Trends in Analytical Chemistry 183 (2025) 118117

Contents lists available at ScienceDirect

Trends in Analytical Chemistry

journal homepage: www.elsevier.com/locate/trac

One class classification (class modelling): State of the art and perspectives

Lorenzo Strani [a], Marina Cocchi [a], Daniele Tanzilli [a,b], Alessandra Biancolillo [c], Federico Marini [d,*], Raffaele Vitale [b]

[a] Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Via Campi 103, 41125, Modena, Italy
[b] Univ. Lille, CNRS, LASIRE (UMR 8516), Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, F-59000, Lille, France
[c] Department of Physical and Chemical Sciences, University of L'Aquila, Via Vetoio, Coppito, 67100, L'Aquila, Italy
[d] Department of Chemistry, Sapienza University of Rome, P.le Aldo Moro 5, Rome, 00185, Italy

ARTICLE INFO

Keywords:
Class modelling
Soft independent modelling of class analogy (SIMCA)
One class-partial least squares (OC-PLS)
Unequal class spaces (UNEQ)
Potential functions (PF)
One class-support vector machines (OC-SVM)
Neural networks (NN)

ABSTRACT

Classification, i.e., the prediction of one or more qualitative attributes of samples based on the measured data, is ubiquitous in chemistry, and, more specifically, in analytical chemistry. Among the possible classification strategies, class modelling techniques, which aim at describing one category at a time, present several advantages over discriminant ones, especially when dealing with asymmetric problems featuring one category of interest being well characterized and representatively sampled and another (made of everything that is not belonging to the first specific group) being under-represented by definition and highly heterogeneous.

In this review, the fundamentals of class modelling are illustrated, together with an overview of the main techniques of this kind proposed in the literature, namely Soft Independent Modelling of Class Analogy (SIMCA), Unequal Class Spaces (UNEQ), Potential Functions (PF), Partial Least Squares (PLS)-based algorithms, One Class-Support Vector Machines (OC-SVM) or Neural Networks (NN)-based strategies.

Abbreviations

^{31}P NMR	Phosphorus-31 Nuclear Magnetic Resonance
Alt-SIMCA	Alternative Soft Independent Modelling of Class Analogy
ANN	Artificial Neural Networks
ATR-FTIR	Attenuated Total Reflection-Fourier-Transform Infrared Spectroscopy
AMS	Ambient Mass Spectrometry
CI-SIMCA	Combined Index Soft Independent Modelling of Class Analogy
CM	Class Modelling
CV	Cross-Validation
DD-SIMCA	Data Driven Soft Independent Modelling of Class Analogy
DD-ComDim	Data Driven Common Dimensions (also known as Common Components and Specific Weights Analysis)
DHS-GC-ToFMS	Dynamical Head Space-Gas Chromatography-Time-of-Flight Mass Spectrometry
FT-MIR	Fourier-Transform Mid-Infrared Spectroscopy
FT-Raman	Fourier-Transform Raman Spectroscopy
GBT	Gradient Boost Tree
HPLC-CAD	High Performance Liquid Chromatography-Charged Aerosol Detector
HS-SPME-GC-MS	Head Space-Solid Phase Micro Extraction-Gas Chromatography-Mass Spectrometry
ICA	Independent Component Analysis
ICP-MS	Inductively Coupled Plasma-Mass Spectrometry
ICP-OES	Inductively Coupled Plasma-Optical Emission Spectrometry
HSI	Hyperspectral Imaging
IR	Infrared Spectroscopy
kNN	k Nearest Neighbors
LF-TD-NMR	Low Field-Time Domain-Nuclear Magnetic Resonance
LDA	Linear Discriminant Analysis
MALDI-ToFMS	Matrix Assisted Laser Desorption Ionization-Time-of-Flight Mass Spectrometry
MCR	Multivariate Curve Resolution
MCR-SIMCA	Multivariate Curve Resolution-Soft Independent Modelling of Class Analogy
MF-ICA	Mean Field-Independent Component Analysis
MIR	Mid-Infrared Spectroscopy
NIR	Near Infrared Spectroscopy
NMR	Nuclear Magnetic Resonance
OC-PLS	One Class-Partial Least Squares
OC-SVM	One Class-Support Vector Machines
OD	Orthogonal Distance
OPLS-DA	Orthogonal Partial Least Squares-Discriminant Analysis
Orig-SIMCA	Original Soft Independent Modelling of Class Analogy
PARAFAC	Parallel Factor Analysis
PC	Principal Component

(*continued on next column*)

(*continued*)

(*continued on next page*)

* Corresponding author.
E-mail address: federico.marini@uniroma1.it (F. Marini).

https://doi.org/10.1016/j.trac.2024.118117
Received 31 May 2024; Received in revised form 31 October 2024; Accepted 23 December 2024
Available online 24 December 2024

(*continued*)

Abbreviations	
PCA	Principal Component Analysis
PCA-kNN	Principal Component Analysis-k Nearest Neighbors
PCA-LDA	Principal Component Analysis-Linear Discriminant Analysis
PCA-SVM	Principal Component Analysis-Support Vector Machines
PF	Potential Functions
PGAA	Prompt Gamma Neutron Activation Analysis
PLS	Partial Least Squares Regression
PLS-CM	Partial Least Squares-Class Modelling
PLS-DA	Partial Least Squares-Discriminant Analysis
PLS-DM	Partial Least Squares-Density Modelling
QDA	Quadratic Discriminant Analysis
RMSECV	Root Mean Square Error in Cross-Validation
RF	Random Forest
SD	Score Distance
SIMCA	Soft Independent Modelling of Class Analogy
Sim-SIMCA	Simple Soft Independent Modelling of Class Analogy
SOMs	Self-Organizing Maps
SO-CovSel-LDA	Sequential and Orthogonalized-Covariance Selection-Linear Discriminant Analysis
SO-PLS-LDA	Sequential and Orthogonalized-Partial Least Squares-Linear Discriminant Analysis
SPORT-DA	Sequential Preprocessing through ORThogonalization-Discriminant Analysis
SVM	Support Vector Machines
TD-NMR	Time Domain-Nuclear Magnetic Resonance
UNEQ	Unequal Class Spaces or Unequal Class Models
UV–Vis	Ultra Violet-Visible Spectroscopy
Vis-NIR	Visible-Near Infrared Spectroscopy
XRF	X-ray Fluorescence

1. Introduction

Classification tasks are widespread in analytical chemistry applications and, more generally, in various and diverse contexts of all sciences [1–5]. The aim of classification is to assign objects to predefined categories (or classes) on the basis of the data collected to characterize them (a set of variables). To do this, the following steps must be taken: i) obtaining a representative sampling of the categories to be modelled; ii) collecting data that contain information suitable for distinguishing the categories; and iii) defining a classification rule to assign the objects to a given category. The basic assumption is that a "class" represents a set of samples that share similar characteristics.

The domain of classification methods is extensive, and they can be categorized according to several criteria. However, a key distinction lies between discriminant classification and class modelling (CM) [6–9]. Considering the objects represented in the original variable space, discriminant classification consists in finding the best boundaries that separate the objects belonging to different classes, whereas CM approaches independently define a boundary for each individual class under consideration, enclosing a specific region of the variable space in which objects belonging to that class are most likely to be found, as shown in Fig. 1.

This difference means that discriminant methods require at least two classes to define a boundary, and an object is uniquely assigned to one of the defined categories. In contrast, CM approaches can handle the so-called asymmetric case, where a single category is represented in the training set – or anyway is of interest – and is the only one that has to be modelled. At the same time, when two or more classes are modelled, an object can be assigned to only one, more than one or none of the defined categories. This reflects on the domain of applicability of the two approaches, e.g., in authentication or quality control the objective is recognizing whether a product is compliant to what declared on its label, or whether it respects definite specifications which translates into a one class problem that cannot be handled by discriminant methods. Despite this, authentication tasks are often formulated as a two-class discrimination problem: the authentic category (class 1), and all the rest (class 2). However, this is an ill-posed question, because class 2 does not fulfil the aforementioned basic assumption, i.e., non-authentic samples are not expected to share common characteristics, but will typically be heterogeneous and will not seize a common, distinct region within the variable space. Moreover, a representative sampling of the alternative category, i.e., the second class, is an unattainable goal. The same difficulties remain if discriminant analysis is applied to contrast different known alternative categories. In fact, there will always be some future inauthentic samples, i.e., differently counterfeited or adulterated, that do not belong to any of the considered classes but that will be nonetheless assigned to one of them, resulting in a wrong decision [10–12]. To cope with these issues, some discriminant techniques have been modified so as to enable the possibility of assigning a sample to multiple classes or to none of the categories in the training set, and this is especially true in the case of Partial Least Squares-Discriminant Analysis (PLS-DA), for which different implementations have been proposed in the literature and are available in freeware toolboxes and commercial software suites. On the other hand, soft versions of PLS-DA [13,14], or two-step classification approaches have also been developed [15].

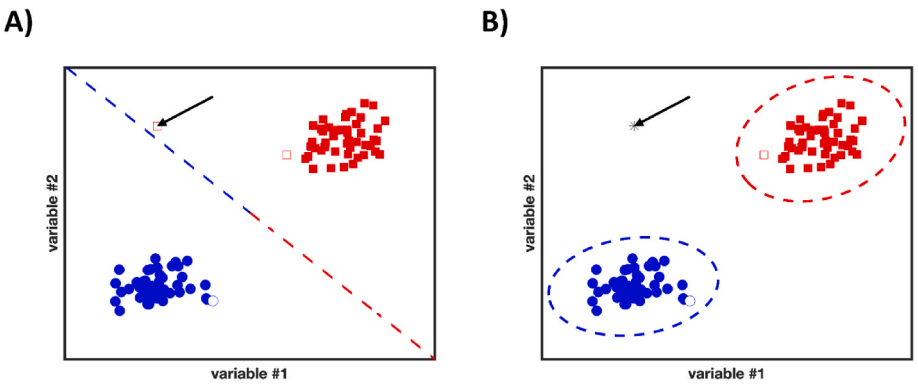

Fig. 1. Schematic representation of the operating principle of A) a discriminant and B) a CM technique in an illustrative example involving two classes of samples (blue dots and red squares). The former defines a global frontier (blue-red dashed line) partitioning the multivariate space of the registered variables into as many subregions as the number of categories represented in the training set and always assigns an object (sample) to one and only one of them. The latter independently estimates a contour for each individual class under study (blue and red dashed line-ellipses), delimiting a specific area where specimens belonging to it are more likely to be found. Notice that empty dots and squares (as well as the black star) denote hypothetical test samples, i.e., samples not taken into account when defining the classification boundaries/rules. Here, the observation lying on the upper left part of the two plots (highlighted by an arrow) would be recognized as a member of the red square category by a discriminant approach but would be rejected by both the independent class models one could possibly construct - this is the reason why such an observation is graphically displayed using two distinct symbols in A) and B). Reproduced from Ref. [6]. with permission from Elsevier. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

However, here it must be stressed that, although being more versatile than their purely discriminant counterparts, these strategies still suffer from some of the drawbacks stated above.

The focus of this work is on the CM methodology, which is less known and applied than discriminant analysis, despite its flexibility and applicability in many fields where the asymmetric case is the norm, such as food authentication, process monitoring, product quality control, drug counterfeiting detection, forensic analysis, medical diagnosis, etc. More in detail, the available CM methods as well as the latest developments and research trends that can be foreseen in the near future related to this particular domain will be here reviewed and discussed.

2. State of the art

Revising the manuscripts published during the last five years including the term "class modelling" in their titles, abstracts and keywords, and limiting the search to scientific areas only, the large majority of the about 160 articles found dealt with food-related issues, more specifically addressing authentication, adulteration or counterfeiting problems. Table 1 reports the applications most related to the fields of analytical chemistry and chemometrics [15–64], highlighting the CM approach used, and for Soft Independent Modelling of Class Analogy (SIMCA) also the particular implementation exploited (see section 2.1 for further details). Its sixth column details if a discriminant approach has also been applied for comparative purposes or whether a two-step strategy was resorted to, i.e., whether first CM was used to assess the product origin/authenticity/integrity and then a discriminant methodology was applied to establish which was the specific category of counterfeiting or adulteration observed [15,41]. It is worth noticing that SIMCA is the prevalent CM approach, especially in its alternative (Alt-SIMCA) and data-driven (DD-SIMCA) implementation, while One Class-Partial Least Squares (OC-PLS) and One Class-Support Vector Machines (OC-SVM) are much less applied. Unequal Class Spaces (UNEQ) and Potential Functions-based (PF) approaches are rare; a reason for that could be they are included only in a few software packages.

Overall, while applications of CM are increasing in recent years the awareness of the different implementations of these methodologies seems limited as well as the use of proper validation approaches.

In the following subsections the basics of the available CM methods are illustrated, focusing on the most widespread, and limiting to the ones truly performing CM (i.e., one-class classification).

2.1. SIMCA

Among the different existing CM methodologies, Soft Independent Modelling of Class Analogy (SIMCA), developed by Svante Wold in 1976 [65,66], is probably the most popular in the field of chemometrics. SIMCA is a fully data-driven approach (i.e., no assumption on the statistical distribution of the collected data is preliminarily made) which, in contrast to standard discrimination strategies, focuses on the similarities among specimens belonging to the individual categories under study rather than on the differences that would permit to distinguish them. More in detail, SIMCA assumes that such similarities can be captured by a Principal Component (PC) representation of the measurements registered for each one of the investigated classes and that, based on these reduced PC representations, one can assess whether new incoming observations belong to one, multiple or none of them.

Practically speaking, the construction of a SIMCA model begins, therefore, with a category-wise Principal Component Analysis (PCA) [67,68] decomposition of the data at hand. Imagine, for example, that a set of J-dimensional spectra or chromatograms has been collected for N samples belonging to a single class or category (e.g., N urine or blood extracts from healthy laboratory mice) and piled into a matrix, say $\mathbf{X}$ (of dimensions $N \times J$), sensibly pretreated (for instance, mean-centered or auto-scaled). $\mathbf{X}$ is, thus, factorized as:

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

where $\mathbf{T}$ ($N \times A$), $\mathbf{P}$ ($J \times A$) and $\mathbf{E}$ ($N \times J$) denote PCA scores, loadings and residuals matrices and A identifies the number of computed PCs. These PCs (that are specifically encoded in the column vectors of $\mathbf{P}$) define a so-called class subspace describing the systematic data variation typical of the individual category taken into account. Clearly, the larger the distance of an observation to this class subspace, the higher the probability that the sample corresponding to such an observation is not a member of the modelled category. In this regard, in SIMCA, two distance metrics are commonly exploited for evaluating whether a new object belongs to the class under study or not: the Orthogonal Distance, *OD*, and the Score Distance, *SD*. *SD* quantifies the dissimilarity of the sample with respect to the distribution of the training objects of the modelled class in the PC subspace, while *OD* accounts for how well the sample is fitted by the class model. Both distances have been defined in different ways across the years, but in the large majority of the implementations, *SD* is calculated as the squared Mahalanobis distance from the origin of the score subspace (also called Hotelling's T^2 statistic), while *OD* corresponds to the sum of squared residuals (i.e., the squared Euclidean distance to the PC subspace, also called Q statistic or squared prediction error, *SPE*). Accordingly, denoting as $\mathbf{x}_{\mathrm{new}}^{\mathrm{T}}$ the measurement vector associated to a generic new sample, its respective *OD* and *SD* values can be estimated based on the following equations:

$$OD_{\mathrm{new}} = \left\|\mathbf{x}_{\mathrm{new}}^{\mathrm{T}}\left(\mathbf{I} - \mathbf{P}\mathbf{P}^{\mathbf{T}}\right)\right\|^2 \tag{2}$$

$$SD_{\mathrm{new}} = \mathbf{x}_{\mathrm{new}}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Lambda}^{-1}\mathbf{P}^{\mathbf{T}}\mathbf{x}_{\mathrm{new}} \tag{3}$$

where $\mathbf{I}$ is a $J \times J$ identity matrix, $\boldsymbol{\Lambda}$ ($A \times A$) is equal to $(N\text{-}1)^{-1}\mathbf{T}^{\mathbf{T}}\mathbf{T}$ and $||\ ||^2$ symbolizes the 2-norm. At this point, OD_{new} and SD_{new} (or some mathematical combinations of them) are compared with characteristic thresholds – generally related to a custom confidence level, 1-α – calculated either by assuming certain statistical distributions for both *OD* and *SD* or empirically from $\mathbf{X}$. In other words, as also shown in Fig. 2, SIMCA delimits a case within the space of the J original variables where samples from the investigated category are likely to be located. Subsequently, if $\mathbf{x}_{\mathrm{new}}$ falls within the boundaries of this case, the new object is recognized as a member of this category[1].

Such boundaries can be marked out in different ways depending on the particular implementation of SIMCA resorted to. Interested readers are addressed to Ref. [6] for a comprehensive survey of the five main SIMCA variants that have been reported in literature so far: the original SIMCA formulation by Wold [65,66], Simple SIMCA (Sim-SIMCA) [69], Alternative SIMCA (Alt-SIMCA) [70], Combined Index SIMCA (CI-SIMCA) [71] and Data Driven SIMCA (DD-SIMCA) [72,73].

When carrying out CM by means of SIMCA, a crucial step is the optimization of the dimensionality or complexity of the SIMCA model itself, A. Two alternative strategies exist for performing such an operation: rigorous and compliant [74]. The former utilizes uniquely data measured for specimens of the modelled category and sets A as the highest dimensionality yielding the true positive[2] rate closest to 1-α. Conversely, the latter estimates A by exploiting both target and non-target category observations and finding the best compromise between true positive and true negative[3] rate. Concerning compliant tuning approaches, it is also worth mentioning that they enable, in principle, the simultaneous optimization of A and α, a solution that was proven effective especially in the presence of strong class overlaps [75].

[1] Notice that, in case multiple classes of samples are coped with, all the procedure described here needs to be iterated for every one of them.

[2] A true positive is an object correctly identified as a member of the category under study.

[3] A true negative is an object correctly identified as a non-member of the category under study.

Table 1
Recent articles related to class modelling (CM) techniques. When not specified otherwise in the fifth column of this table, classification sensitivity, specificity and efficiency were calculated based on the data contained in the calibration set.
Sensitivity = TP/(TP + FN)
TP = True Positives (the amount of objects correctly identified as members of the category under study)
FN = False Negatives (the amount of objects mistakenly identified as non-members of the category under study)
Specificity = TN/(TN + FP)
TN = True Negatives (the amount of objects correctly identified as non-members of the category under study)
FP = False Positives (the amount of objects mistakenly identified as members of the category under study)
Efficiency = (Sensitivity × Specificity)$^{0.5}$

Research objective	Product studied	Analytical technique	CM approach	Model dimensionality selection criterion	Discriminant approach	Reference
Adulteration detection	Oregano	NIR-HSI	Alt-SIMCA	Efficiency in CV	Soft PLS-DA	[16]
	Edible insect flour	IR	Alt-SIMCA	Efficiency in CV Specificity in CV	SPORT-LDA	[17]
	Saffron	HS-SPME-GC-MS	Alt-SIMCA	Efficiency in CV	PLS-DA	[18]
	Honeybush and rooibos tea	XRF	Sim-SIMCA	RMSECV Sensitivity		[19]
	Turmeric powder	ATR-FTIR	Sim-SIMCA OC-SVM	Explained Class Variance		[20]
	Cashew nuts	NIR	SIMCA	RMSECV	PLS-DA	[21]
	Cassava starch	Raman Spectroscopy	OC-SVM SIMCA	Explained Class Variance in CV		[22]
	Olive oil	HPLC-CAD	MCR-SIMCA DD-SIMCA	Sensitivity Specificity Efficiency		[23]
	Olive oil	Raman Spectroscopy	DD-SIMCA	Classification Error in CV	PLS-DA	[24]
	Turmeric	Vis-NIR-HSI	DD-SIMCA on distribution maps from MCR and ICA	Sensitivity		[25]
	Cumin powder	NIR FT-MIR	DD-SIMCA	Sensitivity in External Validation		[26]
	Orange juice	NIR	DD-SIMCA	Sensitivity	Soft PLS-DA PLS-DA GBT Adaboost	[27]
	Goat dairy beverages	NIR	DD-SIMCA	Unspecified		[28]
	Milk	XRF NIR	DD-SIMCA	Sensitivity	PLS-DA SVM	[29]
	Coconut oil	ATR-FTIR	DD-SIMCA	Sensitivity in CV		[30]
	Sanqi powder	NIR	DD-SIMCA	Sensitivity in CV		[31]
	Grape nectar	LF-TD-NMR	SIMCA DD-SIMCA OC-PLS	Sensitivity in CV (unspecified for DD-SIMCA)	PLS-DA	[32]
	Weight loss pills	NIR	OC-PLS	RMSECV		[33]
Authentication	Bell peppers	ICP-OES	Alt-SIMCA	Efficiency in CV		[34]
	Parmigiano Reggiano	Raman Spectroscopy	Alt-SIMCA	Sensitivity in CV Efficiency in CV RMSECV		[35]
	Lime juice	NIR	Alt-SIMCA	CV (unspecified criterion)	PLS-DA	[36]
	Sweet cherries	Physical and Biochemical Parameters	Sim-SIMCA	RMSECV		[37]
	Pork fat	NIR	DD-SIMCA	Sensitivity in Calibration and External Validation Specificity in Calibration and External Validation		[38]
	Fish	NIR DHS-GC-ToFMS	DD-SIMCA	Unspecified	OPLS-DA	[39]
	Oolong tea	AMS	DD-SIMCA	Unspecified	PCA-kNN PCA-LDA PCA-SVM	[40]
	Caterpillar fungus	ATR-FTIR	DD-SIMCA	Sensitivity in External Validation Specificity in External Validation	PLS-DA	[41]
	Saffron	Vis-NIR-HSI	DD-SIMCA MF-ICA	Unspecified	PLS-DA	[42]
	Diesel fuel	NMR TD-NMR	DD-ComDim DD-SIMCA	Matthew's Correlation Coefficient in Calibration and External Validation		[43]
	Materials	PGAA	DD-SIMCA	Sensitivity (compliant approach) Efficiency	RF PLS-DA Soft PLS-DA LDA QDA	[44]
	Antibiotics drugs	NIR	DD-SIMCA	CV (unspecified criterion)		[45]
Geographical/botanical/animal origin determination	Italian chickpeas (geographical)	ICP-OES MIR NIR	Alt-SIMCA	Efficiency in CV	SO-PLS-LDA SO-CovSel-LDA	[46,47]
	Chestnuts (geographical)	NIR	Alt-SIMCA	Efficiency in CV	PLS-DA	[48]

(*continued on next page*)

Table 1 (*continued*)

Research objective	Product studied	Analytical technique	CM approach	Model dimensionality selection criterion	Discriminant approach	Reference
	Lentils (geographical)	ICP-OES	Alt-SIMCA	Efficiency in CV	PLS-DA	[49]
	Saffron (geographical)	ICP-MS	Alt-SIMCA	CV (unspecified criterion)	PLS-DA	[50]
	Canadian honey (geographical)	NMR	Alt-SIMCA	RMSECV	PLS-DA	[51]
	Celery (botanical)	FT-MIR	Alt-SIMCA	Efficiency in CV	SPORT-LDA SO-PLS-LDA	[52]
	Milk (animal)	^{31}P NMR	UNEQ Alt-SIMCA	CV (unspecified criterion)	kNN	[53]
	Herbal tea (geographical/ botanical)	NIR-HSI	Sim-SIMCA	RMSECV	PLS-DA	[15]
	Coffee beans (geographical)	UV–Vis	DD-SIMCA OC-PLS	Sensitivity		[54]
	Brazilian canephora coffee (geographical)	NIR	DD-SIMCA	Explained Class Variance	PLS-DA	[55]
	Strawberries (geographical)	Light isotope determination	DD-SIMCA	Unspecified	OPLS-DA	[56]
	Soybeans (geographical)	Trace element determination	Alt-SIMCA OC-SVM PF	CV (unspecified criterion)	SVM RF kNN	[57]
	Honey (botanical)	MALDI-ToFMS	Sim-SIMCA	Classification Error Rate in CV	PCA-kNN PCA-LDA	[58]
	Gelatin (animal)	Raman Spectroscopy	DD-SIMCA	Sensitivity in CV	PLS-DA	[59]
Medical diagnosis	Metabolic syndrome	FT-MIR	Orig-SIMCA	Unspecified	LDA	[60]
	Colorectal cancer	NIR	OC-PLS	RMSECV		[61]
Non-conformity determination	Pastry dough	NIR	Orig-SIMCA UNEQ	Sensitivity Specificity	Soft PLS-CM	[62]
	Painkillers drugs	Raman Spectroscopy	Orig-SIMCA	CV (unspecified criterion)		[63]
	Embryos for *in-vitro* fertilization	FT-MIR	DD-SIMCA	Explained Class Variance		[64]

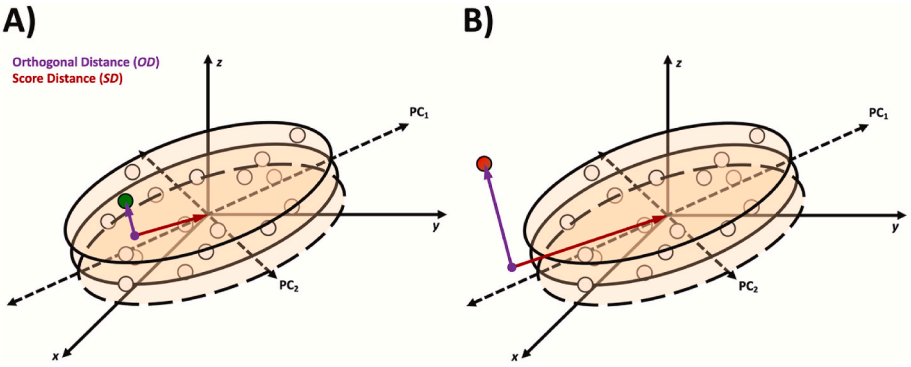

Fig. 2. Schematic representation of the operating principle of SIMCA. A dataset containing the values of three distinct variables (x, y and z) measured for a set of 17 samples (grey dots) belonging to the same class of objects is subjected to a PCA decomposition which yields two different principal components (PC1 and PC2). Based on the estimates of *OD* and *SD* calculated for these samples or by assuming specific statistical distributions for both distance indices, a subregion of the three-dimensional space of the original variables recorded where specimens from the modelled category are more likely to be located is delimited. In A), a new observation (green dot) is found to fall inside this subregion of space and the corresponding object is assigned to such a category. On the other hand, in B), the new observation (red dot) falls outside it and the corresponding object is not assigned to the class under study but rejected as an outlier. Notice that here the outlying observation exhibits abnormal values of both *OD* and *SD*. Reproduced from Ref. [6]. with permission from Elsevier. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

In this regard, it must be underlined that very often the criterion for selecting the number of components of a SIMCA model is not reported, unclear or, wrongly set (e.g., based on the results obtained for the external validation set). At the same time, the number of studies adopting a compliant approach and of those relying on a rigorous strategy are comparable. Most often when DD-SIMCA is applied these criteria are optimized using the calibration set, while when Alt-SIMCA is considered, they are optimized in cross-validation (CV). Besides, explained class variance and root mean square errors in CV (RMSECV) which do not explicitly refer to classification capability are often employed as criteria to determine the number of components.

The original implementations of SIMCA, Alt-SIMCA and DD-SIMCA have also been extended for handling higher-order data sets [76–78], i.e., data that can be organized as three- or higher-dimensional arrays (i.e., tensors), instead of matrices. These data arise from the use of, e.g., fluorescence spectroscopy when the full excitation/emission profile is recorded for each sample, hyphenated techniques, such as chromatography-mass spectrometry, sensory analysis when several attributes are evaluated on the investigated samples by different assessors, or from measuring the same set of variables on the available samples at different time points or conditions. To cope with these data structures, the proposed multi-way versions of SIMCA replace the PCA decomposition step with tensor decompositions carried out by means of approaches such as Parallel Factor Analysis (PARAFAC) or Tucker3 [79] and rely on distance metrics estimated in the resulting multi-way compressed space.

2.2. UNEQ and potential function-based class modelling

Some CM techniques operate by making specific assumptions on the probability density function of the samples belonging to a specific category, so that the definition of the class space is (usually) based on identifying a hypersurface based on the class training set samples which encloses a given probability (usually 95 %). The first of these techniques to be presented in the literature was UNEQ (acronym which was later expanded as Unequal Class Spaces or Unequal Class Models), proposed by Derde and Massart in 1996 [80]. UNEQ can be considered as the modelling analog of quadratic discriminant analysis, as it assumes that, for each category, observations follow a multivariate normal distribution with a class-specific variance-covariance matrix. Training set samples are then used to estimate the barycenter and the variance-covariance matrix for the category distribution, which are, in turn, used to estimate the Mahalanobis distance of each sample to the class centroid and, accordingly, to define the class space. Indeed, defining the class space translates into identifying the hyperellipsoid which encloses the 95 % probability volume around the barycenter of the category which, consecutively, corresponds to setting a threshold to the squared Mahalanobis distance to the centroid. For the prediction of new samples, such threshold is usually estimated through the F distribution. An example of the use of UNEQ is schematized in Fig. 3.

Another technique relying on the explicit calculation of a probability density function for the definition of the class model is the so-called potential function method [81]. The idea behind such an approach is that the probability density function of a generic distribution can be estimated as the superposition of "kernel" probability density functions centered at each of the training samples and usually assumed to be Gaussian or triangular functions [82]. In the context of CM, this translates into the fact that the multivariate probability density function associated to the class distribution $f_c(\mathbf{x})$, can be calculated as the sum of individual analytically defined density functions $\phi_j(\mathbf{x})$ centered around each of the training samples for that category:

$$f_c(\mathbf{x}) = \sum_{j=1}^{N_c} \phi_j(\mathbf{x}) \tag{4}$$

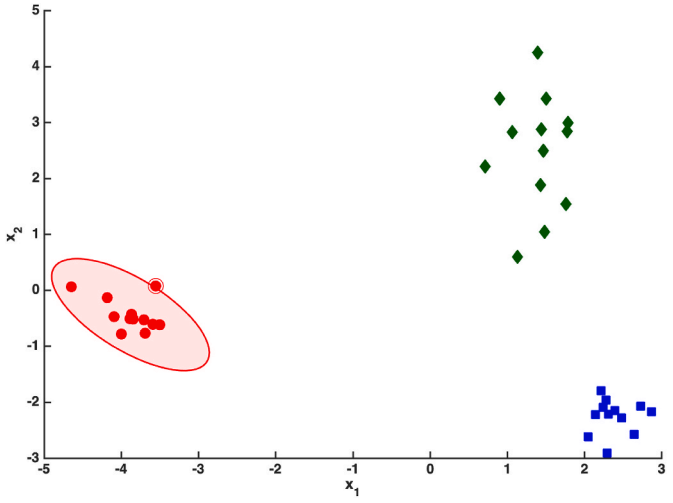

Fig. 3. Example of UNEQ classification. The light red ellipse identifies the class space for the red category (the modelled category) while the thick red contour line represents the class boundary. All samples falling within the ellipse (characterized by a Mahalanobis distance to the class centroid lower than its corresponding threshold value, i.e., the thick red line) are accepted by the model. Accordingly, all the blue and green samples are correctly rejected by the red class model, whereas almost all the red samples (except for the one highlighted with a double circle, which falls outside the class border) are correctly accepted by the red class model. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

where N_c is the number of training samples from the cth category and the kernel functions $\phi_j(\mathbf{x})$ are usually multivariate Gaussians. Once the class probability density function is calculated according to Equation (4), the class space is defined by the hypersurface enclosing a certain probability volume (often 95 %). This is usually accomplished by setting a threshold to the value of the probability density function, based on experimentally determined percentiles of the respective distribution or through the equivalent determinant approach [78]. One of the advantages of the potential function approach is that the kernel functions $\phi_j(\mathbf{x})$ are often parameterized so that, by choosing an adequate value of their parameters (e.g., the width of the Gaussians), it is possible to modulate the shape of the class space. An example is shown in Fig. 4.

2.3. PLS-based class modelling

More recently, CM approaches based on the use of Partial Least Squares regression (PLS) have been proposed in the literature, the first of those being One Class-Partial Least Squares (OC-PLS) [83]. OC-PLS uses the same dummy coding as PLS-DA but, as all CM techniques, assumes that only training set samples from the category to be modelled are available. Accordingly, the starting point is to build a PLS model between a dummy response $\mathbf{y}$ (of dimensions $N_c \times 1$, with N_c being the number of training samples belonging to the class) which is a vector made exclusively of ones and the uncentered class training data $\mathbf{X}_c$ ($N_c \times J$):

$$\mathbf{y} = \mathbf{X}_c\mathbf{b} + \mathbf{e}_c = \mathbf{T}_c\mathbf{q}^{\mathrm{T}} + \mathbf{e}_c \tag{5}$$

where $\mathbf{b}$ ($J \times 1$) and $\mathbf{q}$ ($1 \times A$, with A being the number of latent variables) are the PLS regression coefficients and the $\mathbf{y}$-loadings of the PLS model, while $\mathbf{T}_c$ ($N_c \times A$) and $\mathbf{e}_c$ ($N_c \times 1$) are the scores and the $\mathbf{y}$-residuals estimated for the training set samples, respectively. In its original formulation [83], the cross-validated values of $\mathbf{e}_c$ were used to build confidence intervals for the predicted response around the target value of 1 so that if the predicted response for each new sample falls within the calculated confidence interval the individual is accepted by the model, otherwise it is rejected. Eventually, the method was modified to introduce an acceptance criterion similar to the one adopted in Sim-SIMCA, where, together with the value of the prediction residuals, also the Mahalanobis distance of the sample to the center of the PLS score space is considered [84].

Another CM technique based on PLS was proposed by Oliveri et al. [85], and it is called Partial Least Squares-Density Modelling (PLS-DM). The first characteristic element of PLS-DM is the way it defines the target response for the class training samples: indeed, the response $\mathbf{y}$ (called density) is defined as the sum of the distances of each training individual to its k nearest neighbors (with k being an adjustable parameter). A PLS regression model is then calculated between the density vector and the class training data $\mathbf{X}_c$, and then both the scores and the $\mathbf{X}$-residuals (Q statistics) of the model are used to predict if a sample is accepted or not as a member of the investigated category. In particular, potential functions are applied to the PLS scores to estimate a probability distribution and identify an appropriate class boundary; at the same time, a critical value for the $\mathbf{X}$-residuals is calculated as it is done in SIMCA. Accordingly, a sample is accepted by the model only when both the scores- and the orthogonal distance-based classification criteria are simultaneously satisfied. PLS-DM is more flexible than OC-PLS but at the same time it involves the setting of more adjustable parameters (the number of nearest neighbors k, the PLS model complexity and the width of the potential functions used for estimating the probability distribution within the scores space), so model selection can be more cumbersome. In this respect, Oliveri et al. suggest the use of an exhaustive grid search in the parameter space and rely on the Pareto optimality approach for model selection.

The Partial Least Squares-Class Modelling (PLS-CM) technique described in Ref. [86], although presented as a CM approach, actively

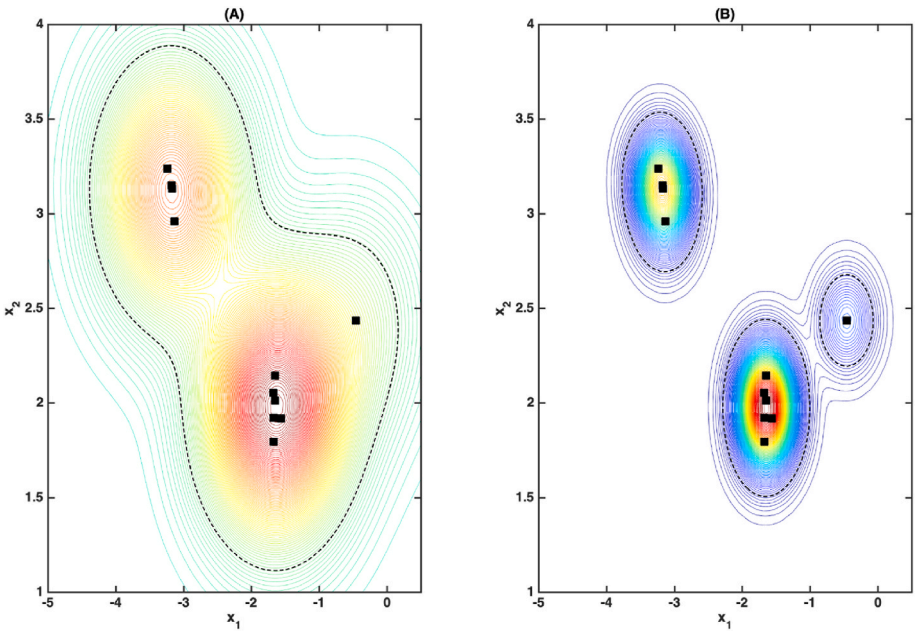

Fig. 4. Example of CM using potential functions. The two panels show the effect of choosing a higher (A) or lower (B) width of the Gaussian kernel functions on the definition of the overall potential density function (whose isodensity levels are represented as contour lines) and on the shape of the class space of the modelled category (enclosed by the black dashed line). The black squares denote the training samples used for building the class model. It is evident how the use of a wider kernel function results in a smoother density and, subsequently, in a less irregular class space.

uses information from class/es other than the one being modelled to define the class subspace of the category of interest; therefore it cannot be strictly considered as a one-class classifier.

2.4. SVM- and ANN-based approaches to class modelling

Differently from what happens in the case of discriminant approaches, since the definition of class spaces involves the identification of bound (closed) regions in the measured variable hyperspace, CM techniques are always non-linear, the extent of non-linearity being moderate for SIMCA, UNEQ and PLS-derived techniques, and tunable in the case of potential functions. In this section, other approaches providing a tunable degree of non-linearity and involving the use of Support Vector Machines (SVM – coupled to non-linear kernel transformations) or Artificial Neural Networks (ANN) will be briefly illustrated and discussed. These methods, especially the ones based on ANN, are still rarely used in the literature (also because they commonly require a larger number of training samples), but they anyway exhibit some interesting features. Undoubtedly, the most popular among these approaches is One Class-Support Vector Machines (OC-SVM) [87]. Briefly, OC-SVM defines the margin problem in a different way with respect to other types of SVM, as it is focused on novelty/outlier detection. Operationally speaking, it searches for the smallest hypersphere enclosing all training samples from a particular class or, in order to make the model not too sensitive to noise, only a certain fraction of them. Of course, if an enlarged feature space is used instead of the original variable one through the kernel trick, the class space can assume a more complex shape than the volume enclosed by the hypersphere.

On the other hand, two different strategies for CM were proposed in the framework of ANN. The first one is based on Kohonen ANN [88]. Kohonen self-organizing maps (SOMs) operate a non-linear topology preserving the projection of samples from the original multivariate space to a discrete 2D space made up of a grid of N_x x N_y positions (neurons) [89]. To turn an unsupervised ANN (the SOM) into a supervised CM tool, the algorithm first augments the class training set with an opportunely chosen set of random vectors, so that only a fraction of the positions in the resulting 2D map are occupied by the samples of the modelled category. Then, a suitable probability distribution for the class under study is calculated as a function of the positions of the samples on the 2D map (usually through a kernel density estimation approach) [79].

The second strategy involves the use of an auto-associator network, which is a neural network architecture characterized by the combination of a set of encoding layers for data compression (returning features that can be considered as sorts of non-linear principal component scores) and a set of decoding layers for data approximation based on the extracted features [90]. When training the network, the same data are used as inputs and target outputs. For CM, each category is described by an auto-associator network, and the corresponding class space is defined according to a distance to the model criterion which, in its first formulation, took into account only the residual standard deviation of the reconstructed input vectors (similarly to the original version of SIMCA) [65,66]. Such a criterion was later modified to account also for the distance to the center of the scores space resulting from the features extracted by the encoding layers [91].

3. Brief comparison of the presented techniques

Based on the overview of the techniques presented in the previous section, some general considerations can be drawn. It was already highlighted how SIMCA is, by far, the most used CM technique. The reasons for this may be sought not only in the fact that, differently from the other ones whose implementations are restricted to custom-written routines or, at most, toolboxes running under programming environments such as Matlab, R or Python, it is the only one to be coded in most commercial software for chemometric data analysis. Indeed, other advantages of SIMCA include its ability to deal straightforwardly with high-dimensional data thanks to the PC compression step, the direct interpretability of the models it yields (that, for instance, can be inspected through a graphical assessment of the corresponding PCA loadings) and its versatility, related not only to the possibility of generalizing it for handling higher-order data structures (e.g., multi-way or multi-block) but also to the different criteria one could rely on for defining the model subspace. This last aspect, though, can also represent a drawback for the less experienced user, as it can be the fact that the

assumption of multivariate normality of the PC scores, which is implicit in the choice of the T^2 statistic, may be insufficient when the classes are highly heterogeneous or the class boundaries are, in general, more irregular. The latter limitation can be partially mitigated by the use of potential functions, which allow to model, at least in principle, any class probability distribution through kernel density estimation. However, for high-dimensional data the use of potential functions requires a preliminary data compression, e.g., through PCA (or PLS, in the case of PLS-DM), so that the model selection phase, which already involves the tuning of the width of the potential functions themselves, involves also the optimization of another metaparameter. On the other hand, UNEQ has the advantages of being firmly rooted in classical statistics, since it is the CM equivalent of quadratic discriminant analysis, and of guaranteeing a straightforward model interpretation. However, as for potential functions, when high-dimensional data are coped with, UNEQ requires an initial data compression step and, as discussed in the case of SIMCA, the multivariate normality assumption may lead to class subspaces which do not match the actual distribution of the samples. SVM with non-linear kernel transformations and ANN exhibit a higher flexibility in terms of the shape of the class space they are able to model but, at the same time, their training requires the optimization of a higher number of metaparameters which is, in general, more cumbersome and less straightforward; moreover, especially in the case of ANN, a higher number of training individuals is also needed. Lastly, both ANN and SVM do not enable a direct interpretation of the resulting models in terms of importance of the original measured variables.

4. Perspectives

Although the history of CM began almost 50 years ago, thanks to the seminal work of Svante Wold and his collaborators, in the last decades new life seems to have been breathed into this particular domain. In fact, the capability of CM approaches to deal with complex multivariate datasets as well as their inherent robustness against effects that can be extremely deleterious when utilizing more classical discriminant approaches (e.g., category unbalancedness, unequal sample size, etc.) have lately attracted much attention from users and practitioners of a wide variety of fields of interest. For this reason, it goes without saying that the job of chemometricians in this sense is not finalized yet. This is especially demonstrated by the significant advances that they have recently made, mainly inspired by the novel challenges that the innovative measurement devices of the modern era of Big Data, Industry 4.0 and Internet of Things currently pose. Just to mention a few, such advances encompass: i) the development of new CM strategies based on alternative approaches of multivariate data analysis like Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS [23]), ii) the integration of non-parametric, semi-parametric or Bayesian decision rules into existing CM solutions [92–94] and iii) the extension of available CM methodologies for probabilistic [95], multi-block [43,96] and multi-way classification [76–78]. Also CM-based weak classifiers have been recently proposed [97]. Still, further un- or partly-explored research lines can be easily envisioned in this context: as examples, adapting algorithms like SIMCA, OC-PLS and PLS-DM for the analysis of non-linear data structures (relying, e.g., on the principle of non-linear kernel transformations [98]) or designing tools for the visualization of the importance or relevance of the recorded variables in SIMCA and UNEQ models [99] (exploiting, for instance, the ideas behind the well-established contribution plots [100] and/or the projection of pseudo-samples [101,102]) may represent intriguing subjects of study. Finally, another aspect that would be worth investigating is the possibility of evaluating the uncertainty associated to the classification of individual samples.

CRediT authorship contribution statement

Lorenzo Strani: Writing – review & editing, Methodology. **Marina Cocchi:** Writing – original draft, Methodology, Conceptualization. **Daniele Tanzilli:** Writing – review & editing, Visualization. **Alessandra Biancolillo:** Writing – original draft, Visualization, Methodology. **Federico Marini:** Writing – original draft, Visualization, Methodology, Conceptualization. **Raffaele Vitale:** Writing – original draft, Visualization, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research has been developed in the frame of the COST Action CA19145 “European Network for Assuring Food Integrity using Non-Destructive Spectral Sensors” (SENSORFINT)

Data availability

No data was used for the research described in the article.

References

[1] N. Cavallini, A. Biancolillo, L. Strani, C. Durante, M. Cocchi, Chapter 5: food forensics, in: Chemom. Methods Forensic Sci, 2023, pp. 90–136. https://books.rsc.org/books/edited-volume/2155/chapter/7824122/Food-Forensics. (Accessed 16 May 2024).
[2] R. Vitale, G. Spinaci, F. Marini, P. Marion, M. Delcroix, A. Vieillard, F. Coudon, O. Devos, C. Ruckebusch, Hierarchical classification and matching of mid-infrared spectra of paint samples for forensic applications, Talanta 243 (2022) 123360, https://doi.org/10.1016/j.talanta.2022.123360.
[3] L. Brunnbauer, Z. Gajarska, H. Lohninger, A. Limbeck, A critical review of recent trends in sample classification using Laser-Induced Breakdown Spectroscopy (LIBS), TrAC Trends Anal. Chem. 159 (2023) 116859, https://doi.org/10.1016/j.trac.2022.116859.
[4] C. Malegori, E. Alladio, P. Oliveri, C. Manis, M. Vincenti, P. Garofano, F. Barni, A. Berti, Identification of invisible biological traces in forensic evidences by hyperspectral NIR imaging combined with chemometrics, Talanta 215 (2020) 120911, https://doi.org/10.1016/j.talanta.2020.120911.
[5] Y.V. Zontov, K.S. Balyklova, A.V. Titova, O.Ye Rodionova, A.L. Pomerantsev, Chemometric aided NIR portable instrument for rapid assessment of medicine quality, J. Pharm. Biomed. Anal. 131 (2016) 87–93, https://doi.org/10.1016/j.jpba.2016.08.008.
[6] R. Vitale, M. Cocchi, A. Biancolillo, C. Ruckebusch, F. Marini, Class modelling by soft independent modelling of class analogy: why, when, how? A tutorial, Anal. Chim. Acta 1270 (2023) 341304, https://doi.org/10.1016/j.aca.2023.341304.
[7] P. Oliveri, C. Malegori, E. Mustorgi, M. Casale, Qualitative pattern recognition in chemistry: theoretical background and practical guidelines, Microchem. J. 162 (2021) 105725, https://doi.org/10.1016/j.microc.2020.105725.
[8] M. Cocchi, A. Biancolillo, F. Marini, in: J. Jaumot, C. Bedia, R. Tauler (Eds.), Data analysis for omic sciences: Methods and applications, Comprehensive Analytical Chemistry, 80, Elsevier, 2018, pp. 265–299, https://doi.org/10.1016/bs.coac.2018.08.006.
[9] O.Ye Rodionova, P. Oliveri, C. Malegori, A.L. Pomerantsev, Chemometrics as an efficient tool for food authentication: golden pillars for building reliable models, Trends Food Sci. Technol. 147 (2024) 104429, https://doi.org/10.1016/j.tifs.2024.104429.
[10] M. Cocchi, Chemometrics for food quality control and authentication, in: Encycl. Anal. Chem., John Wiley & Sons, Ltd, 2017, pp. 1–29, https://doi.org/10.1002/9780470027318.a9579.
[11] O.Ye Rodionova, A.V. Titova, A.L. Pomerantsev, Discriminant analysis is an inappropriate method of authentication, TrAC Trends Anal. Chem. 78 (2016) 17–22, https://doi.org/10.1016/j.trac.2016.01.010.
[12] P. Oliveri, C. Malegori, E. Mustorgi, M. Casale, Application of Chemometrics in the Food Sciences, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, 2nd ed, vol. 4, 2019, pp. 99–111, https://doi.org/10.1016/B978-0-12-409547-2.14748-1.
[13] R. Calvini, G. Orlandi, G. Foca, A. Ulrici, Development of a classification algorithm for efficient handling of multiple classes in sorting systems based on hyperspectral imaging, J. Spectr. Imaging 7 (2018) 1–15, https://doi.org/10.1255/jsi.2018.a13.
[14] A.L. Pomerantsev, O.Ye Rodionova, Multiclass partial least squares discriminant analysis: taking the right way—a critical tutorial, J. Chemom. 32 (2018) e3030, https://doi.org/10.1002/cem.3030.

[15] Z. Małyjurek, D. de Beer, H. van Schoor, J. Colling, E. Joubert, B. Walczak, Class-modelling of overlapping classes. A two-step authentication approach, Anal. Chim. Acta 1191 (2022) 339284, https://doi.org/10.1016/j.aca.2021.339284.
[16] V. Ferrari, R. Calvini, C. Menozzi, A. Ulrici, M. Bragolusi, R. Piro, A. Tata, M. Suman, G. Foca, Addressing adulteration challenges of dried oregano leaves by NIR HyperSpectral Imaging, Chemometr. Intell. Lab. Syst. 249 (2024) 105133, https://doi.org/10.1016/j.chemolab.2024.105133.
[17] M. Foschi, A. D'Addario, A. Antonio D'Archivio, A. Biancolillo, Future foods protection: supervised chemometric approaches for the determination of adulterated insects' flours for human consumption by means of ATR-FTIR spectroscopy, Microchem. J. 183 (2022) 108021, https://doi.org/10.1016/j.microc.2022.108021.
[18] F. Di Donato, A.A. D'Archivio, M.A. Maggi, L. Rossi, Detection of plant-derived adulterants in saffron (crocus sativus L.) by HS-SPME/GC-MS profiling of volatiles and chemometrics, Food Anal. Methods 14 (2021) 784–796, https://doi.org/10.1007/s12161-020-01941-x.
[19] Z. Małyjurek, D. de Beer, E. Joubert, S. Koch, B. Zawisza, B. Walczak, Adulteration detection of natural samples using a class-modelling approach – application to rooibos and honeybush herbal teas, J. Food Compos. Anal. 131 (2024) 106208, https://doi.org/10.1016/j.jfca.2024.106208.
[20] J.I. Ballesteros, L.H.V. Lim, R.B. Lamorena, The feasibility of using ATR-FTIR spectroscopy combined with one-class support vector machine in screening turmeric powders, Vib. Spectrosc. 130 (2024) 103646, https://doi.org/10.1016/j.vibspec.2023.103646.
[21] C.S.W. Miaw, M.L.C. Martins, M.M. Sena, S.V.C. de Souza, Screening method for the detection of other allergenic nuts in cashew nuts using chemometrics and a portable near-infrared spectrophotometer, Food Anal. Methods 15 (2022) 1074–1084, https://doi.org/10.1007/s12161-021-02184-0.
[22] V.G. Kelis Cardoso, R.J. Poppi, Cleaner and faster method to detect adulteration in cassava starch using Raman spectroscopy and one-class support vector machine, Food Control 125 (2021) 107917, https://doi.org/10.1016/j.foodcont.2021.107917.
[23] S.K. Karimvand, A. Pahlevan, S.V. Zade, J.M. Jafari, H. Abdollahi, Multivariate curve resolution-soft independent modelling of class analogy (MCR-SIMCA), Anal. Chim. Acta 1291 (2024) 342205, https://doi.org/10.1016/j.aca.2024.342205.
[24] S. Vali Zade, E. Forooghi, B. Jannat, F. Hashempour-baltork, H. Abdollahi, A combined classification modeling strategy for detection and identification of extra virgin olive oil adulteration using Raman spectroscopy, Chemometr. Intell. Lab. Syst. 240 (2023) 104903, https://doi.org/10.1016/j.chemolab.2023.104903.
[25] F.S. Hashemi-Nasab, S. Talebian, H. Parastar, Multiple adulterants detection in turmeric powder using Vis-SWNIR hyperspectral imaging followed by multivariate curve resolution and classification techniques, Microchem. J. 185 (2023) 108203, https://doi.org/10.1016/j.microc.2022.108203.
[26] J.P. Cruz-Tirado, R.L. de França, M. Tumbajulca, G. Barraza-Jáuregui, D. F. Barbin, R. Siche, Detection of cumin powder adulteration with allergenic nutshells using FT-IR and portable NIRS coupled with chemometrics, J. Food Compos. Anal. 116 (2023) 105044, https://doi.org/10.1016/j.jfca.2022.105044.
[27] S. Ehsani, H. Yazdanpanah, H. Parastar, An innovative screening approach for orange juice authentication using dual portable/handheld NIR spectrometers and chemometrics, Microchem. J. 194 (2023) 109304, https://doi.org/10.1016/j.microc.2023.109304.
[28] J.L.D.P. Teixeira, E.T.D.S. Caramês, D.P. Baptista, M.L. Gigante, J.A.L. Pallone, Adulteration detection in goat dairy beverage through NIR spectroscopy and DD-SIMCA, Food Anal. Methods 15 (2022) 783–791, https://doi.org/10.1007/s12161-021-02151-9.
[29] D. Galvan, C.A. Lelis, L. Effting, F.L. Melquiades, E. Bona, C.A. Conte-Junior, Low-cost spectroscopic devices with multivariate analysis applied to milk authenticity, Microchem. J. 181 (2022) 107746, https://doi.org/10.1016/j.microc.2022.107746.
[30] M.D.G. Neves, R.J. Poppi, Authentication and identification of adulterants in virgin coconut oil using ATR/FTIR in tandem with DD-SIMCA one class modeling, Talanta 219 (2020) 121338, https://doi.org/10.1016/j.talanta.2020.121338.
[31] H. Chen, C. Tan, H. Li, Untargeted identification of adulterated Sanqi powder by near-infrared spectroscopy and one-class model, J. Food Compos. Anal. 88 (2020) 103450, https://doi.org/10.1016/j.jfca.2020.103450.
[32] C.S.W. Miaw, P.M. Santos, A.R.C.S. Silva, A. Gozzi, N.C.C. Guimarães, M. P. Callao, I. Ruisánchez, M.M. Sena, S.V.C. de Souza, Comparison of different multivariate classification methods for the detection of adulterations in grape nectars by using low-field nuclear magnetic resonance, Food Anal. Methods 13 (2020) 108–118, https://doi.org/10.1007/s12161-019-01522-7.
[33] C. Tan, H. Chen, Z. Lin, An improved one-class algorithm combined with NIR spectroscopy for detecting adulterated chemicals in weight-loss pills, Infrared Phys. Technol. 133 (2023) 104817, https://doi.org/10.1016/j.infrared.2023.104817.
[34] F. Di Donato, A. Biancolillo, M. Foschi, V. Di Cecco, L. Di Martino, A. A. D'Archivio, Authentication of typical Italian bell pepper spices by ICP-OES multi-elemental analysis combined with SIMCA class modelling, J. Food Compos. Anal. 115 (2023) 104948, https://doi.org/10.1016/j.jfca.2022.104948.
[35] M. Li Vigni, C. Durante, S. Michelini, M. Nocetti, M. Cocchi, Preliminary assessment of parmigiano reggiano authenticity by handheld Raman spectroscopy, Foods 9 (2020) 1563, https://doi.org/10.3390/foods9111563.
[36] R. Jahani, S. van Ruth, Y. Weesepoel, M. Alewijn, F. Kobarfard, M. Faizi, M. H. Shojaee AliAbadi, A. Mahboubi, A. Nasiri, H. Yazdanpanah, Comparison of portable and benchtop near-infrared spectrometers for the detection of citric acid-adulterated lime juice: a chemometrics approach, Iran. J. Pharm. Res. IJPR 21 (2022) e128372, https://doi.org/10.5812/ijpr-128372.
[37] D. Ceccarelli, F. Antonucci, C. Costa, C. Talento, R. Ciccoritti, An artificial class modelling approach to identify the most largely diffused cultivars of sweet cherry (*Prunus avium* L.) in Italy, Food Chem. 333 (2020) 127515, https://doi.org/10.1016/j.foodchem.2020.127515.
[38] M.P. Totaro, G. Squeo, D. De Angelis, A. Pasqualone, F. Caponio, C. Summo, Application of NIR spectroscopy coupled with DD-SIMCA class modelling for the authentication of pork meat, J. Food Compos. Anal. 118 (2023) 105211, https://doi.org/10.1016/j.jfca.2023.105211.
[39] B. Moser, Z. Jandric, C. Troyer, L. Priemetzhofer, K.J. Domig, H. Jäger, S.P. van den Oever, H.K. Mayer, S. Hann, A. Zitek, Evaluation of spectral handheld devices for freshness assessment of carp and trout fillets in relation to standard methods including non-targeted metabolomics, Food Control 152 (2023) 109835, https://doi.org/10.1016/j.foodcont.2023.109835.
[40] H.R. Tan, L.Y. Chan, H.H. Lee, Y.-Q. Xu, W. Zhou, Rapid authentication of Chinese oolong teas using atmospheric solids analysis probe-mass spectrometry (ASAP-MS) combined with supervised pattern recognition models, Food Control 134 (2022) 108736, https://doi.org/10.1016/j.foodcont.2021.108736.
[41] Y. Li, Q. Bi, W. Wei, C. Yao, J. Zhang, D. Guo, Sequential decision fusion pipeline for the high-throughput species recognition of medicinal caterpillar fungus by using ATR-FTIR, Microchem. J. 179 (2022) 107437, https://doi.org/10.1016/j.microc.2022.107437.
[42] F.S. Hashemi-Nasab, H. Parastar, Vis-NIR hyperspectral imaging coupled with independent component analysis for saffron authentication, Food Chem. 393 (2022) 133450, https://doi.org/10.1016/j.foodchem.2022.133450.
[43] D. Galvan, J.C. de Andrade, C.A. Conte-Junior, M.H.M. Killner, E. Bona, DD-ComDim: a data-driven multiblock approach for one-class classifiers, Chemometr. Intell. Lab. Syst. 233 (2023) 104748, https://doi.org/10.1016/j.chemolab.2022.104748.
[44] N.A. Mahynski, J.I. Monroe, D.A. Sheen, R.L. Paul, H.H. Chen-Mayer, V.K. Shen, Classification and authentication of materials using prompt gamma ray activation analysis, J. Radioanal. Nucl. Chem. 332 (2023) 3259–3271, https://doi.org/10.1007/s10967-023-09024-x.
[45] H. Chen, Z. Lin, C. Tan, Application of near-infrared spectroscopy and class-modeling to antibiotic authentication, Anal. Biochem. 590 (2020) 113514, https://doi.org/10.1016/j.ab.2019.113514.
[46] M. Foschi, A. Biancolillo, F. Marini, F. Cosentino, F. Di Donato, A.A. D'Archivio, Multi-block approach for the characterization and discrimination of Italian chickpeas landraces, Food Control 157 (2024) 110170, https://doi.org/10.1016/j.foodcont.2023.110170.
[47] F. Di Donato, F. Squeo, A. Biancolillo, L. Rossi, A.A. D'Archivio, Characterization of high value Italian chickpeas (*Cicer arietinum* L.) by means of ICP-OES multi-elemental analysis coupled with chemometrics, Food Control 131 (2022) 108451, https://doi.org/10.1016/j.foodcont.2021.108451.
[48] A. Nardecchia, R. Presutto, R. Bucci, F. Marini, A. Biancolillo, Authentication of the geographical origin of "vallerano" chestnut by near infrared spectroscopy coupled with chemometrics, Food Anal. Methods 13 (2020) 1782–1790, https://doi.org/10.1007/s12161-020-01791-7.
[49] M. Foschi, A.A. D'Archivio, L. Rossi, Geographical discrimination and authentication of lentils (*Lens culinaris* Medik.) by ICP-OES elemental analysis and chemometrics, Food Control 118 (2020) 107438, https://doi.org/10.1016/j.foodcont.2020.107438.
[50] M. Perini, S. Pianezze, L. Ziller, M. Ferrante, F. Ferella, S. Nisi, M. Foschi, A. A. D'Archivio, Stable isotope ratio analysis combined with inductively coupled plasma-mass spectrometry for geographical discrimination between Italian and foreign saffron, J. Mass Spectrom. 55 (2020) e4595, https://doi.org/10.1002/jms.4595.
[51] I.W. Burton, M. Kompany-Zareh, S. Haverstock, J. Haché, C.F. Martinez-Farina, P. D. Wentzell, F. Berrué, Analysis and discrimination of Canadian honey using quantitative NMR and multivariate statistical methods, Molecules 28 (2023) 1656, https://doi.org/10.3390/molecules28041656.
[52] A. Biancolillo, M. Foschi, L. D'Alonzo, V. Di Cecco, M. Di Santo, L. Di Martino, A. A. D'Archivio, Green chemometric-assisted characterization of common and black varieties of celery, Molecules 28 (2023) 1181, https://doi.org/10.3390/molecules28031181.
[53] G. Bruschetta, A. Notti, G. Lando, A. Ferlazzo, A promising 31P NMR-multivariate analysis approach for the identification of milk phosphorylated metabolites and for rapid authentication of milk samples, Biochem. Biophys. Rep. 27 (2021) 101087, https://doi.org/10.1016/j.bbrep.2021.101087.
[54] L.B. dos Santos, J. Tarabal, M.M. Sena, M.R. Almeida, UV-Vis spectroscopy and one-class modeling for the authentication of the geographical origin of green coffee beans from Cerrado Mineiro, Brazil, J. Food Compos. Anal. 123 (2023) 105555, https://doi.org/10.1016/j.jfca.2023.105555.
[55] M.R. Baqueta, F. Marini, R.B. Rocha, P. Valderrama, J.A.L. Pallone, Authentication and discrimination of new Brazilian Canephora coffees with geographical indication using a miniaturized near-infrared spectrometer, Food Res. Int. 172 (2023) 113216, https://doi.org/10.1016/j.foodres.2023.113216.
[56] L. Strojnik, D. Potočnik, M. Jagodic Hudobivnik, D. Mazej, B. Japelj, N. Škrk, S. Marolt, D. Heath, N. Ogrinc, Geographical identification of strawberries based on stable isotope ratio and multi-elemental analysis coupled with multivariate statistical analysis: a Slovenian case study, Food Chem. 381 (2022) 132204, https://doi.org/10.1016/j.foodchem.2022.132204.
[57] M.J. Hidalgo, D.C. Fechner, D. Ballabio, E.J. Marchevsky, R.G. Pellerano, Traceability of soybeans produced in Argentina based on their trace element profiles, J. Chemom. 34 (2020) e3252, https://doi.org/10.1002/cem.3252.

[58] R. Brendel, S. Schwolow, N. Gerhardt, J. Schwab, P. Rau, M. Oest, S. Rohn, P. Weller, MIR spectroscopy *versus* MALDI-ToF-MS for authenticity control of honeys from different botanical origins based on soft independent modelling by class analogy (SIMCA) – a clash of techniques? Spectrochim. Acta. A. Mol. Biomol. Spectrosc. 263 (2021) 120225 https://doi.org/10.1016/j.saa.2021.120225.
[59] E. Forooghi, S. Vali Zade, H. Sahebi, H. Abdollahi, N. Sadeghi, B. Jannat, Authentication and discrimination of tissue origin of bovine gelatin using combined supervised pattern recognition strategies, Microchem. J. 187 (2023) 108417, https://doi.org/10.1016/j.microc.2023.108417.
[60] K. Tkachenko, I. Esteban-Díez, J.M. González-Sáiz, P. Pérez-Matute, C. Pizarro, Dual classification approach for the rapid discrimination of metabolic syndrome by FTIR, Biosensors 13 (2023) 15, https://doi.org/10.3390/bios13010015.
[61] H. Chen, Z. Lin, C. Tan, Automatic cancer discrimination based on near-infrared spectrum and class-modeling technique, Vib. Spectrosc. 106 (2020) 102991, https://doi.org/10.1016/j.vibspec.2019.102991.
[62] D. Castro-Reigía, M.C. Ortiz, S. Sanllorente, I. García, L.A. Sarabia, PLS class modelling using error correction output code matrices, entropy and NIR spectroscopy to detect deficiencies in pastry doughs, Chemometr. Intell. Lab. Syst. 246 (2024) 105092, https://doi.org/10.1016/j.chemolab.2024.105092.
[63] J. Omar, A. Boix, F. Ulberth, Raman spectroscopy for quality control and detection of substandard painkillers, Vib. Spectrosc. 111 (2020) 103147, https://doi.org/10.1016/j.vibspec.2020.103147.
[64] C. Beatriz Figoli, M. Garcea, C. Bisioli, V. Tafintseva, V. Shapaval, M.G. Peña, L. Gibbons, F. Althabe, O. Miguel Yantorno, M. Horton, J. Schmitt, P. Lasch, A. Kohler, A. Bosch, A robust metabolomics approach for the evaluation of human embryos from in vitro fertilization, Analyst 146 (2021) 6156–6169, https://doi.org/10.1039/D1AN01191J.
[65] S. Wold, Pattern recognition by means of disjoint principal components models, Pattern Recogn. 8 (1976) 127–139, https://doi.org/10.1016/0031-3203(76)90014-5.
[66] S. Wold, M. Sjöström, SIMCA: a method for analyzing chemical data in terms of similarity and analogy, in: Chemom. Theory Appl, American Chemical Society, 1977, pp. 243–282, https://doi.org/10.1021/bk-1977-0052.ch012.
[67] K.Pearson, LIII. On lines and planes of closest fit to systems of points in space, London, Edinburgh Dublin Phil. Mag. J. Sci. (1901), https://doi.org/10.1080/14786440109462720.
[68] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933) 417–441, https://doi.org/10.1037/h0071325.
[69] C. Albano, W. Dunn, U. Edlund, E. Johansson, B. Nordén, M. Sjöström, S. Wold, Four levels of pattern recognition, Anal. Chim. Acta 103 (1978) 429–443, https://doi.org/10.1016/S0003-2670(01)83107-X.
[70] Eigenvector Research, Inc, SIMCA model builder GUI. https://www.wiki.eigenvector.com/index.php?title=SIMCA_Model_Builder_GUI, 2021.
[71] H.H. Yue, S.J. Qin, Reconstruction-based fault identification using a combined Index, Ind. Eng. Chem. Res. 40 (2001) 4403–4414, https://doi.org/10.1021/ie000141+.
[72] A.L. Pomerantsev, Acceptance areas for multivariate classification derived by projection methods, J. Chemom. 22 (2008) 601–609, https://doi.org/10.1002/cem.1147.
[73] A.L. Pomerantsev, O.Y. Rodionova, Concept and role of extreme objects in PCA/SIMCA, J. Chemom. 28 (2014) 429–438, https://doi.org/10.1002/cem.2506.
[74] O.Ye Rodionova, P. Oliveri, A.L. Pomerantsev, Rigorous and compliant approaches to one-class classification, Chemometr. Intell. Lab. Syst. 159 (2016) 89–96, https://doi.org/10.1016/j.chemolab.2016.10.002.
[75] R. Vitale, F. Marini, C. Ruckebusch, SIMCA modeling for overlapping classes: fixed or optimized decision threshold? Anal. Chem. 90 (2018) 10738–10747, https://doi.org/10.1021/acs.analchem.8b01270.
[76] C. Durante, R. Bro, M. Cocchi, A classification tool for N-way array based on SIMCA methodology, Chemometr. Intell. Lab. Syst. 106 (2011) 73–85, https://doi.org/10.1016/j.chemolab.2010.09.004.
[77] M. Cocchi, M. Li Vigni, C. Durante, in: S. Brown, R. Tauler, B. Walczak (Eds.), 2nd ed.Comprehensive Chemometrics, 3, Elsevier, Oxford, 2020, pp. 701–721, https://doi.org/10.1016/B978-0-12-409547-2.14590-1.
[78] A.P. Pagani, G. Camargo, G.A. Ibañez, A.C. Olivieri, A.L. Pomerantsev, O. Ye Rodionova, Data-driven version of multiway soft independent modeling of class analogy (N-way DD-SIMCA): theory and application, Anal. Chem. 96 (2024) 4845–4853, https://doi.org/10.1021/acs.analchem.3c05096.
[79] A. Smilde, R. Bro, P. Geladi, Multi-way analysis with applications, in: Multiway Anal. Chem. Sci., John Wiley & Sons, 2004, pp. 35–45, https://doi.org/10.1002/0470012110.
[80] M.P. Derde, D.L. Massart, UNEQ: a disjoint modelling technique for pattern recognition based on normal distribution, Anal. Chim. Acta 184 (1986) 33–51, https://doi.org/10.1016/S0003-2670(00)86468-5.
[81] M. Forina, C. Armanino, R. Leardi, G. Drava, A class-modelling technique based on potential functions, J. Chemom. 5 (1991) 435–453, https://doi.org/10.1002/cem.1180050504.
[82] E. Parzen, On estimation of a probability density function and mode, Ann. Math. Stat. 33 (1962) 1065–1076.
[83] L. Xu, H. Fu, N. Jiang, X. Yu, A new class model based on partial least square regression and its applications for identifying authenticity of bezoar samples, Chin. J. Anal. Chem. (2010) 175–180.
[84] L. Xu, S.-M. Yan, C.-B. Cai, X.-P. Yu, One-class partial least squares (OCPLS) classifier, Chemometr. Intell. Lab. Syst. 126 (2013) 1–5, https://doi.org/10.1016/j.chemolab.2013.04.008.
[85] P. Oliveri, M.I. López, M.C. Casolino, I. Ruisánchez, M.P. Callao, L. Medini, S. Lanteri, 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 851 (2014) 30–36, https://doi.org/10.1016/j.aca.2014.09.013.
[86] M.S. Sánchez, M.C. Ortiz, L.A. Sarabia, Two class-modelling techniques that give families of class-models and their relation with the structure of the data, Anal. Bioanal. Chem. 399 (2011) 1941–1950, https://doi.org/10.1007/s00216-010-4291-6.
[87] D.M.J. Tax, R.P.W. Duin, Support vector domain description, Pattern Recogn. Lett. 20 (1999) 1191–1199, https://doi.org/10.1016/S0167-8655(99)00087-2.
[88] F. Marini, J. Zupan, A.L. Magrì, Class-modeling using Kohonen artificial neural networks, Anal. Chim. Acta 544 (2005) 306–314, https://doi.org/10.1016/j.aca.2004.12.026.
[89] T. Kohonen, Self-Organization and Associative Memory, Springer Science & Business Media, 2012.
[90] F. Marini, A.L. Magrì, R. Bucci, Multilayer feed-forward artificial neural networks for class modeling, Chemometr. Intell. Lab. Syst. 88 (2007) 118–124, https://doi.org/10.1016/j.chemolab.2006.07.004.
[91] F. Marini, Non-linear class-modeling using artificial neural networks, in: 2009 Sixth Int. Conf. Fuzzy Syst. Knowl. Discov., IEEE, Tianjin, China, 2009, pp. 271–273, https://doi.org/10.1109/FSKD.2009.805.
[92] A.T. Hermane, S. Pierre-Yves, L. Pierre, H. Philippe, Z. Eric, A probabilistic class-modelling method based on prediction bands for functional spectral data: methodological approach and application to near-infrared spectroscopy, Anal. Chim. Acta 1144 (2021) 130–149, https://doi.org/10.1016/j.aca.2020.11.039.
[93] T.H. Avohou, P.-Y. Sacré, S. Hamla, P. Lebrun, P. Hubert, É. Ziemons, Optimizing the soft independent modeling of class analogy (SIMCA) using statistical prediction regions, Anal. Chim. Acta 1229 (2022) 340339, https://doi.org/10.1016/j.aca.2022.340339.
[94] T.H. Avohou, P.-Y. Sacré, P. Hubert, E. Ziemons, Interpretable one-class classification of Raman spectra using prediction bands estimated by wavelet regression, Anal. Chem. 94 (2022) 4183–4191, https://doi.org/10.1021/acs.analchem.1c04098.
[95] R. Vitale, F. Marini, C. Ruckebusch, A. Smolinska, p-SIMCA: a Non-parametric Probabilistic Version of the SIMCA Classifier, Chimiométrie XXII, Virtual Meeting, 2021.
[96] O. Rodionova, A. Pomerantsev, Multi-block DD-SIMCA as a high-level data fusion tool, Anal. Chim. Acta 1265 (2023) 341328, https://doi.org/10.1016/j.aca.2023.341328.
[97] T. Lemos, J.H. Kalivas, Self-optimized one-class classification using sum of ranking differences combined with a receiver operator characteristic curve, Anal. Chem. 92 (2020) 5354–5361, https://doi.org/10.1021/acs.analchem.0c00017.
[98] B. Schölkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond, MIT Press, 2002.
[99] A. Grandi, Sviluppo di approcci per valutare l'importanza delle variabili in modelli di classe, Master's thesis, Corso di Laurea Magistrale in Chimica Analitica, Università degli Studi di Roma "La Sapienza," (2021).
[100] T. Kourti, J.F. MacGregor, Multivariate SPC methods for process and product monitoring, J. Qual. Technol. 28 (1996) 409–428, https://doi.org/10.1080/00224065.1996.11979699.
[101] R. Vitale, D. Palací-López, H.H.M. Kerkenaar, G.J. Postma, L.M.C. Buydens, A. Ferrer, Kernel-Partial Least Squares regression coupled to pseudo-sample trajectories for the analysis of mixture designs of experiments, Chemometr. Intell. Lab. Syst. 175 (2018) 37–46, https://doi.org/10.1016/j.chemolab.2018.02.002.
[102] L. Blanchet, R. Vitale, R. van Vorstenbosch, G. Stavropoulos, J. Pender, D. Jonkers, F.-J. van Schooten, A. Smolinska, Constructing bi-plots for random forest: tutorial, Anal. Chim. Acta 1131 (2020) 146–155, https://doi.org/10.1016/j.aca.2020.06.043.