*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

# Optimizing the soft independent modeling of class analogy (SIMCA) using statistical prediction regions

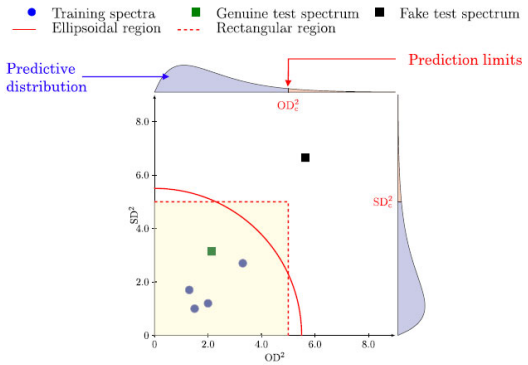
T. Hermane Avohou [a *], Pierre-Yves Sacré [a], Sabrina Hamla [a], Pierre Lebrun [b], Philippe Hubert [a], Eric Ziemons [a]

*a Vibra-Santé Hub, Laboratory of Pharmaceutical and Analytical Chemistry, Department of Pharmacy, CIRM, University of Liège, Avenue Hippocrate 15, 4000, Liège, Belgium*
*b PharmaLex Belgium, Rue Edouard Belin 5, 1435, Mont-St-Guibert, Belgium*

** E-mail address: thavohou@uliege.be*

**Keywords:**
One-class classification, Soft independent modeling of class analogy (SIMCA), Bayesian generalized linear models Prediction intervals, Near-infrared spectroscopy Raman spectroscopy

## Graphical abstract




## Abstract

The ultimate goal of a one-class classifier like the "rigorous" soft independent modeling of class analogy (SIMCA) is to predict with a certain confidence probability, the conformity of future objects with a given reference class. However, the SIMCA model, as currently implemented often suffers from an undercoverage problem, meaning that its observed sensitivity often falls far below the desired theoretical confidence probability, hence undermining its intended use as a predictive tool. To overcome the issue, the most reported strategy in the literature, involves incrementing the nominal confidence probability until the desired sensitivity is obtained in cross- validation. This article proposes a statistical prediction interval-based strategy as an alternative strategy to properly overcome this undercoverage issue. The strategy uses the concept of predictive distributions *sensu stricto* to construct statistical prediction regions for the metrics. Firstly, a procedure based on goodness-of-fit criteria is used to select the best-fitting family of probability models for each metric or its monotonic transformation, among several plausible candidate families of right-skewed probability distributions for positive random variables, including the gamma and the lognormal families. Secondly, assuming the best-fitting distribution, a generalized linear model is fitted to each metric data using the Bayesian method. This method enables to conveniently estimate uncertainties about the parameters of the selected

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

distribution. Propagating these uncertainties to the best-fitting probability model of the metric enables to derive its so-called posterior predictive distribution, which is then used to set its critical limit. Overall, the evaluation of the proposed approach on a diversity of real datasets shows that it yields unbiased and more accurate sensitivities than existing methods which are not based on predictive densities. It can even yield better specificities than the strategy that attempts to improve sensitivities of existing methods by "optimizing" the type 1 error, especially in low sample sizes' contexts.

# 1. Introduction

One-class classification (OCC) models are an important category of multivariate supervised classification models [1–6]. These models predict the conformity of future unknown objects to a given reference class of objects using classification rules defined with a training set of objects belonging exclusively to that class [1]. They are widely used in analytical chemistry to verify the identity or the conformity of food and drug samples to a reference set using vibrational spectroscopic techniques such as the near-infrared (NIR) and Raman spectroscopies [1,4,5]. The most widely used OCC method in chemometrics is undoubtedly the soft-independent modeling of class analogy (SIMCA) [1,5–14]. Briefly, the SIMCA model proceeds as follows. Firstly, the training data are projected onto an optimally reduced principal component (PC) subspace, known as the class model. Secondly, two fundamental classification metrics are computed and used to define the geometric location of each spectrum in an outlyingness metric space. The first metric termed the squared scores' distance ($SD^2$), is the squared (Mahalanobis) distance of the projection of the spectrum from the centroid of the class model. The second metric termed the squared orthogonal distance ($OD^2$), is the residual squared (Euclidean) distance from the spectrum to the class model. Thirdly, assuming a certain theoretical parametric probability distribution for each of the $SD^2$ and $OD^2$, critical values for these metrics are estimated and used to define the boundaries of an acceptance region with a desired nominal confidence or expected coverage probability $1 - \alpha$, where $\alpha \in ]0, 1[$ is the theoretical type 1 error or expected false rejection rate [6–14]. This acceptance region or rule is used to predict the conformity of new incoming spectra with the reference, i.e., to classify the spectra in the reference class. There exist several variants of the SIMCA model, differing in the probability distributions supporting the critical limits of the metrics and, the decision rules and the resulting type of acceptance regions. A comprehensive review of the most popular variants of the SIMCA model is provided by Refs. [9,10,12].

Despite the SIMCA model has been the most prominent OCC method in chemometrics, its current implementations still have some important drawbacks. These drawbacks are related to the estimation of the critical limits of the metrics and undermine their intended use as predictive tools. As first major drawback, each of the existing SIMCA variants *a priori* assumes a specific theoretical probability distribution of each of its metrics. However, this hypothesized probability model may substantially deviate from the empirical distributions of the metric data. As second major drawback, to derive the critical limits of the metrics, the hypothesized distributions are fitted to the observed metric data using the so-called plug-in estimation methods [15]. In other words, point estimates of the distributions' parameters are derived from the training data and plugged-in, without accounting for uncertainties about these estimates, which uncertainties are critical to unbiased predictions and reliable decision-makings [15–17]. Consequently, the resulting acceptance intervals for the metrics are not statistical prediction intervals *stricto sensu* [15,18]. Conceptually, such prediction intervals are the most appropriate for ensuring the targeted coverage probabilities, and consistent with the tasks of predictions of $m$ future values or their mean, or $m$ out of $l$ future values

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

of the metric variable ($l \geq m \geq 1$) [15–18]. Their construction naturally integrates uncertainties about the models' parameters in the definition of their limits, and they enable direct probability statements on the acceptability of single or several future objects [15–18]. For instance, using a prediction interval of confidence 100(1− $\alpha$)% for a single future value, one may state with an expected probability 1 − $\alpha$ ($\alpha \in ]0,1[$), that a single future metric value from the same spectral population will fall inside that interval, which statement is not possible with decision intervals derived from plug-in estimators of the distributions of the metrics [15,18].

Because of the two major drawbacks described above, the current implementations of the SIMCA model often suffer from undercoverage [7,12], meaning that their effective sensitivities may fall short of the desired or nominal 100(1− $\alpha$)% coverage of their acceptance regions, especially in low sample size contexts implying low degrees of freedom and high uncertainties. This issue of undercoverage of the SIMCA methods is well-known in the so-called "rigorous" strategy where the observed sensitivities must be the closest possible to the *a priori* fixed desired nominal confidence probability of 100(1− $\alpha$)% [1,7,11,12]. To overcome the issue, the most common strategy reported in the literature herein termed the ''$\alpha$-optimization'' strategy, involves increasing the nominal confidence probability 1 − $\alpha$ by a positive value $\delta < 1$, so that an approximate proportion 100(1− $\alpha$)% of an internal validation set is accepted [13]. However, this "$\alpha$-optimization" method inevitably results in a discrepancy $\delta$ between the theoretical type 1 error, i.e., $\alpha - \delta$ and the effective false rejection rate of the method, i.e., $\alpha$. Because of this discrepancy, $\alpha$ loses its theoretical meaning of probability of false rejection.

This article proposes a prediction intervals-based approach to "rigorous" (not the "compliant") SIMCA as an alternative strategy to overcome the undercoverage issue of the current implementations and preserve the theoretical meaning of $\alpha$ as probability of false rejection. The approach is based on the concept of statistical prediction intervals for SIMCA metrics, which are derived from Bayesian generalized linear models of these metrics [15,18,19]. Briefly, the approach proceeds as follows. Firstly, a strategy based on goodness-of-fit criteria is used to select for each metric or its monotonic transformations, the best-fit among several candidate families of right-skewed probability models for positive random variables [20–23]. These include the lognormal distribution for each metric and the gamma distribution for each metric, its square-root or reciprocal [24,25]. Secondly, assuming the best-fitting probability model, an intercept-only generalized linear model is estimated with the metric data using the Bayesian method [15,18,19]. Contrary to estimation methods which provides point estimates of models' parameters like the moment method, the Bayesian method of estimation of a parametric probability model enables to conveniently estimate the uncertainties about its parameters given the data at hand. Propagating these uncertainties to the best-fitting model for each metric enables to derive its so-called posterior predictive distribution, which is then easily used to set its critical limit as a 100(1− $\alpha$)% quantile of that distribution [15,18,19]. Overall, the proposed method is more adaptive. It yields unbiased and more accurate sensitivities compared to current benchmark implementations of the SIMCA model. It can even produce better specificities than the "$\alpha$-optimization" strategy that attempts to improve effective sensitivities of classical methods by "optimizing" the type 1 error i.e., $\alpha$.

The rest of the article is organized as follows. Section 2 entitled "Method" presents an overview of existing SIMCA methods and describes the newly proposed prediction interval-based approach to SIMCA. Key concepts and principles for building, optimizing and evaluating "rigorous" SIMCA models are first summarized in Section 2.1. The principal component (PC) model and the fundamental metrics are presented in Section 2.2. A brief mathematical description of two benchmark SIMCA methods, their limitations and opportunities for improvement using the theory of statistical predictions are presented in Section 2.3. Then, our proposed approach using prediction intervals for SIMCA metrics is

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

introduced as an improvement of the classical implementations of the model (Section 2.4). This presentation includes the strategy for choosing the best-fitting family of probability distributions for the metrics (Sections 2.4.1-2.4.2), the Bayesian intercept-only generalized linear model to estimate their critical prediction limits and acceptance regions (Sections 2.4.3-2.4.4). Section 3 entitled "Experimental" describes the datasets and methods to evaluate the performances of the proposed approach. Section 4 entitled "Results" reports the results of the evaluation studies, and Section 5 entitled "Discussion", discusses the advantages, limitations, and possibilities of extensions of the newly proposed approach. Conclusions are presented in Section 6.

# 2. Method

Suppose that $N$ spectra are measured on independent samples of a reference product, by recording their intensities at a set of $K$ wavelengths. Let $\mathbf{x}_i$ ($i$ = 1, …, $N$) of dimension 1 × $K$ be the $i$ th measured, preprocessed and mean-centered spectrum, and $\mathbf{X}$ of dimension $N$× $K$ the matrix containing all mean-centered spectra. Furthermore, suppose that $\mathbf{z}$ of dimension 1 × $K$ is a new incoming spectrum whose compliance with the set $\mathbf{X}$ is tested.

## 2.1. key concepts and principles of "rigorous" SIMCA models' development and evaluation

Given the data matrix $\mathbf{X}$, any "rigorous" (not "compliant") SIMCA method proceeds as follows. Firstly, the type of PC-based classification metric variable denoted $\mathbf{u} = [q, h]$ is defined, where $q$ and $h$ are the $OD^2$ and $SD^2$, respectively. Secondly, each spectrum $\mathbf{x}_i$ is transformed into the metric datum denoted $\mathbf{u}_i = [q_i, h_i]$. Thirdly, a parametric probability model is selected and fitted to the metric data $\{\mathbf{u}_i\}_{i=1}^N$. The estimated model is used to define the critical limits denoted $\mathbf{u}_c = [q_c, h_c]$ and an acceptance area denoted $\mathcal{U}$ that is intended to contain the predicted metric vector denoted $\tilde{\mathbf{u}} = [\tilde{q}, \tilde{h}]$ for any future spectrum of the reference product with a desired prespecified probability 1 − $\alpha$, $\alpha \in ]0, 1]$ being the nominal type 1 error or false rejection rate.

During the step of evaluation of the model performances, if a new incoming test spectrum $\mathbf{z}$ is from the reference product, then the observed probability that its predicted metric $\mathbf{u}_z = [q_z, h_z]$ is inside $\mathcal{U}$, is termed the coverage of $\mathcal{U}$ or sensitivity of the method and is approximated by the true positive rate (TPR in percentage, %). Otherwise, if $\mathbf{z}$ is from a non-reference product, this probability is the complement of the specificity and is approximated by the false positive rate (FPR in percentage, %).

Two important points about the evaluation and comparison of the sensitivity and specificity of two or more models in the "rigorous" SIMCA context, must be stressed:

1. Firstly, regarding sensitivity, as stressed in Ref. [12], the nominal type 1 error $\alpha$ is fixed *a priori* and any unbiased and statistically correct method should ensure a close agreement between the TPR and 1− $\alpha$. Practically, the discrepancy between the TPR and 1− $\alpha$, known as the bias is evaluated as bias = E(TPR) − 100(1− $\alpha$) while the variance is evaluated as variance = $E[(TPR - E(TPR))^2]$. A measure of the total uncertainty or error about the method sensitivity is given by $\sqrt{\text{bias}^2 + \text{variance}}$ [15].

   Evaluating these performance criteria namely the bias and the variance, requires simulating several datasets from a known multivariate distribution of the spectra, generally the multinormal distribution. Alternatively, to avoid

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

relying on such a restrictive assumption, resampling techniques using real data such as Monte Carlo or bootstrap cross-validation techniques can be used.

2. Secondly, regarding specificity in the context of the "rigorous" OCC where all possible non-target classes are theoretically unlimited and cannot not be sampled, a global proxy to compare specificity that can

   be used independently of any non-target test set, is the area of the acceptance region by analogy with the statistical concepts of shortest confidence intervals and minimum volume confidence regions [26]. Of two models calibrated at the same sensitivities, the one with the smallest area of acceptance region should be preferred [26].

Throughout the article, the notation "|" read "given" is used for conditional probabilities; thus, "$u|\theta \sim$ Density($\theta$)" is read "conditional to parameters' vector $\theta$, random quantity $u$ is distributed as the 'Density'"; the notations "$p(\theta)$" and "$p(u|\theta)$" are read "the probability density of $\theta$" and "probability density of $u$ given $\theta$", respectively; the notation $\hat{u}$ refers to an estimate of the quantity $u$ by a classical method while $\tilde{u}$ refers to its prediction by the Bayesian method.

## 2.2. Principal component model and SIMCA metrics

### 2.2.1. Principal component model !

The first step of any SIMCA method involves projecting each spectrum $\mathbf{x}_i$ of the spectral matrix $\mathbf{X}$ onto a principal component (PC) space, i.e.,

$$\mathbf{x}_i = \mathbf{t}_i\mathbf{P}' + \mathbf{e}_i, \cdot \text{for} \cdot i = 1, \cdots, N \quad (1)$$

where $\mathbf{x}_i$ of dimension $1 \times K$ is the $i$ th row of $\mathbf{X}$; $\mathbf{P}$ is the $K \times R$ estimated eigenvectors' or PCs' matrix defining an optimized PC subspace or class model, $R \geq 1$ being the number of PCs which is chosen by a 10-fold cross-validation strategy as the maximum number of PCs yielding a sensitivity greater than $100(1 - \alpha)\%$; $\mathbf{e}_i = [e_{i1}, \cdots, e_{iK}]$ is the $1 \times K$ residuals' vector for $\mathbf{x}_i$; $\mathbf{t}_i = [t_{i1}, \cdots, t_{iR}]$ is the $1 \times R$ scores' vector for $\mathbf{x}_i$; the eigenvalue associated with the $r$ th PC ($r = 1 \cdots, R$) is denoted $\lambda_r \geq 0$. It is estimated as $\hat{\lambda}_r = (N-1)^{-1} \sum_{i=1}^{N} t_{ir}^2$. We denote by $\widehat{\mathbf{\Lambda}} = diag[\hat{\lambda}_1, \ldots, \hat{\lambda}_R]$ the diagonal matrix of the estimated eigenvalues. The model in (1) can be estimated with the *svd()* routine of R statistical software [27].

### 2.2.2. SIMCA metrics !

Based on the PC model in (1), SIMCA proceeds by computing for each spectrum two fundamental classification metrics namely the squared orthogonal distance ($OD^2$) denoted $q$ which measures the residual distance of the spectrum from the optimized PC subspace, and the squared scores' distance ($SD^2$) denoted $h$ which measures the distance of the projection of the spectrum from the centroid of the optimized PC subspace. The $OD^2$ or $q$ is generally computed as squared Euclidean distance [6,8–14], i.e.,

$$q_i = \mathbf{e}_i\mathbf{e}_i', \cdot \text{for} \cdot i = 1, \cdots, N \quad (2)$$

The $SD^2$ or $h$ is generally computed as squared Mahalanobis distance [6,8–14], i.e.,

$$h_i = t_i\widehat{\mathbf{\Lambda}}^{-1}t_i', for\ i = 1, \ldots, N \quad (3)$$

In Equations (2) and (3), $\mathbf{e}_i$, $\mathbf{t}_i$, and $\widehat{\mathbf{\Lambda}}$ are the residuals and scores' vectors, and the estimated eigenvalues' matrix defined in (1), respectively; $q_i$ and $h_i$ denote the $OD^2$ and the $SD^2$ for $\mathbf{x}_i$, respectively.

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

## 2.3. Overview of existing SIMCA methods

There exist several implementations of the "rigorous" SIMCA model differing in the distributional assumptions for the $OD^2$ and the $SD^2$, the way they are combined to define a decision rule and the resulting acceptance regions for new incoming spectra [9,12]. Two benchmark methods reflecting the most commonly used distributions and acceptance regions are considered in this work.

### 2.3.1. The Jackson-Mudholkar SIMCA (JM-SIMCA) method !

***2.3.1.1. Hotelling's T-square and Jackson-Mudholkar** (JM) approximations of the metric variables.* The first benchmark method is termed the Jackson-Mudholkar SIMCA (JM-SIMCA) [12,28] and is coded as Method PG01 in this work. It uses as metrics the $OD^2$ i.e., $q$ in (2) and the $SD^2$ i.e., $h$ in (3). Then, under mild assumptions of multivariate normality of $\mathbf{x}_i$ and thus $\mathbf{t}_i$ and $\mathbf{e}_i$, the distribution of $q$ is estimated by the Jackson-Mudholkar (JM) approximation of the distribution of quadratic forms of normal vectors (see Appendix A.1 and [28] for more details), while $h$ is assumed to follow a Hotelling's $T$-squared distribution, defined by

$$h_i \sim [R(N-1)/(N-R)]\cdot \text{Fisher}(R, N-R) \quad (4)$$

where $N$ and $R$ are the sample size and selected number of PCs defined in (1). These distributions are then used to estimate the critical limits denoted $\hat{q}_c$ and $\hat{h}_c$ for $q$ and $h$, respectively, each at a confidence probability of $100(1-\alpha)\%$ ($\alpha \in ]0,1[$) [12,28].

***2.3.1.2. Decision rule and acceptance area**.* The new spectrum $\mathbf{z}$ is accepted if its predicted metrics satisfy

$$[q_z/\hat{q}_c]^2 + [h_z/\hat{h}_c]^2 \leq 2 \quad (5)$$

resulting in an ellipsoidal acceptance region (segment of an ellipse). In Equation (5), $q_z$ and $h_z$ are the $OD^2$ and $SD^2$ for $\mathbf{z}$ predicted by (2) and (3), respectively; $\hat{q}$ and $\hat{h}$ denote the estimated critical limits for $q$ and $h$, respectively, each at the confidence probability $100(1-\alpha)\%$, with $\alpha \in ]0,1[$.

### 2.3.2. The data-driven SIMCA (DD-SIMCA) method !

***2.3.2.1. Gamma distributions of the metric variables**.* The second method is more recent than the JM-SIMCA in 2.3.1 [9–12]. It is a variant of the well-known data-driven SIMCA (DD-SIMCA) methods [9–12], with a rectangular acceptance region. It is herein coded as PG02.

Like the JM-SMCA in Section 2.3.1, the DD-SIMCA methods use as metrics the $OD^2$ i.e., $q$ in (2) and the $SD^2$ i.e., $h$ in (3) [9–12]. Both metrics are non-negative quadratic forms of multivariate vectors. A possible probability model proposed by Box [29] to approximate the distribution of such quadratic forms is a weighted chi-square distribution (see Theorem 3.1 of [29]), more commonly known as gamma distribution in statistics [24,25]. Assuming this weighted chi-square or gamma approximation and using the method of moments to determine the weights and degrees of freedom parameters, both metrics are scaled so that their scaled transformations are distributed as chi-square distributions (see Appendix A.2 for mathematical details) [9–12], written

$$v_q q_0^{-1} q_i | v_q, q_0 \sim Chisquare(v_q), For\ i = 1, \ldots, N \quad (6)$$

and

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

$$v_h h_0^{-1} h_i | v_h, h_0 \sim Chisquare(v_h), For\ i\ =\ 1, \ldots, N \quad (7)$$

In Equations (6) and (7), $q_i$ and $h_i$ are the OD² and SD² for $\mathbf{x}_i$ in (2) and (3), respectively; the scaling parameters $q_0$ and $h_0$ are the means of the random variables $q$ and $h$, respectively, while $v_q$ and $v_h$ are the degrees of freedom for the chi-square distributions [9–12]. Point estimations of these parameters are then obtained from the metric data using the moment estimators, as

$$\hat{q}_0 = N^{-1} \sum_{i=1}^{N} q_i \ \ and\ \hat{h}_0 = N^{-1} \sum_{i=1}^{N} h_i \quad (8)$$

and

$$\hat{v}_q = 2 \cdot \hat{q}_0^2 \ / \ s_q^2 \cdot \text{and} \cdot \hat{v}_h = 2 \cdot \hat{h}_0^2 \ / \ s_h^2 \quad (9)$$

where $\hat{q}_0$ and $\hat{h}_0$ are the point estimations of the means $q_0$ and $h_0$, and $s_q^2$ and $s_h^2$ are the estimated sample variance for $q$ and $h$, respectively (see Refs. [9–12] for details of these estimations). These point estimations are then plugged-in models (6)–(7) to obtain an estimate of each of these scaled chi-square distributions. In turn, the quantiles of the estimated distributions are used to set upper confidence limits denoted $\hat{q}_c$ and $\hat{h}_c$ for the random quantities $q$ and $h$, respectively, with $\alpha \in ]0,1[$.

Contrary to the Hotelling's *T*-squared and Jackson-Mudholkar approximations of the JM-SIMCA in Section 2.3.1, these scaled chi-square or gamma distributions are not tight to the multivariate normality assumptions of the scores or residuals' vectors, which assumptions are hardly verifiable in practice. Furthermore, their parameters are estimated from the data, contrary to the Fisher or scaled Fisher distributions whose parameters depend only the sample size and the number of selected PCs. Hence, they are intended to be more data-driven and flexible approximations of the distributions of $q$ and $h$ [9–12].

***2.3.2.2. Decision rule and acceptance region.*** Based on the critical limits $\hat{q}_c$ and $\hat{h}_c$ for the two metric variables $q$ and $h$, respectively, several admissible decision rules and acceptance regions may be defined, resulting in several variants of the method [9]. They include among others, the variants with the rectangular or triangular regions [9]. The variant with the rectangular rule is the simplest and is considered in the present work. It is herein termed Method PG02. It accepts a new spectrum **z**, if its metrics $q_z$ and $h_z$ predicted by (2) and (3), are simultaneously inside their acceptance intervals $]0, \hat{q}_c]$ and $]0, \hat{h}_c]$, respectively, each at the confidence probability $100\sqrt{1-\alpha}\%$ ($\alpha \in ]0,1[$) [9]. In other words, to be accepted, **z** must satisfy

$$q_z \leq \hat{q}_c \text{ and } h_z \leq \hat{h}_c \quad (10)$$

and the resulting rectangular region is intended to have the desired confidence probability of $100(1-\alpha)\%$, with $\alpha \in ]0,1]$ [9]. It is important to stress the rationale of the $100\sqrt{1-\alpha}\%$ confidence probability used to estimate each of the quantiles $\hat{q}_c$ and $\hat{h}_c$ Indeed, the rectangular acceptance region implicitly assumes the independence of $q$ and $h$ and its joint confidence probability is just the product of the individual confidence probabilities for $q$ and $h$. Hence, setting these individual confidence probabilities at $100\sqrt{1-\alpha}\%$ for $q$ and $h$ enables to obtain a $100(1-\alpha)\%$ confidence probability for the whole region.

Computationally, the JM-SIMCA (PG01) is implemented with the routine *simca()* of the *mdatools* package [30] of R statistical software [27]. The DD-SIMCA with the rectangular region (PG02) is implemented with the same routine with an extra adaptation code supplied by the developer, as this rule is not directly available in the package.

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

### 2.3.3. Limitations of current SIMCA versions and ways for improvement using prediction intervals

The above-described benchmark methods have two major limitations affecting their predictive performances. As first limitation, they rely on *a priori* fixed assumptions about the probability models of the metrics, which are approximation conjectures. Some of these approximations like the Hotelling's *T*-squared or JM approximations in the JM- SIMCA (Section 2.3.1) rely on a stringent and hardly verifiable Gaussian process assumption of the spectra. In fact, the empirical distributions of the metrics might deviate from these assumed approximations or be better approximated by other distributions. It is a well-known statistical fact that uncertainty (confidence or prediction) intervals derived from biased distributions might be wrong, yielding a potentially biased effective type 1 error (thus, sensitivity) or decreasing the power (i.e., specificity) of the testing procedures.

As second limitation, even if the hypothesized distributions were correct, the acceptance intervals for the metrics estimated by these classical methods rely on point estimates of the parameters of these distributions. These estimates are random quantities whose uncertainties are not accounted for in the acceptance limits. Such intervals are known as plug-in intervals in the statistical or machine learning literature [15]. They are not prediction intervals *sensu stricto* and are prone to undercoverage, meaning that they do not guarantee that the effective TPRs are enough close to, or above the nominal probability content of 100(1 − $\alpha$)% [15,18].

Contrary to these plug-in intervals, prediction intervals *sensu stricto* are more likely to guarantee the targeted coverage, provided that the assumed probability distributions of the metric data are correct [15,18]. These prediction intervals are constructed for each metric variable as the quantile of its so-called predictive distribution, which is defined as the probability distribution of the future metric values, accounting for model parameters' uncertainties [15–19].

The fundamental model to derive a predictive density of any metric variable is clearly defined in the predictive modeling literature, for example in Murphy [15, p. 121–127] and Clarke and Clarke [17, p. 5]. In a nutshell, if $u$ is a metric variable following a probability distribution denoted $p(u|\boldsymbol{\theta})$ with parameters' vector $\boldsymbol{\theta}$ (e.g., mean-variance or shape-rate), and $\mathbf{y}_u$ is the vector of observed metric data, then the fundamental prediction model for a single future value $\tilde{u}$ of $u$ given the observed metric data $\mathbf{y}_u$ is defined as

$$p(\tilde{u}|\mathbf{y}_u) = \int p(\tilde{u}|\boldsymbol{\theta})\boldsymbol{p}(\boldsymbol{\theta}|\mathbf{y}_u)d\boldsymbol{\theta} \quad (11)$$

Simply explained, Eq. (11) means that the predictive distribution of a single future metric value $\tilde{u}$ given the metric data at hand $\mathbf{y}_u$, denoted $p(\tilde{u}|\mathbf{y}_u)$, is obtained by propagating the uncertainty from the probability distribution of the parameters $\boldsymbol{\theta}$ given $\mathbf{y}_u$, denoted $\boldsymbol{p}(\boldsymbol{\theta}|\mathbf{y}_u)$, to the data generating model $p(u|\boldsymbol{\theta})$. The probability distribution of the parameters $\boldsymbol{\theta}$ given $\mathbf{y}_u$ i.e., $\boldsymbol{p}(\boldsymbol{\theta}|\mathbf{y}_u)$, is a critical element of this prediction model in (11). Unlike point estimates such as the moments-based estimates of the model parameters used in the classical methods in Sections 2.3.1 and 2.3.2, it informs about possible values of $\boldsymbol{\theta}$ and their plausibility given the data at hand $\mathbf{y}_u$. Ideally, it is estimated rigorously using the Bayesian method, in which case it is called the posterior distribution of $\boldsymbol{\theta}$ and $p(\tilde{u}|\mathbf{y}_u)$ is termed the posterior predictive distribution [15,18,19] (see Section 2.4.3 for more details). Alternatively, it may be approximated by compliant estimation methods such as bootstrapping techniques [15, 17]

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

## 2.4. Prediction interval-based SIMCA

Our proposed prediction interval-based approach to SIMCA is based on the fundamental prediction model in (11) [15–19]. It proceeds with the following steps. In a first step, the parametric probability or data generating model i.e., $\boldsymbol{p}(u|\boldsymbol{\theta})$ in (11) is chosen as follows. Instead of making *a priori* fixed distributional assumptions on the $OD^2$ and $SD^2$ like the benchmark methods of Sections 2.3.1 to 2.3.2, a goodness-of-fit test [20–23] is used to select among several commonly used candidate families of right-skewed distributions for positive random variables, the best-fit for each of *q* and *h* in (2)-(3), if needed after monotonic (i.e., square-root, reciprocal or logarithmic) transformations for a better fit. This is a reasonable and valid statistical practice, especially in predictive modeling when there is no sound theoretical evidence to support a specific distributional assumption of a random variable [20–22]. This is particularly the case for the PC-based metrics where the current distributional assumptions are mostly approximation conjectures [9–12]. Because of the flexibility it offers, the proposed approach has the advantage of selecting a satisfactorily-fitting probability model for each of *q* and *h* in (2)-(3). Moreover, each of the proposed candidate probability models can be flexibly fitted using the well-established generalized linear models' framework [19].

In a second step, once the best-fitting family of probability model i.e., $\boldsymbol{p}(u|\boldsymbol{\theta})$ is chosen, the model in (11) is estimated with the metric data $\mathbf{y}_u$ using the Bayesian method [15–19]. This enables to derive the so-called posterior predictive distribution of the metric as introduced in Section 2.3.3, which is in turn used to set a critical prediction limit for the future metric values and to construct acceptance regions. These steps are detailed in Sections 2.4.1 to 2.4.4.

### 2.4.1. Candidate families of right-skew probability distributions and variable transformations

The considered candidate families of positively skewed distributions and monotonic transformations for each metric variable i.e., $\boldsymbol{p}(u|\boldsymbol{\theta})$ in (11) include the gamma distribution for $u = q, h, \sqrt{q}, \sqrt{h}, q^{-1}$ or $h^{-1}$, and the lognormal distribution for *u* = *q* or *h*. These distributions can all be flexibly fitted using the well-established generalized linear models' framework [19]. The rationale for their choice is as follows.

***2.4.1.1. The gamma or weighted chi-square distribution.*** The gamma distribution with parameters $\boldsymbol{\theta} = [\gamma,\mu]$ where $\gamma > 0$ is the shape (i.e., the reciprocal of the squared coefficient of variation) parameter, and $\mu > 0$ is the mean parameter (i.e., a rate parameter $\gamma.\mu^{-1}$) [24,25], is depicted on Fig. 1. It is assumed for.

1. The squared distance-based metrics i.e., *u* = *q* or *h* in (2) and (3), or

2. Their square-roots i.e., $u = \sqrt{q}$ or $\sqrt{h}$, or

3. Their reciprocal i.e., $u = q^{-1}$ or $h^{-1}$.

As mentioned in Section 2.3.2, the gamma distribution was proposed by Box [29] to approximate the distributions of non-negative quadratic forms of multivariate vectors i.e., squared distances, and was used in the DD-SIMCA to model both *q* and *h* [9–12]. In Bayesian statistics, it is commonly used as conjugate prior probability model for the reciprocal of the variance also known as precision parameter [19,24]. The proposed square-root and reciprocal transformations of *q* and *h* are monotonic transformations intended to cope with possible deviations from the gamma. They are both power transformations known to attenuate the skewness of right-skewed distributions of positive random variables [24]. Specifically, assuming the gamma model for $u = \sqrt{q}$ or $\sqrt{h}$, is equivalent to fitting the gamma distribution to the distances (not squared i.e., the OD and the SD), while assuming the gamma model for $u = q^{-1}$ or $h^{-1}$ is equivalent to fitting the inverse-gamma model to *q* or *h* [24,25].

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

***2.4.1.2. The lognormal distribution***. The lognormal distribution with mean and variance parameters $\mu$ and $\sigma^2$ [24,25] is assumed for the squared distance-based metrics namely $q$ or $h$ in (2) and (3) i.e., $u$ = log($q$) or log($h$) are each assumed to follow a normal distribution of parameters $\boldsymbol{\theta} = [\mu, \sigma^2]$. This distribution resembles the gamma distribution [24,25] (Fig. 1). It is commonly used in Bayesian analysis as prior model for standard deviations, i.e., square-roots of variances [24,25].

## 2.4.2. Goodness-of-fit tests to identify the best-fitting distribution

To select among the lognormal and the gamma-like models for each metric, the so-called probability plot correlation coefficient (PPCC) test is used [20–22]. The rationale for choosing this test is that, it is conceptually easy to understand, yet as powerful as the well-known competing Anderson-Darling test [22]. Indeed, it combines two fundamentally simple concepts, on the one hand the probability plot to evaluate the agreement between the theoretical quantiles of a hypothesized distribution and the sample quantiles of the data, and on the other hand the correlation coefficient to quantify this agreement. Hence, it enables a comparison of the results by both graphical and numerical evaluations. Its superiority in discriminating between the lognormal and the gamma distributions has been verified in a Monte Carlo power study

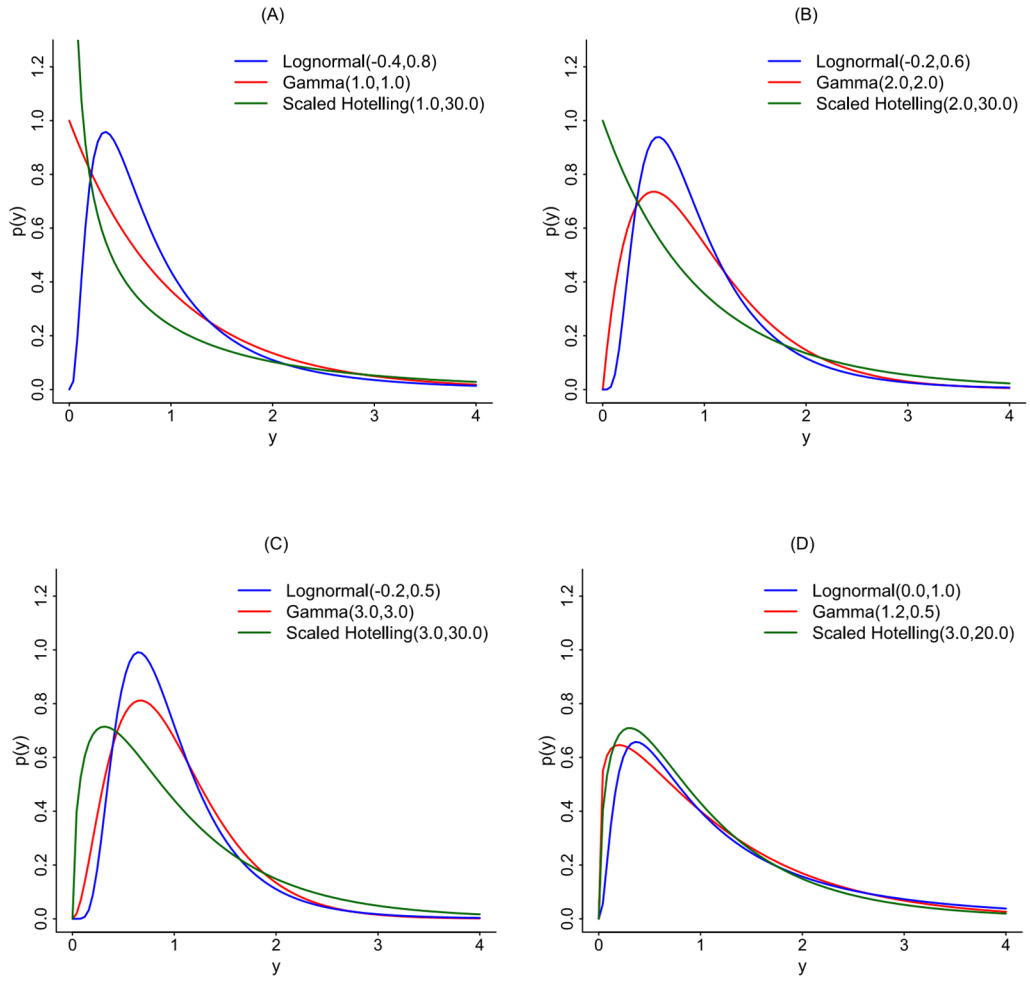



**Fig. 1.** Illustrations of the lognormal, gamma (scaled chi-square) and scaled Hotelling *T*-square (Fisher) distributions used to model the distance metrics. Notes: Lognormal($a$,$b$) = lognormal distribution with mean $a$ and standard deviation $b$ on the logscale; Gamma($a$, $b$) = gamma distribution with shape parameter $a$ and rate parameter $b$; Scaled Hotelling($a$,$b$) = Scaled Hotelling's *T*-square or Fisher distribution with degrees of freedom $a$ and $b$.

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

for various sample sizes (see results of the preliminary power study in Appendix B). A detailed mathematical description of the PPCC test statistic is available in Refs. [20–22,31]. It can be implemented with the routine *gofTest()* of the package *EnvStats* [31] of R software [27].

Based on this PPCC test, the strategy to select the best-fitting distribution among the lognormal and the gamma-like models is as follows. Each of the four candidate distributions is fitted to a random subsample of the metric data and a PPCC test is performed [31]. The null model is rejected if the *p*-values of the test is lower than 0.05. Because *p*-values are random variables [23], the test is repeated on 1000 resamples of the data, and the proportion of acceptances of the null model is used as metric to compare the competing distributions. The candidate distribution having the highest acceptance rate of the null hypothesis is chosen as the best fit, provided its median *p*-value is above 0.05.

### 2.4.3. Bayesian estimation of model parameters and prediction limits for the metrics !

From now on, let $u$ denotes the generic random variable to be modeled, i.e., $u$ is one of the squared distance-based metrics in (2) and (3), or its square-root, or its reciprocal. In other words.

1 $u$ denotes $q$ or $h$ in (2) and (3), if the gamma model for $q$ or $h$ is the best fit; or

2 $u$ denotes $\sqrt{q}$ or $\sqrt{h}$, if the gamma model for $\sqrt{q}$ or $\sqrt{h}$ is the best fit; or

3 $u$ denotes $q^{-1}$ or $h^{-1}$, if the gamma model for $q^{-1}$ or $h^{-1}$ is the best-fit; or

4 $u$ denotes $\log(q)$ or $\log(h)$, if the lognormal model for $q$ or $h$ is the best fit.

Let $u_i$ denotes the value of $u$ for the $i$ th spectrum $\mathbf{x}_i$, and $\mathbf{y}_u = [u_1, \ldots, u_N]$ the $1 \times N$ vector of computed values of the metric variable. Let $\mu$ and $\sigma^2$ be the mean and variance parameters of $u$, respectively, and $\gamma = \mu^2/\sigma^2$ the reciprocal of the squared coefficient of variation of $u$. A predicted future value of $u$ will be denoted $\tilde{u}$, and its critical limit at the confidence $100(1-\alpha_u)\%$ will be denoted $\tilde{u}_c$, where $\alpha_u = \alpha$ if the acceptance region is ellipsoidal and $\alpha_u = 1 - \sqrt{1-\alpha}$ if the acceptance region is rectangular.

***2.4.3.1. Bayesian estimation of model parameters and prediction limits in brief.*** Briefly, the Bayesian method to estimate the parameters' vector $\boldsymbol{\theta}$ of a probability model for the metric data $\mathbf{y}_u$, denoted $p(\mathbf{y}_u|\boldsymbol{\theta})$, uses the Bayes' theorem to derive the so-called posterior probability distribution of $\boldsymbol{\theta}$ given $\mathbf{y}_u$ denoted $p(\boldsymbol{\theta}|\mathbf{y}_u)$, as the product of $p(\mathbf{y}_u|\boldsymbol{\theta})$ and a prior probability distribution of $\boldsymbol{\theta}$ denoted $p(\boldsymbol{\theta})$ [19,32,33], i.e.,

$$p(\boldsymbol{\theta}|\mathbf{y}_u) = p(\mathbf{y}_u|\boldsymbol{\theta}) \times p(\boldsymbol{\theta}) \quad (12)$$

The prior $p(\boldsymbol{\theta})$ encodes the knowledge and uncertainties about $\boldsymbol{\theta}$ before $\mathbf{y}_u$ is observed, while the posterior $p(\boldsymbol{\theta}|\mathbf{y}_u)$ encodes the knowledge and uncertainties about $\boldsymbol{\theta}$ given $\mathbf{y}_u$.

It is important to stress again as in (11) that, $p(\boldsymbol{\theta}|\mathbf{y}_u)$ in (12) is not a point estimates of $\boldsymbol{\theta}$ like in the benchmark methods in Sections 2.3.1 to 2.3.2. Rather, it is a probability distribution of $\boldsymbol{\theta}$ that informs about its possible values and their plausibility given the data at hand $\mathbf{y}_u$. Propagating the parameters' uncertainties from $p(\boldsymbol{\theta}|\mathbf{y}_u)$ to the generating model of a single datum denoted $p(u|\boldsymbol{\theta})$ using the fundamental prediction model in (11), produces the so-called posterior predictive distribution of the metric variable $u$, which is defined as the distribution of any single future value $\tilde{u}$ of the metric variable given the data at hand $\mathbf{y}_u$, and is denoted $p(\tilde{u}|\mathbf{y}_u)$ as in (11). It must be stressed that the posterior predictive distribution depends only on the data at hand $\mathbf{y}_u$, contrary to the distributions derived

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

from plug-in estimators as in the DD-SIMCA method in (6)-(9), which all still depend on a random point estimate of their parameters.

Once the predictive distribution $p(\tilde{u}|\mathbf{y}_u)$ in (11) is estimated, its 100 $(1-\alpha_u)$%-quantile $\tilde{u}_c$ is computed as $\int_0^{\tilde{u}_c} p(\tilde{u}|\,\mathbf{y}_u)\mathrm{d}\tilde{u} = 1-\alpha_u$ and defines an upper prediction limit for the metric variable *u*, with $\alpha_u = \alpha$ if an ellipsoidal acceptance region is involved or $\alpha_u = 1-\sqrt{1-\alpha}$ if a rectangular acceptance region is involved ($\alpha \in ]0,1[$). It can be demonstrated that the quantity $1-\alpha_u$ is the expectation of the probability content of $]0,\tilde{u}_c]$. Hence, the $1-\alpha_u$ is directly interpreted as the average probability that a single future metric value falls inside that interval [18]. Such an interpretation is not possible with intervals constructed by plug-in methods [15] as in the JM-SIMCA and the DD-SIMCA methods in Sections 2.3.1 and 2.3.2.

Once the prediction limit $\tilde{u}_c$ is estimated for *u*, it is back-transformed to obtain the corresponding prediction limits on the original metric scale, if a transformation is applied to *q* or *h* prior to modeling. In other words, the prediction limits of *q* and *h* on the original metric scale denoted $\tilde{q}_c$ and $\tilde{h}_c$, respectively, are estimated as $\tilde{u}_c$, $\exp(\tilde{u}_c)$, $\tilde{u}_c^2$ or $\tilde{u}_c^{-1}$, depending on whether no transformation, a logarithmic, a square-root or a reciprocal transformations, respectively, are applied to *q* or *h* prior to modeling.

***2.4.3.2. Practical and computational aspects.*** Computationally, two strategies are used to estimate the predictive density $p(\tilde{u}|\mathbf{y}_u)$ from the metric data:

1 Firstly, when the lognormal distribution is the best fit for *q* or *h*, then *u* = log(*h*) or log(*q*), and the data generating model $\boldsymbol{p}(u|\boldsymbol{\theta})$ in (11)- (12) is a normal distribution with parameters $\boldsymbol{\theta} = [\mu, \sigma^2]$. The model in (12) can be estimated with a simple intercept-only linear model approach where the prior $p(\mu, \sigma^2)$ is taken to be the well-known non-informative Jeffreys' prior (see Appendix C for details of this linear model estimation) [15,18,19]. Then, it is well-established in statistical textbooks that the predictive distribution $p(\tilde{u}|\mathbf{y}_u)$ in (11) is a non-standardized Student's *t* with *N* − 1 degrees of freedom, written

$$(\hat{u}|\mathbf{y}_u) \sim \text{Student}_{(N-1)}\,[\hat{u}_0, (1+N^{-1})s_u^2 \qquad (13)$$

where $\hat{u}_0 = N^{-1}\sum_{i=1}^{N} u_1$ and $s_u^2 = (N-1)^{-1}\sum_{i=1}^{N}(u_1-\hat{u}_0)^2$ are the sample mean and variance of *u*, respectively [15,18,19]. The critical limit $\tilde{u}_c$ of *u* can simply be computed with the quantile routine *qt()* of R [27]. It is then back-transformed as exp ($\tilde{u}_c$) to obtain the critical limits $\tilde{q}_c$ or $\tilde{h}_c$ on the original metric scale as described in 2.4.3.1.

2 Secondly, when the gamma distribution is the best fit for one of $u = q, h, \sqrt{q}, \sqrt{h}, q^{-1}$ or $h^{-1}$, then the data generating model $p(u|\boldsymbol{\theta})$ in (11)-(12) is a gamma density with shape (i.e., reciprocal of the squared coefficient of variation) parameter $\gamma > 0$ and mean parameter $\mu$, i.e., $\boldsymbol{\theta} = [\gamma,\mu]$. Then, the model in (12) can be estimated with a simple intercept-only generalized linear model approach with the gamma family and the logarithmic link function, assuming a weakly informative prior $p(\gamma,\mu)$ on $\gamma$ and $\mu$ (see Appendix C for details on this generalized linear model estimation) [19]. Contrary to the (log) normal model in (13), there is no closed-form expression for the posterior predictive distribution of a future value $\tilde{u}$ of *u*, i.e., $p(\tilde{u}|\mathbf{y}_u)$ in (11). This predictive distribution is nonetheless validly sampled by a Markov Chain Monte Carlo (MCMC) sampler, using for example the *stan_glm()* and *posterior_predict()* functions of the *rstanarm* package [19,32,33] of R [27]. The critical limit $\tilde{u}_c$ of *u* can be estimated by the quantile of this MCMC sample using the *quantile()* routine of R [27]. It is then back-transformed to obtain

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

the critical limits $\tilde{q}_c$ or $\tilde{h}_c$ on the original metric scale as $\tilde{u}_c^2$ if $u = \sqrt{q}$ or $\sqrt{h}$, and $\tilde{u}_c^{-1}$ if $u = q^{-1}$ or $h^{-1}$. Otherwise, $\tilde{u}_c = \tilde{q}_c$ or $\tilde{h}_c$, if no transformation is applied i.e., $u = q$ or $h$.

The mathematical derivations of the posterior predictive distribution of the metrics under the lognormal and gamma-like distributions are detailed in Ref. [19] and summarized in Appendix C.

### 2.4.4. Decision rules and acceptance regions

Based on the estimated prediction or critical limits $\tilde{q}_c$ and $\tilde{h}_c$, for the metrics $q$ and $h$, two prediction interval-based SIMCA methods differing in the shape of their acceptance regions are proposed. They are the Bayesian counterparts of the benchmark SIMCA methods PG01 and PG02 (Sections 2.3.1-2.3.2). They are as follows:

1. The first method herein termed PI01 uses the ellipsoidal acceptance region defined by the same rule in (5). It is the Bayesian counterpart of the JM-SIMCA method PG01 (Section 2.3.1). Its prediction limits $\tilde{q}_c$ and $\tilde{h}_c$ are each defined at the confidence probability of 100 $(1 - \alpha)$%.
2. The second method herein termed PI02 uses the rectangular acceptance region defined by the same rule in (10). Its critical limits $\tilde{q}_c$ and $\tilde{h}_c$ are each defined at the confidence probability of $100\sqrt{1 - \alpha}$%. It is the Bayesian counterpart of the benchmark DD-SIMCA method PG02 using the rectangular acceptance region (Section 2.3.2).

# 3. Experimental

The statistical properties defined in Section 2.1 and the applicability of the two prediction interval-based SIMCA methods (Section 2.4.4) were evaluated and compared to those of their classical counterparts (Sections 2.3.1-2.3.2) using two validation strategies. These strategies are as follows.

1. Firstly, a Monte Carlo validation strategy was used to evaluate and compare the statistical properties of each prediction interval-based method to those of its classical counterpart with the same type of acceptance rule. These properties include for each method, its bias, its variance, its root mean squared error or total error of sensitivity, and the total area of its acceptance region as a proxy to compare specificity as defined in Section 2.1.
2. Secondly, an external validation strategy was used to demonstrate the applicability of each prediction interval-based method and its practical advantages over its classical counterpart with the same acceptance rule, in two real pharmaceutical quality control settings involving batch testing and release.

Five real spectral datasets involving a variety of applications (food and pharmaceutical), technologies (NIR and Raman), and sample sizes (low to moderate) were used for the evaluations. The first four datasets published in Refs. [2–5] were used for the Monte Carlo studies. They consist of three food NIR dataset [2–4] and one pharmaceutical Raman dataset [5] showing moderate to important overlapping of the classes and low to moderate sample sizes (Fig. 2A–D and 3A-3D). The last dataset was used for the external validation. It is a pharmaceutical Raman data (Fig. 2E) used for the two quality control case studies, each with a different reference product (Fig. 3E and F) [5]. A detailed description of the drug formulations in the pharmaceutical datasets is provided in Appendix D. The datasets and validation strategies are detailed in Sections 3.1 and 3.2, respectively.

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

## 3.1. Datasets, preprocessing and splitting

### 3.1.1. Datasets 1–4 for Monte Carlo studies

The four datasets used to evaluate the statistical properties of the proposed methods in the Monte Carlo studies are as follows.

Dataset 1: Wine NIR spectra [2] (Figs. 2A and 3A). The first dataset comprises 59 Fourier-transformed near-infrared (FT-NIR) spectra of two similar wine appellations, namely the *Barbera d'Alba* herein termed Wine 1 with 23 spectra and the *Dolcetto d'Alba* herein termed Wine 2 with 36 spectra. The dataset was chosen because of its low sample size and the important similarity (overlapping) of the two appellations' spectra (Figs. 2A and 3A). Hence, it reflects many realistic and constrained practical quality control situations where large calibration sample sizes are not affordable. Wine 1 was the reference product. The spectral range considered was 9000–4000 $cm^{-1}$ [2]. All spectra were smoothed with the Savitzky-Golay smoother (polynomial degree = 2, window size = 11, derivative order = 1) [34], and normalized (standard normal variate, SNV). Methods were calibrated with $N$ = 20 random spectra of Wine 1. They were tested with the remaining $n$ = 3 and 36 spectra of Wine 1 and Wine 2, respectively.

Dataset 2: Oil NIR spectra [3] (Figs. 2B and 3B). The second dataset includes 57 FT-NIR spectra of two similar oils, namely the *Chianti Classico* herein termed Oil 1 with 23 spectra and the *Maremma* herein termed Oil 2 with 34 spectra. The two oil classes show a moderate degree of overlap (Figs. 2B and 3B). Oil 1 was the reference product. Like Dataset 1, this dataset also reflects realistic and practical quality control settings constrained to low affordable calibration sample sizes. The spectral range considered was 8900–4400 $cm^{-1}$ [3]. All spectra were smoothed with the Savitzky-Golay smoother (polynomial degree = 2, window size = 11, derivative order = 1) [34], and SNV-normalized. Methods were calibrated with $N$ = 20 random spectra of Oil 1 and tested with the remaining $n$ = 3 and 34 spectra of Oil 1 and Oil 2, respectively.

Dataset 3: Olive NIR spectra [4] (Figs. 2C and 3C). The third dataset consists of 187 FT-NIR spectra of three olive cultivars, namely the *Taggiasca* herein termed Olive 1 with 83 spectra, the *Leccino* herein termed Olive 2 with 59 spectra, and the *Coquillo* herein termed Olive 3 with 45 spectra. Olive 1 and 3 are similar with important overlapping of their spectra (Figs. 2C and 3C). Olive 1 was the reference product. The dataset was chosen to reflect realistic quality control contexts with moderate calibration sample sizes. The spectral range considered was 9000–4200 $cm^{-1}$ [4]. All spectra were preprocessed with the Savitzky-Golay method (polynomial degree = 2, window size = 11, derivative order = 1) and SNV-normalized [34]. Methods were calibrated with $N$ = 62 random spectra of Olive 1, and tested with $n$ = 21, 59 and 45 spectra of Olive 1, 2 and 3, respectively.

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

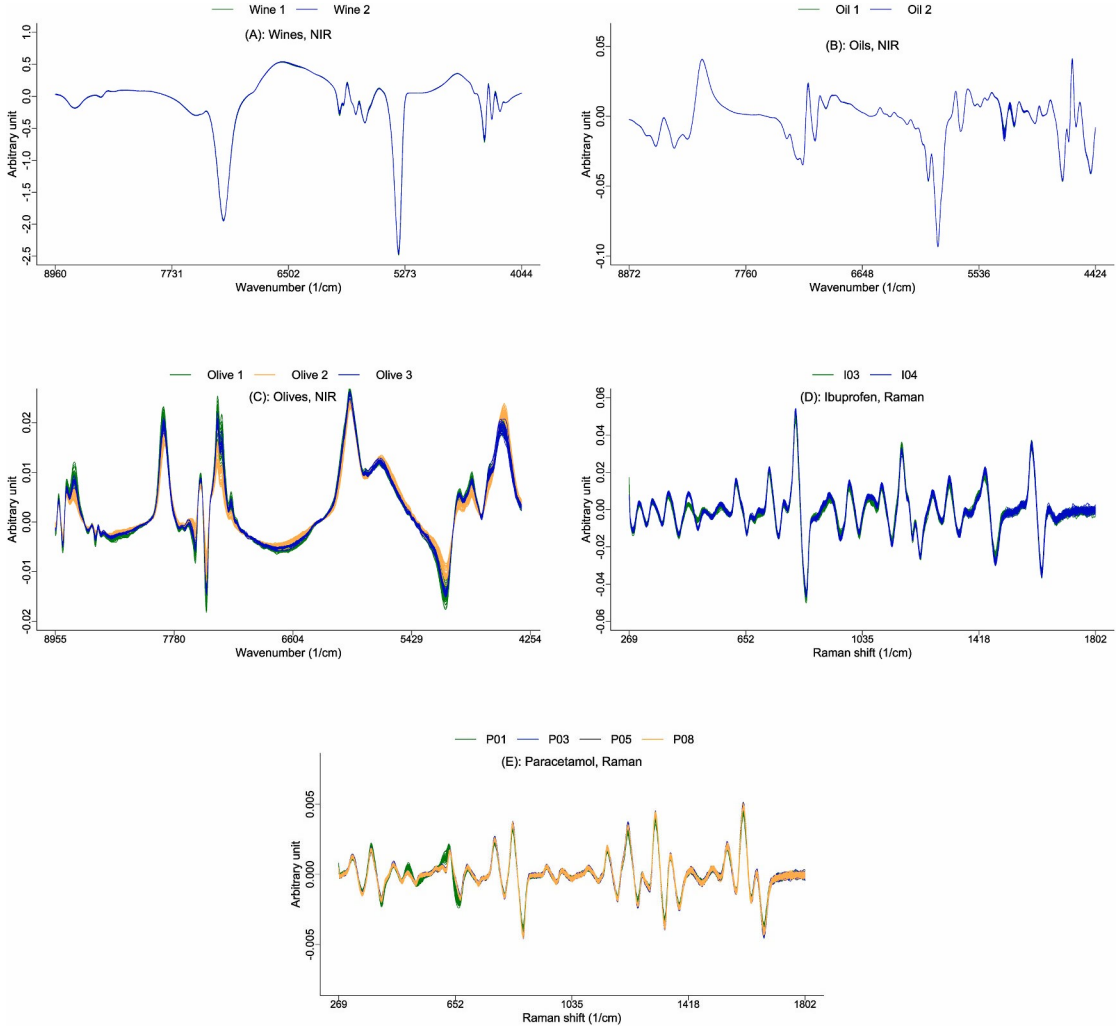



**Fig. 2.** Spectral datasets: (A) Dataset 1, wine NIR spectra; (B) Dataset 2, Oil NIR spectra; (C) Dataset 3, Olive NIR spectra; (D) Dataset 4, ibuprofen Raman spectra; (E) Dataset 5, paracetamol Raman spectra. Note: Paracetamol and ibuprofen formulations in datasets 4 and 5 are described in Appendix D.

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

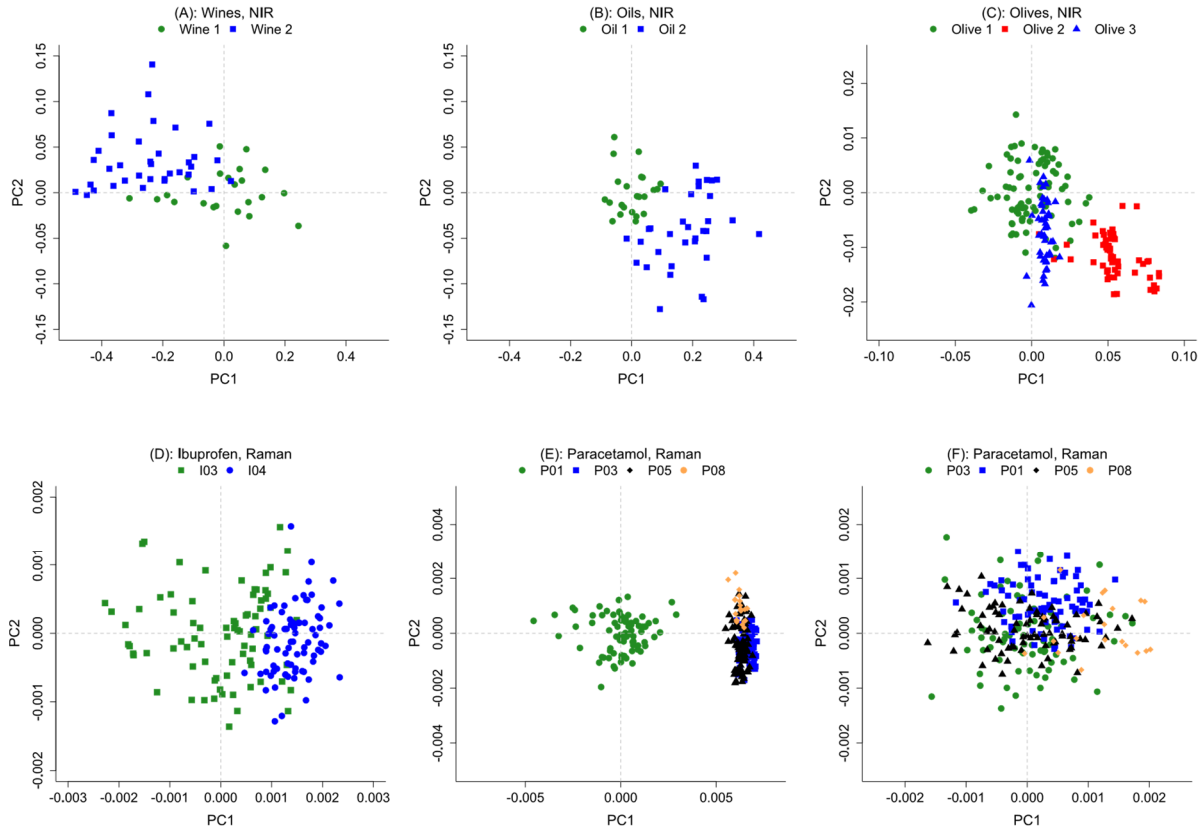



**Fig. 3.** Principal component scores' plots for: (A) wine NIR spectra in Dataset 1, (B) oil NIR spectra in Dataset 2, (C) olive NIR spectra in Dataset 3, (D) ibuprofen Raman spectra in Dataset 4, (E) paracetamol Raman spectra in Dataset 5, model for P01, (F) paracetamol Raman spectra in Dataset 5, model for P03.

Dataset 4: Ibuprofen Raman spectra (Figs. 2D and 3D) [5]. The fourth dataset comprises Raman spectra measured on tablets of two ibuprofen-based formulations. These two formulations differ mainly in the dosage of their active ingredients (Appendix D). They were coded as I03 (the reference, Ibuprofen 0.4g, 80 tablets, 4 batches) and I04 (the non-reference, Ibuprofen 0.6g, 80 tablets, 4 batches) [5]. Spectra were measured using a Raman device namely a Truscan RM® of ThermoFisher Inc. Covering a spectral range of 250–2875 $cm^{-1}$. They were optimally smoothed by the Savitzky-Golay smoother (polynomial degree = 2, window length = 35, derivative order = 1) [34], normalized to unit area (UA) and truncated to the relevant range of 284–1825 $cm^{-1}$ with 1215 variables (Figs. 2D and 3D). Models were calibrated with half of the spectra of I03, i.e., $N = 40$ random spectra randomly sampled from the pooled batches. They were tested with the remaining $n = 40$ and 80 spectra of I03 and I04, respectively.

### 3.1.2. Dataset 5 for external validation (paracetamol Raman spectra)

The fifth dataset was used to demonstrate the applicability of the proposed prediction interval-based methods in the two scenarios of batch testing and release [35], using an external validation strategy. The dataset comprises Raman spectra measured on tablets of four paracetamol-based formulations having a high active ingredient and very low excipients' contents (Fig. 2E, Appendix D). Two of the formulations were the target or reference products, each corresponding to a case study or OCC scenario (Fig. 3E and F). The formulations were coded as P01 (reference 1, Dafalgan 1g, 80 tablets, 4 batches), P03 (reference 2, Paracetamol EG 1g, 80 tablets, 4 batches), P05 (Paracetamol TEVA 1g, 40 tablets, 2 batches), and P08 (Panadol 0.5g, 40 tablets, 2 batches).

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

Particularly, formulations P03 and P05 are very similar and hardly discriminable with Raman spectroscopy as they are both generics with low excipient content. They were chosen to mimic scenarios of hardly detectable quality differences of products. Spectra were measured using the same Truscan RM® Raman device with the same processing as in Dataset 4. Models were calibrated with two batches of each reference i. e., $N$ = 40 spectra of either P01 or P03. The remaining two batches of each reference ($n$ = 40 spectra of either P01 or P03) and all batches of the non-reference formulations were used as test sets.

## 3.2. Evaluation method

The Monte Carlo and the external validation strategies for the evaluation of the statistical properties and demonstration of the applicability of the proposed prediction interval-based methods are detailed in Sections 3.2.1-3.2.2. For these evaluations, all models involving a gamma- like distribution as described in Section 2.4.3.2 were estimated with a valid MCMC sample of size $M$ = 50000. The number of PCs i.e., $R$, as mentioned in Section 2.2, was estimated in a 10-fold cross-validation optimization step, as the maximum number of PCs yielding a sensitivity greater than, or equal to 95%.

### 3.2.1. Monte Carlo validation

The Monte Carlo cross-validation or repeated random subsampling validation proceeded as follows with Datasets 1 to 4. At each iteration, the data splitting was done as described in Section 3.1.1 for each dataset, i.e., a training set including $N$ spectra was randomly sampled without replacement from the reference set and used to calibrate the methods, whose acceptance regions' areas were computed as global proxies to compare their specificities as defined in Section 2.1. The remaining $n$ spectra of either the reference product or the non-reference products were used as test sets to evaluate the TPRs and FPRs, respectively. A total of 500 iterations were used, resulting in 500 random estimates of TPR, FPR and area of the acceptance region for each method.

These random estimates of the figures of merit were summarized and compared as follows. For each of the two types of acceptance rules, a confusion matrix summarizing the means and standard deviations of the TPRs and FPRs of the classical and the prediction interval-based models was computed. For each confusion matrix or acceptance rule, the mean TPRs, FPRs and areas of the acceptance regions of the classical method (PG01 or PG02) and its prediction interval-based counterpart (PI01 or PI02) were compared using the three following strategies.

1. Firstly, the mean TPRs at the nominal 95%-confidence probability of methods PG01 vs PI01 or PG02 vs PI02 were compared. A Student's $t$-test after the logit-transformation of the TPRs was performed at a significance level of 0.05. Moreover, as introduced in Section 2.1, the bias of each method w.r.t. The nominal 95% was calculated as E(TPR)− 95%, and the total error was calculated as $\sqrt{bias^2 + variance}$ [15]. They were used as measures of effects' sizes to compare models. They both provide practical indications of what may be gained in terms of accuracy of the methods' sensitivity.
2. Secondly, when the classes overlap i.e., the models' FPRs ≠ 0, the mean FPRs of methods PG01 vs PI01 or PG02 vs PI02 were compared at three fixed mean TPR values namely: (1) the TPR of the classical method at the nominal 95%-confidence probability, (2) a TPR of 90%, (3) the TPR of the prediction interval-based method at the nominal 95%-confidence probability. At each of these reference TPR values, a Student's $t$-test on the logit-scale was performed at the significance level of 0.05 for each of the three comparisons. A curve of the FPR values as function

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

of the three fixed TPR values was built to visualize and facilitate the comparisons. Of two methods with equivalent TPR values, the one with the lowest FPR should be preferred.

3. Thirdly, the mean areas of the acceptance regions of methods PG01 vs PI01 or PG02 vs PI02 were compared at the mean TPR of 95%. A Student's *t*-test after the logarithm transformation of the areas was performed at the significance level of 0.05 for each comparison. Of two methods with equal TPR values, the one with the smallest area of the acceptance region should be preferred [26].

### 3.2.2. External validation

The external validation proceeded as follows with Dataset 5. As mentioned in Section 3.1.2, two batches of each reference formulation (i.e., $N$ = 40 spectra of either P01 or P03) were used to calibrate the models. The remaining two batches of the reference (i.e., $n$ = 40 spectra of either P01 or P03) and all batches of the non-reference formulations (i.e., $n$ = 40 or 80 spectra) were used as test sets to compute the TPR and FPRs for each model. Both the classical and the prediction interval-based models were first calibrated at a nominal type 1 error $\alpha$ = 0.05. Then, the classical methods were "optimized" using the "$\alpha$-optimization" strategy [13] i.e., by decrementing $\alpha$ in a 10-fold cross-validation step using the calibration batches, until a cross-validated TPR of 95% approximately is achieved.

Then, for each acceptance rule, the following comparisons were performed:

1. Firstly, a *z*-test of comparison of two proportions was used to compare the TPRs of the classical and prediction interval-based methods calibrated at the nominal confidence probability of $1-\alpha = 0.95$. The significance level of the test was 0.05 [35].

2. Secondly, the prediction interval-based and the "$\alpha$-optimized" classical methods were compared for their ability to correctly accept batches of the reference products based on their TPRs, and to reject batches of the non-reference products based on their FPRs. A decision interval to accept or reject batches with the targeted nominal TPR of 95%, denoted DI, was computed as

$$\text{DI} = 0.95 \pm 1.96 \cdot \sqrt{0.95 \cdot (1 - 0.95)/n} \quad (14)$$

where $n$ is the sample size of the test batches, i.e., DI = 88.2% − 100.0% or 90.2%− 99.8%, if $n$ = 40 or 80, respectively [35].

The total areas of the regions produced by the prediction interval- based and the "$\alpha$-optimized" classical methods were also computed as global measures to compare their specificities [26].

# 4. Results

Tables 1–6 report the confusion matrices for the five datasets. The columns entitled "Metrics and probability models" of these tables report for the classical methods the *a priori* fixed probability distributions, and for the prediction interval-based methods the best-fit among the lognormal and gamma-like distributions of the metrics. The proportions of acceptances of the null hypotheses of the goodness-of-fit tests supporting the choice of the best-fitting distributions for each of $q$ and $h$ are reported in Appendix E.

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

## 4.1. Performances in Monte Carlo validation

### 4.1.1. Dataset 1: wine NIR spectra

The prediction interval-based methods i.e., PI01 and PI02 produced almost unbiased TPRs (biases of − 1.0% to +0.3%, Table 1). Conversely, their benchmark counterparts i.e., PG01 and PG02 showed substantial undercoverages (biases of − 22.3% to − 16.7%) due to significantly lower TPR values (Student's *t*-test: *p*-values < 0.05 for each acceptance rule, Table 1). Moreover, the total errors of TPRs of these benchmark methods (PG01 and PG02, 29.4%–40.1%%) were substantially higher than those of PI01 and PI02 (11.5%–14.3%).

Optimizing the classical methods PG01 (ellipsoidal rule) and PG02 (rectangular rule) at average TPRs of 90% and 95% approximately by "optimizing" $\alpha$ using a grid search approach [13], yielded FPR values

**Table 1**
Averages and standard deviations (in brackets) of true positive (Wine 1, *Barbera d'Alba*) and false positive (Wine 2, *Dolcetto d'Alba*) rates of classification of two wine appellations by two classical and two prediction interval-based SIMCA methods with *a priori* fixed 95%-confidence probability.

| Method name | Estimation method | Metrics and probability models | | *R* | Wine 1, TPR (%) | Wine 2, FPR (%) |
|---|---|---|---|---|---|---|
| Ellipsoidal acceptance region | | | | | | |
| PG01 | Moment-JM | $T^2[h]$ | JM[$q$] | 2 | 78.3 (24.2) | 41.4 (5.6) |
| PI01 | Bayesian | Ga[$h$] | Ga[$q^{-1}$] | 2 | 95.3 (11.5) | 53.1 (4.2) |
| Rectangular acceptance region | | | | | | |
| PG02 | Moment | Ga[$h$] | Ga[$q$] | 2 | 72.7 (33.3) | 35.6 (6.1) |
| PI02 | Bayesian | Ga[$h$] | Ga[$q^{-1}$] | 2 | 94.0 (14.3) | 53.2 (3.9) |

*R* = optimized number of PCs; TPR (%) = true positive rate in percentage; FPR (%) = false positive rate in percentage. Distributions: $T^2[\cdot]$ = Hotelling's *T*-square; JM[$\cdot$] = Jackson-Mudholkar approximation; Ga[$\cdot$] = gamma; Log[$\cdot$] = lognormal. Metric variables: $q = OD^2$; $h = SD^2$.

**Table 2**
Averages and standard deviations (in brackets) of true positive (Oil 1, *Chianti Classico*) and false positive (Oil 2, *Maremma*) rates of classification of two oil appellations by two classical and two prediction interval-based SIMCA methods with *a priori* fixed 95%-confidence probability.

| Method name | Estimation method | Metrics and probability models | | *R* | Oil 1, TPR (%) | Oil 2, FPR (%) |
|---|---|---|---|---|---|---|
| Ellipsoidal acceptance region | | | | | | |
| PG01 | Moment-JM | $T^2[h]$ | JM[$q$] | 2 | 87.9 (21.3) | 3.7 (1.4) |
| PI01 | Bayesian | Ga[$h$] | Ga[$\sqrt{q}$] | 2 | 95.1 (14.7) | 6.0 (1.3) |
| Rectangular acceptance region | | | | | | |
| PG02 | Moment | Ga[$h$] | Ga[$q$] | 2 | 78.5 (24.5) | 2.9 (0.9) |
| PI02 | Bayesian | Ga[$h$] | Ga[$\sqrt{q}$] | 2 | 95.8 (14.9) | 6.6 (1.4) |

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

Notes: *R* = optimized number of PCs; TPR (%) = true positive rate in percentage; FPR (%) = false positive rate in percentage.
Distributions: $T^2[\cdot]$ = Hotelling's *T*-square; JM[·] = Jackson-Mudholkar approximation; Ga[·] = gamma; Log[·] = lognormal. Metric variables: $q = OD^2$; $h = SD^2$.

**Table 3**
Averages and standard deviations (in brackets) of true positive (Olive 1, *Taggiasca*) and false positive (Olive 2 & 3) rates of classification of three olive cultivars by two classical and two prediction interval-based SIMCA methods with *a priori* fixed 95%-confidence probability.

| Method name | Estimation method | Metrics and probability models | | *R* | Olive 1, TPR (%) | Olive 2, FPR (%) | Olive 3, FPR (%) |
|---|---|---|---|---|---|---|---|
| Ellipsoidal acceptance region | | | | | | | |
| PG01 | Moment-JM | $T^2[h]$ | JM[*q*] | 2 | 89.7 (7.3) | 0.1 (0.4) | 23.5 (3.9) |
| PI01 | Bayesian | Ga[*h*] | Log[*q*] | 2 | 95.2 (4.9) | 0.6 (0.8) | 28.2 (3.3) |
| Rectangular acceptance region | | | | | | | |
| PG02 | Moment | Ga[*h*] | Ga[*q*] | 2 | 90.4 (7.7) | 0.0 (0.4) | 24.3 (3.6) |
| PI02 | Bayesian | Ga[*h*] | Log[*q*] | 2 | 94.0 (5.5) | 0.1 (0.1) | 27.9 (3.3) |

Notes: R = optimized number of PCs; TPR (%) = true positive rate in percentage; FPR (%) = false positive rate in percentage.
Distributions: $T^2[\cdot]$ = Hotelling's *T*-square; JM[·] = Jackson-Mudholkar approximation; Ga[·] = gamma; Log[·] = lognormal. Metric variables: $q = OD^2$; $h = SD^2$.

that were overly higher than those of their prediction interval-based counterparts PI01 and PI02 (Fig. 4A and B, Student's *t* tests: *p*-values < 0.05 at TPR = 90% and 95% for each acceptance rule). This lower specificity of the "α-optimized" methods PG01 and PG02 is explained by the substantially higher spread of their acceptance regions compared to those of PI01 and PI02, as shown on Fig. 5A–B and 6A (Student's *t* tests: *p*-values < 0.05 for each acceptance rule). Not only did methods PI01 and PI02 produced smaller areas, but the shapes of their regions seemed to better reflect the geometry of the cloud of points in the metrics' space (Fig. 5A–B).

### 4.1.2. Dataset 2: oil NIR spectra

The prediction interval-based methods PI01 and PI02 showed substantially higher TPR values than their benchmark counterparts PG01 and PG02 (Student's *t*-test: *p*-values < 0.05 for each acceptance rule, Table 2). The former were nearly unbiased (biases of +0.1% to +0.8%) with lower total errors (14.7%–14.9%), while the latter were strongly biased with higher total errors (biases of − 7.1% to − 16.5% and total errors of 22.5%–29.3%).

**Table 4**
Averages and standard deviations (in brackets) of true positive rates of classification of a target ibuprofen-based formulation (I03, Ibuprofen EG 0.4g) by two classical and two prediction interval-based SIMCA methods with *a priori* fixed 95%-confidence probability. False positive rates for the non-target drug (I04, Ibuprofen EG 0.6 g) are all zero.

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

| Method name | Estimation method | Metrics and Probability models | | *R* | IO3, TPR (%) |
|---|---|---|---|---|---|
| Ellipsoidal acceptance region | | | | | |
| PG01 | Moment-JM | $T^2[h]$ | JM[*q*] | 8 | 34.4 (9.6) |
| PI01 | Bayesian | Ga[*h*] | Ga[*q*] | 8 | 96.8 (3.3) |
| Rectangular acceptance region | | | | | |
| PG02 | Moment | Ga[*h*] | Ga[*q*] | 4 | 55.3 (14.2) |
| PI02 | Bayesian | Ga[*h*] | Ga[*q*] | 4 | 95.0 (3.7) |

*R* = optimized number of PCs; TPR (%) = true positive rate in percentage; False positive rates for I04 are 0 for all methods.
Distributions: $T^2[\cdot]$ = Hotelling's *T*-square; JM[·] = Jackson-Mudholkar approximation; Ga[·] = gamma; Log[·] = lognormal. Metric variables: $q = OD^2$; $h = SD^2$.

Correcting the TPRs of the benchmark methods PG01 (ellipsoidal rule, Fig. 4C) and PG02 (rectangular rule, Figure 4D) to 90% and 95% approximately by the "$\alpha$-optimization" strategy [13] produced FPRs that were substantially higher than those of PI01 and PI02 at a TPR of 95% (Student's *t* tests: *p*-values < 0.05 at 95% for each type of acceptance region).

Overall, the prediction interval-based methods PI01 and PI02 showed better global specificity as indicated by the substantially smaller areas of their acceptance regions compared to the regions produced by their "$\alpha$-optimized" classical counterparts PG01 and PG02 (Fig. 5C–D and 6B, Student's *t*-test: *p*-values < 0.05). The shapes of their regions also seemed to better reflect the geometry of the cloud of points in the metrics' space (Fig. 5C–D).

### 4.1.3. Dataset 3: olive NIR spectra

With the olive data with moderately large calibration sets, the prediction interval-based methods PI01 and PI02 yielded less biased, more accurate and higher TPR values (biases of − 1.0% to +0.2% and total errors of 4.9%–5.6%) compared to their benchmark counterparts PG01 and PG02 (biases of − 5.3% to − 4.6%, total error of 9.0%–9.5%, Student's *t*-test: *p*-values < 0.05 for each acceptance rule, Table 3).

Regarding specificity, PG01 (23.9%) showed a significantly higher FPR than PI01 (18.8%) for an average TPR of 90% (Fig. 4E, Student's *t*- test: *p*-values < 0.05). This difference vanished at an average TPR of 95% (Fig. 4E, Student's *t*-test: *p*-values > 0.05). There were no statistically significant differences of FPRs between PG02 and PI02 (Fig. 4F, rectangular rule, Student's *t*-test: *p*-values > 0.05).

Fig. 5E and F and 6C illustrate for each acceptance rule, the spread of the regions resulting from the three strategies of models' sensitivity optimization. Methods PI01 and PI02 showed marginally less spread acceptance regions than their "$\alpha$-optimized" classical counterparts PG01 and PG02 (Student's *t*-test: *p*-values < 0.05).

### 4.1.4. Dataset 4: Ibuprofen Raman spectra

The prediction interval-based methods (PI01 and PI02) were less biased (0.0 to +1.8) and with lower total errors (3.7%–3.8%) than their benchmark counterparts (PG01 and PG02, Table 4). The latter showed overly low TPRs (biases of − 60.6% to − 39.9%, Student's *t*-test: *p*-values < 0.05 for each acceptance rule) and higher total errors (42.2%–61.4%). False positive rates with the non-target I04 were zero for all methods. However, when each classical method was calibrated at the same average TPR as its prediction interval-based counterpart using the "$\alpha$-optimization" strategy, PG01 and PG02 showed substantially more spread acceptance regions than their counterparts PI01 and PI02

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

**Table 5**

True positive (P01, Dafalgan 1g) and false positive (P03, Paracetamol EG 1g and P05, Paracetamol TEVA 1g) rates of classification of three paracetamol-based formulations by the classical and the prediction interval-based SIMCA methods.

| Model name | Estimation method | Metrics and probability models | | *R* | *α* | Area (a.u) | P01, TPR (%) | P05, FPR (%) | P03, FPR (%) |
|---|---|---|---|---|---|---|---|---|---|
| Ellipsoidal acceptance region | | | | | | | | | |
| PG01-A | Moment-JM | $T^2[h]$ | JM[*q*] | 6 | 0.0500 | 1209 | 50.0 | 0.0 | 0.0 |
| PG01-B | Moment-JM | $T^2[h]$ | JM[*q*] | 6 | 0.0049 | 4438 | 97.5 | 8.75 | 3.75 |
| PI01 | Bayesian | Ga[*h*] | Ga[*q*] | 6 | 0.0500 | 2686 | 97.5 | 0.0 | 0.0 |
| Rectangular acceptance region | | | | | | | | | |
| PG02-A | Moment | Ga[*h*] | Ga[*q*] | 3 | 0.0500 | 845 | 75.0 | 0.0 | 0.0 |
| PG02-B | Moment | Ga[*h*] | Ga[*q*] | 3 | 0.0022 | 1700 | 97.5 | 0.0 | 0.0 |
| PI02 | Bayesian | Ga[*h*] | Ga[*q*] | 3 | 0.0500 | 1323 | 97.5 | 0.0 | 0.0 |

Notes: *R* = optimized number of PCs; TPR (%) = true positive rate in percentage; False positive rates for P08 are zero for all methods.
Distributions: $T^2[\cdot]$ = Hotelling's *T*-square; JM[·] = Jackson-Mudholkar approximation; Ga[·] = gamma; Log[·] = lognormal. Metric variables: $q = OD^2$; $h = SD^2$.

**Table 6**

True positive (P03, Paracetamol EG 1g) and false positive (P05, Paracetamol TEVA 1g and P08, Panadol 0.5g) rates of classification of three paracetamol formulations by the classical and the prediction interval-based SIMCA methods.

| Model name | Estimation method | Metrics and probability models | | *R* | *α* | Area (a.u) | P03, TPR (%) | P05, FPR (%) | P08, FPR (%) |
|---|---|---|---|---|---|---|---|---|---|
| Ellipsoidal acceptance region | | | | | | | | | |
| PG01-A | Moment-JM | $T^2[h]$ | JM[*q*] | 8 | 0.0500 | 1004 | 27.5 | 10.0 | 0.0 |
| PG01-B | Moment-JM | $T^2[h]$ | JM[*q*] | 8 | 0.0008 | 3963 | 97.5 | 92.5 | 17.5 |
| PI01-A | Bayesian | Ga[*h*] | Ga[*q*] | 8 | 0.0500 | 2153 | 95.0 | 82.5 | 12.5 |
| PI01-B | Bayesian | Ga[*h*] | Ga[*q*] | 8 | 0.0400 | 2315 | 97.5 | 87.5 | 12.5 |
| Rectangular acceptance region | | | | | | | | | |
| PG02-A | Moment | Ga[*h*] | Ga[*q*] | 3 | 0.0500 | 774 | 72.5 | 35.0 | 2.5 |
| PG02-B | Moment | Ga[*h*] | Ga[*q*] | 3 | 0.0004 | 2029 | 100.0 | 97.5 | 15.0 |
| PI02-A | Bayesian | Ga[*h*] | Ga[*q*] | 3 | 0.0500 | 994 | 95.0 | 84.5 | 10.0 |
| PI02-A | Bayesian | Ga[*h*] | Ga[*q*] | 3 | 0.0300 | 1310 | 100.0 | 92.5 | 12.5 |

Notes: R = optimized number of PCs; TPR (%) = true positive rate in percentage; False positive rates for P01 are zero for all methods.
Distributions: $T^2[\cdot]$ = Hotelling's *T*-square; JM= Jackson-Mudholkar approximation; Ga[·] = gamma; Log[·] = lognormal. Metric variables: $q = OD^2$; $h = SD^2$.

(Fig. 5G–H and 6D, Student's *t*-test: *p*-values < 0.05 for each acceptance rule).

To sum up, these Monte Carlo validation studies enabled a systematic evaluation of the newly proposed prediction interval-based approach to "rigorous" SIMCA and compared its performances to classical methods using criteria such as the bias, the variance and the total error of sensitivity, the data-based specificity and global proxy for specificity as measured by the area of the acceptance region. The results demonstrated the reliability of the proposed methodology. It produces unbiased acceptance regions w.r.t. The targeted nominal probability content. The investigated classical methods on the contrary, generally show substantial undercoverage w.r.t. The targeted nominal probability content. Moreover, it has been demonstrated that the "$\alpha$-optimization" approach which attempts to improve the sensitivities of these classical methods by "optimizing" $\alpha$ can worsen their specificity compared to the prediction interval-based methods.

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

## 4.2. Applicability in external validation

### 4.2.1. Case 1, dataset 5, paracetamol Raman spectra, P01 vs P03, P05 and P08

When P01 is the reference, the prediction interval-based methods (PI01 and PI02) correctly recognized the target batches and discriminated the non-target batches, as their TPRs fell inside the decision interval i.e., 88.2%− 100.0%, and their FPRs were zero (Table 5). Furthermore, their TPRs were considerably higher than those of their classical counterparts calibrated at $\alpha$ = 0.05, i.e., PG01-A and PG02-A on Table 5 ($z$-test, $p < 0.05$ for each acceptance rule).

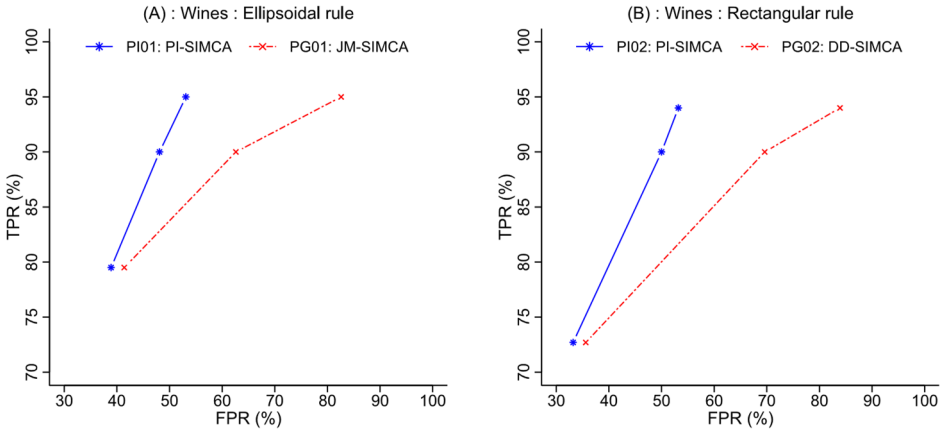
"Optimizing" the classical methods by decrementing $\alpha$ in cross- validation (i.e., methods PG01-B and PG02-B) enabled to achieve equal TPRs with the prediction interval-based methods, with however higher risks of false positives, especially with the ellipsoidal rule as shown on Table 5 and Fig. 7A and B.

### 4.2.2. Case 2, dataset 5, paracetamol Raman spectra, P03 vs P01, P05 and P08

The case when P03 is the reference illustrates how seemingly minor losses of specificity of the "$\alpha$-optimized" classical methods unveiled by the Monte Carlo studies in Section 4.1, can lead to false acceptances of batches in practical quality control situations like batch testing and release (Table 6, Fig. 7C–D). The prediction interval-based methods PI01-A and PI02-A yielded better sensitivities than their classical counterparts PG01-A and PG02-A calibrated at $\alpha$ = 0.05 ($z$-test, $p < 0.05$ for each acceptance rule, Table 6). Moreover, they correctly recognized the targeted batches as their TPRs were inside the decision interval i.e., 88.2% − 100.0%. Likewise, they correctly rejected all non-target batches as their FPRs were outside the decision interval i.e., 88.2% − 100.0%.

"Optimizing" the classical methods in cross-validation by decrementing $\alpha$ until a cross-validated TPR of 95% is achieved (i.e., models PG01-B and PG02-B) enabled to improve these models' sensitivities and to accept the target batches. However, their specificities were lower than those of the prediction interval-based models (PI01-A and PI02-A), and they failed to reject the non-target P05 batches as their FPRs fell inside the decision interval of 88.2% − 100.0% (Table 6).

It is worthwhile to note that for the ellipsoidal region, the prediction interval-based model PI01–B calibrated at $\alpha$ = 0.04 produced a TPR equal to the one of its "$\alpha$-optimized" classical counterpart (i.e., PG01-B). Interestingly, the FPR and area of acceptance region of this prediction interval-based method were still lower than the ones of the



*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

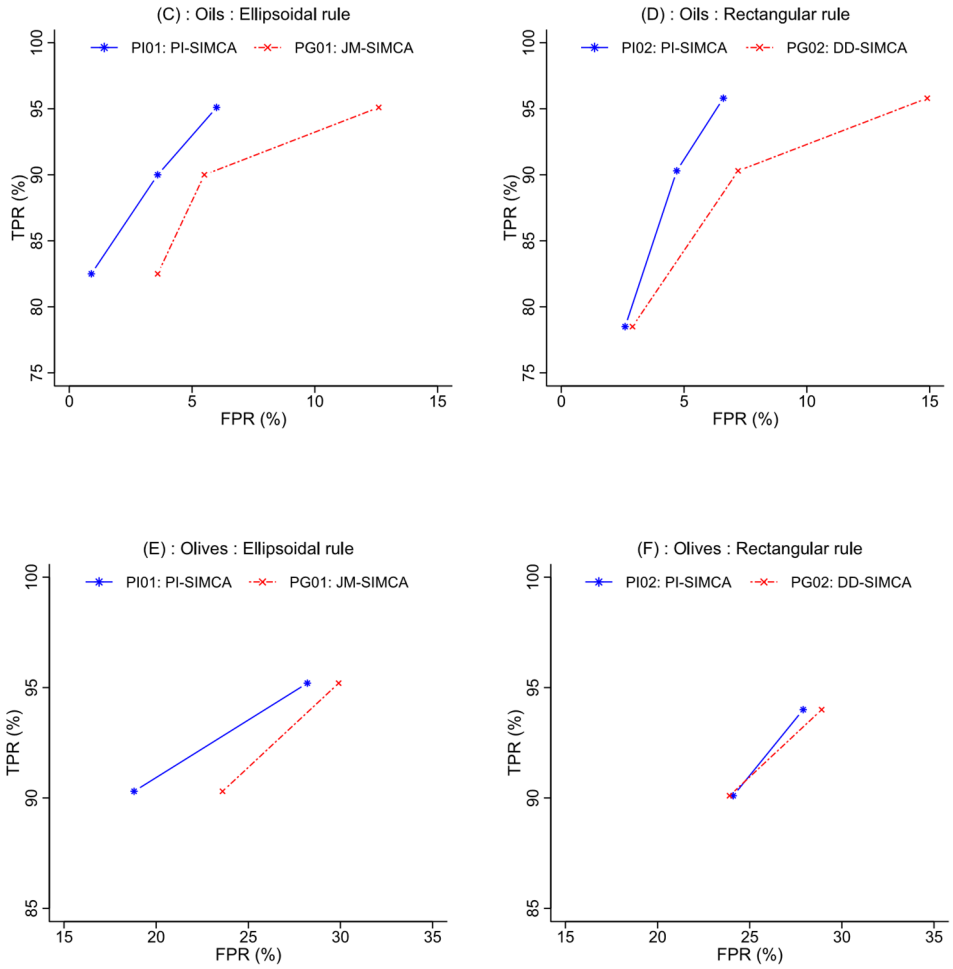



**Fig. 4.** Curves comparing the false positive rates (at fixed true positive rates) of the prediction intervals-based SIMCA methods (PI01 and PI02) and their classical counterparts (PG01 and PG02). The true positive rates (TPR) and false positive rates (FPR) are averages from 500 models on random subsamples of each reference data.

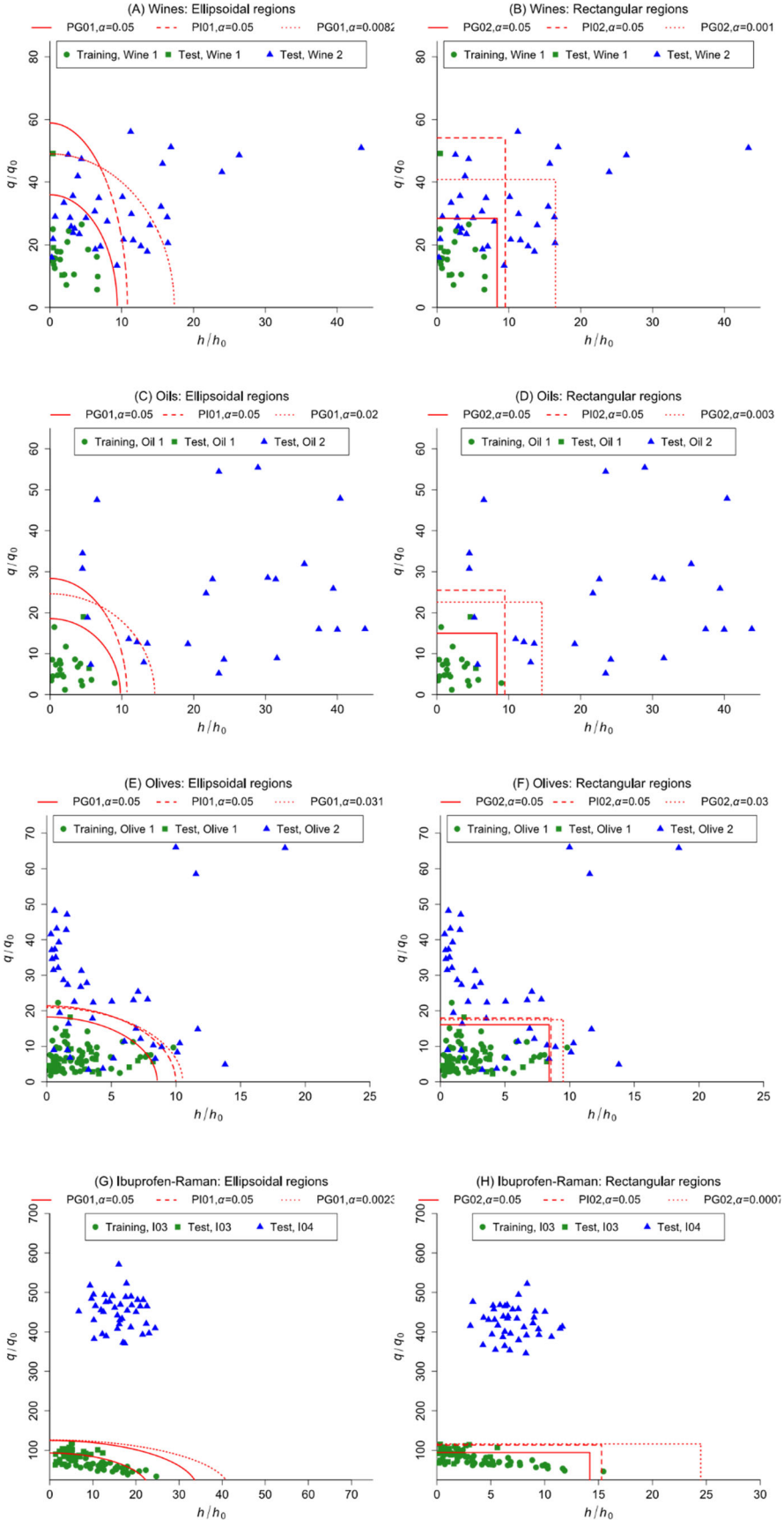
"$\alpha$-optimized" classical method and it still enabled to correctly reject the. non-target P05 batches (Table 6). Similarly, for the rectangular region, the specificity criteria of the prediction interval-based model calibrated at $\alpha = 0.03$ (i.e., PI02–B) were better than the ones of its "$\alpha$-optimized" classical counterpart (i.e., PG02-B) for the same sensitivities at 100% (Table 6)

# 5. Discussion

The ultimate aim of the "rigorous" SIMCA model is to predict

**Fig. 5.** Spread of the acceptance regions of the prediction interval-based SIMCA methods (PI01 and PI02) and their classical counterparts (PG01 and PG02) for Datasets 1–4 in Monte Carlo validation. Notes: (1) The limits are based on one random subsample of each reference set; (2) Classical methods PG01 and PG02 were calibrated at two different type 1 errors, firstly $\alpha = 0.05$, and secondly at a value $\alpha$ yielding an average and a standard deviation of TPR equivalent to those of their prediction interval-based counterparts

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*



*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

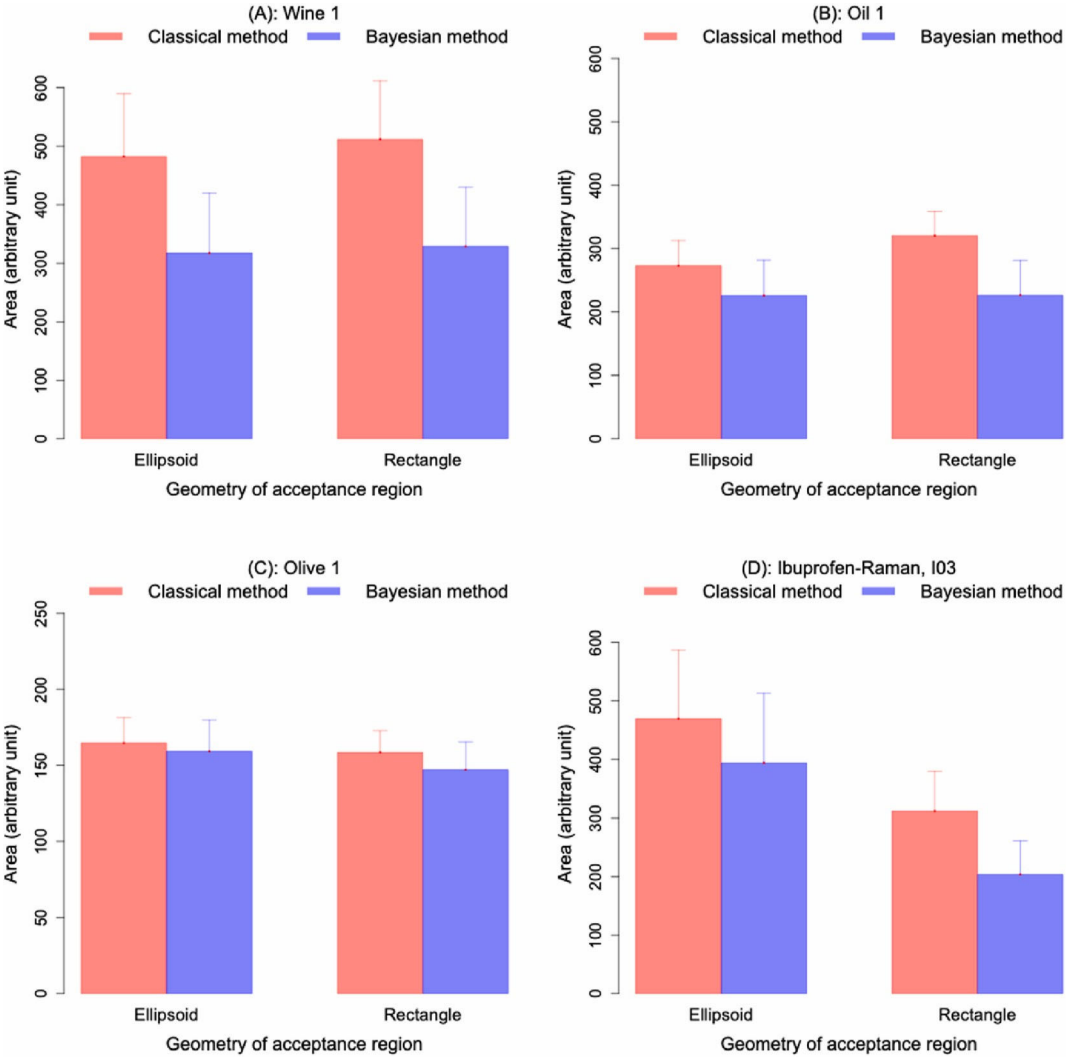



**Fig. 6.** Average areas (plus standard deviations as bars) of the acceptance regions of the prediction interval-based SIMCA (PI01 and PI02) and their classical counterparts (PG01 and PG02) in Monte Carlo validation using Dataset 1–4. Note: Classical methods i.e., PG01 and PG02 are calibrated at average sensitivities of about 95% with the same standard deviations as their prediction interval-based counterparts i.e., PI01 and PI02, by "optimizing". $\alpha$

whether the metrics for a single or several future unknown spectra fall inside a calibrated statistical prediction region with a prespecified probability or expected true acceptance rate $1-\alpha$, $\alpha \in ]0,1[$ being the theoretical type 1 error or expected false rejection rate of the method [10,12]. However, the method as currently implemented may result in low sensitivities and important discrepancies between the expected probability content $1-\alpha$ and the effective sensitivity [7,12]. These discrepancies are clear indications that these classical prediction models are ill-calibrated i.e., the densities used to make predictions do not correctly describe the distribution of future metric values. A more and more common practice to enhance the sensitivities is the strategy termed the "$\alpha$-optimization" strategy in this work which consists in enlarging the acceptance regions by decrementing $\alpha$ by $\delta$ until the desired TPR is achieved in cross-validation [13]. Despite this "$\alpha$-optimization" approach improves models' sensitivities, the discrepancies between its expected and effective false rejection rates i.e., $\alpha - \delta$ and $\alpha$, respectively, persist. Clearly, $\alpha$ loses its theoretical (probabilistic) interpretation.

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

The prediction interval-based approach to SIMCA proposed in this work intends to properly overcome this issue of biased sensitivities. It uses the concept of predictive distributions *sensu stricto* to construct statistical prediction regions for SIMCA metrics [15–19]. These predictive distributions are estimated firstly by selecting appropriate data generating or probability models for the metrics using the well-known strategy of goodness-of-fit tests, and then by fitting these models as Bayesian generalized linear models [18,19].

The evaluation of this newly proposed approach to SIMCA on real datasets shows substantial improvements of the sensitivities, compared to existing classical variants which are not based on predictive densities *sensu stricto*, especially in low sample size contexts. It can even enhance specificities compared to the "$\alpha$-optimization" approach which attempts to improve sensitivities of classical methods by "optimizing" the type 1 error i.e., $\alpha$ [13]. For the wide diversity of data, its performances are stable, suggesting the approach is more likely to guarantee the targeted coverage in most circumstances.

The improvement in sensitivity is because, selecting a satisfactorily fitting probability model reduces the risk of lack-of-fit and integrating model parameters' uncertainties into the definition of the prediction regions naturally enlarges the acceptance regions. Hence, contrary to the "$\alpha$-optimization" strategy, the prediction interval-based SIMCA automatically adjusts its acceptance limits so that its effective sensitivity is close to the nominal confidence probability $1-\alpha$. The potential improvement in specificity compared to the "$\alpha$-optimization" strategy is due to the potentially smaller spread of its acceptance regions, provided the data generating distributions of the metrics are correctly chosen.

This correct choice of the probability models is an important point to be emphasized. It is key to the success of the proposed optimization strategy. The current work used the acceptance rates of the null distributions by goodness-of-fit tests as metrics to screen among several candidate distributions. Alternatively, other criteria like the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) might be investigated [19]. One might also simply use the well-known Q-Q plots or P–P plots. A strategy combining several criteria might also be used to increase the likelihood of selecting a satisfactorily fitting distribution for each metric variable. Other candidate distributions for positive random variables that cannot be fitted using the generalized linear models but can provide a better fit, might also be investigated (e. g., the Weibull distribution).

In practice, it would be reasonable to rigorously compare both the "$\alpha$-optimization" and the prediction interval-based strategies. This may be achieved for example *via* criteria such as the spread of their acceptance regions to determine the most globally specific model for each dataset.

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

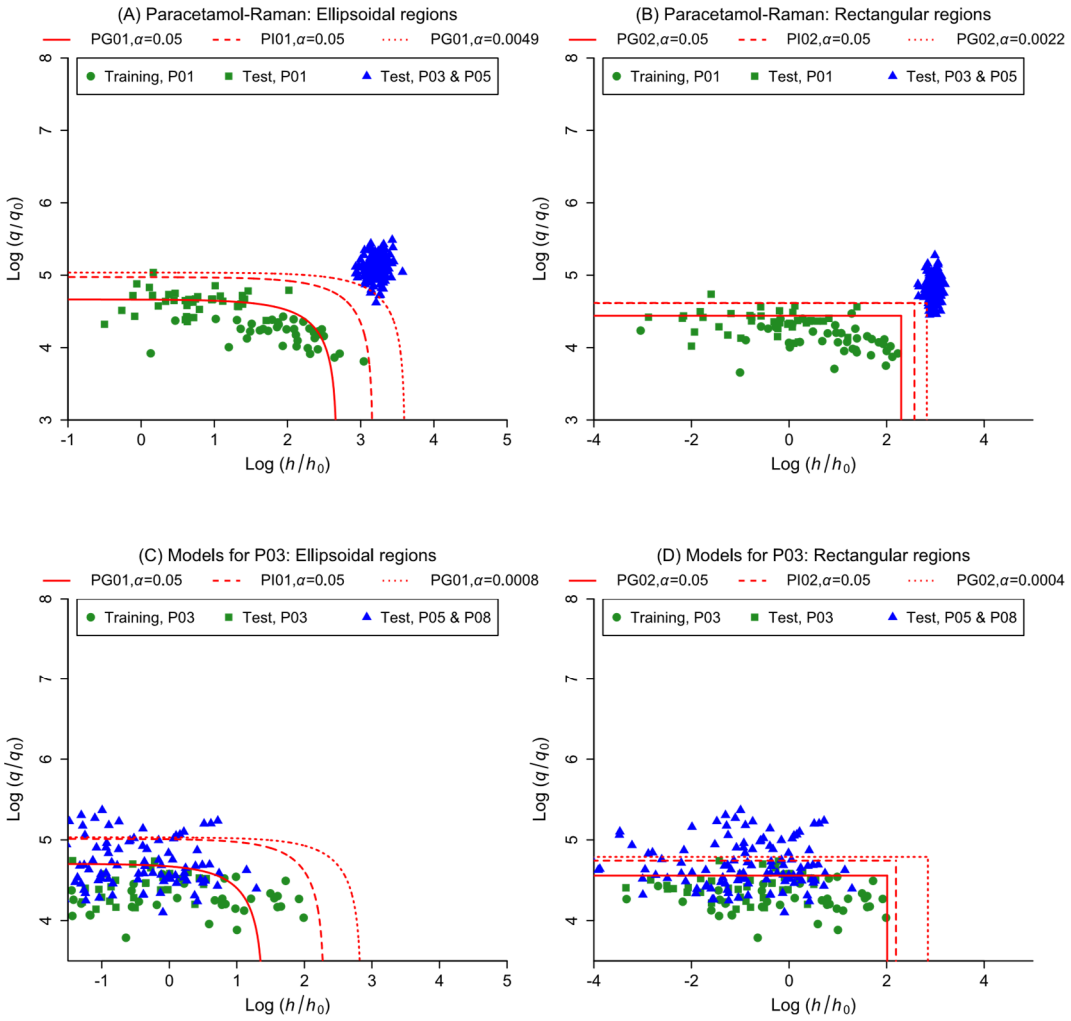



**Fig. 7.** Spread of the acceptance regions of the prediction interval-based SIMCA (PI01 and PI02) and their classical counterparts (PG01 and PG02) for two case studies with Dataset 5 in external validation. Note: Classical methods PG01 and PG02 were calibrated at two different type 1 errors, firstly $\alpha$ = 0.05, and secondly a value $\alpha$ yielding an average TPR of 95% approximately in cross-validation.

Overall, the proposed approach has several advantages. Firstly, it is more consistent with the statistical theory on predictive modeling. Secondly, the statistical properties of the resulting prediction regions are well-known; for example, they are directly interpretable as probabilistic prediction regions i.e., the theoretical confidence probabilities of these regions really reflect their practical sensitivities. Thirdly, once the probability models are chosen, the only parameter to be optimized is the number of PCs, in contrast to the "$\alpha$-optimization" strategy which attempts to optimize not only the number of PCs, but also the confidence probability [13]. Fourthly, selecting satisfactorily fitting distributions for the metrics among several candidates (instead of *a priori* fixed distributions) and estimating these distributions' parameters from the data using the Bayesian approach is more flexible than the existing approaches. Indeed, the data guide the choice of the family of distributions. In addition, the Bayesian method does not provide a point estimate of the model parameters, but it estimates the overall uncertainties about these parameters from the metric data. Fifthly, the proposed framework provides a unified approach to the numerous variants of SIMCA differing in their distributional assumptions of the metrics. Sixthly, it enables to bring the SIMCA to the general framework of

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

generalized linear models so that it can benefit from the numerous appealing prediction procedures already developed for such models, such as the mixed models for modeling inter-batch variability.

A disadvantage concerns the computational cost when MCMC sampling methods are involved to approximate the predictive distributions of the metrics. Indeed, generating a MCMC sample of size 50000 for one metric variable takes on average 3.4 ± 0.2 s for a gamma model on a 2.2 GHz computer without parallelization.

A first possible extension to be investigated might be a more statistically effective integration of inter-batch variability as random effects through the generalized linear mixed models. A second possible extension is the construction of prediction intervals for *m* futures values or their mean, *m* out of *l* future values ($l \geq m \geq 1$), or a tolerance interval for a proportion of the whole spectral population, depending on the sampling schemes to be used during the practical testing phase [18]. A third possible extension is the investigation of other candidate distributions that might provide a better fit, for example the Weibull distribution for the metrics.

# 6. Conclusion

The soft independent modeling of class analogy (SIMCA) is a parametric predictive model and as such, the concept of statistical prediction regions based on predictive densities *sensu stricto* can be used to optimize its acceptance regions. This work used the Bayesian generalized linear models' framework to construct such prediction regions for the SIMCA metrics. The resulting model termed "prediction interval-based SIMCA", can show better performances compared to the existing models. A key to the success of this newly proposed approach to SIMCA is the careful verification of the validity of the distributional assumptions of the metrics.

### Funding

This work was supported by the Wallonia Region of Belgium [Grant N°7517, project Vibra4Fake].

### CRediT authorship contribution statement

**T. Hermane Avohou:** Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing – original draft, Writing - critical review. **Pierre-Yves Sacré:** Investigation, Data curation, Writing - critical review, Supervision, Project administration, Funding acquisition. **Sabrina Hamla:** Writing - critical review. **Pierre Lebrun:** Writing - critical review, Supervision. **Philippe Hubert:** Writing - critical review, Supervision, Project administration, Funding acquisition. **Éric Ziemons:** Writing - critical review, Supervision, Project administration, Funding acquisition.

### Declaration of competing interest

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

### Data availability

The authors do not have permission to share data.

### Acknowledgement

*Published in : Analytica Chimica Acta (2022), vol. 1229, pp. 1–17*
*DOI: 10.1016/j.aca.2022.340339*
*Status : Postprint (Author's version)*

The authors are grateful to: (1) the developer of the *mdatools* package of R for providing the extra codes to implement the DD-SIMCA method with rectangular acceptance region, and (2) all the anonymous reviewers and editors for their insightful comments which helped to significantly improve the manuscript.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2022.340339.

## References

[1] P. Oliveri, Class-modeling in food analytical chemistry Development, sampling, optimization and validation issues – a tutorial, Anal. Chim. Acta 982 (2017) 9–19.
[2] 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 characterization of Italian red wines, Anal. Chim. Acta 668 (2010) 143–148.
[3] P. Oliveri, M. Casale, M.C. Casolino, M.A. Baldo, F.I. Grifi, M. Forina, Comparison between classical and innovative class-modeling techniques for the characterization of a PDO olive oil, Anal. Bioanal. Chim 399 (2011) 2105–2113. [4] P. Oliveri, M.I. Lopez, M.C. Casolino, I. Ruisanchez, M.P. Calao, L. Medini, S. Lanteri, Partial least squares density modeling (PLS-DM) - a new class-modeling strategy applied to the authentication of olives in brine by near-infrared spectroscopy, Anal. Chim. Acta 851 (2014) 30–36.
[5] P.H. Ciza, P.-Y. Sacre, C. Waffo, L. Coïc, T.H. Avohou, J.K. Mbinze, R. Ngono, R. D. Marini, Ph Hubert, E. Ziemons, Comparing the qualitative performances of handheld NIR and Raman spectrophotometers for the detection of falsified pharmaceutical products, Talanta 202 (2019) 469–478.
[6] S. Wold, M. Sjost¨ rom, Chapter 12, SIMCA: a method for analyzing chemical data in¨ terms of similarity and analogy, in: B.R. Kowalski (Ed.), Chemometrics, Theory and Application, vol. 52, American Chemical Society, Washington, DC, 1977, pp. 243–282.
[7] R. De Maesschalck, A. Candolfi, D.L. Massart, S. Heuerding, Decision criteria for soft independent modeling of class analogy applied to near infrared data, Chemometr. Intell. Lab. Syst. 47 (1999) 65–77.
[8] R. Vitale, F. Marini, C. Ruckebusch, SIMCA modeling for overlapping classes: fixed or optimized decision threshold? Anal. Chem. 90 (2018) 10738–10747.
[9] A.L. Pomerantsev, Acceptance areas for multivariate classification derived by projection methods, J. Chemom. 22 (2008) 601–609.
[10] A.L. Pomerantsev, O.Ye. Rodionova, Concept and role of extreme objects in PCA/ SIMCA, J. Chemom. 28 (2014) 429–438.
[11] O.Ye Rodionova, P. Oliveri, A. Pomerantsev, Rigorous and compliant approaches to one-class classification, Chemometr. Intell. Lab. Syst. 159 (2016) 89–96.
[12] A.L. Pomerantsev, O.Ye. Rodionova, Popular decision rules in SIMCA: critical review J, Chem 34 (2020) 429–438.
[13] Z. Malyjurek, R. Vitale, B. Walczak, Different strategies for class model optimization, A comparative study, Talanta 215 (2020) 1–9.
[14] Z. Chen, P. de Boves Harrington, Automatic soft independent modeling for class analogies, Anal. Chim. Acta 1090 (2019) 47–56.
[15] K.P. Murphy, Probabilistic Machine Learning: an Introduction, The MIT Press, Cambridge, 2021, p. 578.
[16] Z. Ghahramani, Probabilistic machine learning and artificial intelligence, Nature 521 (2015) 452–459.
[17] B.S. Clarke, J.S. Clarke, Predictive Statistics: Analysis and Inference beyond Models, Cambridge University Press, Cambridge, 2018, p. 642.
[18] W.Q. Meeker, G.J. Hahn, L.A. Escobar, Statistical Intervals: A Guide for practitioners and Researchers, second ed., John Wiley & Sons Inc, Hoboken, 2017, p. 578.
[19] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, D.B. Rubin, Bayesian Data Analysis, Chapman and Hall/CRC, Boca Raton, 2014, p. 675.
[20] R.A. Lockhart, M.A. Stephens, The probability plot tests of fit based on the correlation coefficient, Handb. Stat. 17 (1998) 453–473.
[21] R.M. Vogel, The probability plot correlation coefficient test for normal, lognormal, and Gumbel distributional hypotheses, Water Resour. Res. 22 (1986) 587–590. [22] T.W. Anderson, Anderson - darling tests of goodness-of-fit, in: M. Lovric (Ed.), International Encyclopedia of Statistical Science, Springer, Berlin, Heidelberg, 2011, pp. 32–54.
[23] D.J. Murdoch, Y.-L. Tsai, J. Adcock, P-values are random variables, Am. Statistician 62 (2008) 242–245.
[24] K. Krishnamoorthy, Handbook of Statistical Distributions with Applications, second ed., Taylor and Francis, Boca Raton, 2016, p. 398.
[25] Stan Development Team, Stan Function Reference Version 2.19, Stan Development Team, 2019, p. 153.
[26] S. Bedbur, J.M. Lennartz, U. Kamps, On minimum volume properties of some confidence regions for multiple multivariate normal means, Stat. Probab. Lett. 158 (2020) 1–4.
[27] R Core Team, R, A Language and Environment for Statistical Computing, 2018. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/ .
[28] J.E. Jackson, J.S. Mudholkar, Control procedures for residuals associated with principal component analysis, Technometrics 21 (1979) 341–349.
[29] G.E.P. 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.
[30] S. Kucheryavskiy, Mdatools - R package for chemometrics, Chemometr. Intell. Lab. Syst. 198 (2020) 1–10.
[31] S.P. Millard, EnvStats an R Package for Environmental Statistics, Springer Science and Business Media, New York, 2013, p. 291.
[32] B. Carpenter, A. Gelman, M.D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, A. Riddell, Stan A probabilistic programming language, J. Stat. Software 76 (2017) 1–32.
[33] B. Goodrich, J. Gabry, I. Ali, S. Brilleman, Rstanarm Bayesian Applied Regression Modeling via Stan, 2020. R package version 2.21.1, https://mc-stan.org/rstanarm.
[34] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639.
[35] D.C. Montgomery, Introduction to Statistical Quality Control, seventh ed., John Wiley and Sons Inc, Hoboken, 2013, p. 754.