Analytica Chimica Acta 1375 (2025) 344575

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Modeling power-based variable selection for rigorous one-class classification with SIMCA

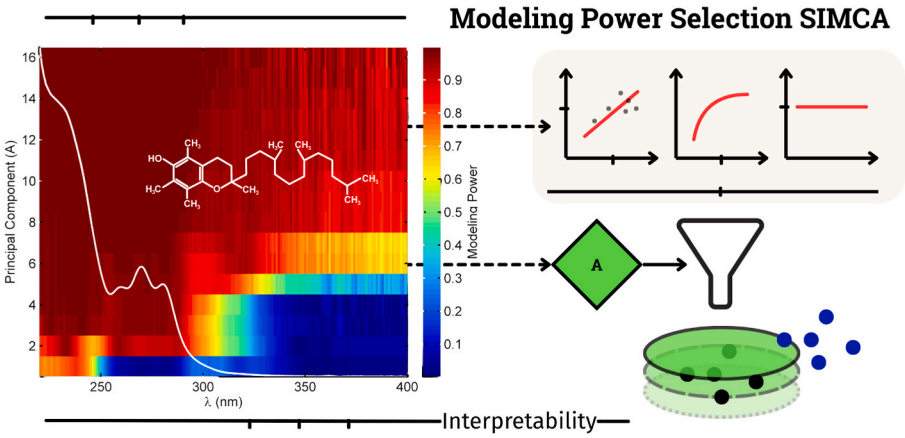
Mateus Pires Schneider [a], Cristina Malegori [b], Paolo Oliveri [b,*], Adriano de Araújo Gomes [a]

[a] Institute of Chemistry, UFRGSAv. Bento Gonçalves, 9500, Porto Alegre, RS, CEP 91591 – 970, Brazil
[b] Dipartimento di Farmacia, DIFAR, Università degli Studi di Genova – UniGeViale Cembrano, 4, Genova, GE, 16148, Italy

HIGHLIGHTS

- A Modeling Power (MP) based algorithm for variable selection in SIMCA is introduced.
- Three criteria are combined: compactness correlation, non-growth rate, MP threshold.
- Interpretable and compact one-class models using only target-class data.
- Classification performance is preserved or improved.
- Selected variables are chemically meaningful and class-representative.

GRAPHICAL ABSTRACT



ARTICLE INFO

Handling editor: Prof. L. Buydens

Keywords:
One-class classification
SIMCA
Variable selection
Modeling power
Chemometrics
Food authentication
Spectroscopy

ABSTRACT

The present study proposes a variable selection algorithm for Soft Independent Modeling of Class Analogy (SIMCA) in the context of rigorous one-class classification (OCC), using Modeling Power (MP) as a central internal metric. The algorithm integrates three complementary criteria: correlation between MP and class compactness, MP non-growth rate across successive principal components, and a minimum MP threshold. These criteria were applied to three experimental datasets: UV–Vis spectra of edible oils, NIR spectra of Argentinean green teas, and HPLC–CAD chromatographic profiles of olive oils. In all cases, the goal was to enhance model parsimony and interpretability without compromising classification performance. The selected variables were chemically meaningful, aligning with known spectral or chromatographic regions associated with key compositional markers. Comparative analyses between the proposed Modeling Power Selector with SIMCA (MPS-SIMCA) and traditional full-spectrum SIMCA showed equivalent or improved classification performance. Additionally, MPS-SIMCA achieved superior model compactness and interpretability, supporting the feasibility of variable selection based solely on internal class structure. This approach offers a robust and interpretable alternative for class modeling in food authentication tasks where only target-class samples are reliably available.

* Corresponding author.
E-mail address: Paolo.Oliveri@unige.it (P. Oliveri).

https://doi.org/10.1016/j.aca.2025.344575
Received 26 June 2025; Received in revised form 25 August 2025; Accepted 26 August 2025
Available online 27 August 2025

1. Introduction

Variable selection is a central topic in multivariate data analysis, particularly in high-dimensional problems often encountered in analytical chemistry and chemometric applications [1–4]. When the number of predictors exceeds the number of samples, the risk of overfitting increases substantially, often compromising model interpretability and generalization. Variable selection aims to reduce model complexity by identifying informative features while discarding noisy or redundant variables, thus enhancing model robustness and interpretability [5,6].

This challenge is especially pronounced in one-class classification (OCC) tasks — also referred to as class-modeling methods —, where only the target class is reliably available, and the absence of non-target well-defined datasets precludes the use of discriminative criteria [1,7,8]. In this setting, traditional variable selection strategies — typically designed for supervised classification problems — may not be applicable as they rely heavily on inter-class variance [5–7]. The situation is further complicated in rigorous OCC, where variable selection and model construction must be performed solely on target-class data, without simulating non-target distributions or pseudo-negatives [9]. In this context, several strategies specifically applicable to OCC have been reported, ranging from filter, wrapper, and embedded methods to hybrid approaches, as systematically reviewed by Schneider et al. [10]. Such strategies are tailored to minimize the mismatch between model sensitivity and the desired confidence level without introducing bias from unknown or unmodeled classes.

Among class-modelling approaches, Soft Independent Modeling of Class Analogy (SIMCA) is particularly suitable for OCC scenarios, where only the target class is reliably available. SIMCA models the structure of the target class using Principal Component Analysis (PCA), defining acceptance boundaries based on score and orthogonal distances derived from the distribution of samples in a reduced latent space [11–14]. However, even within SIMCA, the integration of effective and interpretable variable selection strategies remains an open challenge in the rigorous OCC context.

Recent work by Pomerantsev et al. [15] introduced the LOVE (Leave One Variable Excluded) algorithm, a wrapper-based and interactive variable selection strategy specifically developed for OCC applications using DD-SIMCA. LOVE improves model robustness, compactness, and interpretability by stepwise evaluation of classifier performance after variable or interval exclusion, relying on multiple internal figures of merit such as sensitivity, specificity, and efficiency. In more detail, the LOVE approach was shown to be efficient in enhancing classifier performance, preventing overfitting and improving model stability, fixing outliers without deletion, and leading to a better understanding of relevant factors. While the approach offers valuable insights and allows for flexible, analyst-driven refinement of the model, it is inherently interactive and not designed for full automation. Therefore, despite its significant contributions, there remains a clear need for fully unsupervised and automated variable selection strategies that, while adhering to the constraints of rigorous OCC, operate solely within the internal structure of the target class, defined in this context by the distance of each sample projection to the center of the reduced latent space and by its orthogonal distance to this space. Nonetheless, the development and successful application of LOVE highlight that variable selection in the context of class modeling is a timely and actively evolving research topic.

Modeling Power (MP) — a PCA-derived metric that quantifies how well each variable is represented in the class model — offers a promising route for internal, unsupervised variable selection. MP is computed by comparing the residual standard deviation of a variable in the modeled PCA space to its original standard deviation, offering an interpretable, signal-to-noise-like estimate for each variable [11,13]. Although MP provides a principled basis for objective variable selection and is easily computed in routine calculations within SIMCA models, it has rarely been employed systematically to guide variable reduction, particularly in robust OCC frameworks where external class information is not required.

Aiming to enhance variable selection, model interpretability, and generalization capacity, the present study proposes an algorithm termed Modeling Power Selector for SIMCA (MPS-SIMCA) that combines three MP-derived criteria to select variables in SIMCA–OCC: correlation with class compactness, non-growth rate across successive components, and an absolute minimum threshold. By focusing exclusively on the internal structure of the target class, the approach ensures that variable selection is objective, data-driven, and independent of any information about non-target samples. The method is evaluated on three datasets — two spectroscopic (UV–Vis and NIR) and one chromatographic — using food-related applications as case studies to assess classification performance, interpretability, and robustness. Its effectiveness is assessed by comparing MPS–SIMCA to the traditional full-spectrum SIMCA, focusing on analytical figures of merit (AFOM), including sensitivity, specificity, and model compactness as key indicators of classification quality. While food authentication provides a real-world application context, the methodological contributions of this work are broadly applicable to any OCC scenario involving high-dimensional data and the need for unsupervised variable selection.

2. Proposed algorithm

This study proposes a novel variable selection algorithm tailored for Soft Independent Modeling of Class Analogy (SIMCA) within the framework of OCC. The algorithm integrates Modeling Power (MP) as its primary criterion, leveraging its capacity to quantify variable contribution to class modeling across principal components. Beyond absolute MP, the approach incorporates complementary metrics derived from MP behavior itself, including the correlation between MP and class compactness, the growth of MP across principal components, and the evaluation of minimum MP thresholds. Rather than guaranteeing an inherent improvement in rigorous OCC performance, this variable selection strategy aims to balance classification accuracy with enhanced chemical interpretability of the spectral data. It represents an exploratory attempt to refine SIMCA models by prioritizing variables that consistently support class definition while potentially reducing dimensionality and improving the interpretability of selected features in chemometric analyses.

2.1. Modeling power calculation

Following the construction of the SIMCA model for the target class, the variable selection algorithm evaluates Modeling Power (MP) as a central metric for identifying relevant variables. Assuming X as the matrix of training data and dimensioned with j variables, the MP quantifies how effectively the variance of each variable is explained by the principal components retained in the class model. As shown in Equation (1), for a given variable j, its residual standard deviation after projection onto the model is denoted as $s_{j,res}$, and its original standard deviation (i.e. before projection) is denoted as $s_{j,x}$ [11,13].

The MP is calculated as:

$$MP_j = 1 - \frac{s_{j,res}}{s_{j,x}} \tag{1}$$

This ratio expresses the remaining unexplained variation (residual noise) relative to the total variation of the variable. A value of MP close to 1 indicates that the variable is well captured by the model, whereas values near 0 suggest poor representation and potential irrelevance to the modeled class structure.

The residual deviation $s_{j,res}$ is computed from the residual matrix E, which contains the differences between the original centered data X and its projection onto the retained PCs. Each entry e_{ij} in E represents the residual for sample i and variable j. The term $s_{j,x}$ is derived from the

standard deviation of the original centered data matrix for variable j. The use of standard deviation offers practical and interpretative advantages: it preserves the units of the original variables, allows MP to be directly interpreted as a dimensionless proportion of total variability explained by the model, and ensures values are bounded between 0 and 1. Alternative formulations based on variance ratios would alter the numerical scale and interpretation of MP, potentially affecting its comparability across variables and studies.

2.2. Principal component selection strategies

In the context of rigorous OCC, the selection of the optimal number of principal components (A) in the present study is performed exclusively using samples from the target class, and strictly within the training dataset. This restriction — particularly the use of target-class data only — is fundamental to avoid introducing any information from non-target samples, whether real or simulated, thus preserving the unsupervised nature of the modeling process and preventing potential bias.

The two strategies employed for selecting the optimal number of components are both based on the analysis of the Coefficient of Variation (CV) of Modeling Power (MP) across components. For each principal component, the CV is computed as the ratio between the standard deviation and the mean of the MP values across all variables. One strategy, referred to as the Exponential Fit Criterion, identifies the point of stabilization in MP variability by analyzing the slope of an exponential decay fit to the CV curve. The alternative approach, the CV-Mean Criterion, selects the component whose CV is closest to the overall mean CV across all components — capturing a balance between explained variance and residual variability.

2.2.1. Exponential Fit Criterion

This approach estimates the optimal number A by modeling CV of the Modeling Power (MP) across components using an exponential decay function. The CV at each component is computed as the ratio between the standard deviation and the mean of MP values across all variables. As shown in Equation (2), optimal A is identified as the point ($|CV'_A(x)|$) where the absolute slope of the fitted curve falls below a predefined threshold (δ_{PC}), indicating that further components provide diminishing contributions to MP variability.

$$\left|CV'_A(x)\right| < \delta_{PC} \tag{2}$$

2.2.2. CV-mean criterion

As an alternative, this method selects the component whose CV value is closest to the overall mean CV across all components, following a rationale similar to the Kaiser criterion [16]. Instead of relying on eigenvalues, as demonstrated in Equation (3), it focuses on variability in Modeling Power (MP), assuming that components near the average CV reflect a balance between information retention and model simplicity.

$$A = arg\ min_i |CV_A(i) - \overline{C}V_A| \tag{3}$$

Both strategies are computed and displayed in the user interface, allowing for a transparent comparison. By default, the algorithm recommends the lower of the two estimates as the optimal number of components. However, the analyst may override this suggestion to explore alternative model complexities. Future developments aim to further automate this process, enhancing reproducibility and minimizing manual intervention.

2.3. Variable filtering criteria

2.3.1. Compactness contribution

The compactness index (CI) reflects the variability of the samples' score distances within the principal component space, providing a measure of class homogeneity at each retained component. A lower CI indicates that the samples are more tightly clustered along that component, suggesting a more cohesive class structure. It is calculated as the standard deviation of the score distances (SD) across the training samples at each individual component, as shown in Equation (4).

$$CI_i = std(SD) \tag{4}$$

To assess how each variable's Modeling Power (MP) profile relates to class compactness across components—capturing monotonic relationships without assuming linearity or normality—the Spearman correlation coefficient (r_j) is computed between the vector of MP values for variable j and the vector of compactness indices across the retained components [17]. This is expressed mathematically in Equation (5):

$$r_j = 1 - \frac{6\sum_{a=1}^{A}\left[R\left(MP_{a,j}\right) - R(CI_a)\right]^2}{A\left(A^2-1\right)} \tag{5}$$

In this formulation, r_j is calculated as a function of the ranks of its $MP_{a,j}$ values and the corresponding CI_a at each principal component a, with A being the total number of components considered. The operator $R(\cdot)$ represents the rank transformation applied independently to each vector. The squared differences between the ranks quantify the degree of agreement between the variable's representativeness in the model and the class compactness. A strong positive or negative value of r_j indicates a consistent trend between the two, while values near zero suggest no monotonic relationship.

A threshold for variable selection is then defined in Equation (6), retaining variables whose correlation exceeds the threshold adjusted by a sensitivity factor, as indicated in Equation (7).

$$threshold = median(r) + (r_0 \times \sigma(r)) \tag{6}$$

$$r_j > threshold \tag{7}$$

2.3.2. Modeling power (MP) non-growth rate

The MP growth metric (MPg) quantifies the evolution of each variable's MP across the selected number of principal components (A). It is calculated for each variable j as the difference between the MP at component A and the MP at component 1, divided by A (Equation (8)). Since an increase in MP is generally expected as more components are retained, a threshold is defined to identify variables with disproportionately high growth, calculated as the median plus a sensitivity factor (ζ_{MP}) times the standard deviation of MPg (Equation (9)). Variables exceeding this threshold are excluded from selection, as they may represent features that only become relevant at later components, potentially reflecting noise or overfitting. To the best of our knowledge, this variable selection strategy based on MP growth has not been previously reported in the literature. No similar methodology was identified during the development of this work, suggesting that the approach is being proposed here for the first time.

$$MPg_j = \frac{MP_{A,j} - MP_{1,j}}{A} \tag{8}$$

$$threshold = median(MPg) + (\zeta_{MP} \times \sigma(MPg)) \tag{9}$$

To align this criterion with others and facilitate interpretation, MPg is transformed. First, its absolute value is normalized to a [0,1] scale (Equation (10)), then expressed as its complement, termed the MP non-growth rate, $MPng$ (Equation (11)). This transformation ensures that higher MP non-growth rate values correspond to variables with lower growth, enabling intuitive graphical comparison across selection criteria.

$$MPg_{norm,j} = \frac{\left|MPg_j\right| - \min\left(|MPg|\right)}{\max(|MPg|) - \min\left(|MPg|\right)} \tag{10}$$

$$MPng = 1 - MPg_{norm,j} \tag{11}$$

2.3.3. Minimum modeling power threshold

Modeling Power (MP) calculation was previously introduced in Equation (1). Following Wold and Sjöström [11], variables exhibiting a modeling power lower than 0.3 are considered to contribute poorly to the definition of the class space and may be excluded from the model without compromising interpretability or parsimony. The threshold can be adapted depending on the dataset characteristics, as suggested by Ferreira [13], who reinforced that a MP close to 1 indicates that the variable is well explained by the retained principal components, whereas values below 0.5 suggest a poor contribution to class modeling. This threshold serves as a basic filter, ensuring that only variables meeting a minimal modeling quality are retained for further analysis.

2.4. Summary of workflow

The workflow of the proposed variable selection process is summarized in Fig. 1. Initially, a SIMCA model is constructed using the full set of variables, and the Modeling Power (MP) is calculated for each variable across principal components. An optimal number of components (A) is then determined using criteria based on the coefficient of variation of MP. Subsequently, three filtering criteria are computed: (i) the correlation between MP and class compactness, (ii) the MP non-growth rate, and (iii) the minimum MP threshold. These criteria are combined to select variables that contribute both to variance explanation and to maintaining class compactness while avoiding variables that become relevant only at later components. A new SIMCA model is then constructed using the selected variables, and performance metrics are extracted to evaluate model quality. This workflow integrates statistical thresholds and interactive decisions to balance model interpretability and classification performance.

3. Material and methods

3.1. Case studies

Three experimental datasets were used in the present study to evaluate and compare classification models, each simulating a realistic one-class classification problem involving a single target class and several non-target samples. The first dataset contains UV–Vis absorption spectra (220–400 nm) of edible oils, previously acquired by Pontes [18]. Sunflower oil (SNF, n = 28) was here defined as the target class, while the non-target samples (n = 15) included soybean, corn, and canola oils. The second dataset consists of NIR reflectance spectra (1001–2500 nm) of tea samples, originally provided by Diniz and previously studied using both NIR and UV–Vis techniques [19,20]. Argentinean green tea (AGT, n = 20) was defined as the target class of the present study, and the non-target set (n = 24) comprised Brazilian green teas and black teas from Argentina and Sri Lanka.

The third dataset, obtained from the UCPH Chemometrics Data Sets platform [21] and previously used for olive oil authentication [22], includes 120 HPLC chromatograms acquired with Charged Aerosol Detection (CAD). The dataset consists of 71 olive oils (target class – OLO) and 49 non-olive oils (e.g., sunflower, corn, soybean, grapeseed, and blends). It is important to note that, unlike spectroscopic detectors such as DAD (Diode Array Detector), CAD provides a first-order multivariate chromatographic signal, represented by the elution time versus intensity, with no spectral dimension. Therefore, each sample is represented by a single vector of intensity values across the elution time, and no preprocessing related to spectral data is required. Each chromatogram covers approximately 45 min, comprising circa 4000 variables per sample. All the above case studies were selected due to the chemical similarity between classes and the spectroscopic or chromatographic complexity of the data, providing scenarios for testing one-class classification strategies.

3.2. Chemometric procedures and data analysis

Owing to the distinct spectral characteristics of each dataset, tailored preprocessing strategies were applied. For the UV–Vis data of vegetable oils, no preprocessing was performed due to the high signal quality and low baseline drift. The NIR spectra of teas underwent Standard Normal Variate (SNV) transformation and baseline correction to mitigate scatter and drift effects [23].

For the HPLC-CAD chromatographic dataset, the first 8 and the last 10 min were removed from each chromatogram to eliminate regions lacking relevant chemical information. A second-order polynomial baseline correction was then applied individually to each sample. The resulting chromatograms were mean-centered across variables prior to normalization by the Euclidean norm, in order to preserve the relative variance structure before scaling sample magnitudes. This approach was adapted and extended from previously proposed preprocessing routines [22], slightly shifting the authors' procedure, where the normalization preceded mean centering. The adaptation was intentional and aimed to

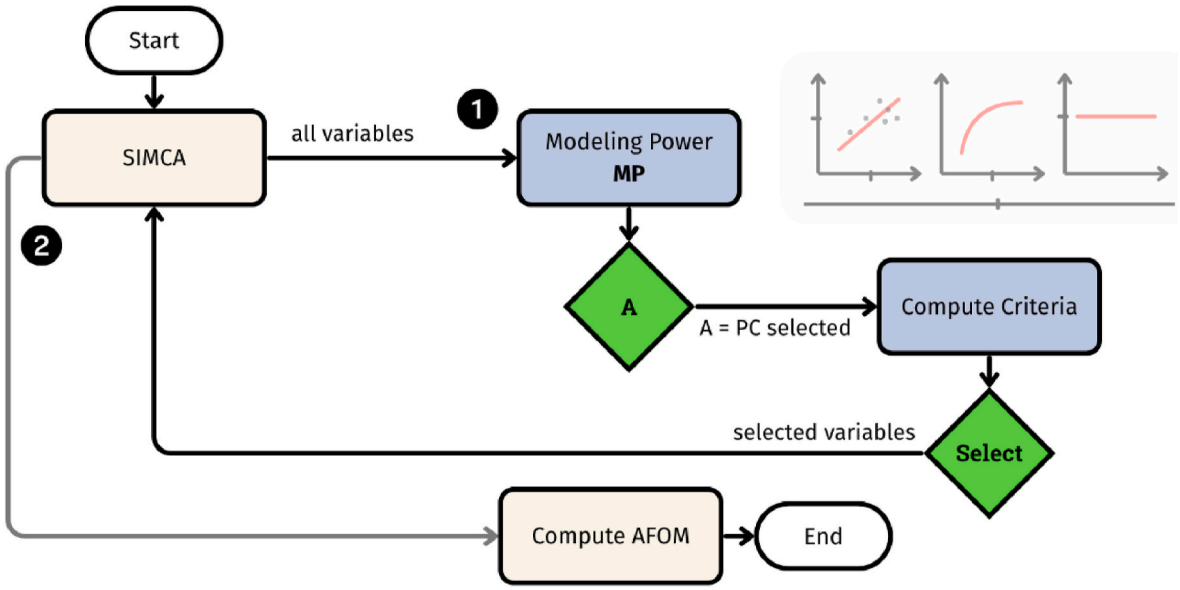


Fig. 1. Flowchart of the modeling power selection process for SIMCA.

enhance comparability across chromatograms while maintaining variance patterns relevant for subsequent multivariate analysis.

In all case studies, samples from the target class were split into training and test sets using the Kennard–Stone algorithm [24], adopting a 70:30 ratio. A pseudo-validation set, with dimensions matching the training set, was generated using the Procrustes Cross-Validation (PCV) strategy [25]. For this purpose, the number of principal components (PCs) was deliberately set to 20, ensuring that the PC-set would be valid for any subsequent modeling approach using a number of components A $\leq$ 20, as recommended by Kucheryavskiy et al. [25]. Moreover, in our implementation, the PCV procedure employed a leave-one-out (LOO) splitting scheme during its internal cross-validation step, exclusively for the purpose of generating the pseudo-validation set. This choice was motivated by the relatively small size of the datasets and aimed to maximize the variability captured in the PV set while ensuring its correspondence with the global model. The PV set was then used as an independent validation set for assessing model robustness. All datasets were mean-centered across variables (without scaling.

The performance of conventional SIMCA [11–14] was compared to that of the proposed variable-selection-based approach in both scenarios. In both cases, the number of PCs was defined based on the coefficient of variation of the modeling power (MP) index. The significance levels for $\alpha 1$ (0.05) and $\alpha 2$ (0.01) followed established guidelines [26], defining the acceptance threshold for the target class and the rejection rate of outliers, respectively.

Classification outcomes for the test set were evaluated using the confusion matrix, from which the following analytical figures of merit (AFOM) were calculated: Sensitivity (SEN), Specificity (SPC), Precision (PRC), F-score (FSR), Efficiency (EFF), and the Matthews Correlation Coefficient (MCC). These figures were selected to provide a comprehensive view of model performance, balancing detection accuracy, rejection ability, and overall classification reliability. As discussed by Ballabio et al. [27], no single index is sufficient to fully describe the classification ability of a model, since each one captures different aspects of performance and may behave differently depending on the classification scenario. Therefore, a combined set of complementary measures was adopted to ensure a more robust and unbiased assessment of classification performance.

4. Results and discussion

The performance of the proposed MPS-SIMCA algorithm was benchmarked against conventional SIMCA using three independent datasets — two spectroscopic and one chromatographic. Models were constructed exclusively from target-class training data, and variable selection was performed using the three MP-based criteria previously described. The number of principal components was defined based on internal evaluation of MP variability, and classification performance was assessed through acceptance plots and figures of merit derived from confusion matrices.

4.1. Modeling power and PC selection

4.1.1. UV–vis dataset – edible oils

Fig. 2a presents the heatmap of Modeling Power (MP) across PCs, overlaid with the average spectrum of the target class (sunflower oil). The heatmap reveals that the lowest-order component (PC1) concentrates high MP values in the region between 200 and 250 nm, indicating that these spectral variables play a central role in class modeling. PCs 2 through 4 extend the influence of informative variables up to approximately 325 nm, while preserving elevated MP values in the lower spectral range, suggesting consistency in variable importance throughout early components. As the number of components increases, MP values progressively rise across the entire spectral domain, with nearly all variables reaching MP values close to 1. This behavior suggests that beyond PC6, all variables begin to contribute uniformly to the model, which may indicate a tendency toward overfitting due to the nonselective modeling of variance from all spectral regions.

Fig. 2b displays the CV of variables' MP values across PCs (blue curve), with the CV values on the y-axis and the PC index on the x-axis. As a direct consequence of the MP behavior observed in Fig. 2a, lower-order PCs show greater heterogeneity in variable importance—some variables exhibit very high MP while others remain low—resulting in high CV values. From PC6 onwards, the CV curve stabilizes, reflecting the uniformity in MP distribution as more components are included. The red curve represents the fitted exponential function used to guide the selection of the optimal number of components. The green dot at PC5 corresponds to the exponential criterion, where the slope of the tangent to the curve drops below a defined threshold, indicating stabilization. The pink dot at PC4 marks the last PC before the CV falls below the mean of all computed values (CV-Mean criterion). The black "x" denotes the number of PCs ultimately selected by the analyst to construct the SIMCA and MPS-SIMCA models, based on both criteria and model interpretability.

4.1.2. NIR dataset – tea samples

Fig. 2c presents the heatmap of Modeling Power (MP) across PCs, overlaid with the mean NIR spectrum of the target class (AGT). In the four lowest-order PCs, a diverse distribution of MP values is observed across the spectral variables. Notably, PCs 1 and 2 highlight spectral regions with consistently high MP values—particularly between 1000 and 1100 nm, 1400–1800 nm, and 2200–2500 nm. Similar patterns are present in PCs 3 and 4, albeit with less pronounced contrast. From PC5 onward, the MP distribution begins to stabilize, with a consistent number of high-MP variables observed across subsequent components. Unlike the UV–Vis dataset, however, this does not clearly indicate overfitting, as intermediate MP values persist even in higher PCs, suggesting that spectral contribution remains somewhat selective despite the broader variance coverage.

Fig. 2d displays the variation in the coefficient of variation (CV) of MP across PCs, with the CV values shown in blue along the y-axis and the PC indices along the x-axis. As observed in Fig. 2c, the lowest-order PCs combine highly informative variables with others of limited relevance, resulting in elevated CV values for PCs 1 through 4. From PC6 onward, the curve shows stabilization in MP variability. The red line represents the fitted exponential curve that models the CV decay across components. In contrast to the UV–Vis dataset, the NIR data present a slower exponential decay, reflecting the more gradual convergence in variable importance during the early PCs. This slower decay influenced the exponential slope criterion, whose suggestion—marked by the green dot at PC1—would involve the risk of underfitting data. The pink dot at PC3 identifies the last component before the CV drops below the overall mean, representing a more robust alternative. Ultimately, the analyst opted for a parsimonious choice, marked by the black "x" at PC2, to balance variance modeling and avoid over- or underfitting in the resulting SIMCA and MPS-SIMCA models.

4.1.3. HPLC – CAD dataset – olive oil samples

Fig. 2e presents the heatmap of Modeling Power (MP) across successive PCs, overlaid with the mean HPLC-CAD chromatogram of the olive oil training set. As observed in the previous case studies, the MP values tend to increase uniformly across variables as more components are retained. In particular, PCs 1 to 5 reveal a heterogeneous distribution of MP values, with several variables exhibiting low modeling power in the early components and reaching high MP values only after PC5. This trend suggests that variable importance becomes progressively more uniform with increasing model complexity — a behavior that may signal a risk of overfitting beyond PC10 due to excessive inclusion of weakly informative variables.

Fig. 2f displays the curve of the coefficient of variation (CV) of the modeling power (MP) values across principal components (PCs) for the HPLC-CAD dataset of olive oils. As observed in the previous case studies,

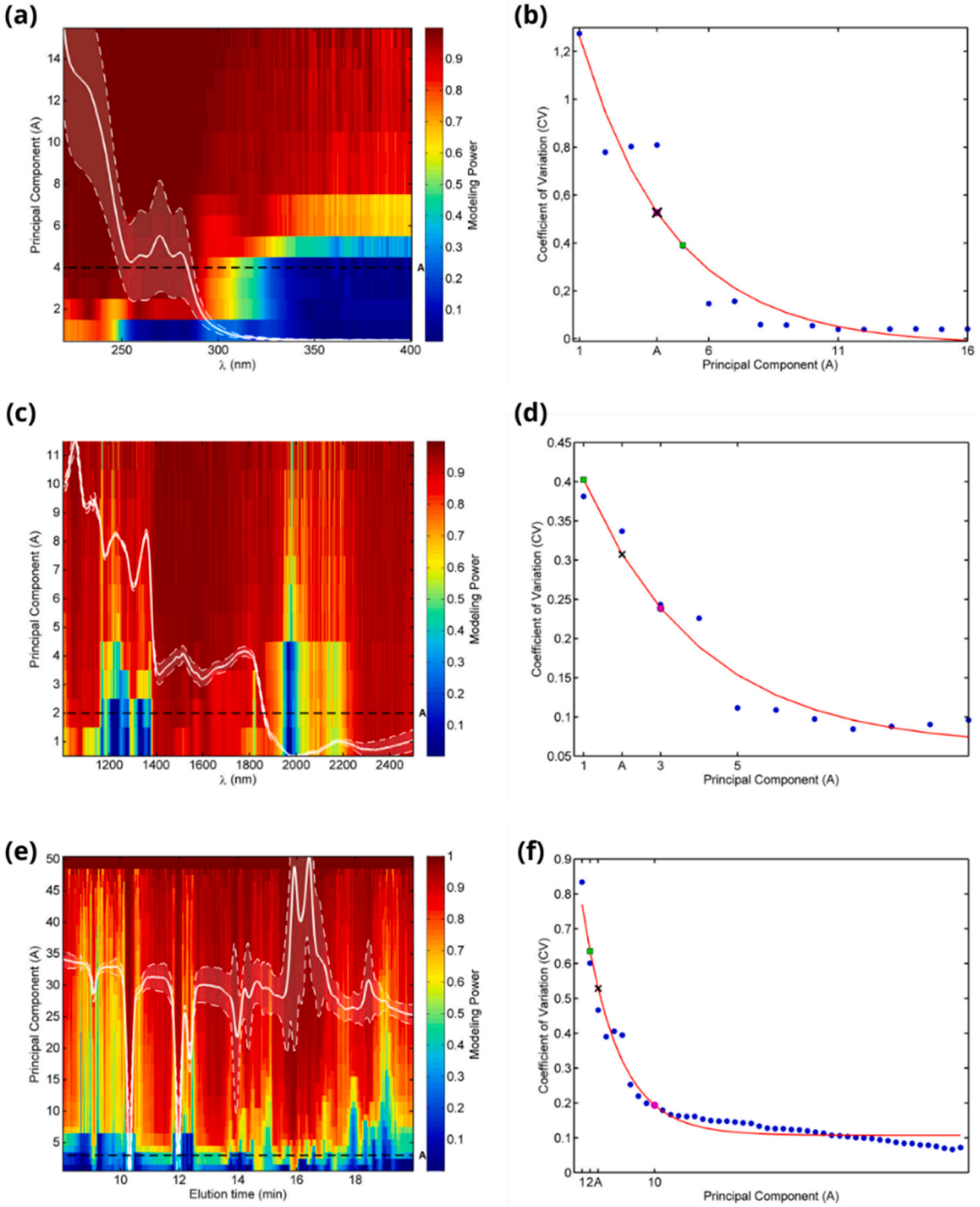


Fig. 2. Heatmaps of Modeling Power (MP) across principal components (PCs) (left column) and corresponding coefficient of variation (CV) profiles (right column) for each dataset: (a–b) Edible Oils (target: SNF); (c–d) Tea Samples (target: AGT); (e–f) Olive Oil (target: OLO). In the heatmaps (a, c, e), the MP of each variable at each PC is displayed, overlaid with the mean spectrum/chromatogram of the target class, with standard deviation of the data, and horizontal dash line indicating A, the number of PC chosen to work with. In the CV plots (b, d, f), blue dots represent the CV of variables' MP at each PC, the red line is the fitted exponential curve, the green square indicates the PC selected by the exponential slope criterion, the pink circle marks the last PC before the mean CV, and the black "x" denotes the final number of PCs selected by the analyst. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

the CV decreases as more components are retained, indicating a progressive homogenization in the importance of retention time variables—consistent with the uniform MP behavior discussed earlier.

According to the exponential slope criterion (green square), the optimal number of components is suggested to be PC2. In contrast, the CV-mean criterion (pink circle), which identifies the last component before the CV falls below the overall mean, suggests PC10. This discrepancy arises from the flattening of the CV curve beyond PC10, where several components exhibit similarly low CV values, thereby raising the average and delaying the mean-based cutoff point.

Considering both criteria and the goal of preserving model parsimony while avoiding overfitting, the final number of components selected by the analyst was PC3, as indicated by the black "x". This choice balances early stabilization in MP variability with the need to retain enough structure to capture the main TAG-related features described previously.

4.2. Visualization of selected variables

Application of the three previously defined criteria for variable selection based on Modeling Power (MP) revealed chemically meaningful regions in the UV–Vis spectra of sunflower oil. In Fig. 3, variables selected according to each criterion are shown in blue, while those simultaneously meeting all three criteria are highlighted in red. Based on the correlation between MP and the compactness index, 89 out of the 181 original variables (49.2 %) were selected, as shown in Fig. 3a. These variables are primarily located in the regions from 220 to 234 nm, 241–264 nm, 289–317 nm, and 352–367 nm. These wavelengths correspond to π→π* electronic transitions in conjugated systems and oxidized pigments [28], indicating a strong contribution to class compactness and structure. The MP non-growth rate criterion retained 133 variables (73.5 %), as presented in Fig. 3b, spanning from 220 to 248 nm and from 297 to 400 nm. The intermediate region from 249 to 296 nm was excluded, suggesting that these variables contribute primarily to higher PCs and may introduce instability or noise into the model. The fixed absolute threshold of MP, Fig. 3c, selected 90 variables (49.7 %), covering the range from 220 to 309 nm. This pattern aligns with the MP behavior observed in Fig. 2a, where variables beyond 310 nm exhibited MP values below 0.6 in PC4.

A total of 25 variables (13.8 %) satisfied all three criteria simultaneously. These wavelengths are concentrated in specific and chemically relevant regions: 220–234 nm, 241–247 nm, and 297–309 nm. The first two intervals encompass transitions typical of conjugated dienes and early-stage oxidation products [29], whereas the latter includes bands associated with natural antioxidants such as tocopherols, known markers of oxidative stability and botanical origin in vegetable oils [30]. As illustrated in Fig. 4a, the final set of selected variables reinforces the convergence of the criteria and the chemometric significance of the selection process, enhancing both interpretability and robustness of the class model.

When applied to the NIR dataset of tea samples, the variable selection method based on MP was equally effective in identifying

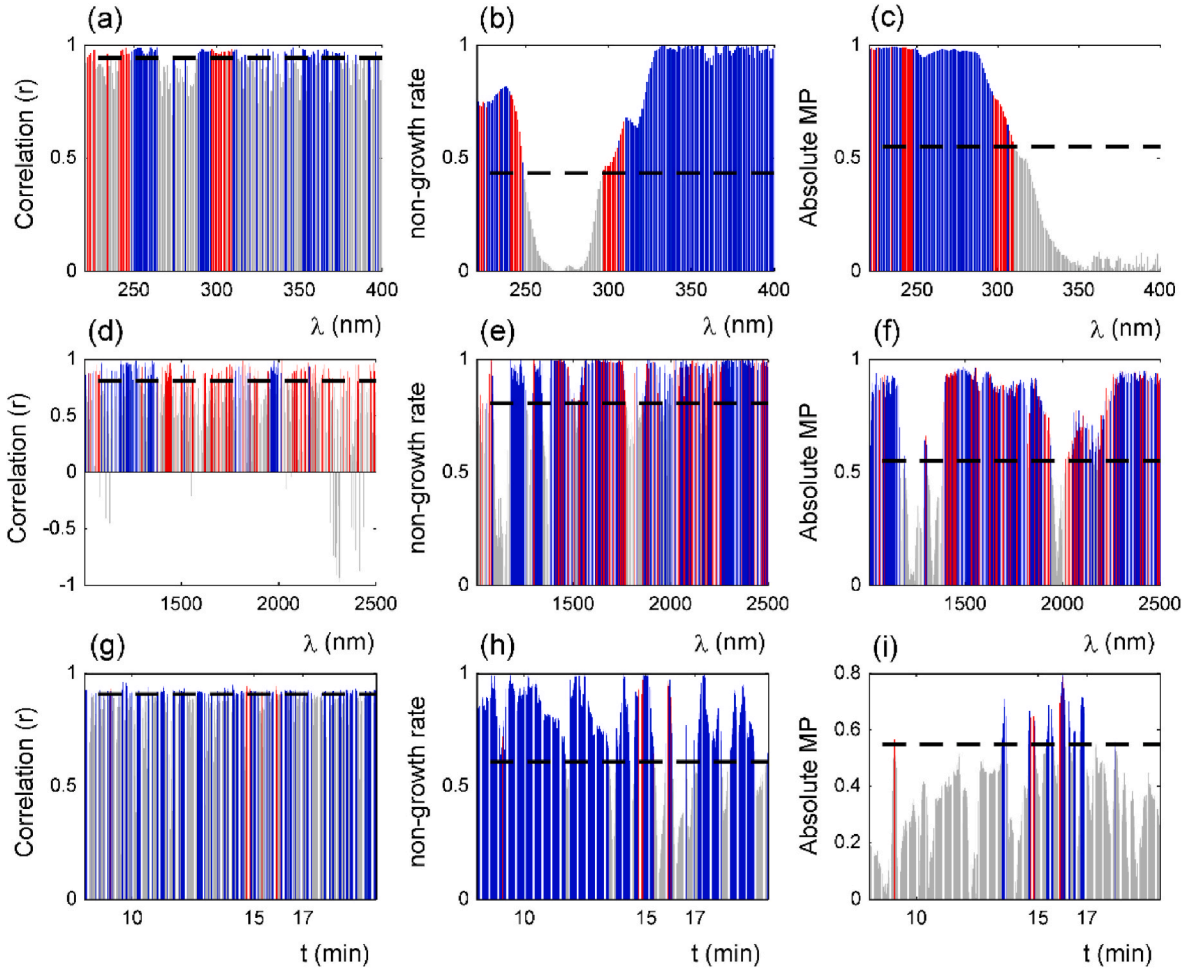


Fig. 3. MP-based variable selection. (a–c) UV–Vis (SNF); (d–f) NIR (AGT); (g–i) HPLC–CAD (OLO). Blue: variables selected by the criterium. Red: variables selected by all criteria. (a,d,g): MP–compactness correlation; (b,e,h): exclusion by MP growth (gray); (c,f,i): fixed MP threshold. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

informative regions. Using the correlation between MP and the compactness index of the target class as a first criterion, 733 variables (48.9 %) were selected for their contribution to class cohesion (Fig. 3d). The second criterion, based on the non-growth rate of MP across successive PCs, retained 1171 variables (78.1 %), reflecting a stability pattern in their modeling contribution (Fig. 3e). The third criterion led to the selection of 1206 variables (80.4 %), emphasizing descriptors with consistently strong modeling capacity (Fig. 3f), as also shown in Fig. 2c.

A total of 411 variables (27.4 %) satisfied all three criteria simultaneously (Fig. 4b), defining regions that are chemically meaningful for the characterization of green tea. These variables are located in spectral intervals rich in absorption features related to key bioactive constituents, including catechins, theanine, caffeine, and free amino acids. Moreover, the selected regions also capture structural signatures of the tea leaf matrix. The predominance of informative bands in the spectral ranges 1000–1100 nm, 1300–1500 nm, and 1900–2100 nm reinforces the relevance of both phenolic aromatic content and the hydric–structural profile for class modeling [31–33].

When applied to the HPLC-CAD dataset of olive oil, the MP-based variable selection strategy demonstrated a strong capacity to isolate chromatographic regions of high class-representativeness. As shown in Fig. 3g–i, 47.5 % of the original variables were retained based on the Spearman correlation between MP and the compactness index, highlighting regions that contribute to internal cohesion of the olive oil class. The MP non-growth criterion preserved 78.3 % of the variables, indicating a broad base of variables with stable modeling contribution across PCs. In contrast, the absolute MP threshold criterion was more selective, retaining 551 variables (13.8 %) with strong individual representation in the final PCA model. A total of 104 variables (2.6 %) simultaneously satisfied all three criteria.

These final selected variables (shown in Fig. 4c) are concentrated in well-defined chromatographic windows, notably around retention times of 9, 15–16, and 17 min. The region near 9 min likely captures early-eluting TAGs or minor polar constituents characteristic of specific olive oil cultivars or processing profiles. The 15–17 min region, on the other hand, aligns with the elution of major triacylglycerols—such as triolein (OOO), olein–palmitin (OOP), and olein–linolein (OOL)—which are key compositional markers of olive oil. These TAGs play a critical role in authenticity assessment, as their relative abundance and profiles distinguish olive oil from common adulterants like sunflower, soybean, and grapeseed oils [22].

4.3. SIMCA acceptance regions

Fig. 5 illustrates the acceptance regions defined by the proposed MPS-SIMCA model using SD and OD, for the training (Fig. 5a, d and 5g), validation (Fig. 5b, e and 5h), and test sets (Fig. 5c, f and 5i) applied to the edible oil (UV–Vis), tea (NIR), and olive oil (HPLC – CAD) datasets, respectively. In the training phase, green squares represent correctly accepted target-class samples (TP), yellow squares denote extreme samples still considered TP, and red squares indicate target-class samples wrongly rejected as outliers (FN). For the UV–Vis oil data (Fig. 5a), two of the 18 training samples were classified as extremes, while no outliers were identified. In contrast, the NIR tea dataset (Fig. 5d) showed one outlier among 13 training samples. Moreover, olive oil data (Fig. 5g), presented 42 TP and 8 FN, out of 50 training samples. These results suggest that the variable selection strategy retained compact and representative variables, preserving internal class cohesion during model construction.

During internal validation using pseudo-samples generated by PCV, the acceptance region plots (Fig. 5b, e and 5h) reveal the model's ability to generalize within the target class. Again, green squares indicate TP samples, while pink squares correspond to FN samples (target class samples incorrectly rejected). In the oil dataset (Fig. 5b), of the 18 pseudo-samples were misclassified as outliers, while 13 were correctly accepted. For the tea dataset (Fig. 5e), 4 out of 13 pseudo-samples were classified as outliers. In the olive oil dataset (Fig. 5h), among 50, 14 were classified as FN meanwhile 36 as TP. These results highlight the increased challenge of maintaining model sensitivity during internal validation, especially for datasets with limited and heterogeneous sample structures.

Fig. 5c, f and 5i present the model performance on the test sets, including both target and non-target samples. In these panels, green squares denote TP, pink squares denote FN, red squares represent correctly rejected non-target samples (TN), and blue squares indicate non-target samples incorrectly accepted by the model (FP). For the oil dataset (Fig. 5c), the model achieved 10 TP and 15 TN, with no FN or FP. For the tea dataset (Fig. 5f), out of 31 test samples, 6 were correctly accepted as TP and 1 was a FN among the 7 target samples; among the 24 non-target samples, 22 were correctly rejected (TN) and 2 were FP. In the olive oil dataset (Fig. 5i), all 21 target and 49 non-target samples were correctly classified (TP and TN, respectively). These findings are summarized into performance metrics (Sensitivity, Specificity, Precision, F-Score, and MCC), which are reported in Table 1, where MPS-SIMCA is compared to SIMCA with no variable selection.

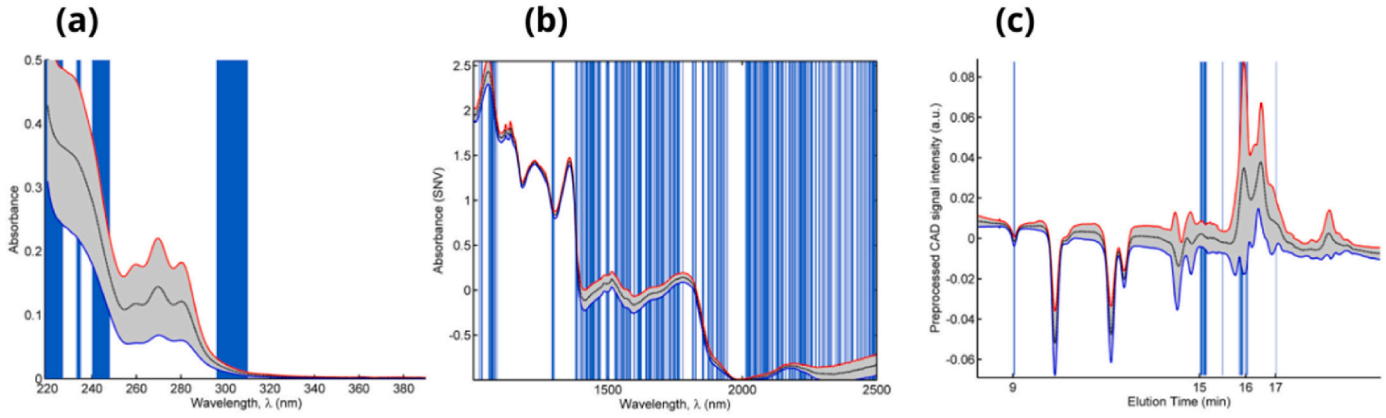


Fig. 4. Final set of selected variables (vertical blue lines) on the training set of the target class for each dataset: (a) UV–Vis spectrum of edible oils (target: SNF); (b) NIR spectrum of tea samples (target: AGT); (c) HPLC–CAD chromatogram of olive oil (target: OLO). The dashed line represents the mean spectrum/chromatogram, and the red and blue lines indicate the upper and lower standard deviation limits relative to the mean. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

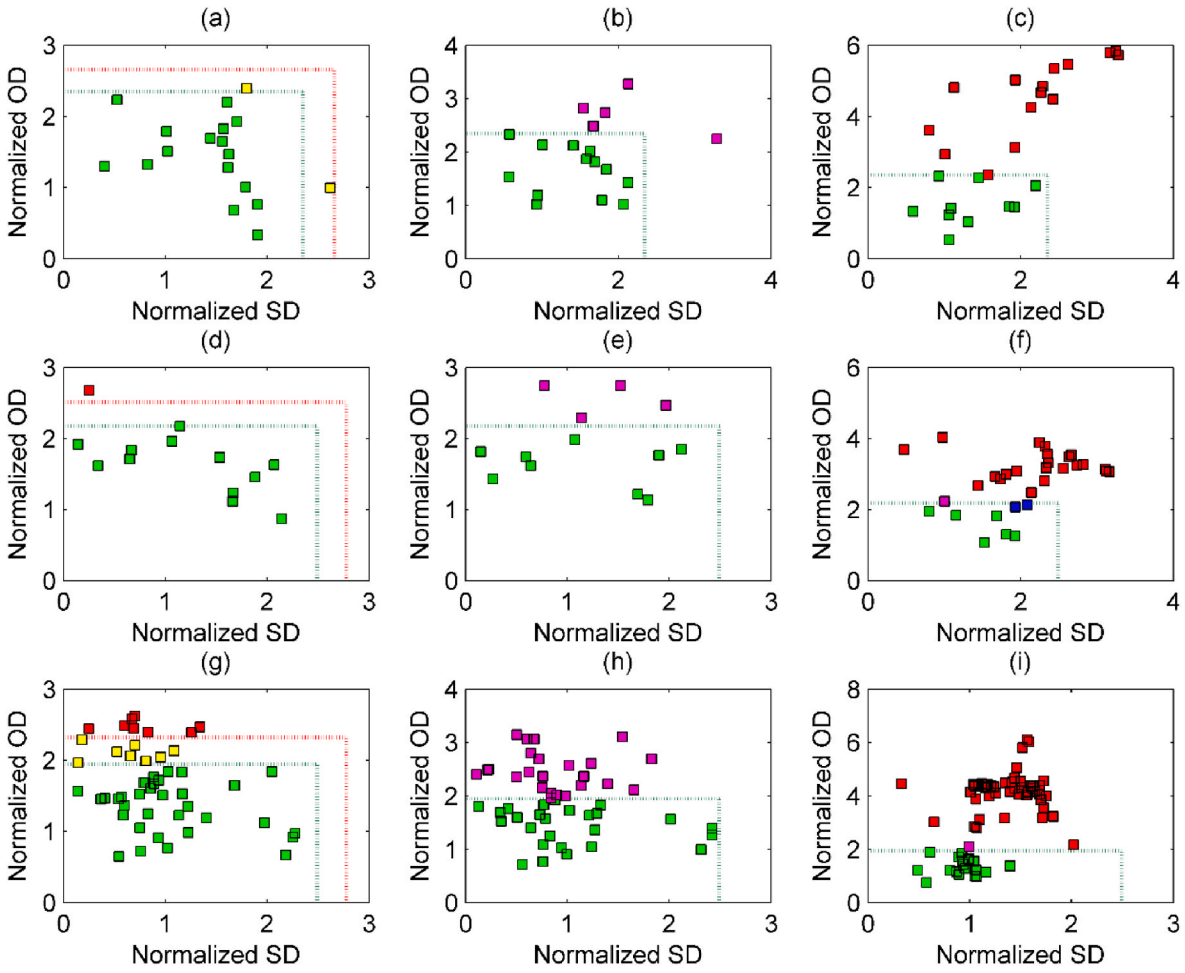


Fig. 5. Acceptance plots (Score Distance vs Orthogonal Distance) for SIMCA models with variable selection (MPS-SIMCA), applied to edible oils UV–Vis data (Fig. 5a–c), tea NIR data (Fig. 5d–f), and edible oils HPLS – CAD data (Fig. 5g–i). (a), (d), and (g): training sets; (b), (e), and (h): Procrustes-based validation sets; (c), (f) and (i): external test sets. Green squares: TP; yellow squares: extreme also TP; pink squares: FN; red squares: TN; blue squares: FP. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

4.4. Classification performance and statistical comparison

Table 1 summarizes the classification performance of SIMCA and MPS-SIMCA across studied datasets using multiple evaluation metrics. For the UV–Vis dataset (edible oils), training sensitivity was slightly reduced when variable selection was applied (SIMCA = 94 %, MPS-SIMCA = 89 %), and validation sensitivity also decreased slightly (SIMCA = 78 %, MPS-SIMCA = 72 %). However, in the test set, MPS-SIMCA achieved the maximum sensitivity score (100 %) compared to 90 % for the full model. Specificity was 100 % for both approaches. As a result, MPS-SIMCA reached 100 % efficiency, precision, and F-score, while SIMCA achieved 95 %, 100 %, and 95 %, respectively. The Matthews Correlation Coefficient (MCC) also improved from 0.92 (SIMCA) to 1.00 (MPS-SIMCA), indicating complete agreement between predicted and true classes.

For the NIR tea dataset, the training sensitivity improved from 85 % (SIMCA) to 92 % (MPS-SIMCA), while validation sensitivity remained unchanged at 85 % for both models. In the test set, both models achieved 100 % sensitivity. However, MPS-SIMCA improved specificity from 83 % to 87 %, efficiency from 91 % to 93 %, precision from 64 % to 70 %, and F-score from 78 % to 82 %. The MCC increased from 0.73 (SIMCA) to 0.82 (MPS-SIMCA), indicating better overall classification consistency.

For the chromatographic dataset based on HPLC-CAD, both models also achieved perfect classification on the test set, with 100 % sensitivity, specificity, precision, efficiency, F-score, and MCC. However,

Table 1
Classification performance metrics for training, validation, and test sets on SIMCA with all variables and for MPS-SIMCA (SIMCA with variable selection - VS).

Dataset	Method	A	Train	Validation	Test					
			SEN	SEN	SEN	SPC	EFF	PRC	FSC	MCC
Oil	All	4	94	78	90	100	95	100	95	0.92
	VS	4	89	72	100	100	100	100	100	1.00
Tea	All	2	85	85	100	83	91	64	78	0.73
	VS	2	92	85	100	87	93	70	82	0.78
Olive	All	3	94	90	100	100	100	100	100	1.00
	VS	3	84	72	100	100	100	100	100	1.00

SEN: Sensitivity (%); SPC: Specificity (%); EFF: Efficiency (%); PRC: Precision (%); FSC: F-Score (%); MCC: Matthews Correlation Coefficient.

differences were observed in the internal modeling steps. The conventional SIMCA model, using all variables, reached 94 % sensitivity in training and 90 % in validation, with a compactness index of 3.47. In contrast, MPS-SIMCA, constructed with only 104 selected variables and three principal components, exhibited reduced sensitivity in training (84 %) and validation (72 %), but yielded a more compact class model (compactness = 2.08). These results reflect a common trade-off in variable selection: improved interpretability and model compactness at the cost of minor reductions in training sensitivity, without compromising performance on external data.

To compare the classification performance of the full SIMCA model with its variable-selection-based version (MPS-SIMCA), statistical tests were applied to the contingency tables derived from the test sets of the first two case studies. For the UV–Vis dataset (vegetable oils), the test set included 10 target samples (sunflower) and 15 non-target samples. Fisher's exact test [34,35] indicated no statistically significant difference between the models (p-value = 1.00). SIMCA correctly classified 24 out of 25 samples, while MPS-SIMCA achieved perfect classification (25 out of 25). Similarly, McNemar's test [36], with continuity correction, revealed no significant sample-level disagreement (p-value = 1.00). The 2 × 2 McNemar table showed agreement on 24 samples (a = 24), with one case correctly reclassified by MPS-SIMCA (c = 1), and no other disagreements (b = 0, d = 0).

For the NIR tea dataset, which involved 7 target samples (AGT) and 24 non-target samples, the same statistical tests were applied. Fisher's exact test again yielded a p-value of 1.00, indicating no significant difference. SIMCA misclassified four samples, while MPS-SIMCA improved slightly with three misclassifications. McNemar's test confirmed this result (p-value = 1.00), with agreement on 27 samples (a = 27), one correction by MPS-SIMCA (c = 1), and no disagreement in the opposite direction (b = 0, d = 3).

For the HPLC-CAD dataset (olive oil authentication), statistical comparison was not necessary, as both models achieved perfect classification on the test set. Since SIMCA and MPS-SIMCA correctly classified all target and non-target samples, there were no classification errors to assess statistically. As such, no p-value could be computed, and both models can be considered equivalent in predictive performance for this dataset.

In summary, on the datasets tested in the present study, variable selection led to slightly improved or comparable results in all cases, with differences that were not statistically significant, differently from what was reported for other published strategies [15]. Nonetheless, these findings indicate that the proposed MPS-SIMCA strategy preserves the predictive capabilities of the original model, offering gains in interpretability and model compactness without compromising classification reliability.

It is worth noting that all datasets were mean-centered prior to modeling, which emphasizes variables with higher absolute variance, particularly in the first principal component. While such preprocessing is appropriate for spectroscopic and chromatographic analyses, it does not inherently guarantee optimal variable selection for classification. This potential limitation is addressed by the structure of the proposed method: variable selection is guided by the Modeling Power index calculated across all retained components, not just PC1. Furthermore, three additional criteria are applied to identify variables that consistently contribute to class compactness throughout the latent structure. This approach prevents overreliance on variance alone and supports the selection of chemically interpretable spectral or chromatographic regions, even when collinear or partially redundant. For example, in the NIR dataset, SNV preprocessing was applied before mean-centering to correct for scattering effects, demonstrating the method's flexibility to integrate standard spectral correction techniques.

Although the datasets employed in this study are based on two-dimensional structures, the proposed variable selection strategy may be further investigated to extend its applicability to multi-way data structures, exploring how the Modeling Power-based criterion could be adapted to such frameworks. Several studies have demonstrated the effectiveness of multi-way SIMCA approaches for class modeling in higher-order datasets, such as MCR-SIMCA [37], N-way SIMCA [38], and N-way DD-SIMCA [39]. This extension could involve integration with tensor decomposition techniques or structured unfolding strategies, potentially broadening the applicability of the proposed method to techniques such as fluorescence spectroscopy, which yields excitation–emission matrices for each analyzed sample.

Since Modeling Power is derived from the structure of the latent variable space and reflects the contribution of each variable to class compactness, analogous metrics could be computed from tensor decompositions (e.g., PARAFAC, Tucker models) or from mode-wise analyses following structured unfolding. A multi-way version of the Modeling Power index could, for instance, quantify variable relevance across multiple modes, guiding the selection of meaningful subspaces in multi-way SIMCA frameworks. These adaptations, however, require dedicated mathematical formulations and are left as future developments.

Finally, although the proposed variable selection method was evaluated on datasets characterized by high collinearity, its behavior in high-dimensional contexts with predominantly independent variables (e.g., omics data) remains an open question. In such cases—where redundancy is limited and variables act more independently—the Modeling Power index may distribute relevance more uniformly. This could reduce its discriminative ability and increase the risk of including variables only weakly associated with class compactness. Future studies could explore adaptations of the Modeling Power criterion, such as incorporating sparsity-inducing constraints or weighting schemes, to improve its performance in low-redundancy, high-dimensional settings.

5. Conclusion

The present study proposed and evaluated a Modeling Power-based variable selection algorithm tailored for rigorous one-class classification using SIMCA (MPS-SIMCA). By integrating three complementary criteria—correlation with class compactness, modeling power non-growth rate, and an absolute threshold—the algorithm enabled the identification of spectrally and chemically meaningful variables without requiring external class information. Initially applied to two distinct datasets involving UV–Vis and NIR spectroscopic data, MPS-SIMCA achieved comparable or improved classification performance relative to conventional SIMCA models that use the full spectral range.

The method was further tested using a third dataset composed of HPLC-CAD chromatographic profiles of olive oils. In this case, MPS-SIMCA preserved 100 % sensitivity and specificity on the external test set, while substantially reducing the number of variables and improving class compactness. Although training and validation sensitivities were modestly reduced compared to the full-variable model, the retention time regions selected reinforce the algorithm's ability to retain critical features for classification while enhancing model interpretability.

Overall, MPS-SIMCA demonstrates that variable selection grounded in internal model criteria can effectively support the development of compact, interpretable, and reliable one-class models—essential attributes for practical implementation in routine analytical and quality control workflows. Future work may extend this approach to other class-modeling methods and automate parameter selection to increase reproducibility and applicability across broader domains of analytical chemistry.

CRediT authorship contribution statement

Mateus Pires Schneider: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Cristina Malegori:** Writing – review & editing, Visualization, Supervision. **Paolo Oliveri:** Writing – review & editing, Supervision, Project administration, Methodology,

Conceptualization. **Adriano de Araújo Gomes:** Writing – review & editing, Supervision, Resources, Methodology, Conceptualization.

Permission for third party material

All figures presented in this manuscript were created by the first author using original routines developed by the authors in MATLAB R2014a. The software environment was accessed under an institutional license granted to UFRGS. No third-party material requiring permission was used in the preparation of this work.

Funding

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Finance Code 001, through the CAPES/PRINT program (Process n° 88887.877611/2023-00).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Financial support provided by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) is reported.

Acknowledgments

The authors would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Finance Code 001, under the International Cooperation Program CAPES/PRINT – Call n° 41/2017, for the doctoral mobility scholarship (Process n° 88887.877611/2023-00) awarded to Mateus P. Schneider, which enabled the research stay at the Università degli Studi di Genova.

Data availability

Data will be made available on request.

References

[1] A.A. Gomes, S.M. Azcarate, P.H.G.D. Diniz, D.D.S. Fernandes, G. Veras, Variable selection in the chemometric treatment of food data: a tutorial review, Food Chem. 370 (2022) 131072, https://doi.org/10.1016/j.foodchem.2021.131072.
[2] O.Y. Rodionova, P. Oliveri, C. Malegori, A.L. Pomerantsev, Chemometrics as an efficient tool for food authentication: golden pillars for building reliable models, Trends Food Sci. Technol. 147 (2024) 104429, https://doi.org/10.1016/j.tifs.2024.104429.
[3] A. Tugnolo, V. Giovenzana, C. Malegori, P. Oliveri, A. Casson, M. Curatitoli, R. Guidetti, R. Beghi, A reliable tool based on near-infrared spectroscopy for the monitoring of moisture content in roasted and ground coffee: a comparative study with thermogravimetric analysis, Food Control 130 (2021) 108312, https://doi.org/10.1016/j.foodcont.2021.108312.
[4] V. Belchior, B.G. Botelho, A.S. Franca, Comparison of spectroscopy-based methods and chemometrics to confirm classification of specialty coffees, Foods 11 (2022) 1655, https://doi.org/10.3390/foods11111655.
[5] E.I. George, The variable selection problem, J. Am. Stat. Assoc. 95 (2000) 1304–1308, https://doi.org/10.1080/01621459.2000.10474336.
[6] C.M. Andersen, R. Bro, Variable selection in regression—a tutorial, J. Chemom. 24 (2010) 728–737, https://doi.org/10.1002/cem.1360.
[7] P. Oliveri, Class-modelling in food analytical chemistry: development, sampling, optimisation and validation issues – a tutorial, Anal. Chim. Acta 982 (2017) 9–19, https://doi.org/10.1016/j.aca.2017.05.013.
[8] O.Y. Rodionova, A.V. Titova, A.L. Pomerantsev, Discriminant analysis is an inappropriate method of authentication, TrAC Trends Anal. Chem. 78 (2016) 17–22, https://doi.org/10.1016/j.trac.2016.01.010.
[9] O.Y. Rodionova, P. Oliveri, A.L. Pomerantsev, Rigorous and compliant approaches to one-class classification, Chemometr. Intell. Lab. Syst. 159 (2016) 89–96, https://doi.org/10.1016/j.chemolab.2016.10.002.
[10] M.P. Schneider, C. Malegori, A.A. Gomes, P. Oliveri, Enhancing one-class classification performance through variable selection: a review based on advanced literature search approaches, Chemometr. Intell. Lab. Syst. 265 (2025) 105491, https://doi.org/10.1016/j.chemolab.2025.105491.
[11] 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, ACS Symp. Ser., vol. 52, American Chemical Society, Washington, DC, 1977, pp. 243–282, https://doi.org/10.1021/bk-1977-0052.ch012.
[12] R. Vitale, M. Cocchi, A. Biancolillo, C. Ruckebusch, F. Marini, Class modelling by soft independent modelling of class analogy: why, when, how? A tutorial, Anal. Chim. Acta 1270 (2023) 341304, https://doi.org/10.1016/j.aca.2023.341304.
[13] M.M.C. Ferreira, *Quimiometria: Conceitos, Métodos E Aplicações* [Chemometrics: Concepts, Methods and Applications], first ed., UNICAMP, Campinas, 2015 https://doi.org/10.7476/9788526814714.
[14] A.L. Pomerantsev, O.Y. Rodionova, Soft independent modeling by class analogy, in: second ed., in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Compr. Chemom, vol. 3, Elsevier, Oxford, 2020, pp. 605–623, https://doi.org/10.1016/B978-0-12-409547-2.14588-3.
[15] A.L. Pomerantsev, S. Kucheryavskiy, O.Y. Rodionova, Variable selection for one class classifiers. Introduction of LOVE, Anal. Chim. Acta 1368 (2025) 344302, https://doi.org/10.1016/j.aca.2025.344302.
[16] H.F. Kaiser, The application of electronic computers to factor analysis, Educ. Psychol. Meas. 20 (1960) 141–151, https://doi.org/10.1177/001316446002000116.
[17] C. Spearman, The proof and measurement of association between two things, Am. J. Psychol. 15 (1904) 72–101, https://doi.org/10.2307/1412159.
[18] M.J.C. Pontes, *Algoritmo das Projeções Sucessivas para Seleção de Variáveis Espectrais em Problemas de Classificação* [Successive Projections Algorithm for Variable Selection in Spectral Classification Problems]. Universidade Federal da Paraíba, João Pessoa, 2009. Ph.D. thesis, https://www.ufpb.br/ppgq/contents/documentos/teses-e-dissertacoes/teses/2009/Tese_Marcio_J_C_Pontes.pdf.
[19] P.H.G.D. Diniz, A.A. Gomes, M.F. Pistonesi, B.S.F. Band, M.C.U. Araújo, Simultaneous classification of teas according to their varieties and geographical origins by using NIR spectroscopy and SPA-LDA, Food Anal. Methods 7 (2014) 1712–1718, https://doi.org/10.1007/s12161-014-9809-7.
[20] P.H.G.D. Diniz, M.F. Barbosa, K.D.T.M. Milanez, M.F. Pistonesi, M.C.U. Araújo, Using UV–Vis spectroscopy for simultaneous geographical and varietal classification of tea infusions simulating a home-made tea cup, Food Chem. 192 (2016) 374–379, https://doi.org/10.1016/j.foodchem.2015.07.022.
[21] UCPH Chemometrics Group, Olive oil data – UCPH chemometrics data sets. https://ucphchemometrics.com/olive/. (Accessed 23 June 2025).
[22] P. de la Mata-Espinosa, J.M. Bosque-Sendra, R. Bro, L. Cuadros-Rodríguez, Discriminating olive and non-olive oils using HPLC-CAD and chemometrics, Anal. Bioanal. Chem. 399 (2011) 2083–2092, https://doi.org/10.1007/s00216-010-4366-4.
[23] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989) 772–777, https://doi.org/10.1366/0003702894202201.
[24] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148, https://doi.org/10.1080/00401706.1969.10490666.
[25] S. Kucheryavskiy, O. Rodionova, A. Pomerantsev, Procrustes cross-validation of multivariate regression models, Anal. Chim. Acta 1255 (2023) 341096, https://doi.org/10.1016/j.aca.2023.341096.
[26] S. Kucheryavskiy, O. Rodionova, A. Pomerantsev, A comprehensive tutorial on data-driven SIMCA: theory and implementation in web, J. Chemom. 38 (2024) e3556, https://doi.org/10.1002/cem.3556.
[27] D. Ballabio, F. Grisoni, R. Todeschini, Multivariate comparison of classification performance measures, Chemometr. Intell. Lab. Syst. 174 (2018) 33–44, https://doi.org/10.1016/j.chemolab.2017.12.004.
[28] D.L. Pavia, G.M. Lampman, G.S. Kriz, Introduction to Spectroscopy, third ed., Brooks/Cole–Thomson Learning, Belmont, CA, 2001, pp. 363–387. Ch. 7.
[29] D. Grigoriadou, M.Z. Tsimidou, Quality control and storage studies of virgin olive oil: exploiting UV spectrophotometry potential, Eur. J. Lipid Sci. Technol. 108 (2006) 61–69, https://doi.org/10.1002/ejlt.200500204.
[30] R.P. Gonçalves, P.H. Março, P. Valderrama, Thermal edible oil evaluation by UV–Vis spectroscopy and chemometrics, Food Chem. 163 (2014) 83–86, https://doi.org/10.1016/j.foodchem.2014.04.109.
[31] D.C. Fechner, R.A. Martinez, M.J. Hidalgo, A.A. Gomes, R.G. Pellerano, H. C. Goicoechea, Geographic authentication of Argentinian teas by combining one-class models and discriminant methods for modeling near infrared spectra, Chemometr. Intell. Lab. Syst. 251 (2024) 105156, https://doi.org/10.1016/j.chemolab.2024.105156.
[32] Y. Liu, K. Pan, Z. Liu, Y. Dai, X. Duan, M. Wang, Q. Shen, Simultaneous determination of four catechins in black tea via NIR spectroscopy and feature wavelength selection: a novel approach, Sensors 24 (2024) 3362, https://doi.org/10.3390/s24113362.
[33] M. Zhang, T. Zhang, Y. Wang, X. Duan, L. Pu, Y. Zhang, Q. Li, Y. Liu, Accurate prediction of tea catechin content with near-infrared spectroscopy by deep learning based on channel and spatial attention mechanisms, Chemosensors 12 (2024) 184, https://doi.org/10.3390/chemosensors12090184.
[34] R.A. Fisher, On the interpretation of χ^2 from contingency tables, and the calculation of P, J. Roy. Stat. Soc. 85 (1922) 87–94, https://doi.org/10.2307/2340521.
[35] F. Yates, Contingency tables involving small numbers and the χ^2 test, J. R. Stat. Soc. Ser. B Methodol. 1 (1934) 217–235, https://doi.org/10.1093/oxfordjournals.jrsssb.a005989.
[36] Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentages, Psychometrika 12 (1947) 153–157, https://doi.org/10.1007/BF02295996.

[37] S.K. Karimvand, A. Pahlevan, S. Vali Zade, J.M. Jafari, H. Abdollahi, Multivariate curve resolution-soft independent modelling of class analogy (MCR-SIMCA), Anal. Chim. Acta 1291 (2024) 342205, https://doi.org/10.1016/j.aca.2024.342205.
[38] C. Durante, R. Bro, M. Cocchi, A classification tool for N-way array based on SIMCA methodology, Chemometr. Intell. Lab. Syst. 106 (2011) 73–85, https://doi.org/10.1016/j.chemolab.2010.11.002.
[39] A.P. Pagani, G. Camargo, G.A. Ibañez, A.C. Olivieri, A.L. Pomerantsev, O. Ye Rodionova, Data driven version of multi-way soft independent modeling of class analogy (N-way DD-SIMCA). Theory and application, Anal. Chem. 96 (2024) 4845–4853, https://doi.org/10.1021/acs.analchem.3c05311.