, Variable importance in latent variable regression models, J. Chemometr. 28 (2014) 615–622.
[58] O.M. Kvalheim, Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots, J. Chemometr. 24 (2010) 496–504.
[59] A. Biancolillo, K.H. Liland, I. Måge, T. Næs, R. Bro, Variable selection in multi-block regression, Chemom. Intel. Lab. Syst. 156 (2016) 89–101.
[60] B. Krakowska, I. Stanimirova, J. Orzel, M. Daszykowski, Detection of discoloration in diesel fuel based on gas chromatographic fingerprints, Anal. Bioanal. Chem. 407 (2015) 1159–1170.
[61] M. Farrés, S. Platikanov, S. Tsakovski, R. Tauler, Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation, J. Chemometr. 29 (2015) 528–536.
[62] T.N. Tran, N. Lee, L.M.C. Buydens, L. Blanchet, Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC), Chemom. Intel. Lab. Syst. 138 (2014) 153–160.
[63] C.A. Esquerre, A.A. Gowen, A.O. Gorman, G. Downey, C.P.O. Donnell, Evaluation of ensemble Monte Carlo variable selection for identification of metabolite markers on NMR data, Anal. Chim. Acta 964 (2017) 45–54.
[64] E. Marengo, E. Robotti, M. Bobba, P.G. Righetti, Evaluation of the variables characterized by significant discriminating power in the application of SIMCA classification method to proteomic studies, J. Proteome Res. 7 (2008) 2789–2796.
[65] R. Leardi, Genetic algorithms, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, vol. 1, Elsevier, Amsterdam, The Netherlands, 2009, pp. 631–654.
[66] V. Centner, Multivariate approaches: UVE-PLS, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, vol. 3, Elsevier, Amsterdam, The Netherlands, 2009, pp. 609–618.
[67] L.N. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, Appl. Spectrosc. 54 (2000) 413–419.

[68] F. Savorani, M.A. Rasmussen, Å. Rinnan, S.B. Engelsen, Interval-based chemometric methods in NMR foodomics, in: F. Marini (Ed.), Chemometrics in Food Chemistry, Elsevier, Oxford, UK, 2013, pp. 449–486.
[69] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning, Springer, Heidelberg, Germany, 2013.
[70] H. Chun, S. Keleş, Sparse partial least squares regression for simultaneous dimension reduction and variable selection, J. R. Stat. Soc. Series B Stat. Methodol. 72 (2010) 3–25.
[71] K.A. Lê Cao, S. Boitard, P. Besse, Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems, BMC Bioinformatics 12 (2011) 253.
[72] M.A. Rasmussen, R. Bro, A tutorial on the Lasso approach to sparse modeling, Chemom. Intel. Lab. Syst. 119 (2012) 21–31.
[73] I.T. Jolliffe, N.T. Trendafilov, M. Uddin, A modified principal component technique based on the LASSO, J. Comput. Graph. Stat. 12 (2003) 531–547.
[74] D.V. Nguyen, D.M. Rocke, Multi-class cancer classification via partial least squares with gene expression profiles, Bioinformatics 18 (2002) 1216–1226.
[75] K.A. Lê Cao, D. Rossouw, C. Robert-Granié, P. Besse, A sparse PLS for variable selection when integrating omics data, Stat. Appl. Genet. Mol. Biol. 7 (2008) 35.
[76] E. Acar, G. Gürdeniz, M.A. Rasmussen, D. Rago, L.O. Dragsted, R. Bro, in: Coupled matrix factorization with sparse factors to identify potential biomarkers in metabolomics, Proceedings of the 12th IEEE International Conference on Data Mining Workshops, ICDMW, 2012, pp. 1–8.
[77] Y. Liu, S.D. Brown, Wavelet multiscale regression from the perspective of data fusion: new conceptual approaches, Anal. Bioanal. Chem. 380 (2004) 445–452.
[78] B.K. Lavine, C.G. White, T. Ding, M.M. Gaye, D.E. Clemmer, Wavelet based classification of MALDI-IMS-MS spectra of serum N-linked glycans from normal controls and patients diagnosed with Barrett's esophagus, high grade dysplasia, and esophageal adenocarcinoma, Chemom. Intel. Lab. Syst. 176 (2018) 74–81.
[79] B.K. Alsberg, Parsimonious multiscale classification models, J. Chemometr. 14 (2000) 529–539.
[80] B.K. Alsberg, A.M. Woodward, M.K. Winson, J.J. Rowland, D.B. Kell, Variable selection in wavelet regression models, Anal. Chim. Acta 368 (1998) 29–44.
[81] W. Cai, Y. Li, X. Shao, A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra, Chemom. Intel. Lab. Syst. 90 (2008) 188–194.
[82] F. Westad, F. Marini, Validation of chemometric models: a tutorial, Anal. Chim. Acta 893 (2015) 14–24.
[83] R. Harshmann, "How can I know if it's real?" A catalogue of diagnostics for use with three-mode factor analysis and multidimensional scaling, in: H.G. Law, C.W. Snyder Jr., J. Hattie, R.P. McDonald (Eds.), Research Methods for Multimode Data Analysis, Praeger, New York, NY, 1984, pp. 566–591.
[84] S. Smit, M.J. van Breemen, H.C.J. Hoefsloot, A.K. Smilde, J.M.F.G. Aerts, C.G. de Koster, Assessing the statistical validity of proteomics based biomarkers, Anal. Chim. Acta 592 (2007) 210–217.
[85] E. Szymanska, E. Saccenti, A.K. Smilde, J. Westerhuis, Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies, Metabolomics 8 (Suppl. 1) (2012) 3–16.
[86] J.A. Westerhuis, E.J.J. van Velzen, H.C.J. Hoefsloot, A.K. Smilde, Discriminant Q2 (DQ2) for improved discrimination in PLSDA models, Metabolomics 4 (2008) 293–296.
[87] E. Anderssen, K. Dyrstad, F. Westad, H. Martens, Reducing over-optimism in variable selection by cross-model validation, Chemom. Intel. Lab. Syst. 84 (2006) 69–74.