This is the peer reviewd version of the followng article:

Class modelling by soft independent modelling of class analogy: why, when, how? A tutorial / Vitale, Raffaele; Cocchi, Marina; Biancolillo, Alessandra; Ruckebusch, Cyril; Marini, Federico. - In: ANALYTICA CHIMICA ACTA. - ISSN 0003-2670. - 1270:(2023), pp. 341304-341359. [10.1016/j.aca.2023.341304]

Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

09/05/2024 07:24

(Article begins on next page)

Journal Pre-proof

Class modelling by soft independent modelling of class analogy: why, when, how? A tutorial

Raffaele Vitale, Marina Cocchi, Alessandra Biancolillo, Cyril Ruckebusch, Federico Marini

PII: S0003-2670(23)00525-1

DOI: https://doi.org/10.1016/j.aca.2023.341304

Reference: ACA 341304

To appear in: *Analytica Chimica Acta*

Received Date: 12 January 2023

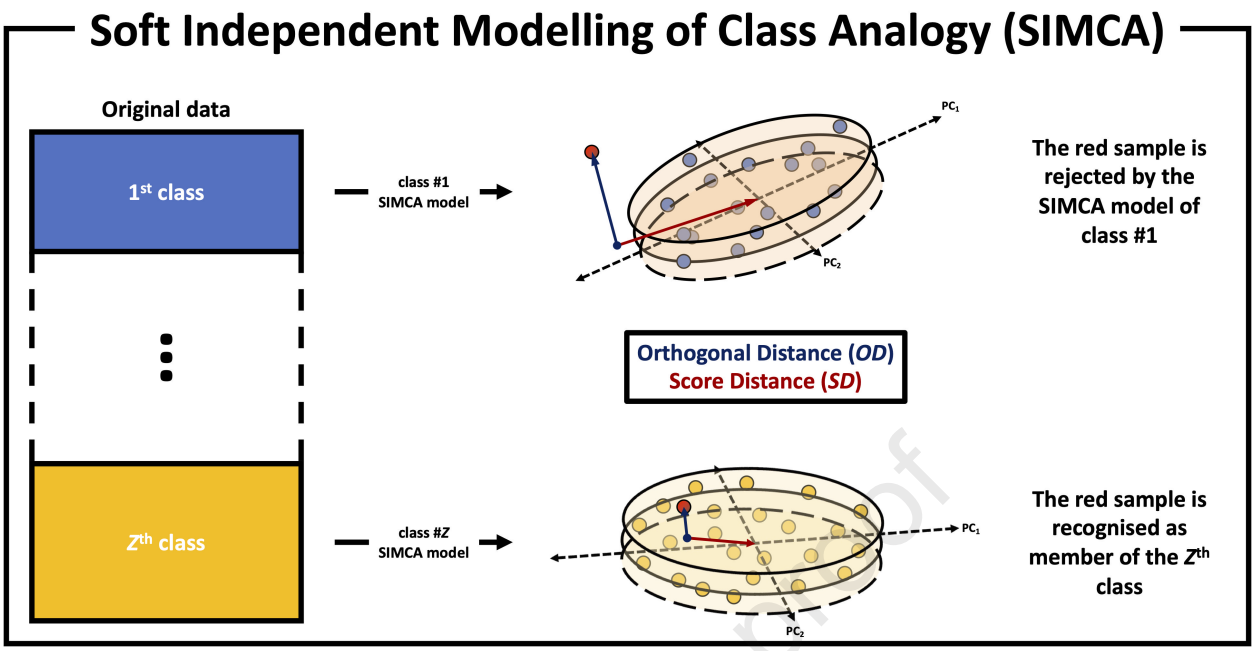
Revised Date: 27 April 2023

Accepted Date: 28 April 2023

Please cite this article as: R. Vitale, M. Cocchi, A. Biancolillo, C. Ruckebusch, F. Marini, Class modelling by soft independent modelling of class analogy: why, when, how? A tutorial, *Analytica Chimica Acta* (2023), doi: https://doi.org/10.1016/j.aca.2023.341304.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2023 Published by Elsevier B.V.



Journal Pre-proof

Journal Pre-proof

Class modelling by Soft Independent Modelling of Class Analogy: why, when, how? A tutorial

Raffaele Vitale[a,*], Marina Cocchi[b], Alessandra Biancolillo[c], Cyril Ruckebusch[a], Federico Marini[d]

[a] *U. Lille, CNRS, LASIRE, Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, Cité Scientifique, F-59000 Lille, France*
[b] *Dipartimento di Scienze Chimiche e Geologiche, Università degli Studi di Modena e Reggio Emilia, Via Giuseppe Campi 103, 41125 Modena, Italy*
[c] *Dipartimento di Scienze Fisiche e Chimiche, Università degli Studi dell'Aquila, Via Vetoio (Coppito 2, Edificio "Angelo Camillo De Meis"), 67100 Coppito, Italy*
[d] *Dipartimento di Chimica, Università degli Studi di Roma "La Sapienza", Piazzale Aldo Moro 5, 00185 Roma, Italy*

Abstract

This article contains a comprehensive tutorial on classification by means of Soft Independent Modelling of Class Analogy (SIMCA). Such a tutorial was conceived in an attempt to offer pragmatic guidelines for a sensible and correct utilisation of this tool as well as answers to three basic questions: "*why employing SIMCA?*", "*when employing SIMCA?*" and "*how employing/not employing SIMCA?*". With this purpose in mind, the following points are here addressed: i) the mathematical and statistical fundamentals of the SIMCA approach are presented; ii) distinct variants of the original SIMCA algorithm are thoroughly described and compared in two different case-studies; iii) a flowchart outlining how to fine-tune the parameters of a SIMCA model for achieving an optimal performance is provided; iv) figures of merit and graphical tools for SIMCA model assessment are illustrated and v) computational details and rational suggestions about SIMCA model validation are given. Moreover, a novel Matlab toolbox, which encompasses routines and functions for running and contrasting all the aforementioned SIMCA versions is also made available.

Keywords: class modelling (CM), Soft Independent Modelling of Class Analogy (SIMCA), Principal Component Analysis (PCA), Orthogonal Distance (OD), Score Distance (SD)

[*]Corresponding author:
Telephone number: +33769476654
Email address: raffaele.vitale@univ-lille.fr (Raffaele Vitale)

Journal Pre-proof

1. Introduction

Classification problems are ubiquitous in analytical chemistry. Food product authentication and quality assessment [1–5], pharmaceutical counterfeit detection [6–9] and forensic trace characterisation [10–12] are just few scenarios in which practitioners may utilise statistical and machine learning approaches in order to distinguish objects based on the specific types of data recorded for them. In spite of their intrinsic methodological diversity, though, in chemometrics, such approaches are frequently categorised in two broad families: those performing *discrimination* and those based on the principles of the so-called *class modelling* (CM) [13]. The difference between these two groups of strategies can be easily visualised through the following example (see also Figure 1 for a graphical illustration): suppose one has collected a given number of samples (for instance, blood extracts) belonging to two particular classes (being withdrawn from healthy and diseased patients) and has measured for them the values of two characteristic variables (*e.g.*, chemical or physical parameters) with the goal of discerning such classes. If one represents these

Figure 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 recognised as 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).

two groups of specimens in a bivariate plot as the ones in Figures 1a and 1b, statistical classification translates into trying to define boundaries or frontiers within the graphed space dividing as efficiently as possible the two clusters of blue and red points. To do this, discriminant methods strictly partition the space of the measured variables into as many subregions as the number of

categories of objects in the training set and, therefore, always assign each data point to one and only one class (that within whose boundaries it falls). Conversely, CM approaches (also known as *one-class classifiers*) independently define a frontier for each individual category under study, enclosing a specific region of the variable (hyper-)space where specimens belonging to it are more likely to be found. As a consequence, if more than one class is modelled, samples can be recognised as members of none, one, or multiple modelled categories, which makes the application of this family of methodologies well-suited when the investigated categories are expected to constitute only a part of those that could be potentially encountered and explored.

Among the various existing CM techniques, the first ever appeared in literature and probably the most popular and widespread in chemometrics is Soft Independent Modelling of Class Analogy (SIMCA), originally developed by Svante Wold in 1976 [14, 15]. The words defining its acronym accurately summarise its main features and characteristics:

- *soft* indicates that the method is fully data-driven and no *a priori* assumption on the distribution of the collected data is made;
- *independent* means that every class of objects under study is treated individually and separately, contrarily to how a standard discrimination strategy would operate;
- *modelling of class analogy* implies that SIMCA focuses on the similarities among the samples belonging to the individually investigated category rather than on the differences that would allow distinguishing it from the others, once again in contrast to how discriminant techniques operate.

More specifically, regarding this last point, SIMCA assumes that the systematic information associated to these similarities can be captured by a Principal Component (PC) representation (of appropriate dimensionality) of the data collected for the individually modelled class and that the assessment of whether new observations belong to the modelled class can be carried out according to statistical measures/indices estimated based on such a reduced PC representation.

As recently discrimination seems to be very frequently overused and, unfortunately, misused, while the benefits of methods like SIMCA are quite often underestimated or overlooked [16, 17],

Journal Pre-proof

this tutorial has been conceived to shed better light on three fundamental aspects: why to employ SIMCA, when to employ SIMCA and how to/not to employ SIMCA. Concretely, the main idea behind it is to provide potential users with an extensive survey of the operating principles of SIMCA, the visualisation tools one can resort to for effectively reporting outcomes resulting from its utilisation, the circumstances in which coping with classification problems by means of SIMCA would be ideal in the light of the specific objectives the data analysis phase aims at and the pros and cons that SIMCA can exhibit over standard discrimination methods depending on the particular scenario one could face. For all these purposes and considering the motivations of this work, the paper features the following structure:

- Section 2 is devoted to the description of the SIMCA algorithm and of all its existing computational variants. Specific attention will be paid to the various ways of setting SIMCA-based classification rules, which is actually the principal distinctive attribute of all such variants. In addition, a short historical introduction of how this technique evolved and improved across the years is also given;

- Section 3 includes the illustration of two case-studies where all the aforementioned variants of SIMCA were applied so as to highlight possible similarities and/or diverging behaviours among them. In this regard, it has to be noticed that a novel Matlab (MathWorks, Inc., Natick, United States of America) toolbox, which encompasses routines and functions for all these different SIMCA versions and that automatically carries out performance comparisons among them, was made available to the interested readers and can be found at `https://github.com/RomeChemometrics/Simca`;

- Section 4 discusses the implications of using SIMCA as well as the advantages and disadvantages they may show over discriminant strategies in an attempt to ease the understanding of when and why SIMCA might be more suitable than discrimination;

- Section 5 holds final concluding remarks.

2. The principles of SIMCA modelling

Building and assessing a SIMCA classification model encompasses 4 main steps, which will be thoroughly described in the following subsections:

1. class-wise data decomposition by Principal Component Analysis (PCA [18, 19]);
2. decision or assignment rule definition;
3. SIMCA model parameter optimisation;
4. SIMCA model validation.

2.1. Class-wise PCA data decomposition

Imagine that a series of J-dimensional measurement vectors (for example, spectra or chromatograms) has been collected for a set of N samples belonging to a single class or category (*e.g.*, N blood specimens from healthy individuals) and piled into a matrix, say $\mathbf{X}$ (of size $N \times J$), sensibly preprocessed or pretreated (for instance, mean-centred or auto-scaled). The first computational step SIMCA carries out is the decomposition of $\mathbf{X}$ according to the well-known PCA bilinear approximation, which can be executed by means of algorithms like Singular Value Decomposition (SVD [20]) or Non-linear Iterative PArtial Least Squares (NIPALS [21, 22]) and presents the following model structure:

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

where $\mathbf{T}$ ($N \times A$), $\mathbf{P}$ ($J \times A$) and $\mathbf{E}$ ($N \times J$) are the scores, loadings and residuals arrays resulting from the factorisation of $\mathbf{X}$, while A identifies the number of computed PCs. These PCs (encoded in the columns of $\mathbf{P}$) define what is commonly known as a *class subspace*, typical of the individual category considered. As intuition would suggest, the higher the distance of an observation to this subspace, the higher the probability that the respective sample does not belong to the particular class under study. In SIMCA, two distance metrics are frequently exploited to evaluate whether a new object is member of the investigated category or not: the squared Euclidean distance from its corresponding measurement vector to its projection onto the aforementioned class subspace (also known as Orthogonal Distance, OD) and the squared Mahalanobis distance between this projection and the origin of the PC subspace (also known as Score Distance, SD). Denoting such

a measurement vector as $\mathbf{x}_{\text{new}}^{\text{T}}$ $(1 \times J)$, the OD and SD values associated to it can be estimated based on the following equations:

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

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

with $\mathbf{I}$ $(J \times J)$ being an identity matrix, $\mathbf{\Lambda}$ $(A \times A)$ being equal to $\frac{\mathbf{T}^{\text{T}}\mathbf{T}}{N-1}$ and $\| \|^2$ symbolising the 2-norm. The way this assessment is specifically conducted will be detailed in the next section. Finally, it is important to 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.

2.2. SIMCA decision/classification rule definition

Once the collected data have been decomposed as described before, a criterion needs to be established to decide whether new measurement vectors fit the class subspace or not, *i.e.*, to determine whether incoming samples are either accepted or rejected by the model of the investigated category and, thus, can be considered as its members or as outliers. For this purpose, the values of OD_{new} and SD_{new} (or of some arithmetic combinations of them) are generally compared with characteristic thresholds - corresponding to a user-specified confidence level, $(1 - \alpha)$ - estimated either theoretically (*i.e.*, by assuming specific statistical distributions for both OD and SD) or empirically/heuristically from $\mathbf{X}$. In other words, as already stressed in Section 1, SIMCA delimits a volume or *case* within the space of the J original variables where samples from the category at hand are more likely to be located. Subsequently, if $\mathbf{x}_{\text{new}}$ is found to fall inside this case, the new object is accepted as a member of the corresponding category (see Figure 2 for a schematic representation). The way the boundaries of such a case are marked out determines the SIMCA decision or classification rule and actually connotes the main distinctive feature of the 5 algorithmic variants of this methodology discussed here [23]: the original SIMCA formulation by Wold [14, 15], Simple SIMCA (Sim-SIMCA [24]), Alternative SIMCA (Alt-SIMCA [25]), Combined Index SIMCA (CI-SIMCA [26, 27]) and Data Driven SIMCA (DD-SIMCA [28, 29]).

Journal Pre-proof

Figure 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 (PC_1 and PC_2). 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 and rejected as an outlier. Notice that here the outlying observation exhibits abnormal values of both OD and SD.

2.2.1. Original SIMCA

For classification purposes, the original implementation of SIMCA [14] only focuses on the orthogonal distance of a new object from the class model subspace, slightly redefined with respect to Equation 3 and expressed as:

$$s_{\text{new}} = \sqrt{\frac{OD_{\text{new}}}{J - A}} \tag{4}$$

More specifically, it carries out a Fisher's F test with an appropriate number of degrees of freedom to compare such a value with the average distance from the same subspace of the training observations estimated as:

$$s = \frac{\sqrt{\sum_{n=1}^{N} \sum_{j=1}^{J} \frac{e_{nj}^2}{J-A} \frac{N}{N-A-1}}}{N} \tag{5}$$

where e_{nj} denotes the (n, j)-th element of $\mathbf{E}$. A new specimen is, therefore, accepted by the category under study if:

$$s_{\text{new}}^2 \leq s^2 F^{-1}\left(\alpha, J - A, (J - A)(N - A - 1)\right) \tag{6}$$

with $F^{-1}\left(\alpha, J - A, (J - A)(N - A - 1)\right)$ being the $(1 - \alpha)$ quantile of the Fisher's F distribution with $J - A$ and $(J - A)(N - A - 1)$ degrees of freedom. This first formulation of SIMCA was almost immediately amended to incorporate also a measure of distance within the class subspace [15] (see [30–32] for further details). Nonetheless, both these SIMCA formulations are seldom exploited nowadays although they can be found available in several software suites, not always

under the name "SIMCA"[i]. For this reason, they will be excluded from the performance assessment studies and the comparisons that will follow this section of the article.

2.2.2. Simple SIMCA (Sim-SIMCA)

In Sim-SIMCA [24], a new specimen is accepted by the modelled class if the following two conditions are simultaneously fulfilled:

$$OD_{\text{new}} \leq OD_{\text{crit}} \tag{7}$$

$$SD_{\text{new}} \leq SD_{\text{crit}} \tag{8}$$

OD_{crit} and SD_{crit} denote critical values for OD and SD, respectively. A comprehensive survey on how such critical values can be calculated is provided in Section 2.2.6.

2.2.3. Alternative SIMCA (Alt-SIMCA)

In Alt-SIMCA [25], OD_{new} and SD_{new} are combined in a single statistical index known as *reduced distance* and expressed as:

$$d_{\text{new}} = \sqrt{\left(\frac{OD_{\text{new}}}{OD_{\text{crit}}}\right)^2 + \left(\frac{SD_{\text{new}}}{SD_{\text{crit}}}\right)^2} \tag{9}$$

A new object is accepted by the category under study if and only if:

$$d_{\text{new}} \leq \sqrt{2} \tag{10}$$

The two terms summed under the square root in Equation 9, in fact, equal 1 when $OD_{\text{new}} = OD_{\text{crit}}$ and $SD_{\text{new}} = SD_{\text{crit}}$, that is to say when both statistics assume values identical to their corresponding decision thresholds. Nevertheless, it is worth noticing that this particular classification

[i]The original formulation of SIMCA (renamed to "PCA-Class") is implemented in the software SIMCA© developed and commercialised by Sartorius AG (Göttingen, Germany). It is also included in the packages Aspen Unscrambler™ (Aspen Technology, Inc., Bedford, United States of America) and PARVUS [33, 34]. It is worth stressing that, in SIMCA©, the default distance metric on which the class membership assessment is based is called *DModXPS+* and that, in Aspen Unscrambler™, the decision rule is slightly different from the one initially proposed by Wold.

rule may yield a certain flexibility when it comes to assessing the class membership of observations falling close to the class boundaries: such observations, indeed, might exhibit, *e.g.*, values of SD_{new} exceeding SD_{crit} and be anyway recognised as member of the investigated category in case $OD_{\text{new}} \ll OD_{\text{crit}}$.

2.2.4. Combined Index SIMCA (CI-SIMCA)

Compared to Alt-SIMCA, CI-SIMCA [26, 27] fuses differently OD_{new} and SD_{new}:

$$c_{\text{new}} = \frac{OD_{\text{new}}}{OD_{\text{crit}}} + \frac{SD_{\text{new}}}{SD_{\text{crit}}} \tag{11}$$

The rationale behind this *combined index* lies in the fact that OD and SD are already quadratic distance metrics, thus, one does not need to square them again before their summation [26, 27]. In this case, a new sample is accepted by the investigated class if:

$$c_{\text{new}} \leq c_{\text{crit}} \tag{12}$$

Here, c_{crit} represents the critical value of the c statistic estimated as:

$$c_{\text{crit}} = g\chi^{-2}(\alpha, h) \tag{13}$$

where $\chi^{-2}(\alpha, h)$ denotes the $(1-\alpha)$ quantile of the χ^2 distribution with h degrees of freedom. The parameters g and h are defined as:

$$g = \frac{\frac{A}{SD_{\text{crit}}^2} + \frac{\theta_2}{OD_{\text{crit}}^2}}{\frac{A}{SD_{\text{crit}}} + \frac{\theta_1}{OD_{\text{crit}}}} \tag{14}$$

$$h = \frac{\left(\frac{A}{SD_{\text{crit}}} + \frac{\theta_1}{OD_{\text{crit}}}\right)^2}{\frac{A}{SD_{\text{crit}}^2} + \frac{\theta_2}{OD_{\text{crit}}^2}} \tag{15}$$

with:

$$\theta_l = \sum_{a=A+1}^{\text{rank}(\mathbf{X})} \lambda_a^l \tag{16}$$

and λ_a being the a-th eigenvalue resulting from the PCA factorisation of $\mathbf{X}$.

2.2.5. Data Driven SIMCA (DD-SIMCA)

In DD-SIMCA [28, 29], a weighted sum of OD_{new} and SD_{new} (also known as *full distance*) is calculated as:

$$f_{\text{new}} = L_{OD}\frac{OD_{\text{new}}}{OD_0} + L_{SD}\frac{SD_{\text{new}}}{SD_0} \tag{17}$$

with $L_{OD} = \frac{2OD_0^2}{s_{OD}^2}$, $L_{SD} = \frac{2SD_0^2}{s_{SD}^2}$ and OD_0/SD_0 and s_{OD}^2/s_{SD}^2 being the mean and the variance of the OD/SD values computed for the objects belonging to the training set, $\mathbf{X}$ - alternatively, especially in the presence of outliers in $\mathbf{X}$, robust estimators of central tendency and variation can also be resorted to. A specimen is, therefore, assigned to the modelled class if:

$$f_{\text{new}} \leq f_{\text{crit}} \tag{18}$$

where f_{crit} denotes a f-statistic threshold estimated as:

$$f_{\text{crit}} = \chi^{-2}(\alpha, L_{OD} + L_{SD}) \tag{19}$$

2.2.6. OD and SD threshold value estimation

Since the very first version of SIMCA was developed, several have been the criteria proposed to estimate OD_{crit} and SD_{crit}. Here, a comprehensive description of the different options documented in literature and implemented in the Matlab code released together with this article is provided. For OD, such an estimation can be carried out:

- calculating the $(1-\alpha)$ percentile of the OD values observed for the training observations. This is a non-parametric approach particularly effective in cases where the sample size, N, is relatively large;
- based on Box's theory [35, 36] as:

$$OD_{\text{crit}} = g_{\text{B}}\chi^{-2}(\alpha, h_{\text{B}}) \tag{20}$$

where:

$$g_{\text{B}} = \frac{\theta_2}{\theta_1} \tag{21}$$

and

$$h_{\text{B}} = \frac{\theta_1^2}{\theta_2} \tag{22}$$

Journal Pre-proof

while θ_l has the same meaning as in Equation 16;

- according to the approximation suggested by Jackson and Mudholkar [37]:

$$OD_{\text{crit}} = \theta_1 \left[1 + \frac{c^{-1}(\alpha)\sqrt{2\theta_2 h_0^2}}{\theta_1} + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{\frac{1}{h_0}} \tag{23}$$

with:

$$h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2} \tag{24}$$

and $c^{-1}(\alpha)$ being the deviate corresponding to the upper $(1-\alpha)$ quantile of the normal distribution with zero-mean and unit-variance;

- as reported in [28, 29, 36]:

$$OD_{\text{crit}} = \frac{s_{OD}^2}{2OD_0} \chi^{-2}(\alpha, L_{OD}) \tag{25}$$

For SD, the estimation can instead be performed:

- computing non-parametrically the $(1-\alpha)$ percentile of the SD values observed for the training observations;

- based on the definition of the Fisher's F distribution [38] as:

$$SD_{\text{crit}} = \frac{A\left(N^2 - 1\right)}{N(N-A)} F^{-1}(\alpha, A, N-A) \tag{26}$$

- according to an approximation, originally proposed by Massart in the context of SIMCA [39, 40], assuming that N is high enough for the data-driven estimation of both sample mean and covariance to be exact:

$$SD_{\text{crit}} = \frac{A(N-1)}{(N-A)} F^{-1}(\alpha, A, N-A) \tag{27}$$

- calculating the $(1-\alpha)$ quantile of the χ^2 distribution with A degrees of freedom [35];

- as in [28, 29]:

$$SD_{\text{crit}} = \frac{s_{SD}^2}{2SD_0} \chi^{-2}(\alpha, L_{SD}) \tag{28}$$

Mind that, in DD-SIMCA, OD_{crit} and SD_{crit} are retrieved exclusively as in Equations 25 and 28.

2.3. SIMCA model parameter optimisation

One of the most critical aspects when tackling class modelling problems by means of SIMCA is the adjustment of the SIMCA class model itself, namely the optimisation of its complexity, A. Varying the number of PCs extracted from $\mathbf{X}$, indeed, can significantly affect the performance of this method. Generally speaking, two alternative approaches exist for carrying out such an adjustment operation, which are commonly defined as *rigorous* and *compliant*, respectively [41]. The former, in strict line with the philosophy that originally inspired SIMCA and many other related techniques, exploits only objects belonging to the modelled category so as to guarantee that the actual confidence level corresponds to the one imposed *a priori* by the operator, $(1 - \alpha)$. For this purpose, the observations of the training set are classified according to the specific decision rule chosen and A is determined as the highest dimensionality yielding the value of classification sensitivity closest to $1 - \alpha$. Sensitivity is also known as *true positive rate*, measures how well target class samples are recognised as such and is commonly expressed as:

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{29}$$

where TP and FN stand for True Positives (the amount of objects correctly identified as members of the category under study) and False Negatives (the amount of objects mistakenly identified as non-members of the category under study), respectively.

On the other hand, when a compliant strategy is adopted, A is set utilising both target and non-target category observations and trying to find the best compromise between classification sensitivity and specificity. Specificity, also called *true negative rate*, represents how many non-target class samples are rejected by the model constructed for the investigated category and is calculated as:

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{30}$$

with TN and FP standing for True Negatives (the amount of objects correctly identified as non-members of the category under study) and False Positives (the amount of objects mistakenly identified as members of the category under study), respectively. In other words, a compliant approach would usually aim at optimising the classification efficiency yielded by a SIMCA model, which is

equal to the geometric mean of classification sensitivity and specificity:

$$\text{efficiency} = \sqrt{\text{sensitivity} \times \text{specificity}} \tag{31}$$

It is worth noticing that, in order to avoid data overfitting, sensitivity, specificity and/or efficiency values should not be computed based on the OD and SD distance indices directly resulting from the full factorisation of $\mathbf{X}$ (training phase), but rather applying appropriate resampling methodologies like cross-validation [42, 43] or bootstrapping [44, 45]. Recent studies have highlighted that any of such methodologies is capable of providing accurate estimates of A, especially when N is large [17, 46].

Concerning compliant parameter tuning, it should be mentioned that a novel approach for the simultaneous adjustment of A and α has lately been proposed in an attempt to render SIMCA classification models more robust in the presence of abnormally high within-category variability and/or strong category overlaps [47]. In similar scenarios, in fact, confidence levels fixed *a priori* may not be adequate solutions to adopt. This approach relies on the principles of Receiver Operating Characteristic (ROC) curves [48, 49] that display the TP rate *vs* the (1 − TN) rate returned by a binary classifier as α varies. In a nutshell, the complexity of the class model is selected as the number of PCs maximising the area under the ROC curve (a statistic reflecting the general classification quality) obtained through a generic cross-validatory procedure, while α is taken so as to minimise the Euclidean distance to its top-left corner which, also here, begets the best compromise between classification sensitivity and specificity.

2.4. SIMCA model validation

Last but not least, as for any type of prediction or discrimination models, SIMCA ones need to undergo a proper validation step prior to their full implementation in a laboratory or in industry. To this end, *external* observations - *i.e.*, data items not taken into account during the training phase (so as to reduce the risk of drawing overoptimistic conclusions), but for which the class membership is known - are required in order to assess the validity, the reliability, the generalisability and the robustness of the SIMCA models constructed. Operationally, validation implies evaluating - according to the same figures of merit described in Section 2.3 - the performance of such models

on these external observations that, in principle, should reflect as close as possible the distribution of future incoming objects and be representative of all potential sources of expected variability (*e.g.*, manufacturing campaign, measurement experimental conditions, *etc.*). If, as it is common in most real-world scenarios, an additional dataset with similar characteristics (a so-called *test set*) cannot be collected due to particular limitations hampering further sampling procedures, all the available recordings could be split into two blocks to be utilised for calibration and validation tasks, respectively. Given, though, the specific features such blocks should exhibit (they should span the largest possible amount of overall data variation), random splitting schemes may result in suboptimal outcomes. Targeted selection approaches should instead be adopted: examples are the Kennard-Stone [50] and the Duplex [51, 52] algorithms or techniques based on the principles of D-optimal design of experiments [53, 54].

3. Two illustrative case-studies: the beer and the cell datasets

To give an example of its working principle and of the way the results returned by its application can be reported and visualised, SIMCA was here utilised to process two real-world datasets chosen so as to mimic a scenario of reduced sample size and a scenario of large sample size. The former consists of 20 near-infrared (NIR) spectral profiles of an Italian variety of craft beer, "Birra Reale", and 40 of other non-craft beers primarily registered for authentication purposes [55]. The latter contains the measures of 52 light intensity and shape descriptors (rectangularity, convexity, *etc.*) extracted from phase-contrast microscopy images of 3035 bacterial cells of 5 diverse morphological classes (deformed: 693 instances, long: 564 instances, normal: 893 instances, round: 498 instances, and small: 387 instances) and originally collected to monitor the possible onset of distinctive structural alterations of the cellular membrane under the effect of an antimicrobial drug [56]. More in detail, the SIMCA variants described in Sections 2.2.2, 2.2.3, 2.2.4 and 2.2.5 coupled to all the possible combinations of OD and SD threshold value estimation approaches (see Section 2.2.6) and both SIMCA model optimisation strategies (rigorous and compliant - see Section 2.3) were tested and compared for modelling all the categories of specimens under study in the two different cases. For a more comprehensive assessment of the performance of the tested

methodologies, a repeated double cross-validation procedure encompassing the following 4 computational steps was employed:

1. the available data were randomly partitioned into a training and a test set carrying 70% and 30% of the total amount of objects belonging to every individual category considered;
2. for each of these categories, the training set was further iteratively divided into two separate blocks of measurements according to a venetian blind cross-validation splitting scheme. The first was used to calibrate a SIMCA model of a certain dimensionality and to calculate the corresponding OD and SD critical limits; the second together with the entire ensemble of training observations of the other (non-target) classes were projected onto its subspace for an estimation of the classification sensitivity and specificity it could yield, respectively. This so-called *internal loop* was run for a number of principal components ranging from 1 to 10 and until all the target-category training specimens were left out from the first data block at least once;
3. SIMCA model complexity was tuned according to either the rigorous or the compliant criterion discussed in Section 2.3;
4. the class membership of the samples of the test set was predicted via the final optimised SIMCA model.

Such a procedure was replicated 300 times so as to retrieve empirical distributions for the sensitivity, specificity and efficiency values obtained in external validation.
Prior to conducting any modelling operation, the beer spectra were preliminarily corrected by means of the Standard Normal Variate (SNV [57]) transform and afterwards mean-centred, while, given their heterogeneous nature, the morphological descriptors of the cell dataset were autoscaled.

3.1. SIMCA model performance assessment

The performance of the trained SIMCA models was evaluated based on the figures of merit described in Section 2.3 estimated in external validation, *i.e.* when assessing the class membership of the objects belonging to the 300 test sets generated iteratively all along the progression of the

Figure 3. Beer dataset - "Birra Reale" class model - Overall classification sensitivity yielded in external validation by all the tested variants of SIMCA. Panel A) refers to Sim-SIMCA. Panel B) refers to Alt-SIMCA. Panel C) refers to CI-SIMCA. Panel D) refers to DD-SIMCA. The represented confidence intervals are delimited by the 2.5-th and 97.5-th percentile of the empirical distribution of the corresponding figures of merit retrieved through the repeated double-cross validation strategy described in Section 3. Bar colour-coding - for A), B) and C) - and x-axis labels relate to the approaches used for estimating the threshold values for OD and SD, respectively.

Figure 4. Beer dataset - "Birra Reale" class model - Overall classification specificity yielded in external validation by all the tested variants of SIMCA. Panel A) refers to Sim-SIMCA. Panel B) refers to Alt-SIMCA. Panel C) refers to CI-SIMCA. Panel D) refers to DD-SIMCA. The represented confidence intervals are delimited by the 2.5-th and 97.5-th percentile of the empirical distribution of the corresponding figures of merit retrieved through the repeated double-cross validation strategy described in Section 3. Bar colour-coding - for A), B) and C) - and x-axis labels relate to the approaches used for estimating the threshold values for OD and SD, respectively.

Figure 5. Beer dataset - "Birra Reale" class model - Overall classification efficiency yielded in external validation by all the tested variants of SIMCA. Panel A) refers to Sim-SIMCA. Panel B) refers to Alt-SIMCA. Panel C) refers to CI-SIMCA. Panel D) refers to DD-SIMCA. The represented confidence intervals are delimited by the 2.5-th and 97.5-th percentile of the empirical distribution of the corresponding figures of merit retrieved through the repeated double-cross validation strategy described in Section 3. Bar colour-coding - for A), B) and C) - and x-axis labels relate to the approaches used for estimating the threshold values for OD and SD, respectively.

Figure 6. Cell dataset - Normal cell class model - Overall classification sensitivity yielded in external validation by all the tested variants of SIMCA. Panel A) refers to Sim-SIMCA. Panel B) refers to Alt-SIMCA. Panel C) refers to CI-SIMCA. Panel D) refers to DD-SIMCA. The represented confidence intervals are delimited by the 2.5-th and 97.5-th percentile of the empirical distribution of the corresponding figures of merit retrieved through the repeated double-cross validation strategy described in Section 3. Bar colour-coding - for A), B) and C) - and x-axis labels relate to the approaches used for estimating the threshold values for OD and SD, respectively.

algorithmic procedure. Figures 3-8 summarise the outcomes for only one of the categories of samples underlying each dataset at hand (for the sake of brevity and simplicity): the "Birra Reale" and the normal cell category, respectively. Clearly, the differences among approaches become more pronounced if a lower sample size is handled. Overall, it can be said that in this particular case compliant parameter optimisation approaches permitted to achieve higher efficiency values by targeting a more balanced compromise between classification sensitivity and specificity. This was mainly made possible, though, by the fact that the non-target classes taken into account during the SIMCA model adjustment phase were exactly the same as those from which the non-target class observations of the test set were drawn. If such a condition does not hold and a poor matching exists between these two groups of non-target categories, a compliant technique may lead to an undesired bias and result to be suboptimal [17, 58]. In addition, within rigorous and compliant tuning methodologies, small performance variations could be observed except for Sim-SIMCA and Alt-SIMCA which led to the largest discrepancy between sensitivity and specificity when the critical threshold for OD was estimated based on Box's theory or through the approximation suggested by Jackson and Mudholkar. On the other hand, discrepancies are less significant when a larger amount of observations is coped with. This is somehow expected if one thinks that, in such a contingency, i) the similarity among the statistical distributions which both OD and SD are assumed to follow may increase (see, for example, [59]) and ii) parametric and non-parametric estimates of OD_{crit} and SD_{crit} might converge.

3.2. Result representation

The most immediate and direct way to display the outcomes resulting from the application of a SIMCA classification model is to depict the estimates of OD and SD or their combination (depending on the adopted decision rule) together with their respective critical thresholds corresponding to the confidence level set in the specific case-study at hand. An example of such a representation is provided in Figure 9. Notice that for Sim-SIMCA, since OD and SD are not mathematically fused into an individual statistical index as in all the other SIMCA variants described in this article, a chart of the maximum values between $\frac{OD}{OD_{\text{crit}}}$ and $\frac{SD}{SD_{\text{crit}}}$ - $\max\left\{\frac{OD}{OD_{\text{crit}}}, \frac{SD}{SD_{\text{crit}}}\right\}$ - can alternatively be utilised for assessing whether a sample can be considered a member or not of

Figure 7. Cell dataset - Normal cell class model - Overall classification specificity yielded in external validation by all the tested variants of SIMCA. Panel A) refers to Sim-SIMCA. Panel B) refers to Alt-SIMCA. Panel C) refers to CI-SIMCA. Panel D) refers to DD-SIMCA. The represented confidence intervals are delimited by the 2.5-th and 97.5-th percentile of the empirical distribution of the corresponding figures of merit retrieved through the repeated double-cross validation strategy described in Section 3. Bar colour-coding - for A), B) and C) - and x-axis labels relate to the approaches used for estimating the threshold values for OD and SD, respectively.

Figure 8. Cell dataset - Normal cell class model - Overall classification efficiency yielded in external validation by all the tested variants of SIMCA. Panel A) refers to Sim-SIMCA. Panel B) refers to Alt-SIMCA. Panel C) refers to CI-SIMCA. Panel D) refers to DD-SIMCA. The represented confidence intervals are delimited by the 2.5-th and 97.5-th percentile of the empirical distribution of the corresponding figures of merit retrieved through the repeated double-cross validation strategy described in Section 3. Bar colour-coding - for A), B) and C) - and x-axis labels relate to the approaches used for estimating the threshold values for OD and SD, respectively.

Figure 9. Cell dataset - Normal cell class model - A) OD, B) SD and C) $\max\left\{\frac{OD}{OD_{\text{crit}}}, \frac{SD}{SD_{\text{crit}}}\right\}$ charts returned by Sim-SIMCA for the external test set samples. D) d, E) c and F) f charts yielded by Alt-SIMCA, CI-SIMCA and DD-SIMCA, respectively, for the external test set samples. The dashed lines denote the decision thresholds corresponding to each one of these distance metrics. In Sim-SIMCA, an object is rejected as an outlier if for it either OD or SD is found to be larger than its associated decision threshold estimated as detailed in Section 2.2.6. Alternatively, an object is rejected as an outlier if $\max\left\{\frac{OD}{OD_{\text{crit}}}, \frac{SD}{SD_{\text{crit}}}\right\}$ is found to be larger than 1. In Alt-SIMCA, CI-SIMCA and DD-SIMCA, an object is rejected as an outlier if for it d, c or f is found to be larger than its associated decision threshold estimated as detailed in Sections 2.2.3, 2.2.4 and 2.2.5. Results are displayed only for the models leading to the highest overall classification efficiency. Axes were rescaled for an enhanced visualisation.

Figure 9. Continuation.

the category of interest: only those objects exhibiting $\max\left\{\frac{OD}{OD_{\text{crit}}}, \frac{SD}{SD_{\text{crit}}}\right\}$ larger than 1 are rejected as outliers.

When multiple target classes are simultaneously investigated, in order to determine the degree of potential confusion between any pair of them, bivariate extensions of Figures 9C, 9D and 9E the so-called Coomans plots [60], can be constructed. For a given set of specimens, a Coomans plot (see, for instance, Figure 10) shows the values of the joint distances d (for Alt-SIMCA), c (for CI-SIMCA) or f (for DD-SIMCA) from two distinct class model subspaces simultaneously. According to the illustration in Figure 10, it is clear that:

- a sample is deemed belonging only to the first category under study if the symbol associated to it falls in the bottom-right area of the graph;
- a sample is deemed belonging only to the second category under study if the symbol associated to it falls in the top-left area of the graph;
- a sample is deemed belonging to none of the two categories under study if the symbol associated to it falls in the top-right area of the graph;
- a sample is deemed belonging to both the categories under study if the symbol associated to it falls in the bottom-left area of the graph;

It goes without saying that for the same reason highlighted before, a Coomans plot for Sim-SIMCA needs to be built using the maxima between $\frac{OD}{OD_{\text{crit}}}$ and $\frac{SD}{SD_{\text{crit}}}$.

An additional strategy (rather frequently exploited in literature) for the visualisation of the results yielded by a one-class SIMCA modelling approach implies displaying, either in linear or logarithmic scale, the values of the ratios $\frac{OD}{OD_{\text{crit}}}$ and $\frac{SD}{SD_{\text{crit}}}$ (or $\frac{OD}{OD_0}$ and $\frac{SD}{SD_0}$ when DD-SIMCA is concerned) for a certain group of observations (see, *e.g.*, Figure 11). Based on a similar graph, all the symbols found to be located within the acceptance subregion (bottom-left) relate to objects recognised as member of the explored category and *vice versa*. This acceptance subregion is delimited by a frontier that is estimated differently depending on the SIMCA variant resorted to:

- in Sim-SIMCA, the acceptance subregion is a square with unit-length sides;

Figure 10. Cell dataset - Normal cell class model *vs* small cell class model - Coomans plots (in logarithmic scale) yielded by A) Sim-SIMCA, B) Alt-SIMCA, C) CI-SIMCA and D) DD-SIMCA for the external test set samples. The dashed lines denote the decision thresholds corresponding to the max $\left\{\frac{OD}{OD_{\text{crit}}}, \frac{SD}{SD_{\text{crit}}}\right\}$, d, c and f statistics, respectively. Notice that every graph is partitioned into four distinct subregions and that i) a specimen is deemed belonging only to the normal cell class if the symbol associated to it falls in the bottom-right one, ii) a specimen is deemed belonging only to the small cell class if the symbol associated to it falls in the top-left one; iii) a specimen is deemed belonging to none of the two categories under study if the symbol associated to it falls in the top-right one, and iv) a specimen is deemed belonging to both the categories under study if the symbol associated to it falls in the bottom-left one. Results are displayed for models adjusted as those Figure 9 refers to. Axes were rescaled for an enhanced visualisation.

Figure 10. Continuation.

Figure 11. Cell dataset - Normal cell class model - $\frac{OD}{OD_{\text{crit}}}$ *vs* $\frac{SD}{SD_{\text{crit}}}$ plot yielded by A) Sim-SIMCA, B) Alt-SIMCA and C) CI-SIMCA for the external test set samples. D) $\frac{OD}{OD_0}$ *vs* $\frac{SD}{SD_0}$ plot returned by DD-SIMCA for the external test set samples. Notice that a sample is rejected as an outlier if the symbol associated to it falls outside the acceptance area delimited by the dashed lines (whose geometry and extension depend on the specific classification rule adopted - see Section 3.2 for further details). Results are displayed only for the models Figure 9 refers to. Axes were rescaled for an enhanced visualisation.

Figure 11. Continuation.

- in Alt-SIMCA, the frontier of the acceptance subregion is the circular arc satisfying the relation:

$$\sqrt{\left(\frac{OD}{OD_{\text{crit}}}\right)^2 + \left(\frac{SD}{SD_{\text{crit}}}\right)^2} = \sqrt{2} \tag{32}$$

  which defines a curved line;

- in CI-SIMCA, the acceptance subregion frontier is defined as the geometrical locus of all points satisfying the relation:

$$\frac{OD}{OD_{\text{crit}}} + \frac{SD}{SD_{\text{crit}}} = c_{\text{crit}} = g\chi^{-2}(\alpha, h) \tag{33}$$

  which defines a straight line;

- in DD-SIMCA, the acceptance subregion frontier is defined as the geometrical locus of all points satisfying the relation:

$$L_{OD}\frac{OD}{OD_0} + L_{SD}\frac{SD}{SD_0} = f_{\text{crit}} = \chi^{-2}(\alpha, L_{OD} + L_{SD}) \tag{34}$$

  which also defines a straight line.

As a concluding remark, it is worth stressing here that all the graphical tools employed in this section enable the assessment of what is also known as the *local* specificity of a SIMCA class model, *i.e.* its specificity with respect to particular individual non-target categories.

4. Discussion

This article was conceived in an attempt to provide practitioners with pragmatic guidelines for a sensible and correct utilisation of SIMCA as well as with answers to four basic questions: why performing CM by means of SIMCA? In which circumstances is SIMCA a suitable option for tackling classification tasks? How applying SIMCA in a classification scenario? How not applying SIMCA in a classification scenario?

Journal Pre-proof

4.1. Why applying SIMCA?

As already stressed in Section 1, SIMCA and, more generally, CM approaches are grounded on an operating principle that renders them unique with respect to standard and better-established discrimination techniques. CM relies, in fact, on the training of individual models for each class of samples under study (with the possibility of modelling even a single class, if needed), which enables the definition of boundaries or frontiers for these categories whose nature technically only depends on the features of the objects strictly belonging to them. As a result, new incoming specimens can be recognised not only as members of one class or another, but also as confused, *i.e*., likely to belong to two or more classes, or as not coming from any of the modelled categories. In other words, rather than setting a classification problem in a discriminant way as "*is a specimen coming from class A, class B or class C?*", CM translates it into "*is a specimen coming from class A or not/class B or not/class C or not?*" [61, 62]. And this yields several benefits in many application scenarios that will be briefly discussed in the next subsection. Additionally and more concretely, SIMCA guarantees a further advantage over alternative CM strategies originally developed in the machine learning domain (for example, One-Class Support Vector Machines - OC-SVM [63, 64]): being a *white box* latent variable-based methodology, it theoretically ensures full interpretability of the systematic patterns of data variability characteristic of the modelled categories.

4.2. When applying SIMCA?

SIMCA constitutes an ideal solution for addressing and solving classification problems when the interest is focused only on one or few categories of objects. Food [65, 66] and pharmaceutical authentication [67, 68] as well as industrial process monitoring [69] are just some of the potential scenarios where attention is paid solely to single classes of specimens, *i.e*. added-value or high-quality products and *in-control* time periods, respectively. Furthermore, readers can easily envision that, in similar circumstances, it is also particularly complicated to plan sampling campaigns for the collection of measurement observations representative of all the possible categories one could actually observe in reality (*e.g*., all possible adulterated versions of a drug, all possible manufacturing faults, *etc*.). And in such situations, as also underlined in Section 1, the application of a discriminant approach might lead to an unavoidable bias since each object under study will

always be assigned to one of the classes taken into account during the model training phase even if they represent only a reduced ensemble of those from which such an object can ideally originate [70]. This is the main reason why Rodionova *et al*. have recently stated that *discriminant analysis is an inappropriate method of authentication* [16] and can be replaced by a rational utilisation of SIMCA.

4.3. How applying SIMCA?

In this article, 4 variants of the original SIMCA algorithm proposed by Wold in 1976 were thoroughly described. All these variants share the same modelling principle (a class-wise PCA decomposition of the data at hand), but are based on distinct decision rules when it comes to assigning an object to the investigated category/categories or not. In Section 3.1, it was basically shown that such decision rules yield virtually indistinguishable outcomes when the sample size of the training set for the modelled category/categories is relatively high (as when the cell dataset was analysed) and that an appropriate tuning of the SIMCA model dimensionality together with a sensible choice of the OD and SD reference distributions can significantly enhance the classification performance while reducing the differences they may exhibit when such a sample size decreases: it can be said, for example, that when the beer spectra were processed, compliant adjustment approaches resulted in relatively higher percentages of classification efficiency especially when OD_{crit} was estimated according to Equations 20 and 23 and when Alt-SIMCA and CI-SIMCA were concerned. Such an improvement, instead, was found to be less pronounced for Sim-SIMCA. Anyway, regardless of the specific conclusions that can be drawn in these two particular case-studies, interested readers are provided with a fully functional Matlab code capable of running all the aforementioned variants of SIMCA and by which possibly carrying out comprehensive comparative studies in diverse case-studies typical of their own domain of expertise.

4.4. How not applying SIMCA?

As should already be clear to the reader at this point of the article, typical CM problems call for the collection and analysis of samples of one or more well-known and well-defined categories of interest and (possibly) samples belonging to a plethora of ill-represented classes that do not

Journal Pre-proof

necessarily constitute all the others from which specimens can originate. This strictly implies that often such sets of available non-target class objects only span a reduced portion of the variability which all those to be potentially assessed in the future may actually exhibit. It has also been highlighted that, in similar circumstances, discriminant approaches will not be capable of returning optimal classification outcomes. Furthermore, even if operational attempts are made in order to enhance sample representativeness and cover a larger amount of sources of sample variability and/or a higher number of sample categories, discrimination models might easily become overparametrised, *i.e.*, too complex and less robust when it comes to characterising new incoming observations.

Nonetheless, although resorting to discriminant techniques in CM scenarios seems to be one of the most frequent mistakes users commit, another practice that is common among chemometricians and analytical chemists and that in certain cases can be procedurally questionable relates to the utilisation of CM strategies like SIMCA for tackling discrimination tasks: indeed, notwithstanding that this practice may offer a complementary perspective on the investigated issues, by forcing, for example, SIMCA to assign each single specimen under study to only one or at least one of the considered classes (according, *e.g.*, to a minimum reduced distance or combined index criterion) one might risk to literally denature its essence and significantly jeopardise its flexibility. Conversely, a widespread misunderstanding to be absolutely avoided in situations where SIMCA and, thus, CM are concerned is gathering the ill-represented classes mentioned before into an individual one and, afterwards, model it as a whole. In such a way, in fact, a severe bias could be induced in the obtained outcomes which could conceivably lead to draw skewed and distorted conclusions. It should now also be evident how many times, rather than blindly choosing the classification method yielding the best predictive performance, it is more critical and important to select the most pertinent one (and apply it in the most pertinent fashion) for answering the specific scientific questions at hand.

5. Conclusions and perspectives

SIMCA is a statistical classification approach into which nowadays new life seems to have been breathed. Its inherent capability of coping with multivariate datasets together with its robust-

ness against the potential unbalancedness affecting the size of the different categories of samples investigated - deriving from the fact that each one of them can be treated separately and independently - has lately attracted much attention from scientists of disparate fields of interest. This tutorial was basically developed for this particular reason and with in mind the specific aim of easing the access of non-expert users to such a tool. Yet, the job of chemometricians in this regard is far from being definitely finished. Novel research lines, indeed, can be easily envisioned in the context of SIMCA: just to mention a few, extending the SIMCA algorithm for the analysis of non-linear data structures (based, *e.g.*, on the principle of kernel transformations [71]) may constitute an intriguing and challenging subject of study, while advanced tools for the visualisation of the importance or relevance of the recorded variables in SIMCA models still need to be designed and implemented [72]. Concerning this aspect, it must be noticed that, in [15], Wold and Sjöström had already defined measures for the *discriminant* and *modelling* power of a variable, but both these measures have to be somehow adapted for being possibly exploited in the framework of the most recent SIMCA variants. Alternatively, graphical representations such as the well-established contributions plots [73] could be resorted to, but, to the best of the authors' knowledge, they are not readily suitable for dealing with joint distance indices like d, c or f.

It is worth noticing, in addition, that some of the different SIMCA versions described in Section 2 (namely, the original one, Alt-SIMCA and DD-SIMCA) have been already adapted to handle multi-way arrays [31, 32] through a modification of their computational procedure that replaces the PCA decomposition step with a Parallel Factor Analysis (PARAFAC [74–76]) or Tucker3 [77–79] factorisation. Such multi-way SIMCA approaches will be soon made available in the Matlab toolbox provided together with this tutorial, which, hopefully, will serve as an inspiration also for further advances in this sense.

References

[1] M. Casale, P. Oliveri, C. Armanino, S. Lanteri, M. Forina, NIR and UV-vis spectroscopy, artificial nose and tongue: comparison of four fingerprinting techniques for the characterisation of Italian red wines, Anal. Chim. Acta 668 (2010) 143–148.

[2] M. Bevilacqua, R. Bucci, A. Magrì, A. Magrì, F. Marini, Tracing the origin of extra virgin olive oils by infrared spectroscopy and chemometrics: a case study, Anal. Chim. Acta 717 (2012) 39–51.

[3] R. Vitale, M. Bevilacqua, R. Bucci, A. Magrì, A. Magrì, F. Marini, A rapid and non-invasive method for authenticating the origin of pistachio samples by NIR spectroscopy and chemometrics, Chemometr. Intell. Lab. 121 (2013) 90–99.

[4] 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.

[5] F. Di Donato, A. Biancolillo, A. Ferretti, A. D'Archivio, F. Marini, Near infrared spectroscopy coupled to chemometrics for the authentication of donkey milk, J. Food Compos. Anal. In Press (2022) 105017.

[6] E. Deconinck, P. Sacré, P. Courselle, J. De Beer, Chemometrics and chromatographic fingerprints to discriminate and classify counterfeit medicines containing PDE-5 inhibitors, Talanta 100 (2012) 123–133.

[7] E. Deconinck, P. Sacré, D. Coomans, J. De Beer, Classification trees based on infrared spectroscopic data to discriminate between genuine and counterfeit medicines, J. Pharmaceut. Biomed. 57 (2012) 68–75.

[8] O. Rodionova, K. Balyklova, A. Titova, A. Pomerantsev, Quantitative risk assessment in classification of drugs with identical API content, J. Pharmaceut. Biomed. 98 (186-192) (2014).

[9] Y. Zontov, K. Balyklova, A. Titova, O. Rodionova, A. Pomerantsev, Chemometric aided NIR portable instrument for rapid assessment of medicine quality, J. Pharmaceut. Biomed. 131 (2016) 87–93.

[10] S. Steffen, M. Otto, L. Niewoehner, M. Barth, Z. Brożek-Mucha, J. Biegstraaten, R. Horváth, Chemometric classification of gunshot residues based on energy dispersive X-ray microanalysis and inductively coupled plasma analysis with mass-spectrometric detection, Spectrochim. Acta B 62 (2007) 1028–1036.

[11] 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.

[12] 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.

[13] F. Marini, Classification methods in chemometrics, Curr. Anal. Chem. 6 (2010) 72–79.

[14] S. Wold, Pattern recognition by means of disjoint principal components models, Pattern Recogn. 8 (1976) 127–139.

[15] S. Wold, M. Sjöström, SIMCA: a method for analyzing chemical data in terms of similarity and analogy, in: B. Kowalski (Ed.), Chemometrics: Theory and Application, Vol. 52, American Chemical Society, Washington D.C., USA, 1977, pp. 243–282.

[16] O. Rodionova, A. Titova, A. Pomerantsev, Discriminant analysis is an inappropriate method of authentication, Trend. Anal. Chem. 78 (2016) 17–22.

Journal Pre-proof

[17] Z. Małyjurek, R. Vitale, B. Walczak, Different strategies for class model optimization. a comparative study, Talanta 215 (2020) 120912.

[18] K. Pearson, On lines and planes of closest fit to systems of points in space, Philos. Mag. 2 (1901) 559–572.

[19] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933) 417–441.

[20] C. Eckart, G. Young, The approximation of one matrix by another of lower rank, Psychometrika 1 (1936) 211–218.

[21] H. Wold, Estimation of principal components and related models by iterative least squares, in: P. Krishnajah (Ed.), Multivariate Analysis, Academic Press, Inc., New York, United States of America, 1966, pp. 391–420.

[22] H. Wold, Path models with latent variables: the NIPALS approach, in: H. Blalock, A. Aganbegian, F. Borodkin, R. Boudon, V. Capecchi (Eds.), Quantitative Sociology: International Perspectives on Mathematical and Statistical Modeling, Academic Press, Inc., New York, United States of America, 1975, pp. 307–357.

[23] A. Pomerantsev, O. Rodionova, Popular decision rules in SIMCA: critical review, J. Chemometr. 34 (2020) e3250.

[24] 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.

[25] Eigenvector Research, Inc. SIMCA Model Builder GUI, `https://www.wiki.eigenvector.com/index.php?title=SIMCA_Model_Builder_GUI` [online, cited November 2nd 2022].

[26] H. Yue, S. Qin, Reconstruction-based fault identification using a combined index, Ind. Eng. Chem. Res. 40 (2001) 4403–4414.

[27] S. Qin, Statistical process monitoring: basics and beyond, J. Chemometr. 17 (2003) 480–502.

[28] A. Pomerantsev, Acceptance areas for multivariate classification derived by projection methods, J. Chemometr. 22 (2008) 601–609.

[29] A. Pomerantsev, O. Rodionova, Concept and role of extreme objects in PCA/SIMCA, J. Chemometr. 28 (2013) 429–438.

[30] S. De Luca, R. Bucci, A. Magrì, F. Marini, Class modeling techniques in chemometrics: theory and applications, in: R. Meyers (Ed.), Encyclopedia of Analytical Chemistry, John Wiley & Sons, Ltd., Chichester, United Kingdom, 2018, pp. 1–24.

[31] C. Durante, R. Bro, M. Cocchi, A classification tool for *N*-way array based on SIMCA methodology, Chemometr. Intell. Lab. 106 (2011) 73–85.

[32] M. Cocchi, M. Li Vigni, C. Durante, Multi-way classification, in: S. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, Elsevier, B.V., Amsterdam, The Netherlands, 2020, pp. 701–721.

[33] M. Forina, PARVUS, Trend. Anal. Chem. 3 (1984) 38–39.

[34] B. Vandeginste, PARVUS: An extendable package of programs for data exploration, classification and corre-

Journal Pre-proof

lation, M. Forina, R. Leardi, C. Armanino and S. Lanteri, Elsevier, Amsterdam, 1988, Price: US $645 ISBN 0-444-43012-1, J. Chemometr. 4 (1990) 191–193.

[35] G. Box, Some theorems on quadratic forms applied in the study of analysis of variance problems: effect of inequality of variance in one-way classification, Ann. Math. Stat. 25 (1954) 290–302.

[36] P. Nomikos, J. MacGregor, Multivariate SPC charts for monitoring batch processes, Technometrics 37 (1995) 41–59.

[37] J. Jackson, G. Mudholkar, Control procedures for residuals associated to Principal Component Analysis, Technometrics 21 (1979) 341–349.

[38] N. Tracy, J. Young, R. Mason, Multivariate control charts for individual observations, J. Qual. Technol. 24 (1992) 88–95.

[39] J. Jackson, A User's Guide to Principal Components, John Wiley & Sons, Inc., Hoboken, United States of America, 1991.

[40] R. De Maesschalck, A. Candolfi, D. Massart, S. Heuerding, Decision criteria for soft independent modelling of class analogy applied to near infrared data, Chemometr. Intell. Lab. 47 (1999) 65–77.

[41] O. Rodionova, P. Oliveri, A. Pomerantsev, Rigorous and compliant approaches to one-class classification, Chemometr. Intell. Lab. 159 (2016) 89–96.

[42] M. Stone, Cross-validatory choice and assessment of statistical predictions, J. R. Stat. Soc. B Met. 36 (1974) 111–133.

[43] D. Allen, The relationship between variable selection and data augmentation and a method for prediction, Technometrics 16 (1974) 125–127.

[44] B. Efron, Bootstrap methods: another look at the jackknife, Ann. Stat. 7 (1979) 1–26.

[45] R. Wehrens, H. Putter, L. Buydens, The bootstrap: a tutorial, Chemometr. Intell. Lab. 54 (2000) 35–52.

[46] V. Carboni, Metodi di modellamento di classe: recenti sviluppi, Master's thesis, Corso di Laurea Magistrale in Scienze Chimiche, Dipartimento di Scienze Chimiche e Geologiche, Università degli Studi di Modena e Reggio Emilia (2019/2020).

[47] R. Vitale, F. Marini, C. Ruckebusch, SIMCA modeling for overlapping classes: fixed or optimized decision threshold?, Anal. Chem. 90 (2018) 10738–10747.

[48] J. Swets, Measuring the accuracy of diagnostic systems, Science 240 (1988) 1285–1293.

[49] C. Brown, H. Davis, Receiver operating characteristics curves and related decision measures: a tutorial, Chemometr. Intell. Lab. 80 (2006) 24–38.

[50] R. Kennard, L. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148.

[51] R. Snee, Validation of regression models: methods and examples, Technometrics 19 (1977) 415–428.

[52] M. Daszykowski, B. Walczak, D. Massart, Representative subset selection, Anal. Chim. Acta 468 (2002) 91–103.

Journal Pre-proof

[53] W. Wu, B. Walczak, D. Massart, S. Heuerding, F. Erni, I. Last, K. Prebble, Artificial neural networks in classification of NIR spectral data: design of the training set, Chemometr. Intell. Lab. 33 (1996) 35–46.

[54] P. Goos, B. Jones, Optimal Design of Experiments: A Case Study Approach, John Wiley & Sons, Ltd., Chichester, United Kingdom, 2011.

[55] A. Biancolillo, R. Bucci, A. Magrì, A. Magrì, F. Marini, Data-fusion for multiplatform characterization of an Italian craft beer aimed at its authentication, Anal. Chim. Acta 820 (2014) 23–31.

[56] T. Zahir, R. Camacho, R. Vitale, C. Ruckebusch, J. Hofkens, M. Fauvart, J. Michiels, High-throughput time-resolved morphology screening in bacteria reveals phenotypic responses to antibiotics, Commun. Biol. 2 (2019) 269.

[57] R. Barnes, M. Dhanoa, S. Lister, Standard Normal Variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989) 772–777.

[58] A. Pomerantsev, O. Rodionova, New trends in qualitative analysis: performance, optimization, and validation in multi-class and soft models, Trend. Anal. Chem. 143 (2021) 116372.

[59] R. Brereton, The *F* distribution and its relationship to the chi squared and *t* distributions, J. Chemometr. 29 (2015) 582–586.

[60] D. Coomans, I. Broeckaert, M. Derde, A. Tassin, D. Massart, S. Wold, Use of a microcomputer for the definition of multivariate confidence regions in medical diagnosis based on clinical laboratory profiles, Comput. Biomed. Res. 17 (1984) 1–14.

[61] M. Forina, P. Oliveri, S. Lanteri, M. Casale, Class-modeling techniques, classic and new, for old and new problems, Chemometr. Intell. Lab. 93 (2008) 132–148.

[62] P. Oliveri, Class-modelling in food analytical chemistry: development, sampling, optimisation and validation issues - A tutorial, Anal. Chim. Acta 982 (2017) 9–19.

[63] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, J. Platt, Support vector method for novelty detection, in: S. Solla, T. Leen, K. Müller (Eds.), Advances in Neural Information Processing Systems, Vol. 12, MIT Press, Cambridge, United States of America, 1999, pp. 582–588.

[64] D. Tax, R. Duin, Support vector data description, Mach. Learn. 54 (2004) 45–66.

[65] P. Oliveri, G. Downey, Multivariate class modeling for the verification of food-authenticity claims, Trend. Anal. Chem. 35 (2012) 74–86.

[66] A. Biancolillo, F. Marini, C. Ruckebusch, R. Vitale, Chemometric strategies for spectroscopy-based food authentication, Appl. Sci. Basel 10 (2020) 6544.

[67] B. Krakowska, D. Custers, E. Deconinck, M. Daszykowski, Chemometrics and the identification of counterfeit medicines - A review, J. Pharmaceut. Biomed. 127 (2016) 112–122.

[68] D. Custers, P. Courselle, S. Apers, E. Deconinck, Chemometrical analysis of fingerprints for the detection of counterfeit and falsified medicines, Rev. Anal. Chem. 35 (2016) 145–168.

[69] A. Ferrer, Multivariate Statistical Process Control based on Principal Component Analysis (MSPC-PCA): some reflections and a case study in an autobody assembly process, Qual. Eng. 19 (2007) 311–325.

[70] M. Cocchi, Chemometrics for food quality control and authentication, in: R. Meyers (Ed.), Encyclopedia of Analytical Chemistry, John Wiley & Sons, Ltd., Chichester, United Kingdom, 2017, pp. 1–29.

[71] B. Schölkopf, A. Smola, Learning with Kernels, MIT Press, Cambridge, United States of America, 2002.

[72] 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, Dipartimento di Chimica, Università degli Studi di Roma "La Sapienza" (2020/2021).

[73] T. Kourti, J. MacGregor, Multivariate SPC methods for process and product monitoring, J. Qual. Technol. 28 (1996) 409–428.

[74] R. Harshman, Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multimodal factor analysis, in: UCLA Working Papers in Phonetics, Vol. 16, 1970, pp. 1–84.

[75] J. Carroll, J. Chang, Analysis of individual differences in multidimensional scaling via an *N*-way generalization of "Eckart-Young" decomposition, Psychometrika 35 (1970) 283–319.

[76] R. Bro, PARAFAC. Tutorial and applications, Chemometr. Intell. Lab. 38 (1997) 149–171.

[77] L. Tucker, Some mathematical notes on three-mode factor analysis, Psychometrika 31 (1966) 279–311.

[78] R. Henrion, N-way principal component analysis. Theory, algorithms and applications, Chemometr. Intell. Lab. 25 (1994) 1–23.

[79] A. Smilde, R. Bro, P. Geladi, Multi-way Analysis: Applications in the Chemical Sciences, John Wiley & Sons, Ltd., Chichester, United Kingdom, 2004.

[80] Sartorius AG. SIMCA© webpage, `https://www.sartorius.com/en/products/process-analytical-technology/data-analytics-software/mvda-software/simca` [online, cited April 20th 2023].

[81] Aspen Technology, Inc. Aspen Unscrambler™ webpage, `https://www.aspentech.com/en/products/apm/aspen-unscrambler` [online, cited April 20th 2023].

[82] A. Olivieri, Introduction to Multivariate Calibration - A Practical Approach, Springer Nature Switzerland AG, Cham, Switzerland, 2018.

[83] Eigenvector Research, Inc. PLS_Toolbox webpage, `https://eigenvector.com/software/pls-toolbox` [online, cited April 20th 2023].

[84] Y. Zontov. DD-SIMCA Tool repository, `https://github.com/yzontov/dd-simca` [online, cited April 20th 2023].

[85] Y. Zontov, O. Rodionova, S. Kucheryavskiy, A. Pomerantsev, DD-SIMCA - A MATLAB GUI tool for data driven SIMCA approach, Chemometr. Intell. Lab. 167 (2017) 23–28.

[86] Milano Chemometrics and QSAR Research Group. Classification Toolbox for Matlab webpage, `https:`

Journal Pre-proof

//michem.unimib.it/download/matlab-toolboxes/classification-toolbox-for-matlab [online, cited April 20th 2023].

[87] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Anal. Methods 5 (2013) 3790–3798.

[88] S. Kucheryavskiy. mdatools webpage, https://cran.r-project.org/web/packages/mdatools [online, cited April 20th 2023].

[89] S. Kucheryavskiy, mdatools - R package for chemometrics, Chemometr. Intell. Lab. 198 (2020) 103937.

[90] V. Todorov. RSIMCA webpage, https://rdrr.io/cran/rrcovHD/man/RSimca.html [online, cited April 20th 2023].

[91] K. Vanden Branden, M. Hubert, Robust classification in high dimensions based on the SIMCA method, Chemometr. Intell. Lab. 79 (2005) 10–21.

[92] V. Todorov, P. Filzmoser, Software tools for robust analysis of high-dimensional data, Austrian J. Stat. 43 (2014) 255–266.

Appendix

In order to help the reader find the needle in the haystack of all the SIMCA variants described here, Table 1 contains a list of the most popular commercial and non-commercial software tools and computational packages - currently available and developed prior to the one provided together with this tutorial - that enable SIMCA modelling as well as details on the particular SIMCA version they incorporate.

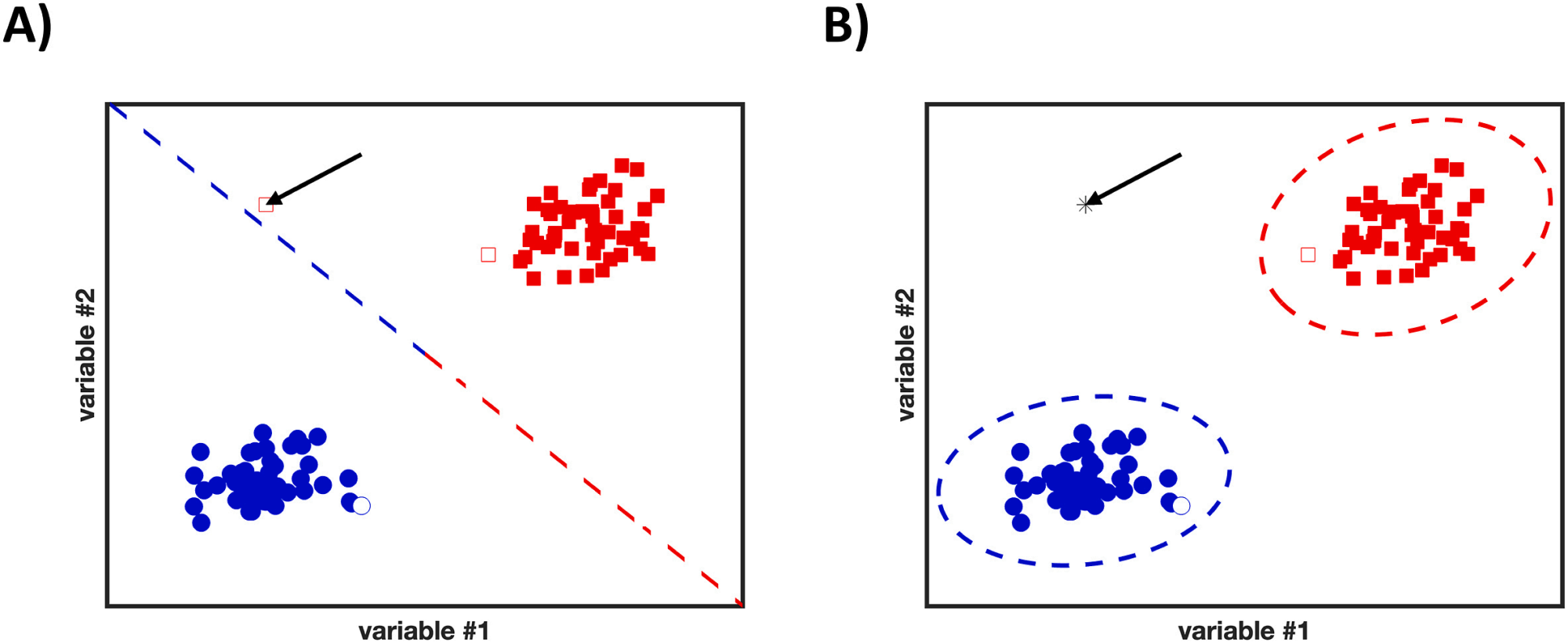
Journal Pre-proof

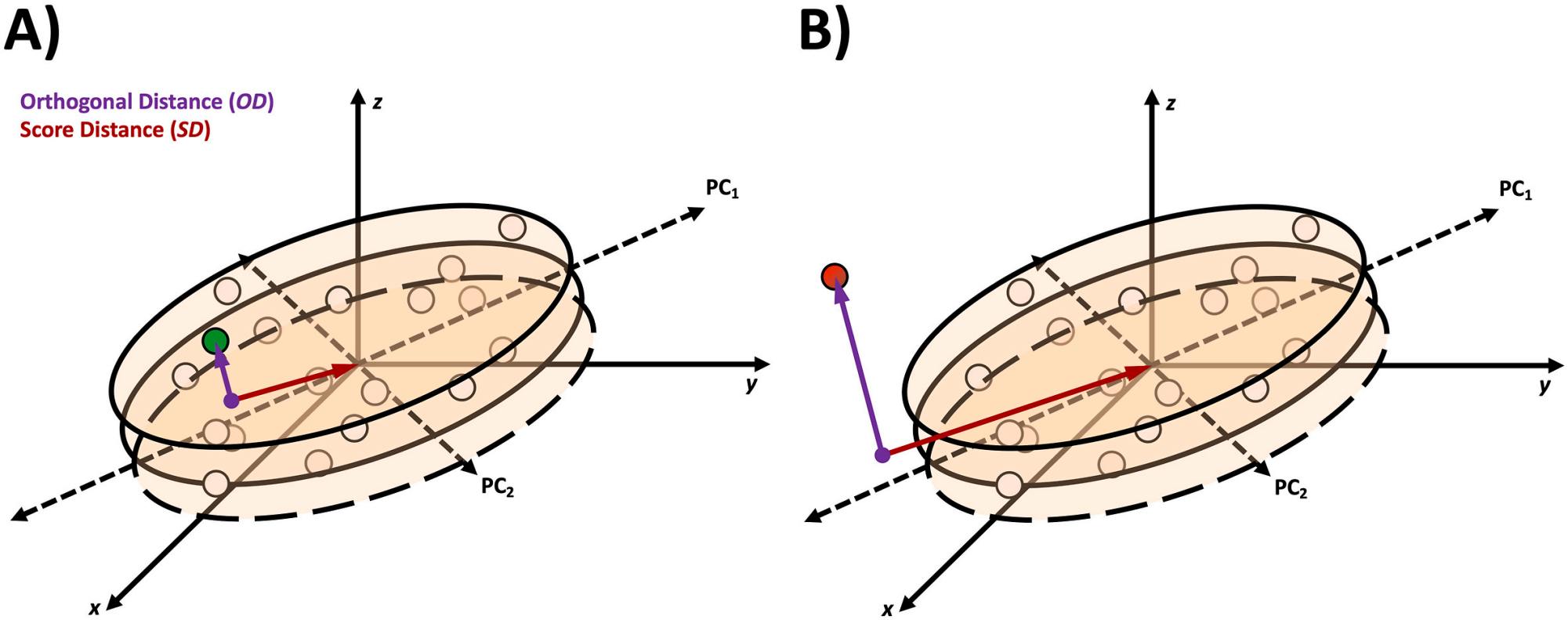
Journal Pre-proof

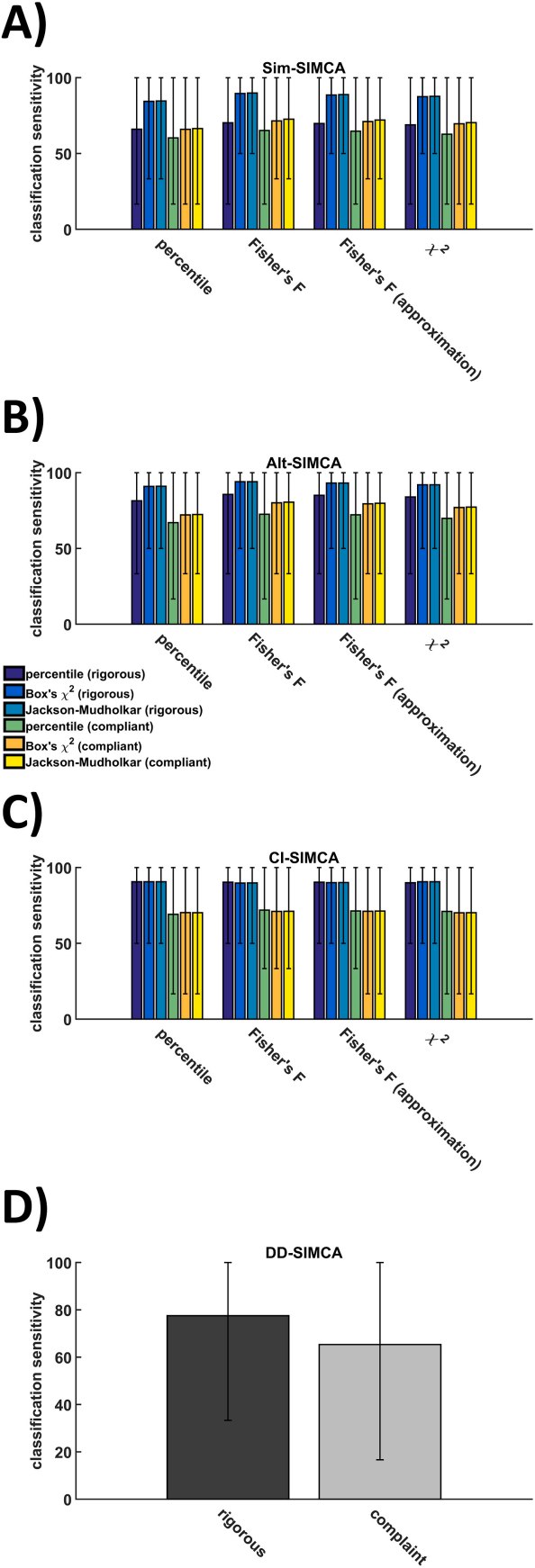
Table 1. List of the most popular commercial and non-commercial software tools and computational packages - currently available and developed prior to the one provided together with this tutorial - that enable SIMCA modelling. The table also contains information on the particular SIMCA version they incorporate.

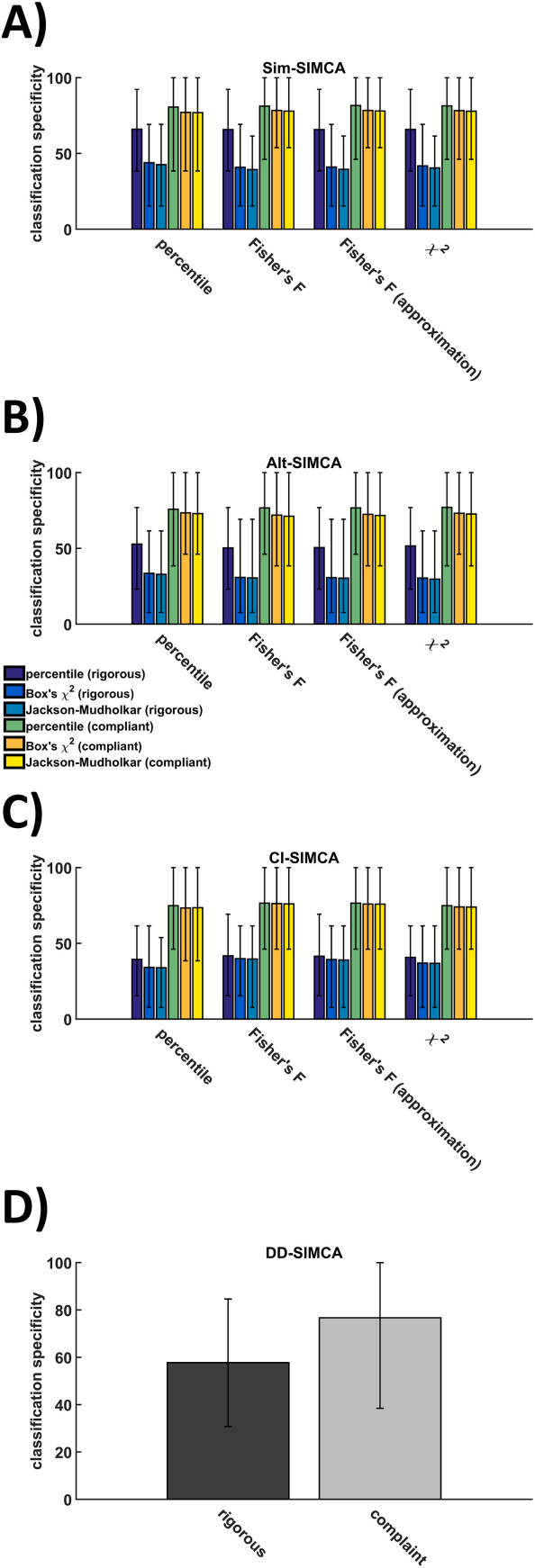
Software package	Software license	Operating environment(s)	SIMCA version(s)	Notes
SIMCA© (Sartorius AG, Göttingen, Germany) [80]	Commercial	Standalone	Original SIMCA	Two different classification rules are implemented in SIMCA©: the first is based on a measure of OD called $DModXPS$ and defined as in Equation 4, the second on a distance metric denominated $DModXPS+$ which combines both OD and SD estimates. It is worth noticing that here SD is expressed as in [15]. $DModXPS$ and $DModXPS+$ (normalised by s - see Equation 5) are assumed to follow a Fisher's F distribution with a number of degrees of freedom redefined with respect to Equation 6. SIMCA© does not automatically adjust the SIMCA model dimensionality based on measures of classification errors estimated in cross-validation.
Aspen Unscrambler™ (Aspen Technology, Inc., Bedford, United States of America) [81]	Commercial	Standalone	Original SIMCA	In Aspen Unscrambler™, a sample is accepted by a given class model if its respective OD and SD values are found to be both lower than their corresponding statistical thresholds. It is worth noticing that here SD is calculated as a measure of leverage [82]. Aspen Unscrambler™ does not automatically adjust the SIMCA model dimensionality based on measures of classification errors estimated in cross-validation.
PLS_Toolbox (Eigenvector Research, Inc., Manson, United States of America) [83]	Commercial	Standalone/Matlab	Alt-SIMCA	In PLS_Toolbox, samples can be classified based on i) their respective OD values, ii) their respective SD values or iii) a combination of their respective OD and SD values. Statistical thresholds for OD are estimated as per Box's theory or the approximation suggested by Jackson and Mudholkar. Statistical thresholds for SD are estimated as in Equation 27. Such thresholds are then exploited to calculate class belonging probabilities for all the investigated specimens which are afterwards optionally assigned (in a discriminant fashion) to the different categories under study according to a maximum-probability or a probability-higher-than-a-threshold criterion. PLS_Toolbox does not automatically adjust the SIMCA model dimensionality based on measures of classification errors estimated in cross-validation.
DD-SIMCA Tool [84, 85]	Non-commercial	Matlab	DD-SIMCA	The DD-SIMCA Tool also incorporates a robust version of DD-SIMCA. It does not automatically adjust the SIMCA model dimensionality based on measures of classification errors estimated in cross-validation.
Classification Toolbox for Matlab [86, 87]	Non-commercial	Matlab	Alt-SIMCA	OD and SD statistical thresholds as well as the SIMCA model dimensionality are simultaneously tuned through a compliant cross-validatory procedure similar to the one described in [47].
mdatools [88, 89]	Non-commercial	R	Alt-SIMCA and DD-SIMCA	-
RSIMCA [90–92]	Non-commercial	R	Alt-SIMCA	RSIMCA incorporates a robust version of Alt-SIMCA.

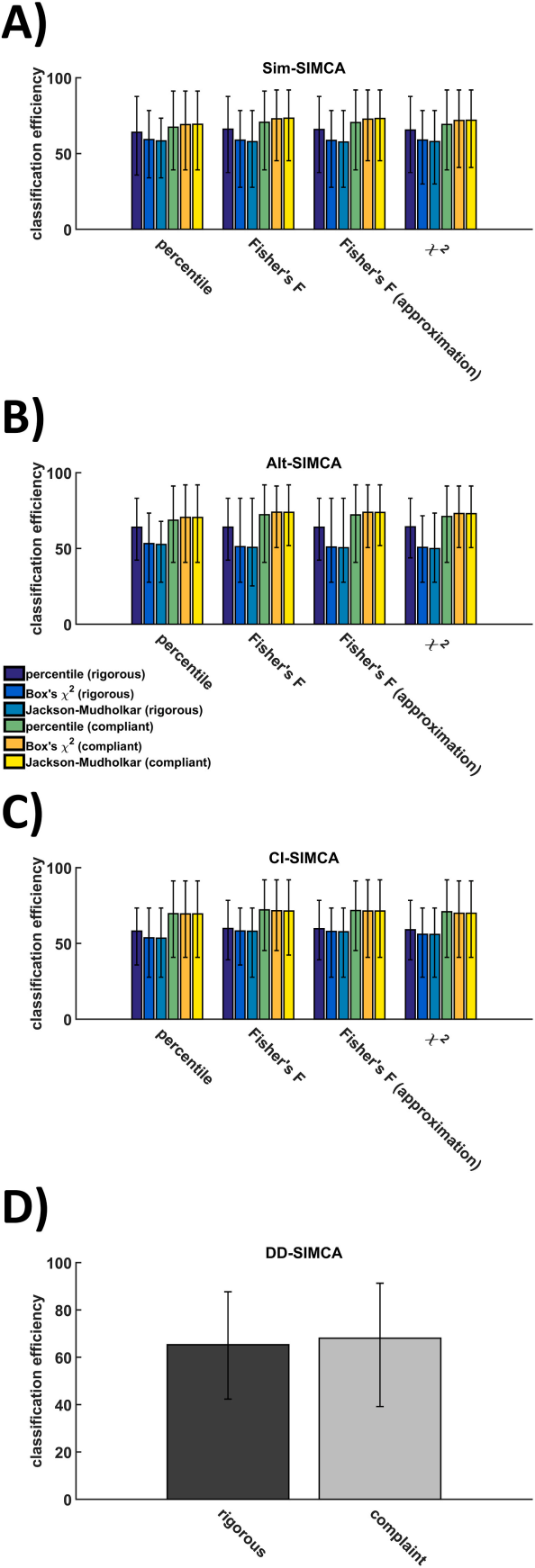
Journal Pre-proof

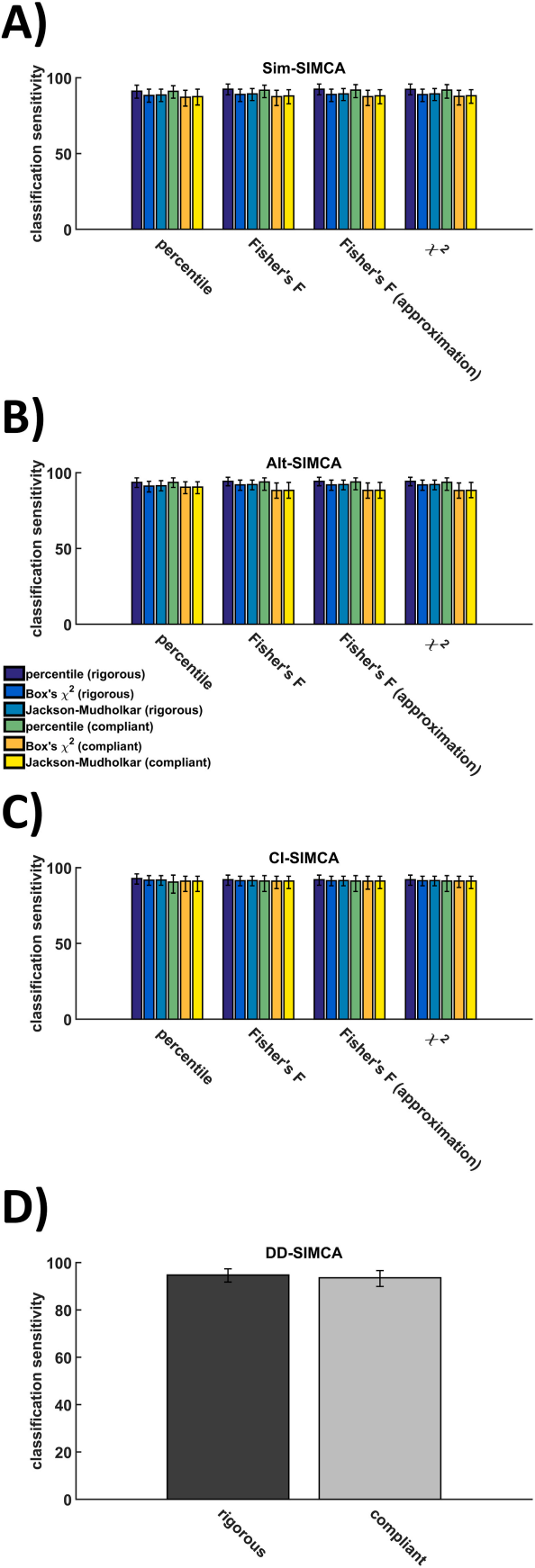

A)
variable #2
variable #1
B)
variable #2
variable #1


A)
Orthogonal Distance (OD)
Score Distance (SD)
z
PC1
y
PC2
x
B)
z
PC1
y
PC2
x

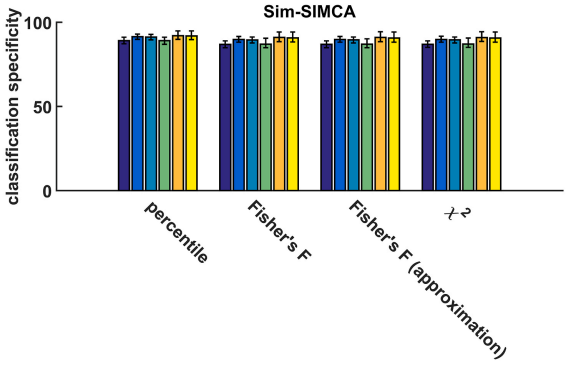

A)
Sim-SIMCA
classification sensitivity
100
50
0
percentile
Fisher's F
Fisher's F (approximation)
χ²
B)
Alt-SIMCA
classification sensitivity
100
50
0
percentile
Fisher's F
Fisher's F (approximation)
χ²
percentile (rigorous)
Box's χ² (rigorous)
Jackson-Mudholkar (rigorous)
percentile (compliant)
Box's χ² (compliant)
Jackson-Mudholkar (compliant)
C)
CI-SIMCA
classification sensitivity
100
50
0
percentile
Fisher's F
Fisher's F (approximation)
χ²
D)
DD-SIMCA
classification sensitivity
100
80
60
40
20
0
rigorous
complaint


A)
Sim-SIMCA
classification specificity
100
50
0
percentile
Fisher's F
Fisher's F (approximation)
χ^2
B)
Alt-SIMCA
classification specificity
100
50
0
percentile
Fisher's F
Fisher's F (approximation)
χ^2
percentile (rigorous)
Box's χ^2 (rigorous)
Jackson-Mudholkar (rigorous)
percentile (compliant)
Box's χ^2 (compliant)
Jackson-Mudholkar (compliant)
C)
CI-SIMCA
classification specificity
100
50
0
percentile
Fisher's F
Fisher's F (approximation)
χ^2
D)
DD-SIMCA
classification specificity
100
80
60
40
20
0
rigorous
complaint

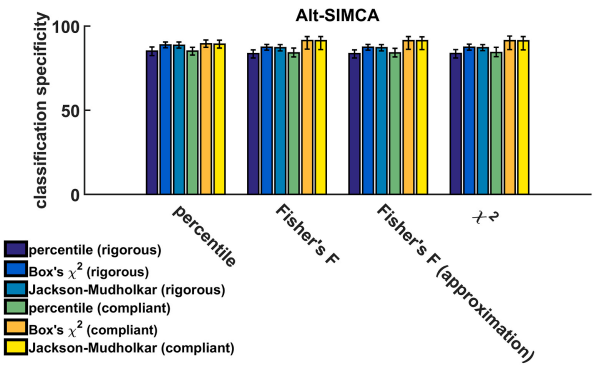

A)
Sim-SIMCA
classification efficiency
100
50
0
percentile
Fisher's F
Fisher's F (approximation)
χ^2
B)
Alt-SIMCA
classification efficiency
100
50
0
percentile
Fisher's F
Fisher's F (approximation)
χ^2
percentile (rigorous)
Box's χ^2 (rigorous)
Jackson-Mudholkar (rigorous)
percentile (compliant)
Box's χ^2 (compliant)
Jackson-Mudholkar (compliant)
C)
CI-SIMCA
classification efficiency
100
50
0
percentile
Fisher's F
Fisher's F (approximation)
χ^2
D)
DD-SIMCA
classification efficiency
100
80
60
40
20
0
rigorous
complaint


A)
Sim-SIMCA
classification sensitivity
100
50
0
percentile
Fisher's F
Fisher's F (approximation)
χ^2
B)
Alt-SIMCA
classification sensitivity
100
50
0
percentile
Fisher's F
Fisher's F (approximation)
χ^2
percentile (rigorous)
Box's χ^2 (rigorous)
Jackson-Mudholkar (rigorous)
percentile (compliant)
Box's χ^2 (compliant)
Jackson-Mudholkar (compliant)
C)
CI-SIMCA
classification sensitivity
100
50
0
percentile
Fisher's F
Fisher's F (approximation)
χ^2
D)
DD-SIMCA
classification sensitivity
100
80
60
40
20
0
rigorous
compliant

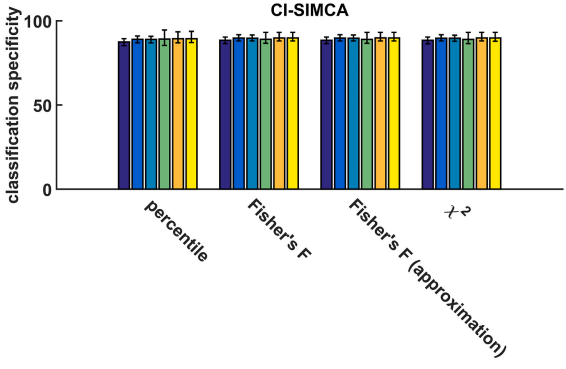
A)


Sim-SIMCA
classification specificity
100
50
0
percentile
Fisher's F
Fisher's F (approximation)
χ^2

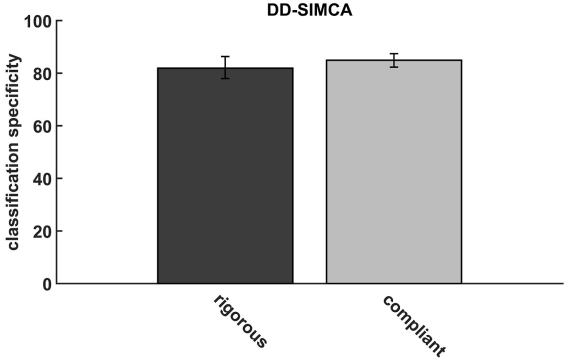
B)

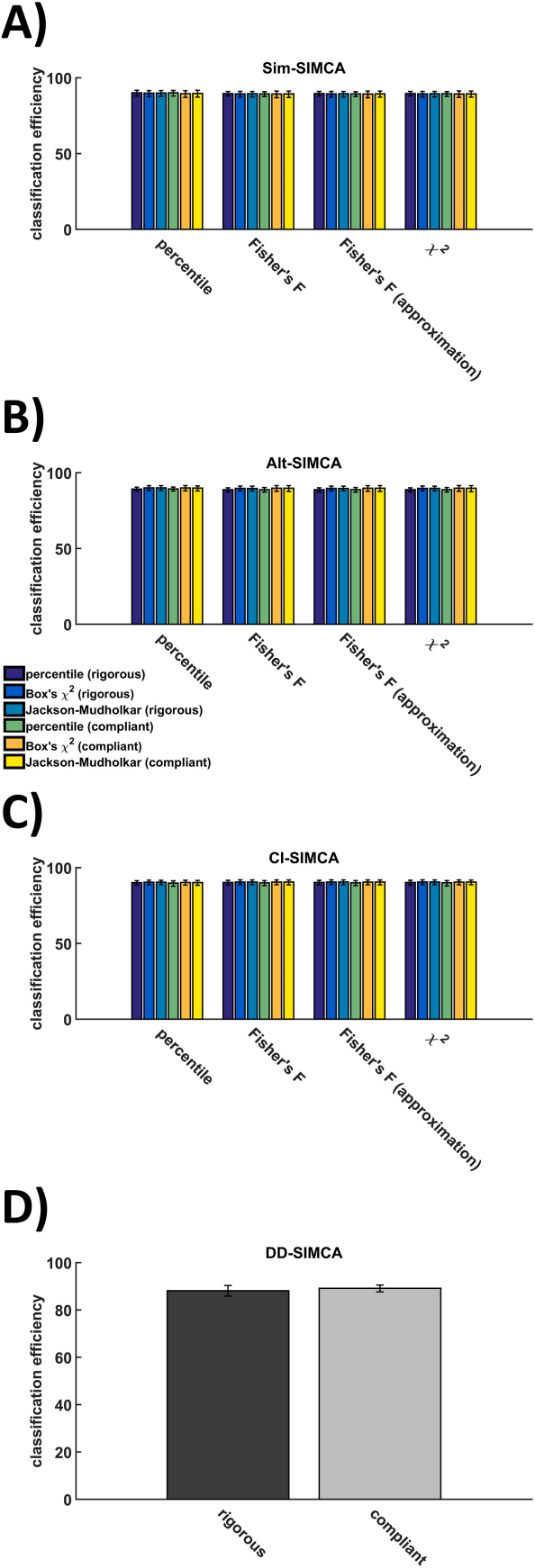

Alt-SIMCA
classification specificity
100
50
0
percentile
Fisher's F
Fisher's F (approximation)
χ^2
percentile (rigorous)
Box's χ^2 (rigorous)
Jackson-Mudholkar (rigorous)
percentile (compliant)
Box's χ^2 (compliant)
Jackson-Mudholkar (compliant)

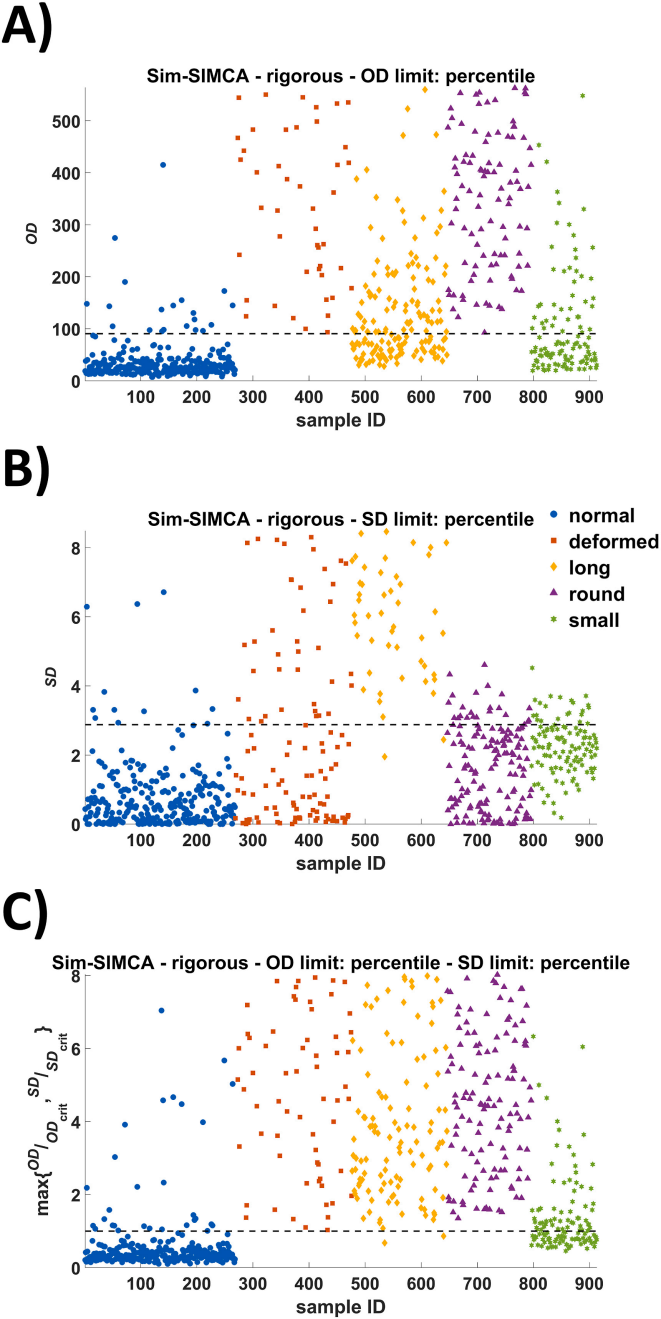
C)


CI-SIMCA
classification specificity
100
50
0
percentile
Fisher's F
Fisher's F (approximation)
χ^2

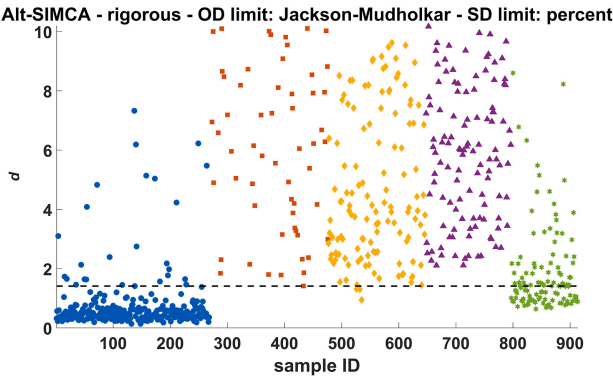
D)


DD-SIMCA
classification specificity
100
80
60
40
20
0
rigorous
compliant

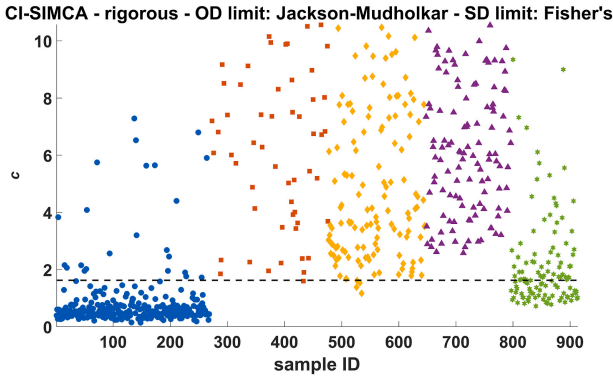

A)
Sim-SIMCA
classification efficiency
100
50
0
percentile
Fisher's F
Fisher's F (approximation)
χ^2
B)
Alt-SIMCA
classification efficiency
100
50
0
percentile
Fisher's F
Fisher's F (approximation)
χ^2
percentile (rigorous)
Box's χ^2 (rigorous)
Jackson-Mudholkar (rigorous)
percentile (compliant)
Box's χ^2 (compliant)
Jackson-Mudholkar (compliant)
C)
CI-SIMCA
classification efficiency
100
50
0
percentile
Fisher's F
Fisher's F (approximation)
χ^2
D)
DD-SIMCA
classification efficiency
100
80
60
40
20
0
rigorous
compliant


A)
Sim-SIMCA - rigorous - OD limit: percentile
OD
sample ID
B)
Sim-SIMCA - rigorous - SD limit: percentile
normal
deformed
long
round
small
SD
sample ID
C)
Sim-SIMCA - rigorous - OD limit: percentile - SD limit: percentile
max{OD/ODcrit, SD/SDcrit}
sample ID

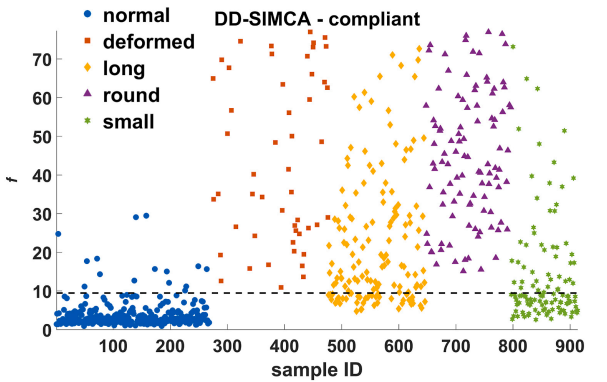
D)


Alt-SIMCA - rigorous - OD limit: Jackson-Mudholkar - SD limit: percentile
d
sample ID

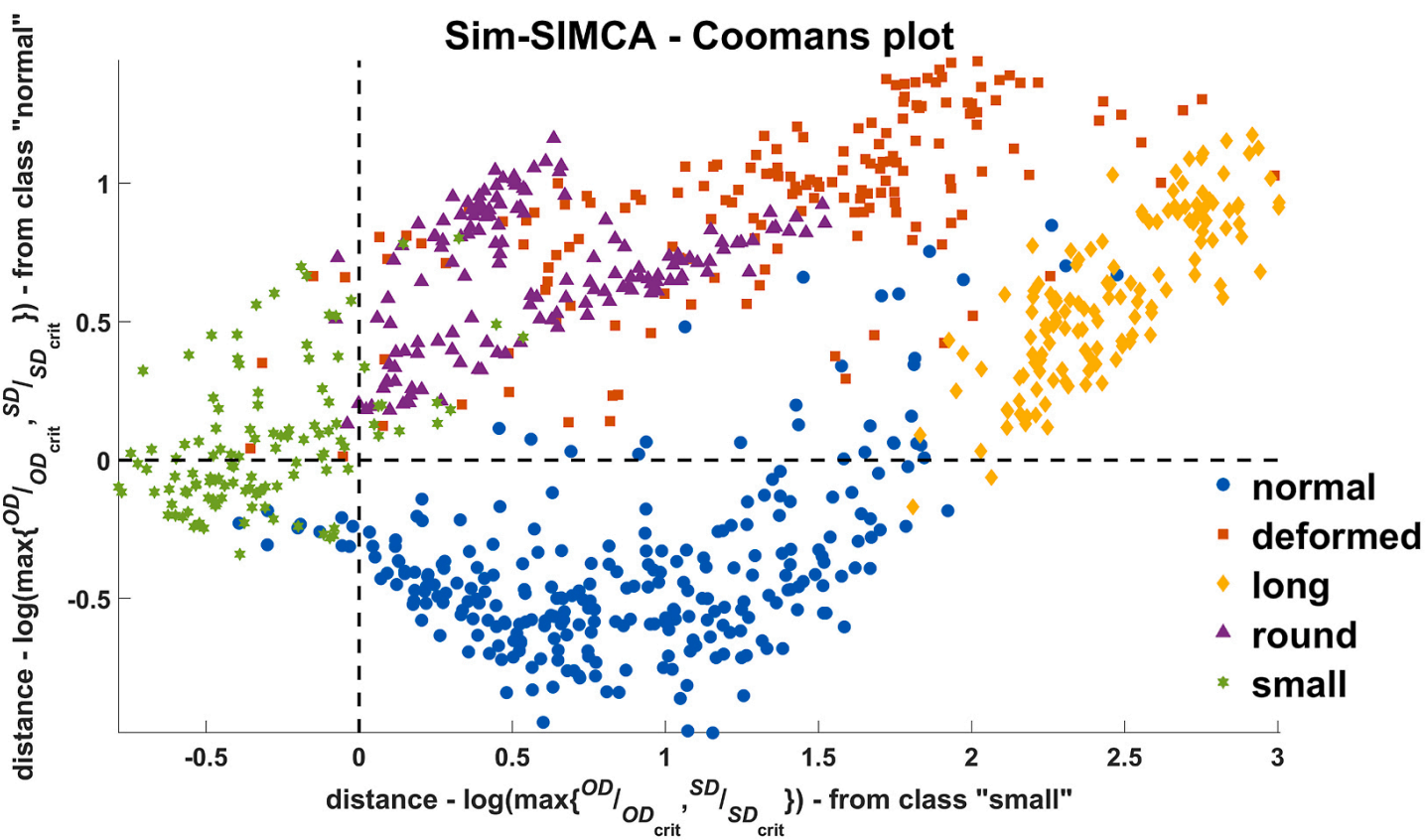
E)


CI-SIMCA - rigorous - OD limit: Jackson-Mudholkar - SD limit: Fisher's F
c
sample ID

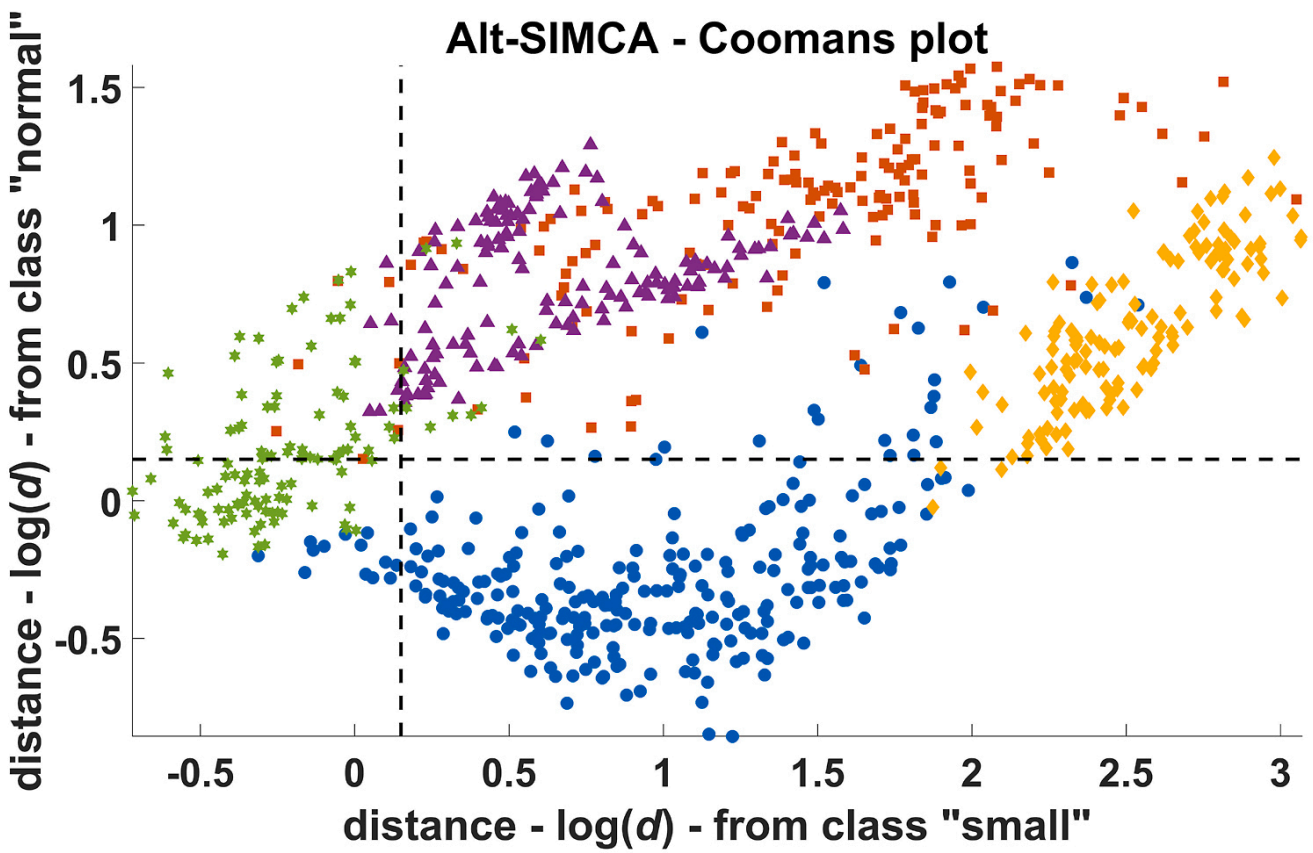
F)


normal
deformed
long
round
small
DD-SIMCA - compliant
f
sample ID

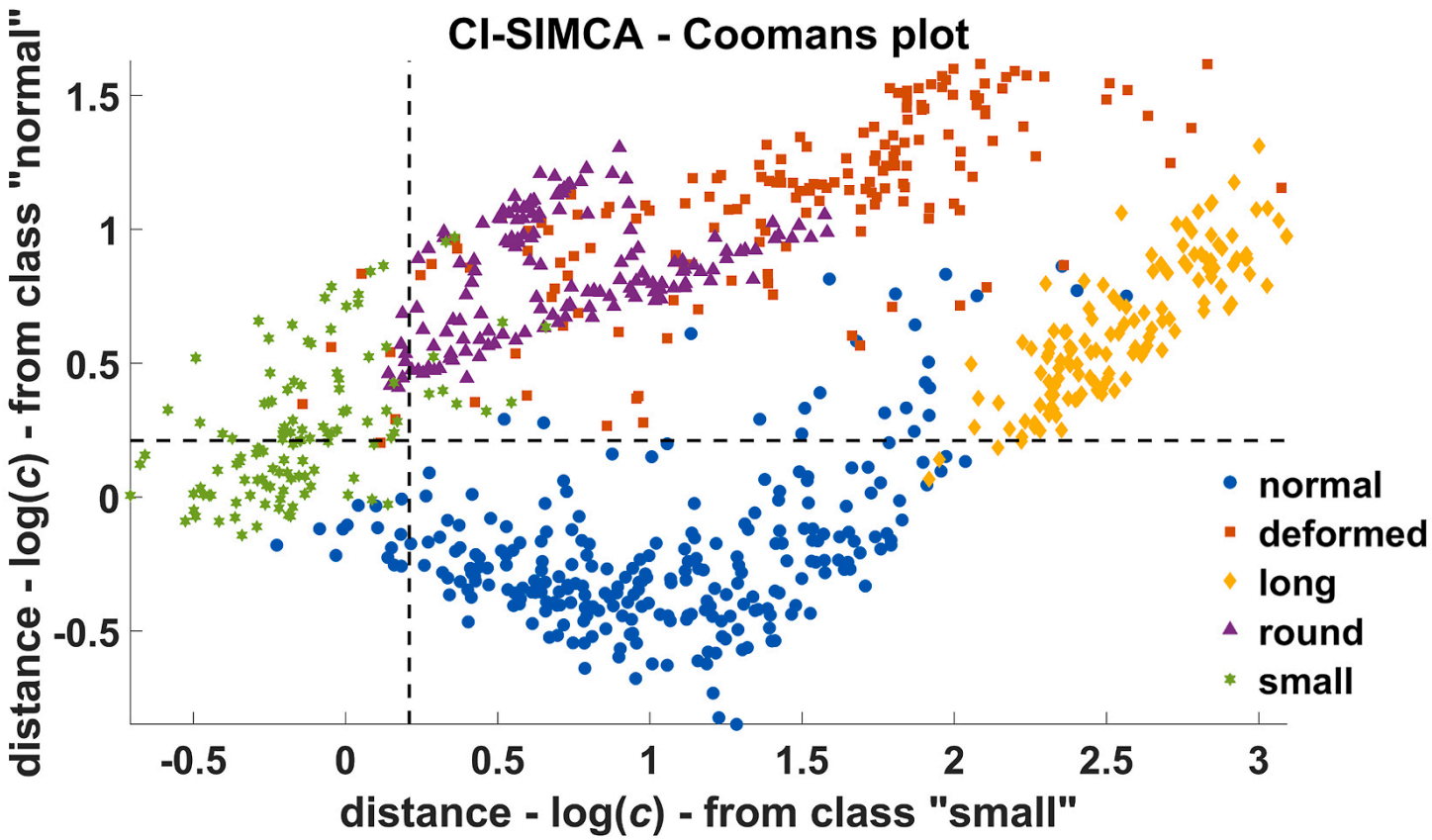
A)


Sim-SIMCA - Coomans plot
distance - log(max{$^{OD}/_{OD_{crit}}$, $^{SD}/_{SD_{crit}}$}) - from class "normal"
distance - log(max{$^{OD}/_{OD_{crit}}$, $^{SD}/_{SD_{crit}}$}) - from class "small"
normal
deformed
long
round
small

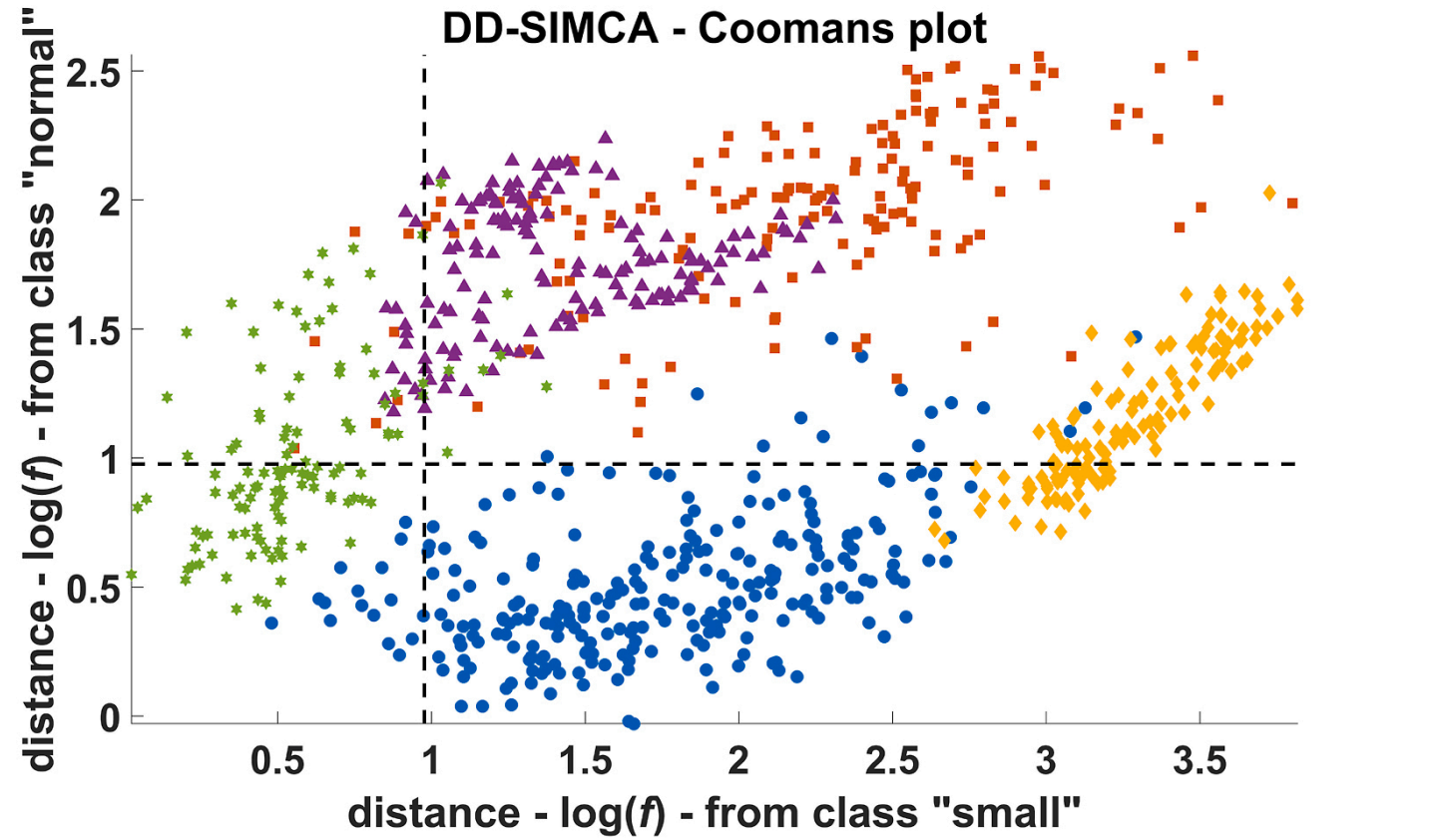
B)


Alt-SIMCA - Coomans plot
distance - log(d) - from class "normal"
distance - log(d) - from class "small"

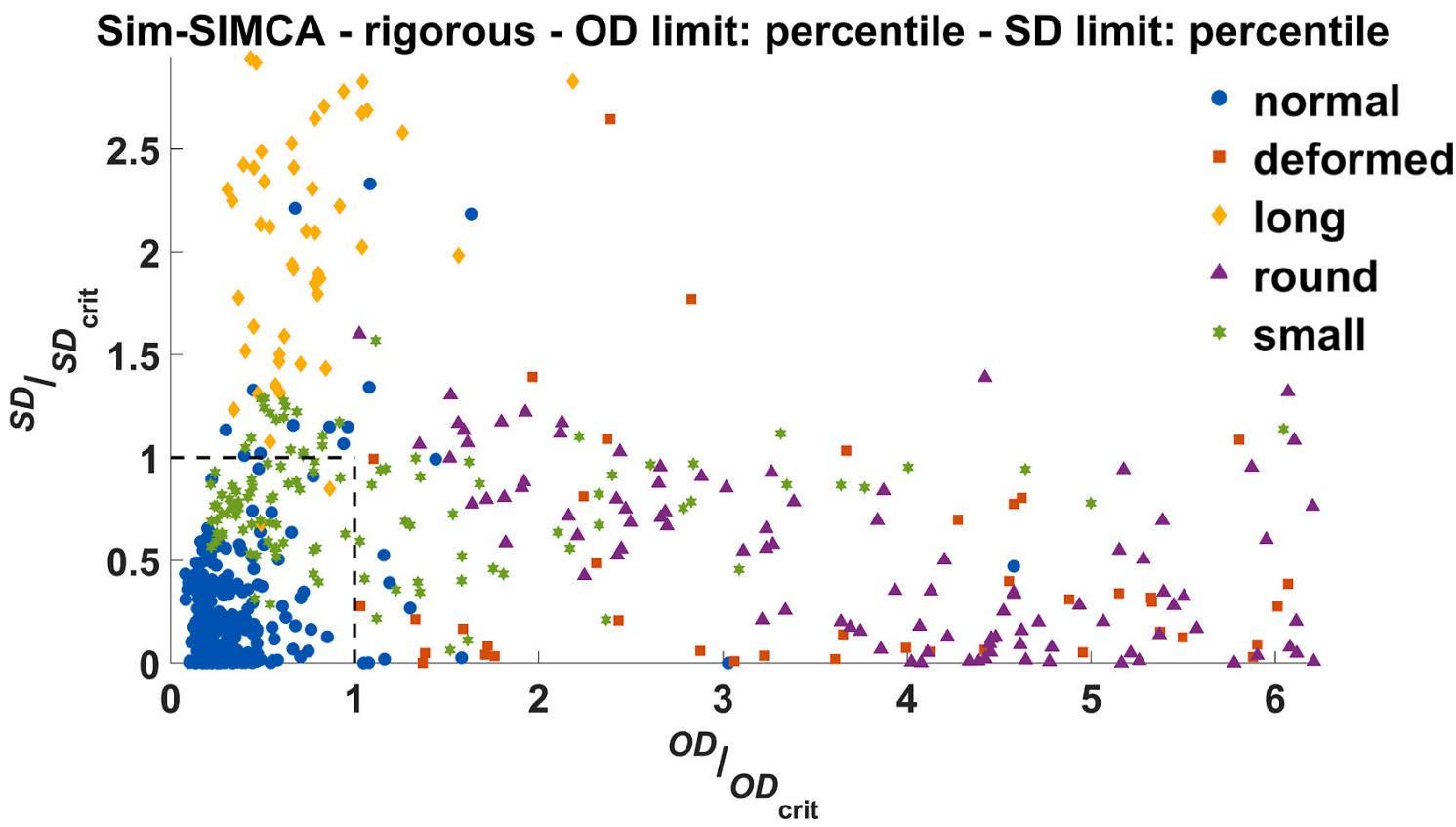
C)


CI-SIMCA - Coomans plot
distance - log(c) - from class "normal"
distance - log(c) - from class "small"
1.5
1
0.5
0
-0.5
-0.5
0
0.5
1
1.5
2
2.5
3
normal
deformed
long
round
small

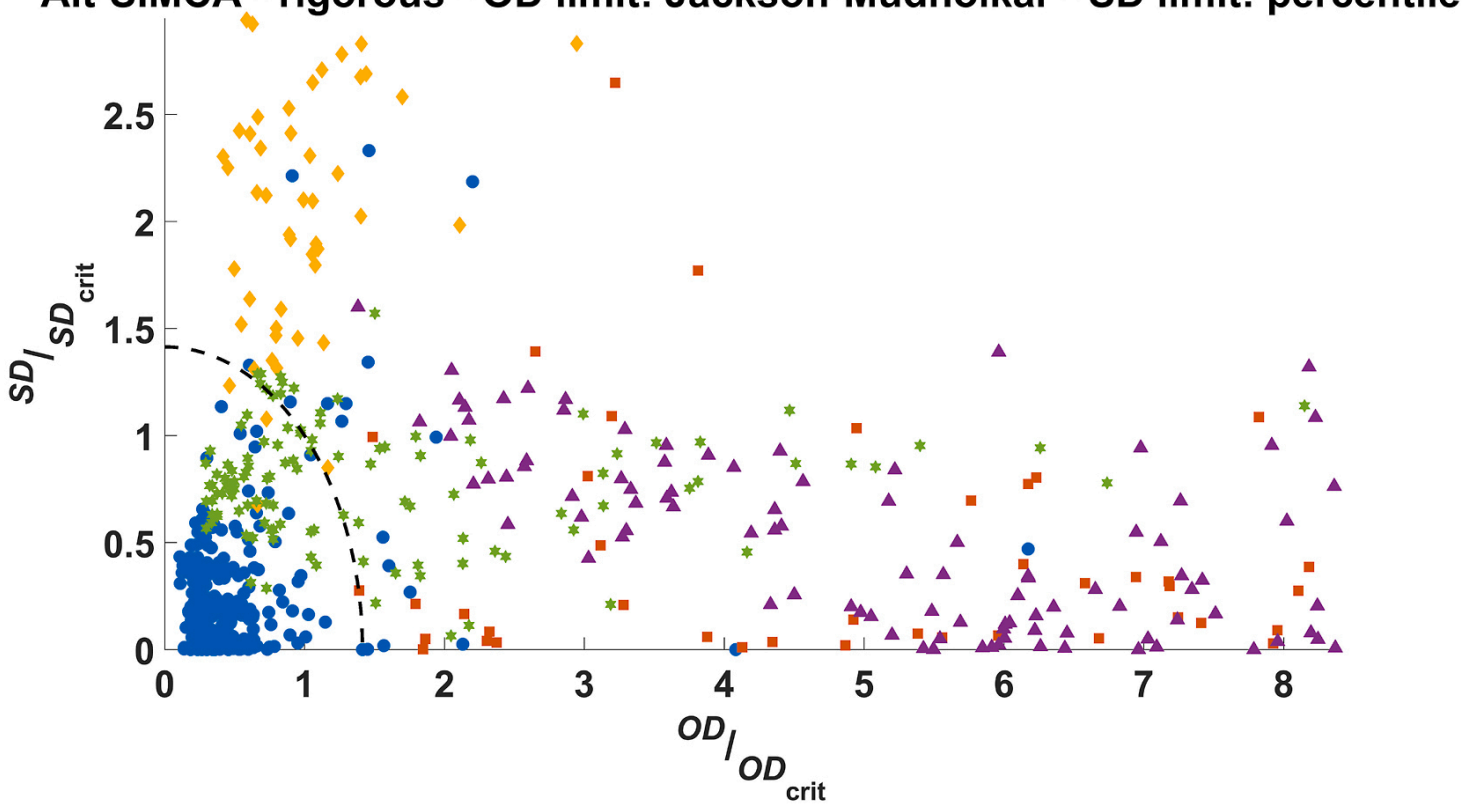
D)


DD-SIMCA - Coomans plot
distance - log(f) - from class "normal"
distance - log(f) - from class "small"
2.5
2
1.5
1
0.5
0
0.5
1
1.5
2
2.5
3
3.5

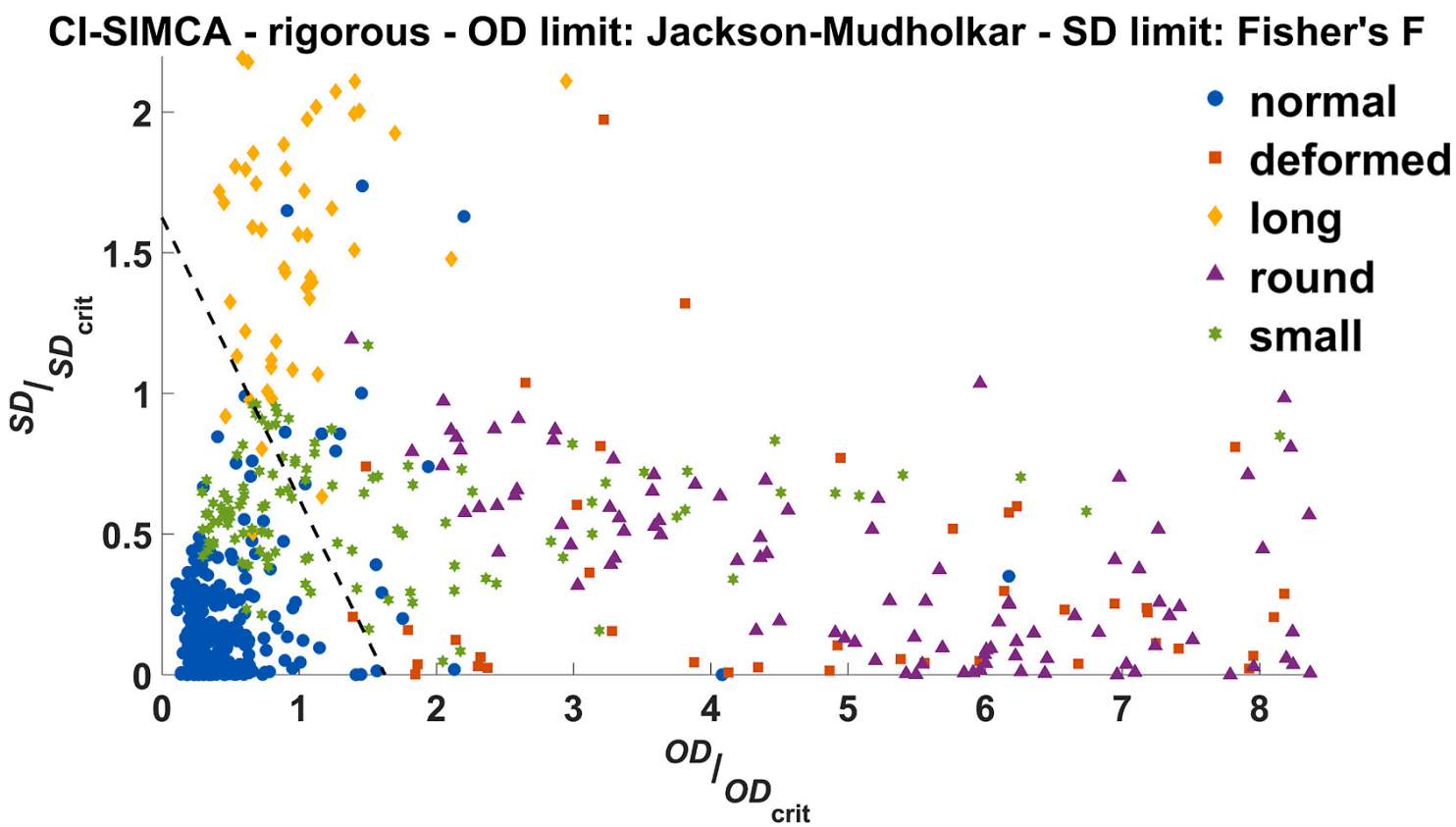
A)


Sim-SIMCA - rigorous - OD limit: percentile - SD limit: percentile
normal
deformed
long
round
small
SD / SD crit
OD / OD crit
0
0.5
1
1.5
2
2.5
0
1
2
3
4
5
6

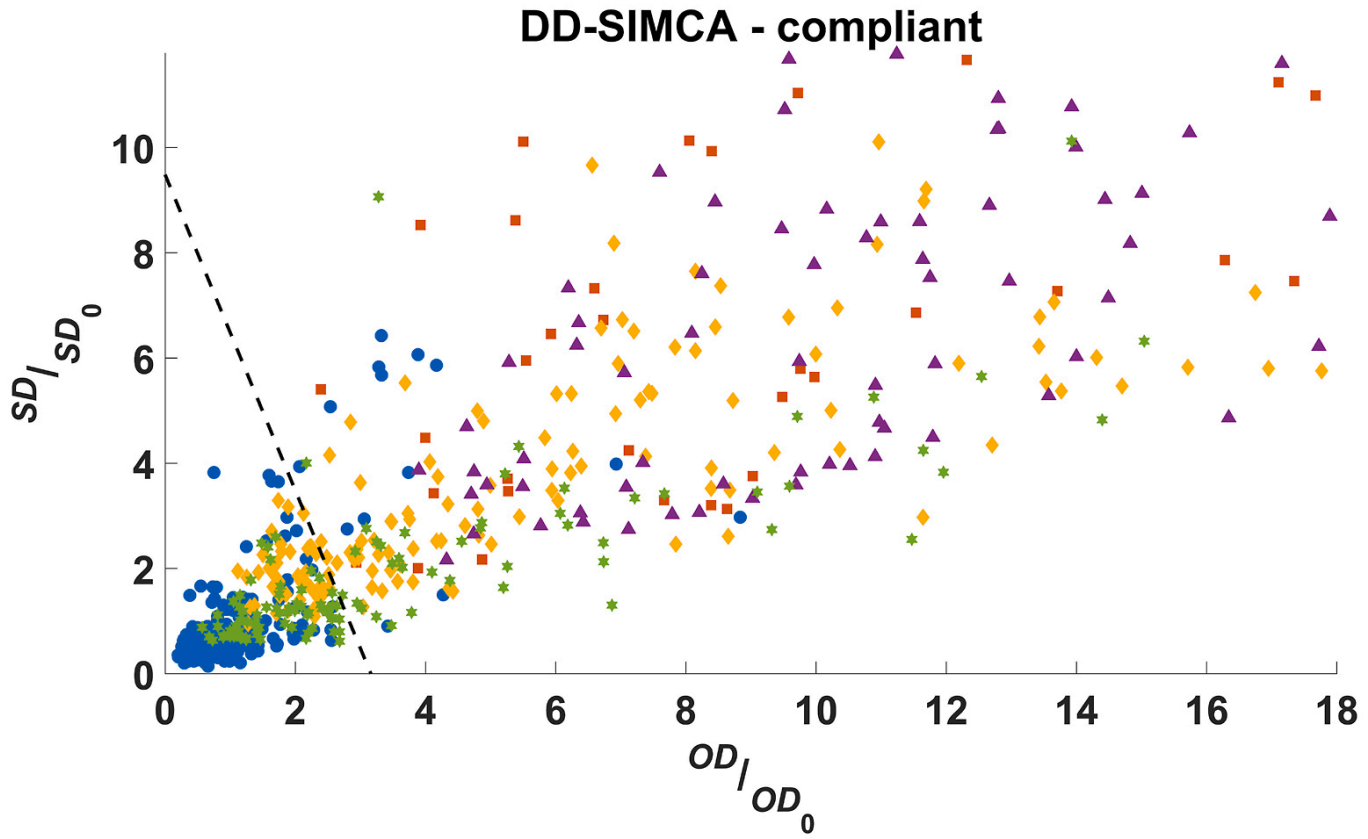
B)


Alt-SIMCA - rigorous - OD limit: Jackson-Mudholkar - SD limit: percentile
SD / SD crit
OD / OD crit
0
0.5
1
1.5
2
2.5
0
1
2
3
4
5
6
7
8

C)


CI-SIMCA - rigorous - OD limit: Jackson-Mudholkar - SD limit: Fisher's F
normal
deformed
long
round
small
SD / SD crit
OD / OD crit
0
0.5
1
1.5
2
0
1
2
3
4
5
6
7
8

D)


DD-SIMCA - compliant
SD / SD 0
OD / OD 0
0
2
4
6
8
10
0
2
4
6
8
10
12
14
16
18

Highlights

- A tutorial on Soft Independent Modelling of Class Analogy (SIMCA)
- Practical guidelines about why, when and how/how not employing SIMCA
- A comparison among four distinct variants of the original SIMCA algorithm
- A flowchart on how to fine-tune SIMCA model parameters
- Rational suggestions about SIMCA model assessment and validation

Journal Pre-proof

Alessandra Biancolillo obtained her Ph.D. in Spectroscopy and Chemometrics at the University of Copenhagen (Denmark) in 2016, working on a joint project with the Norwegian Research Institute of Fishery and Food (NOFIMA). She had post-doc positions at the University of Rome "La Sapienza" (Italy) and at the IRSTEA institute in Montpellier (France). From 2019, she is Senior Researcher at the University of L'Aquila (Italy). In 2022, she was awarded the Young Researcher Prize of the Italian Chemometric Group. Her main research topics cover multi-block data analysis, classification method development, and feature selection.

Journal Pre-proof

Cyril Ruckebusch is Full Professor and member of the "Dynamics, Nanoscopy and Chemometrics" (DyNaChem) team of the Laboratory of Advanced Spectroscopy, Dynamics, Reactivity and Environmental Studies (LASIRE – https://lasir.cnrs.fr/dynachem/) of the University of Lille (France). He has published 130 research papers covering many aspects related to the application of chemometrics in time-resolved spectroscopy and spectral imaging. His current research focuses on developing new workflows for data acquisition and unmixing in hyperspectral microscopy imaging. He is Associate Editor-in-Chief of Journal of Chemometrics and member of the editorial board of Analytica Chimica Acta.

Journal Pre-proof

Federico Marini is Full Professor of Analytical Chemistry at the University of Rome "La Sapienza". In 2006, he was awarded the Young Researcher Prize of the Italian Chemical Society and in 2012 he won the Chemometrics and Intelligent Laboratory Systems Award *for his achievements in chemometrics*. He has been visiting researcher in various Universities (Copenhagen, Stellenbosch, Silesia, Lille). His research activity is focused on all aspects of chemometrics, ranging from the application of existing methods to real world problems in different fields to the design and development of novel algorithms. He is author of more than 240 papers in international journals, and he edited and co-authored the book "Chemometrics in Food Chemistry" (Elsevier). He is editor of Chemometrics and Intelligent Laboratory Systems and Frontiers in Analytical Science and member of the editorial boards of Analytica Chimica Acta, Journal of Chemometrics, Journal of Near-Infrared Spectroscopy, Journal of Spectral Imaging, and Food Control and serves as Associate Editor for Chemometrics in Encyclopedia of Analytical Chemistry (Wiley). He is currently the leader of the study group in chemometrics of DAC-EuChemS.

Journal Pre-proof

Journal Pre-proof

Marina Cocchi is Full Professor in Analytical Chemistry and Chemometrics at the Department of Chemical and Geological Sciences of the University of Modena and Reggio Emilia (Italy). She holds a degree and a Ph.D. in Chemical Sciences from the University of Modena and she teaches chemometrics at undergraduate and graduate levels. During her Ph.D., she worked with Prof. Svante Wold on the development of chemometric approaches for 3D QSAR. She has published more than 100 papers in international journals and books covering a range of topics embracing multivariate, multi-way and multiset methods, data fusion, 2D wavelet transform in multivariate image analysis for fault detection and pattern recognition, algorithms for feature selection in wavelet domain, multivariate statistical process control, food authenticity, and chemical fingerprinting by spectroscopy and chromatography. She has been member of the board of the Italian Chemometric Group from 2001 to 2015, acting as president between 2007 and 2011. She has been member of the editorial board of Chemometrics and Intelligent Laboratory Systems since 2010 and member of the Advisory Board for "Comprehensive Chemometrics" (Elsevier). She is the editor of the book "Data Fusion: Methods and Applications" (Elsevier).

Journal Pre-proof

Raffaele Vitale is Associate Professor and member of the "Dynamics, Nanoscopy and Chemometrics" (DyNaChem) team of the Laboratory of Advanced Spectroscopy, Dynamics, Reactivity and Environmental Studies (LASIRE – https://lasir.cnrs.fr/dynachem/) of the University of Lille (France). He received his M.Sc. in Analytical Chemistry at the University of Rome "La Sapienza" (Italy) and his Ph.D. in Statistics and Optimisation at the Technical University of Valencia (Spain). Raffaele is author of around 50 peer-reviewed publications and has been granted with several awards including the Siemens Process Analytics Prize for Young Scientist in 2017, the III Jean-Pierre Huvenne Award for the Best Ph.D. thesis in Chemometrics in 2019 and the XVI European Network for Business and Industrial Statistics Young Statistician Award in 2020. Currently, his work is mainly focused on the development and application of multivariate statistical approaches for the analysis of hyperspectral and optical microscopy images. He is member of the editorial board of Chemometrics and Intelligent Laboratory Systems.

Journal Pre-proof

The authors declare no conflict of interest.

Journal Pre-proof