Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

A PROBABILISTIC CLASS-MODELLING METHOD BASED ON PREDICTION BANDS FOR FUNCTIONAL SPECTRAL DATA: METHODOLOGICAL APPROACH AND APPLICATION TO NEAR-INFRARED SPECTROSCOPY

Avohou T. Hermane
University of Liege (ULiege), CIRM, Vibra-Santé Hub, Laboratory of Pharmaceutical Analytical Chemistry, Department of Pharmacy, Avenue Hippocrate 15, 4000 Liège, Belgium

Sacré Pierre-Yves
University of Liege (ULiege), CIRM, Vibra-Santé Hub, Laboratory of Pharmaceutical Analytical Chemistry, Department of Pharmacy, Avenue Hippocrate 15, 4000 Liège, Belgium

Lebrun Pierre
PharmaLex Belgium, Rue Edouard Belin 5, 1435, Mont-St-Guibert, Belgium

Hubert Philippe
University of Liege (ULiege), CIRM, Vibra-Santé Hub, Laboratory of Pharmaceutical Analytical Chemistry, Department of Pharmacy, Avenue Hippocrate 15, 4000 Liège, Belgium

Ziemons Eric
University of Liege (ULiege), CIRM, Vibra-Santé Hub, Laboratory of Pharmaceutical Analytical Chemistry, Department of Pharmacy, Avenue Hippocrate 15, 4000 Liège, Belgium

Keywords: Class-modelling; Functional data analysis; Bayesian chemometrics; Spectral predictive; distribution; Prediction band; Depth statistic; Multivariate data analysis

Abstract

Class-modelling methods aim to predict the conformity of new unknown samples with a single target class, using statistical decision rules built exclusively with objects of that class. This article introduces a novel class-modelling method for spectral data. The method uses the concept of $\beta\%$-prediction band for functional data to classify spectra. The band is defined by an upper and a lower limiting spectra which delimit critical trajectories for $\beta\%$ of future spectra of the target class. It is constructed in three main steps: firstly, a naïve bootstrap sample of calibration spectra is projected onto a parsimonious principal component (PC) basis and their scores are estimated. The posterior predictive distribution of the scores on each PC is estimated using a Bayesian zero-mean normal model. This procedure is repeated on naïve bootstrap estimations of the PCs to obtain the predictive distribution of the scores. These enable to account for all modelling uncertainties including the random deviation of scores from their zero-mean on each PC, uncertainty in the variance of scores (eigenvalue) on each PC, and uncertainty in the PC estimations. Secondly, the predicted scores are back-transformed to the original signal scale to obtain the predictive distribution of future spectra. Thirdly, the predicted spectra are ranked to select the $\beta\%$ most central ones as typical set, whose ranges of variation are used to

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

construct the simultaneous limits of the band. Once the band is constructed, reconstructions of future unknown test spectra by bootstrap PC models are projected onto it, and the extent to which they overlap with it is used to decide their acceptance or rejection. The statistical properties and classification performances of the proposed prediction band are evaluated on real near-infrared datasets and compared to the well-known soft- independent modelling of class analogy (SIMCA) model. The results of the evaluation provide evidence that the proposed prediction band possesses satisfactory predictive performances. It even outperforms the SIMCA while offering attractive advantages like risk-management and straightforward physical interpretability of outlyingness patterns of tested spectra.

1. Introduction

Class-modelling (CM) is an important family of semi-supervised classification methods, fundamental to multivariate pattern recognition in analytical chemistry [1,2]. These methods primarily aim at predicting the belonging of future unknown objects to a given class, called the target class. These classifications are based on statistical decision rules built using a set of objects belonging exclusively to that class. Mathematically, this is achieved firstly by building a prediction space of acceptable variation of some attributes for regular or typical objects of the target class, and secondly by verifying whether the same attributes for any future unknown object comply with the predicted space or not. Hence, conceptually, CM methods resemble multivariate outliers or extreme objects detection methods [2e6], and are based on reference spectrum data only, as with auto-encoders. If there is more than one class, a model is built for each, independently of the others. As a result, the prediction spaces of two or more classes may overlap and new objects may be assigned to two or more classes [7,8]. The principal motivation of using CM methods, instead of discriminant classification methods, lies in the fact that they do not require information about objects belonging to non-target classes to define decision rules. Thus, they are more appropriate for authentication and verification issues where non-target or alternative classes are generally not known [2,5].

In an ever-growing number of CM applications, measurements on objects are one-dimensional functional spectral data. This means that, the observation for each object consists of a set of signal values measured along a possibly infinite sequence of (ordered) values of a spectral variable (e.g. NIR wavelengths or Raman shifts) [9]. These observations are assumed to be values of an underlying random smooth curve, which are measured at a finite grid of points, possibly with noise. Because of this continuum of the spectral variable, functional data differ from standard multivariate data and specific statistical methods that account for this peculiarity may be devised for them. Such methods are known as functional data analysis (FDA) and include for example functional principal component analysis and functional regression models [9e11]. Typical examples of functional data in chemistry include spectroscopic curves such as near-infrared (NIR) and Raman spectra [9,11]. For instance, the studies depicted in the present paper aims at identifying drugs having similar compositions based on their NIR spectral signature taken through their unopened blister.

There exists a small number of CM methods used to analyze spectroscopic data in chemometrics, all of which are exclusively multivariate methods in nature, meaning that they do not account for the

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

intrinsic functional nature of the spectrum. Briefly, these methods generally proceed first by selecting a parsimonious projection subspace (e.g. a principal components' subspace) to represent the dataset as a cloud of points. Second, a univariate outlyingness metric is estimated for each observed spectrum in that subspace (e.g. the Mahalanobis distance). Third, assuming a probability distribution of the estimated outlyingness metric, a statistical confidence region is determined as acceptance space for testing unknown future spectra [1,6]. Typical representatives are: the soft independent modelling of class analogy (SIMCA) method which uses as outlyingness metric the so-called orthogonal distance (Q-statistic) and score distance (T^2-statistic) or a combination of both [5,12,13]; and the unequal class modelling (UNEQ) method which uses as outlyingness metric the Mahalanobis distance of each object from the centroid of the class [14]. These methods are well-established among the chemometric community, SIMCA being the most prominent one [1e3,6].

The present article aims at introducing a novel approach of class-modelling for spectroscopic data, especially NIR data. The proposed approach uses as classification rule, the concept of prediction band for a random spectrum, which extends the well-known concept of two-sided prediction interval for a random variable to random curves. In this approach, each measured spectrum is considered as an observation of an underlying random smooth spectrum. Then, a Bayesian zero-mean model of principal component (PC) scores combined with a naïve bootstrap of the PC decomposition are used to estimate the predictive distribution of future spectra given the observed spectra. This distribution predicts the expected trajectories of the spectral population of the target class that might be observed in the future given the observed dataset. It accounts for all types of modelling uncertainties including uncertainties in the mean spectrum and the PC decomposition quantities involved in estimating the spectrum-level deviation from the mean spectrum (i.e. eigenvalues, eigenfunctions and number of eigenfunctions). Eventually, a statistical prediction band is defined to include the most central (deepest) spectra of the predictive distribution, using the concept of functional depth statistics [15]. This band defines critical boundaries for acceptable trajectories for future spectra of the target class. It is delimited by an upper limiting spectrum and a lower limiting spectrum, which are defined so that the band contains on average, a pre-specified proportion of future spectra of the target class, the most central ones. If the reconstructions of an unknown tested spectrum are inside this band, entirely or partly, with a certain probabilistic or risk criterion, it is accepted as conforming to the target class. Otherwise, it is considered as an outlier. Moreover, the band may be used to derive further information about the outlyingness patterns of spectra of target and non-target classes, by computing for any tested spectrum interpretable outlyingness metrics such as the proportion of points outside the band. This information may be used to further tune classification rules, similarly to decision rules used in control charts, say tolerating a small number of points outside the band.

Compared to the existing CM methods like the SIMCA, the proposed approach has several attractive advantages. First and foremost, because it models the whole spectrum, it enables to analyze the outlyingness pattern of tested unknown spectra alongside the spectral variables, hence enabling more interpretable classification results, e.g. identifying the subset of spectral variables where departure from reference limits is observed. Second, it is fully predictive as it accounts for all relevant modelling uncertainties. Hence, it is more correctly interpretable as probabilistic prediction region for a single future spectrum of the target class. Overall, the proposed approach provides a comprehensive

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

assessment of outlyingness behaviour of a tested spectrum while being fully predictive, risk-oriented and practically implementable.

The article is organized as follows. Section 2, details the statistical concepts and techniques upon which the proposed CM method is based. The concept of prediction band as classification rule is explained (Section 2.1) and the workflow of the method to construct it is briefly and schematically presented (Section 2.2). Then, Section 2.3 to 2.10 detail the statistical models and estimation techniques involved in this workflow including the functional class-model (Sections 2.3-2.4), and its estimation and the prediction of the future curves (Sections 2.5-2.8), the construction of the band limits (Section 2.9) and the proposed probabilistic procedures to test the conformance of unknown spectra (Section 2.10). Section 3 presents the methodologies to evaluate the statistical properties and performances of the prediction band, and to compare them to those of the benchmark SIMCA model using four real NIR datasets. Section 4 presents the results of the evaluation studies and discusses the advantages and possible limitations of the prediction band, compared to the SIMCA model. Conclusions and opportunities for research and improvement are presented in Section 5.

2. Method

2.1. CONCEPT OF PREDICTION BAND FOR A SINGLE FUTURE SPECTRUM

The concept of prediction band for a next future spectrum is the foundation of the classification rule of the proposed CM method. This band is simply a conceptual and mathematical extension of the concept of two-sided prediction interval for a univariate random variable to a functional variable. To recall, a prediction interval for a single future observation with confidence level $\beta\%$, denoted $\beta\%$-prediction interval, is a statistical interval with an upper and a lower limits, inside which a future observation of the same population would fall with a confidence $\beta\%$ [16]. It is also interpreted as a $\beta\%$-expectation tolerance interval, which is an interval inside which a prespecified proportion $\beta\%$ of the sampled population falls on average [17]. Such statements are not possible with a confidence interval, which provides an interval estimate of a population parameter (e.g. the mean or the standard deviation). These intervals are used to define reference regions in analytical conformity testing problems such as the well-known total analytical error profile [18,19]. Several statistical approaches have been proposed to construct such intervals, including the Bayesian approach [17,20]. In this approach, the interval is computed using quantiles of the so-called posterior predictive distribution of the outcome of interest, to select its most central range [17,20]. This is the probability distribution of a future outcome that accounts for both model error and uncertainty about unknown model parameters [20].

By analogy with a two-sided prediction interval for a single future univariate observation, this work defines a prediction band for a spectral curve as a statistical band, delimited by an upper and a lower limiting spectra, and inside which one can state with a confidence level $\beta\%$ that a single future spectrum from of the same spectral population falls. However, because there might exist substantial uncertainty in the estimations of an unknown tested spectrum to be tested, these estimations may fall entirely or partly inside the band, i.e. they may overlap with the band partly or entirely. This

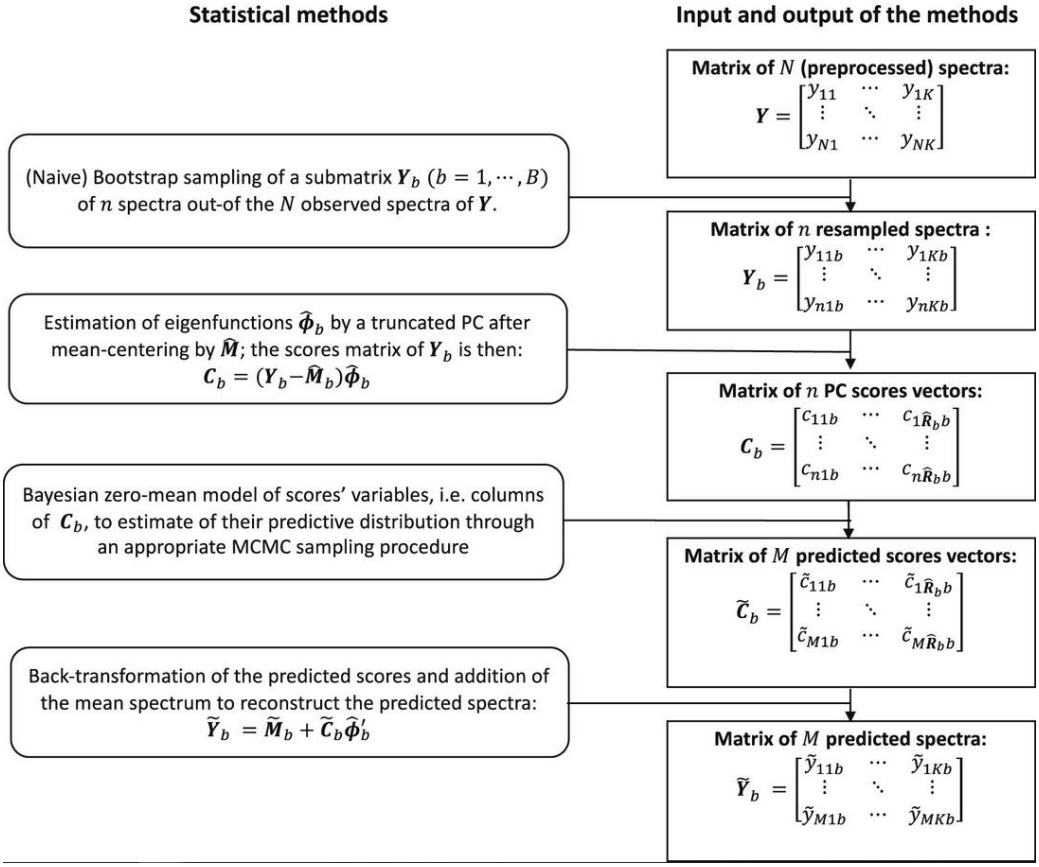
Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

uncertainty in the estimations of unknown spectra to be tested is also considered to define probabilistic procedures for testing their conformity.

2.2. WORKFLOW OF THE METHOD TO BUILD A β%-PREDICTION BAND

The methodology used to construct a β%-prediction band combines two major statistical techniques: firstly, a Bayesian zero-mean model on mean-centered principal component (PC) scores combined with a naïve bootstrap of the PC decomposition [21]. This technique is used to derive the predictive distribution of a future spectrum given the observed dataset. The workflow of this technique is summarized in Fig.1. Secondly, a non-parametric centrality or depth statistics [15] is used to perform a center-outwards ranking of the predicted spectra, hence enabling to choose the most central ones as typical and to set the limits of the band.

Fig.1. Statistical workflow to compute the predictive distribution of future spectra of the target class. The whole workflow is repeated on naïve bootstrap submatrices of the spectral matrix to account for uncertainties in the mean spectrum, the eigenfunctions, and the number of eigenfunctions. Notations: $\boldsymbol{Y}$ is the $N \times K$ spectral matrix, where N is the number of spectra and K is the number of spectral variables; $\boldsymbol{Y}_b$ is an $n \times K$ submatrix of $\boldsymbol{Y}$ at bootstrap iteration b $(b = 1, \ldots, B)$, where $n < N$ is the number of randomly sampled spectra at each bootstrap iteration; $\hat{\varphi}_b$ is $K \times \hat{R}_b$ the matrix of values of eigenfunctions at the K spectral variables and $\widehat{\varphi'}_b$ is its transpose, where $\hat{R}_b$ is an estimated truncation point of the PCs at bootstrap iteration b; $\boldsymbol{C}_b$ is an $n \times \hat{R}_b$ matrix of PCs' scores; $\widetilde{\boldsymbol{C}}_b$ and $\widetilde{\boldsymbol{Y}}_b$ are respectively $M \times \hat{R}_b$ and $M \times K$ matrices of predicted PCs scores and predicted spectra, where M is the total number of predictions at each bootstrap iteration.



The Bayesian zero-mean model on PC scores involves as a first step, projecting each preprocessed and mean-centered spectrum onto a truncated principal component (PC) basis, which may be either the

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

standard PC or the functional PC bases. The core idea of functional PC analysis is to estimate smoothed eigenfunctions, enabling a proper control of overfitting while preserving all the properties of standard PC analysis, including decorrelation of the signal, energy conservation and sparsity [21]. In a second step, the scores on the selected PCs are modelled using a Bayesian zero-mean normal model, and their posterior predictive distribution accounting for uncertainties in the eigenvalues given the PC decomposition is derived. This posterior predictive distribution is then back-transformed to the original signal space to obtain the posterior predictive distribution of future spectra given the PC decomposition and the spectral data (Fig. 1). These modelling steps are repeated on naïve boostrap submatrices of the dataset to account for uncertainties in the PC decomposition quantities (mean spectrum, eigenfunctions and number of eigenfunctions). This bootstrap is necessary as modelling uncertainties about the PC decomposition within a Bayesian Markov Chain Monte Carlo (MCMC) sampler is not trivial. This enables to obtain an appropriate estimation of the predictive distribution of future spectra given the observed dataset. This distribution predicts possible trajectories of the spectral population accounting for all modelling uncertainties including the uncertainty about the mean spectrum and uncertainties about the PC decomposition quantities involved in estimating the spectrum-level deviation from the mean spectrum (eigenvalues, eigenfunctions and number of eigenfunctions).

The depth concept generalizes the univariate concepts of rank and order statistics to functional data [15]. It enables a center-outwards non-parametric ranking of the predicted spectra [15], the selection of the most central ones and the definition of the limits of the prediction band based on the trajectories of most central or deepest predicted spectra.

All these modelling procedures are detailed in Sections 2.3-2.9. The algorithms and codes to implement them are available on request.

2.3. MODEL OF THE RANDOM SPECTRUM VARIATION

Assume that $y(t)$ is a random spectrum characterizing some analytical features of a given target product, where is the spectral variable (e.g. wavelength or Raman shift with a possibly infinite number of values). $y(t)$ is assumed to be of finite-energy, i.e. the integral of its squared absolute value is finite. This assumption is necessary for basis representations of $y(t)$.

The generic class-model that accounts for the random variation of $y(t)$ is a functional mean-model, written

$$y(t) = \mu(t) + u(t) \quad (1)$$

and its truncated Karhunen-Loève decomposition [21] is written

$$y(t) \approx \mu(t) + \sum_{r=1}^{R} c_r \varphi_r(t) \quad (2)$$

where $\mu(t)$ is the expected spectrum; $\mu(t)$ is the random deviation of $y(t)$ from $\mu(t)$; $\varphi_r(t)(r = 1, \dots, R)$ is the r-th orthonormal eigenfunctions of the covariance surface of $y(t)$ with associated eigenvalues λ_r ; $R \geq 1$ is the basis truncation point, i.e. number of eigenfunctions selected out of a possibly infinite set to approximate $y(t)$; c_r $(r = 1, \dots, R)$ is the score of $y(t)$ associated with

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

eigenfunction $\varphi_r(t)$; c_r is a zero-mean random variable with variance λ_r, uncorrelated with any other score variable $c_{r'}$ $(r \neq r')$; hence, it may be modelled and predicted in subsequent analyses by assuming a convenient probabilistic model [21,22]. If $y(t)$ is gaussian, then c_r is independent of $c_{r'}$ $(r \neq r')$ and normally distributed,

$$c_r \sim \mathrm{N}(0, \lambda_r) \tag{3}$$

In Equation (3), λ_r is the unknown model parameter to be estimated if $\mu(t)$, $\varphi_r(t)$ and R are fixed or known [23]. In fact, $\mu(t)$, $\varphi_r(t)$and R are also unknown random quantities [22] and hence are considered as hyperparameters upon which the model in Equation (3) is conditioned [22,23]. Hence, they must be estimated from discretized and noisy observations of $y(t)$, and uncertainties in their estimations can reasonably be accounted for by a naïve bootstrap procedure, i.e. by performing PC decompositions on bootstrap samples of the observations [23].

2.4. DESIGN AND BOOTSTRAP MODEL FOR NOISY OBSERVATIONS

Suppose that N independent and noisy realizations of $y(t)$ are observed at discretized values of t . In other words, N spectra (indexed by $i = 1, \dots, N$) are independently sampled from the target class, and the intensities of each spectrum is measured at K ordered and equally spaced values $t_k (k = 1, \dots, K)$ of t, with homoscedastic measurement noise. Let $\boldsymbol{t} = [t_1, \dots, t_K]$ denotes the $1 \times K$ vector of the equispaced values of t, and $\boldsymbol{y}_i = [y_{i1}, \dots, y_{iK}]$ the $1\ x\ K$ vector containing intensities of i-th spectrum. Denote by $\boldsymbol{Y}$ the $N\ x\ K$ signal matrix containing the N spectra $\boldsymbol{y}_i$ $(i = 1, \dots, N)$that are stacked.

To account for uncertainties in the estimations of $\mu(t)$, $\varphi_r(t)$ and R in (2), one can build B bootstrap submatrices (indexed by $b = 1, \dots, B$), each of size $n \times K$, by randomly resampling n out of the N rows of $\boldsymbol{Y}$. Let $\boldsymbol{Y}_b\ (b = 1, \dots, B)$ be one of these bootstrap submatrices, and $\boldsymbol{y}_{ib} = [y_{ib}(t_1), \dots, y_{ib}(t_k)] = [y_{i1b}, \dots, y_{iKb}]$be the i-th row of $\boldsymbol{Y}_b$, with $i = 1, \dots, n$. Denote by R_b the optimal number of meaningful eigenfunctions to approximate $\boldsymbol{y}_{ib}$ given $\boldsymbol{Y}_b$, with $R_b \leq$ minimum (n, K).

Then, given any bootstrap submatrix $\boldsymbol{Y}_b$, the model in (1) becomes

$$y_{ib}(t_k) = \mu_b(t_k) + u_{ib}(t_k) + e_{ib}(t_k) \tag{4}$$

and its principal component (PC) decomposition is

$$y_{ib}(t_k) \approx \mu_b(t_k) + \sum_{r=1}^{R_b} c_{irb}\varphi_{rb}(t_k) + e_{ib}(t_k) \tag{5}$$

Where $\mu_b(t_k)$ is the value of the mean spectrum at t_k; $u_{ib}(t_k)$ is the i-th spectrum random deviation from $\mu_b(t_k)$ at t_k; $e_{ib}(t_k)$ is the white noise at t_k with the assumption that $e_{ib}(t_k) \sim N(0, \sigma_b^2)$, where $\sigma_b^2 > 0$ is the noise variance; $\varphi_{rb}(n)(r = 1, \dots, R_b)$ are the set of R_b orthonormal eigenfunctions with associated eigenvalues λ_{rb} satisfying $\lambda_{1b} \geq \dots \geq \lambda_{R_b b}$ and c_{irb} the corresponding scores for the i-th spectrum; hence, c_{irb} $(r = 1, \dots, R_b)$ are random variables independent across i because of the sampling design and uncorrelated across r , with zero-mean and variancesλ_{rb} . If $y(t)$ is a gaussian process, then

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

$$c_{irb} \sim \mathrm{N}\left(0, \lambda_{rb}\right) \cdot \mathrm{with} \cdot r = 1, \cdots, R_b \tag{6}$$

or more compactly

$$\boldsymbol{c}_{ib} \sim \mathrm{N}\left(0_b, \boldsymbol{\Lambda}_b\right) \tag{7}$$

where 0_b is an $1 \times R_b$ zero-vector and $\boldsymbol{c}_{ib}$ is $1 \times R_b$ the vector of scores for the i-th spectrum; $\boldsymbol{\Lambda}_b$ is an $R_b \times R_b$ diagonal covariance matrix whose elements are the true but unknown eigenvalues $\lambda_{1b}, \ldots, \lambda_{R_b b}$. It is worth to note from (4), (5) that the noise terms $e_{ib}(t_k)$ can be computed from the eigenfunctions and scores as $e_{ib}\left(t_k\right) = \sum_{r=R_b+1}^{n} c_{irb}\varphi_{rb}\left(t_k\right)$. Hence, once the eigendecomposition is estimated and identifiable (unique), then the separation of the signal $u_{ib}(t_k)$ and the noise $e_{ib}(t_k)$ is also identifiable (unique).

2.5. ESTIMATION OF THE MEAN SPECTRUM, EIGENFUNCTIONS AND THE NUMBER EIGENFUNCTIONS

Given the submatrix $\boldsymbol{Y}_b$, first, the mean spectrum is estimated by computing the sample mean at each spectral point t_k in $\boldsymbol{t}$ as $\widehat{\mu}_b\left(t_k\right) = n^{-1}\sum_{i=1}^{n} y_{ib}\left(t_k\right)$ or more compactly as

$$\widehat{\boldsymbol{\mu}}_b = n^{-1} 1_n' \boldsymbol{Y}_b \tag{8}$$

where $\widehat{\boldsymbol{\mu}}_b = \left[\widehat{\mu}_b\left(t_1\right), \cdots, \widehat{\mu}_b\left(t_K\right)\right] = \left[\widehat{\mu}_{1b}, \cdots, \widehat{\mu}_{Kb}\right]$ is the $1 \times K$ vector of estimated means; 1_n is the $n \times 1$ vector whose elements are all 1; n and $\boldsymbol{Y}_b$ are defined as in Section 2.4.

Second, the empirical covariance matrix is computed from the centered submatrix as

$$\widehat{\boldsymbol{\Sigma}}_b = n^{-1}\left[\boldsymbol{Y}_b - \widehat{\boldsymbol{M}}_b\right]'\left[\boldsymbol{Y}_b - \widehat{\boldsymbol{M}}_b\right] \tag{9}$$

where $\widehat{\boldsymbol{\Sigma}}_b$ is the $K \times K$ empirical covariance matrix; $\widehat{\mathbf{M}}_b$ is a $n \times K$ means matrix obtained by stacking n times $\widehat{\boldsymbol{\mu}}_b$; n and $\boldsymbol{Y}_b$ are defined in Section 2.4.

Third, the values of the eigenfunctions at t_k, denoted $\varphi_{rb}(t_k)$ and the associated eigenvalues λ_{rb} are estimated either by a standard PCA approach or by a functional PCA approach [21]. Standard PCA of spectral data is a well-known technique to the chemometric community. It aims at discovering the dominant modes of variations of the spectra. It proceeds either by eigendecomposition of $\widehat{\boldsymbol{\Sigma}}_b$ or singular values decomposition of $\boldsymbol{Y}_b - \widehat{\boldsymbol{M}}_b$, to derive the (unsmoothed) values of the eigenfunctions at t_k, denoted $\hat{\varphi}_{rb}\left(t_k\right)$ for $k = 1, \ldots, K$ and the associated eigenvalues $\hat{\lambda}_{rb}$ of $\widehat{\boldsymbol{\Sigma}}_b$ [21]. Unlike the standard PCA, functional PCA takes a further step by estimating smoothed values of the eigenfunctions at t_k, also denoted $\hat{\varphi}_{rb}\left(t_k\right)$, and the associated eigenvalues, $\hat{\lambda}_{rb}$ [21,24]. Indeed, eigenfunctions estimated through standard PCA, for example by singular values decomposition of $\boldsymbol{Y}_b - \widehat{\boldsymbol{M}}_b$, might exhibit excessive variability, especially in low sample size and noisy spectra contexts. Hence, regularization through smoothing might be needed to prevent overfitting [21,24]. A simple and direct approach to functional PCA, is the so-called two-step functional PCA that consists in smoothing the PCs

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

estimated by singular values decomposition using the well-known smoothing splines technique, and then rescaling the smoothed PCs so that their quadratic integral equals 1 [21,24].

Fourth, the number R_b of meaningful eigenfunctions to retain, i.e. the truncation point of the eigenvalues, is estimated by the rank $\hat{R}_b$ of the centered data matrix $\boldsymbol{Y}_b - \widehat{\boldsymbol{M}}_b$. Indeed, under the assumption of the model in (4) that $e_{ib}(t_k) \sim N(0, \sigma_b^2)$, the rank of $\boldsymbol{Y}_b - \widehat{\boldsymbol{M}}_b$ can be estimated as the number of PCs minimizing the cross-validation prediction error [25]. However, this method is rather computationally intensive. Instead, a generalized cross-validation approach might be used as a faster alternative to approximate R_b[25]. However, it is rather unstable and generally results in an overestimation of the number of meaningful PCs [25]. More recently, Gavish and Donoho [26] proposed a principled and mathematically proven rule to estimate R_b. They established that, if the assumption that $e_{ib}(t_k) \sim N(0, \sigma_b^2)$ holds, the optimal threshold below which eigenvalues can be set to zero, is

$$\hat{\lambda}_o = 2.858 \times median\left(\sqrt{\hat{\lambda}_{1b}}, \cdots, \sqrt{\hat{\lambda}_{nb}}\right) \cdot \text{if} \cdot n = K \tag{10}$$

or

$$\hat{\lambda}_o = \omega(\kappa_0) \times median\left(\sqrt{\hat{\lambda}_{1b}}, \cdots, \sqrt{\hat{\lambda}_{nb}}\right) \cdot \text{if} \cdot n < K \tag{11}$$

where $\hat{\lambda}_o > 0$ is the optimal eigenvalues' hard-thresholding or truncation point; $\omega(\kappa_0)$ is a function of $\kappa_0 = n / K$; n, K are defined as in Section 2.4 and $\hat{\lambda}_{rb}$ (r = 1,..., n) are the estimated eigenvalues. The analytic form of $\omega(\kappa_0)$ is not available, but it can be approximated numerically [27]. Moreover, when a high-precision value of $\omega(\kappa_0)$ is not needed, one can use the approximation $\omega(\kappa_0) = 0.56\kappa_0^3 + 0.95\kappa_0^2 + 1.82\kappa_0 + 1.43$. This thresholding rule has been demonstrated to adapt to the unknown rank and the unknown noise level in an optimal manner and is more and more recommended to recover unknown rectangular signal matrices in the presence of unknown noise [[26], [27], [28]]. In the present study, it has been used to estimate R_b by the rank $\hat{R}_b$ of $\boldsymbol{Y}_b - \widehat{\boldsymbol{M}}_b$, assuming the number of components to extract is a priori completely unknown.

The above described standard PCA and functional PCA as well as the method to select the number of meaningful PCs are implemented with the *fpca2s* function of the *refund* package [24,28] of R statistical software environment [29], which outputs as results of the analysis:

- $\hat{R}_b$, the estimated number meaningful PCs;
- $\widehat{\boldsymbol{\mu}}_b = [\hat{\mu}_{1b}, \cdots, \hat{\mu}_{Kb}]$, the vector of the estimated mean signal values;
- $\hat{\varphi}_b = \left[\hat{\varphi}'_{1b}, \cdots, \hat{\varphi}'_{\hat{R}_b b}\right]$, the $K \times \hat{R}_b$ matrix containing values of either the unsmoothed or smoothed eigenfunctions at the K spectral points, i.e. $\hat{\varphi}_{rb} = [\hat{\varphi}_{rb}(t_1), \cdots, \hat{\varphi}_{rb}(t_K)]$
- $\hat{\boldsymbol{\Lambda}}_b = \text{diag}\left(\hat{\lambda}_{1b}, \cdots, \hat{\lambda}_{\hat{R}_b b}\right)$, the $\hat{R}_b \times \hat{R}_b$ diagonal matrix of estimated eigenvalues.

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

2.6. ESTIMATION OF THE SCORES

Given the values of the eigenfunctions, $\hat{\varphi}_b$, the scores of the i-th mean-centered spectrum are estimated either by its projection onto eigenfunctions estimated by standard PCA, as

$$\boldsymbol{c}_{ib} = (\boldsymbol{y}_{ib} - \hat{\boldsymbol{\mu}}_b)\,\hat{\varphi}_b \tag{12}$$

or by a least square estimator whose design matrix is the eigenfunctions' matrix estimated by functional PCA [24], as

$$\boldsymbol{c}'_{ib} = \left(\hat{\varphi}'_b\hat{\varphi}_b\right)^{-1}\hat{\varphi}'_b(\boldsymbol{y}_{ib} - \hat{\boldsymbol{\mu}}_b)' \tag{13}$$

where $\boldsymbol{c}_{ib} = \left[c_{i1b}, \cdots, c_{i\hat{R}_b b}\right]$ is the $1 \times \hat{R}_b$ vector of estimated scores for the i-th spectrum; $\boldsymbol{y}_{ib}$ is defined as in Section 2.4; $\hat{\boldsymbol{\mu}}_b$ and $\hat{\varphi}_b$ are defined as in Section 2.5.

2.7. ESTIMATION OF THE POSTERIOR PREDICTIVE DISTRIBUTION OF THE SCORES

Given the values of the eigenfunctions, $\hat{\varphi}_b$, the scores $\boldsymbol{c}_{ib}$ are random vectors, and under the assumptions of models in Equations (3), (4) that $y(t)$ is a gaussian process, they are modelled with a multivariate zero-mean normal model as

$$\boldsymbol{c}_{ib} \sim \text{Normal}\,(0_b, \boldsymbol{\Lambda}_b) \tag{14}$$

where $\boldsymbol{c}_{ib}$ is the scores' vector of the i-th spectrum, defined as in Equation (12) or (13); 0_b is an $1 \times \hat{R}_b$ zero-vector; $\boldsymbol{\Lambda}_b$ is an $\hat{R}_b \times \hat{R}_b$ diagonal covariance matrix whose elements are the model parameters, i.e. the true but unknown eigenvalues λ_{1b}, …, $\lambda_{\hat{R}_b b}$.

The model in Equation (14) is fitted using a Bayesian approach to enable to properly predict the scores, accounting conveniently for the uncertainty about the model parameter $\boldsymbol{\Lambda}_b$ through the so-called posterior predictive distribution of future scores' vectors [20]. In a nutshell, the Bayesian approach to the model in (14) with unknown parameter $\boldsymbol{\Lambda}_b$ uses the Bayes theorem to estimate the probability distribution of $\boldsymbol{\Lambda}_b$ given the observed scores and eigenfunctions (known as posterior distribution of $\boldsymbol{\Lambda}_b$), as the product of the scores' likelihood and a prior probability density of $\boldsymbol{\Lambda}_b$ that encodes any prior knowledge about the unknown eigenvalues [20]. This posterior distribution accounts for uncertainties about $\boldsymbol{\Lambda}_b$ given the observed scores and estimated eigenfunctions. Propagating that uncertainty to the scores in Equation (14) enables to derive the posterior predictive distribution of the scores' vector of a future spectrum given the observed scores and estimated eigenfunctions.

A multivariate normal likelihood of scores vectors combined with independent non-informative Jeffreys' priors on the inverse of the eigenvalues (i.e. on the reciprocal of each diagonal element of $\boldsymbol{\Lambda}_b$) to encode a 'complete' lack of prior knowledge about $\boldsymbol{\Lambda}_b$ may be used [20]. It is well-established in statistical textbooks [20] that the analytic expression of the posterior predictive distribution of the scores' vector, denoted $\tilde{\boldsymbol{c}}_b$, of any future spectrum from the same spectral population is a product of independent univariate Student-t distributions each with n degrees of freedom, written as

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

$$(\tilde{\boldsymbol{c}}_b|\boldsymbol{C}_b,\widehat{\varphi}_b) \sim \prod_{r=1}^{\widehat{R}_b} \text{Student}_n\left(0,\widehat{\lambda}_{rb}\right) \quad (15)$$

where $\tilde{\boldsymbol{c}}_b = \left[\tilde{c}_{1b},\cdots,\tilde{c}_{\widehat{R}_b b},\right]$ is a $1 \times \widehat{R}_b$ scores' vector for a future spectrum; $\boldsymbol{C}_b = [\boldsymbol{c}'_{1b},\cdots,\boldsymbol{c}'_{nb}]' = \left[\boldsymbol{c}_b^{(1)},\cdots,\boldsymbol{c}_b^{\left(\widehat{R}_b\right)}\right]$ of dimension $n \times \widehat{R}_b$ is the scores' matrix, with rows $\boldsymbol{c}_{ib}$ of size $1 \times \widehat{R}_b$ and columns $\boldsymbol{c}_b^{(r)}$ of size $n \times 1$; $\widehat{\lambda}_{rb}$ is the r-th diagonal element the eigenvalues' matrix defined in Section 2.5 and estimated as $\widehat{\boldsymbol{\Lambda}}_b = \text{diag}\left(\widehat{\lambda}_{1b},\cdots,\widehat{\lambda}_{\widehat{R}_b b}\right) = n^{-1}\boldsymbol{C}'_b\boldsymbol{C}_b$ is the $\widehat{R}_b \times \widehat{R}_b$; the symbolism $X \mid Z \sim P$ means variable X given Z is distributed as P . Note that because of independence of the Student-t distributions in (15), each score variable can be predicted independently from the others, by a univariate Student's *t* distribution with *n* degrees of freedom as

$$\left(\tilde{c}_{rb}\middle|\boldsymbol{c}_b^{(r)},\widehat{\varphi}_{rb}\right) \sim \text{Student}_n\left(0,\widehat{\lambda}_{rb}\right) \quad (16)$$

Where $\tilde{c}_{rb}$ is the predicted score for the r-th PC; $\boldsymbol{c}_b^{(r)}$ is the r-th column of the scores' matrix $\boldsymbol{C}_b$ as defined in (15); $\widehat{\varphi}_{rb}$ of size $1 \times K$ is the r-th PC as defined in Section 2.5; $\widehat{\lambda}_{rb} = n^{-1}\boldsymbol{c}_b^{(r)\prime}\boldsymbol{c}_b^{(r)}$ is the eigenvalue or variance of scores on the r-th PC. Monte Carlo (MC) samples of the posterior predictive distribution in (15) or (16) can be quickly generated in any statistical computing environment, e.g. by the rt function in R [29].

Non-informative Jeffreys' priors on reciprocal of eigenvalues might have detrimental effects on inferences and predictions, especially if true eigenvalues are low or close to zero, or the sample size is too small [20]. Alternative choices of priors include the independent non-informative uniform priors denoted Uniform(0,+∞), vague priors (e.g. a half-normal distribution with variance 100, denoted Normal+(0,100)) and weakly informative priors (e.g. a half-normal distribution with variance 1, denoted Normal$_+$(0,1)) on the square-root of the eigenvalues, i.e. the square-root of the diagonal elements of $\boldsymbol{\Lambda}_b$ in (14) [20]. For each of these priors on square-root of eigenvalues, the resulting posterior predictive distributions of $\tilde{\boldsymbol{c}}_b$ is not analytically tractable, but can be validly approximated using well-known Markov Chain Monte Carlo (MCMC) procedures such as the Hamiltonian Monte Carlo (HMC) available in *Stan* programming language [20,30].

As in any Bayesian analysis, the model must be thoroughly criticized and validated using recommended techniques such as the diagnosis of the quality of MCMC chains (convergence and mixing) for each element of $\boldsymbol{\Lambda}_b$ if an MCMC procedure is used to approximate the posterior distribution of $\boldsymbol{\Lambda}_b$ and the posterior predictive check [20,30]. Especially, the posterior predictive check is a crucial indicator of the adequacy of the Bayesian prediction model; it has to provide strong evidence that the predictions of scores on each PC are consistent with the observed scores. Simple graphical visualizations (e.g. boxplots) comparing the distributions of both predicted and observed scores may be used [20,30] (see an example in Appendix A.2).

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

2.8. ESTIMATION OF THE PREDICTIVE DISTRIBUTION OF A FUTURE SPECTRUM

Suppose that, given the values of the eigenfunctions, $\hat{\varphi}_b$, M predicted scores vectors, each of size $1 \times \hat{R}_b$ and denoted $\tilde{\boldsymbol{c}}_{mb}$ (m = 1,…,M) have been drawn from the posterior predictive distribution of future scores by a MC sampling from (15), (16) or a HMC sampling, and stacked resulting in a matrix of predicted scores denoted $\widetilde{\boldsymbol{C}}_b$ of dimension $M \times \hat{R}_b$. Then, the posterior predictive distribution of any future spectrum $\tilde{\boldsymbol{y}}$ given $\hat{\varphi}_b$ and $\boldsymbol{Y}_b$, can be approximated by first back-transforming the predicted scores matrix, $\widetilde{\boldsymbol{C}}_b$, to the original signal scale and then adding the mean matrix, $\widetilde{\boldsymbol{M}}_b$, i.e.

$$\widetilde{\boldsymbol{Y}}_b = \widetilde{\boldsymbol{M}}_b + \widetilde{\boldsymbol{C}}_b \hat{\varphi}_b' \qquad (17)$$

Where $\widetilde{\boldsymbol{Y}}_b$ is the $M \times K$ matrix of predicted spectra at bootstrap iterations b ; $\widetilde{\boldsymbol{M}}_b$ is the $M \times K$ mean matrix obtained by stacking M times $\tilde{\boldsymbol{\mu}}_b$.

Eventually, the predictive distribution of $\tilde{\boldsymbol{y}}$ given $\boldsymbol{Y}$, accounting for uncertainties about the mean, the eigenfunctions and the number eigenfunctions and the eigenvalues, is approximated by pooling together the predictions from the B boostrap iterations, i.e.

$$\widetilde{\boldsymbol{Y}} = \left[\widetilde{\boldsymbol{Y}}_1', \cdots, \widetilde{\boldsymbol{Y}}_B'\right]' \qquad (18)$$

where $\widetilde{\boldsymbol{Y}}$ is the $(BnM) \times K$ matrix of predicted trajectories for $\tilde{\boldsymbol{y}}$. This predictive distribution simulates plausible trajectories of future spectra from the same spectral population that might be expected given the observed spectral dataset.

2.9. ESTIMATION OF THE PREDICTION BAND LIMITS

Let $\tilde{\boldsymbol{y}}_m$ of size $1 \times K$ denotes the m-th row of the predicted spectral matrix $\widetilde{\boldsymbol{Y}}$ and $\tilde{\boldsymbol{y}}_k$ of size $(BnM) \times 1$ denotes the k-th column of $\widetilde{\boldsymbol{Y}}$. Once, the matrix of predicted spectra $\widetilde{\boldsymbol{Y}}$ is obtained, two types of β%-prediction bands may be constructed using the most central regions of the predicted spectra. On one hand, a pointwise band is the simplest. It is built by computing at each spectral point t_k, the (100 - β)/2 and (100 + β)/2 quantiles of $\tilde{\boldsymbol{y}}_k$, the k-th column of $\widetilde{\boldsymbol{Y}}$. However, it is well-known that the resulting pointwise band limits do not necessarily guarantee the intended nominal coverage of $\beta\%$, due to multiplicity testing over a potentially large set of spectral variables t_k [20,23]. On the other hand, a simultaneous band is more likely to guarantee the intended nominal coverage of $\beta\%$ [20,23]. Our proposed approach to construct such a band is to use the non-parametric concept of functional band depth [15,31] to perform a center-outwards ranking of the predicted spectra. Then, the boundaries of the $\beta\%$ most central spectra are used as limits of the prediction band. Specifically, the so-called Modified Band Depth (MBD) can be used [15]. Briefly, the MBD of any predicted spectrum $\tilde{\boldsymbol{y}}_m$ w.r.t the predicted spectra, $\widetilde{\boldsymbol{Y}}$, denoted $\tilde{d}_m$, is defined as the average proportion of spectral points at which $\tilde{\boldsymbol{y}}_m$ falls inside bands delimited by all pairs of spectra of $\widetilde{\boldsymbol{Y}}$ [15,31,32]. It is calculated as follows: first, a pair of spectra (rows) is selected from $\widetilde{\boldsymbol{Y}}$, say $\tilde{\boldsymbol{y}}_{m1}$ and $\tilde{\boldsymbol{y}}_{m2}$, with $1 < m1 < m2 < (M \cdot B)$; second, the proportion of spectral points where $\tilde{\boldsymbol{y}}_m$ is inside $\tilde{\boldsymbol{y}}_{m1}$ and $\tilde{\boldsymbol{y}}_{m2}$ is calculated; third, this proportion is calculated for all pairs of spectra of $\widetilde{\boldsymbol{Y}}$ and then averaged. Mathematically, this is written

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

$$\tilde{d}_m = K^{-1}\binom{M \cdot B}{2}^{-1} \sum_{m_1<m_2} \sum_{k=1}^{K} I\left(\tilde{y}_{m_1 k} \leq \tilde{y}_{mk} \leq \tilde{y}_{m_2 k}\right) \quad (19)$$

where $\tilde{y}_{m1k}$, $\tilde{y}_{m1k}$ and $\tilde{y}_{mk}$ are the k-th elements of $1\ x\ K$ the predicted row-vectors $\tilde{\boldsymbol{y}}_{m1}$, $\tilde{\boldsymbol{y}}_{m2}$ and $\tilde{\boldsymbol{y}}_m$ of $\tilde{\boldsymbol{Y}}$ respectively. The MBD values $\tilde{d}_m$, can be obtained for all predicted spectra $\tilde{\boldsymbol{y}}_m$ $(m = 1, \ldots, M \cdot B)$, resulting in an $(M \cdot B)$-vector of depth values, denoted $\tilde{\boldsymbol{d}}$. This is done efficiently in a few seconds with the package *roahd* [31,32] in R software [29]. Based on these depth values, the predicted spectra are ranked from the deepest having the highest depth value to the most outlying having the lowest depth value w.r.t $\tilde{\boldsymbol{Y}}$ [15,31,32]. Then, the $\beta\%$ (say 95%) deepest predicted spectra are selected as typical or regular spectra and the band defined by their range of variation is used as prediction band.

Let $\tilde{\boldsymbol{y}}_l = [\tilde{y}_{l1}, \ldots, \tilde{y}_{lK}]$ and $\tilde{\boldsymbol{y}}_u = [\tilde{y}_{u1}, \ldots, \tilde{y}_{uK}]$ be respectively the lower and upper $1 \times K$ boundary-vectors of the obtained band. They may be interpreted as critical or limiting trajectories for typical spectra of the target class.

2.10. TEST OF CONFORMITY OF UNKNOWN SPECTRA

Let $z(t)$ be a new unknown spectrum to be tested and $\boldsymbol{z}$ its values at $\boldsymbol{t}$, i.e. $\boldsymbol{z} = [z(t_1), \cdots, z(t_K)] = [z_1, \cdots, z_K]$. The scores of $\boldsymbol{z}$ are predicted in each of the B eigenfunctions-basis to account for uncertainties in its predictions. These scores are computed in the case of the standard PCA as

$$\tilde{\boldsymbol{c}}_{zb} = (\boldsymbol{z} - \hat{\boldsymbol{\mu}}_b)\,\hat{\varphi}_b \quad (20)$$

or in the case of functional PCA using the least square estimator as

$$\tilde{\boldsymbol{c}}'_{zb} = \left(\hat{\varphi}'_b \hat{\varphi}_b\right)^{-1} \hat{\varphi}'_b (\boldsymbol{z} - \hat{\boldsymbol{\mu}}_b)' \quad (21)$$

Where $\tilde{\boldsymbol{c}}_{zb}$ is the $1 \times \hat{R}_b$ vector of predicted scores for $\boldsymbol{z}$; $\hat{\boldsymbol{\mu}}_b$ and $\hat{\varphi}_b$ are defined as in Sections 2.5.

$\boldsymbol{z}$ is then reconstructed in each eigenfunctions-basis resulting in B predictions accounting for the uncertainty in its reconstruction as

$$\tilde{\boldsymbol{z}}_b = \hat{\boldsymbol{\mu}}_b + \tilde{\boldsymbol{c}}_{zb}\hat{\varphi}'_b \quad (22)$$

where $\tilde{\boldsymbol{z}}_b = [\tilde{z}_b(t_1), \cdots, \tilde{z}_b(t_K)] = [\tilde{z}_{1b}, \cdots, \tilde{z}_{Kb}]$ is the $1 \times K$ vector of reconstructed values $\tilde{z}_{kb}$ of $\boldsymbol{z}$ at t_k $(k = 1, \ldots, K)$ given $\hat{\varphi}_b (b = 1, \ldots, B)$.

Because of these uncertainties in the reconstruction of $\boldsymbol{z}$, three probabilistic testing procedures are proposed to decide whether $\boldsymbol{z}$ is inside the band limits or not. These three procedures aim at evaluating how likely the reconstructions of $\boldsymbol{z}$, $\tilde{z}_b (b = 1, \ldots, B)$, overlap with the $\beta\%$-prediction band. The first and most rigorous procedure computes a bootstrap probability of overlapping with the band, denoted π_1, as the proportion of times $\hat{\boldsymbol{z}}_b$ is entirely inside the prediction band, as

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

$$\pi_1 = B^{-1} \sum_{b=1}^{B} I\left[\tilde{y}_{lk} \leq \tilde{z}_{kb} \leq \tilde{y}_{uk}, \text{for all } k = 1, \cdots, K\right] \quad (23)$$

where $I[\cdot]$ is the indicator function; $\tilde{z}_{kb}$ is the reconstructed value of $\boldsymbol{z}$ at t_k in eigenfunctions-basis b as defined as in Equation (22); $\tilde{y}_{lk}$ and $\tilde{y}_{uk}$ are the lower and upper band limits at t_k as defined in Section 2.9. $z(t)$ is accepted as the conforming to $y(t)$, if π_1 is greater than or equal to a threshold, say κ_1, with $0 < \kappa_1 \leq 1$. For example, $\kappa_1 = 0.50$ means that $z(t)$ is accepted if at least 50% of its reconstructions in the B eigenfunctions-basis are entirely inside the band.

The second procedure consists in computing a probability of conformance at each spectral point t_k, denoted π_{2k} $(k = 1, \ldots, K)$, as the proportion of times when $\tilde{z}_{kb}$ $(b = 1, \ldots, B)$ is between $\tilde{y}_{lk}$ and $\tilde{y}_{uk}$, as

$$\pi_{2k} = B^{-1} \sum_{b=1}^{B} I\left(\tilde{y}_{lk} \leq \tilde{z}_{kb} \leq \tilde{y}_{uk}\right), \text{.for } k = 1, \cdots, K \quad (24)$$

where $I[\cdot]$ is the indicator function; $\tilde{z}_{kb}$ is the reconstructed value of $\boldsymbol{z}$ at t_k in eigenfunctions-basis b as defined as in Equation (22); $\tilde{y}_{lk}$ and $\tilde{y}_{uk}$ are the lower and upper band limits at t_k as defined in Section 2.9. $z(t)$ is accepted as the conforming to $y(t)$ if π_{2k} is greater than a threshold, say κ_2, at all spectral points t_1,..., t_K, with $0 < \kappa_2 \leq 1$. For example, $\kappa_2 = 0.50$ means that $z(t)$ is accepted if the pointwise medians of its reconstructions at each point t_k fall inside the band. This procedure is less rigorous than the previous one and might be used to set different probabilities of acceptance at different spectral points.

The third procedure consists in computing the average proportion of spectral points, denoted π_3, where $\tilde{z}_b$ is inside the band as

$$\pi_3 = B^{-1} K^{-1} \sum_{b=1}^{B} \sum_{k=1}^{K} I\left(\tilde{y}_{lk} \leq \tilde{z}_{kb} \leq \tilde{y}_{uk}\right) \quad (25)$$

where $I[\cdot]$ is the indicator function; $\tilde{z}_{kb}$ is the reconstructed value of $\boldsymbol{z}$ at t_k in eigenfunctions-basis b as defined as in Equation (22); $\tilde{y}_{lk}$ and $\tilde{y}_{uk}$ are the lower and upper band limits at t_k as defined in Section 2.9. $z(t)$ is accepted as the conforming to $y(t)$ if π_3 is greater than a threshold, say $(K - \kappa_3)/K$, where κ_3 is the average number of tolerable spectral points where $\tilde{z}_b$ $(b = 1, \ldots, B)$ is outside the band. This procedure is similar to decision rules used to enhance decision-making in control charts.

3. Experimental section

3.1. DATASETS AND PREPROCESSING

The evaluation of classification performances of the proposed β%-prediction band was done in two studies, with four real NIR datasets. These four datasets were measured with two different pieces of portable device (Table 1) on two categories of model-formulations namely five paracetamol-based formulations and five ibuprofen-based formulations (Table 2). Table 1 describes the two pieces of device. The first piece of device, denoted Device 1, outputs a signal with 100 points while the second

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

piece of device, denoted Device 2, outputs a signal at higher resolution with 256 points. Table 2 describes the four datasets including the drug formulations involved, the numbers of batches and samples' sizes per piece of device and formulation. The two categories of model-formulations were studied by Ciza and coworkers as two challenging situations in vibrational spectroscopy [33]. They differ in the ratio of active pharmaceutical ingredient (API) to excipients contents denoted API-excipients ratio. Firstly, paracetamol-based formulations were all tablets having high dosages of paracetamol and relatively low dosages of other compounds (high API-excipients ratio) [33]. Secondly, ibuprofen-based formulations were also tablets but with more balanced API-excipients ratio. They differs among themselves in their dosages of API, but also in their colors (e.g. white, pink) and coating nature (e.g. sugar coating) [33]. The drug samples were collected from local pharmacies in Belgium. All spectra were measured on tablets through their original blister [33]. The first two datasets were measured with the first piece of device (Device 1, Table 1) on the five paracetamol-based formulations (Dataset 1, illustrated on Fig. 2A–E) and the five ibuprofen-based formulations (Dataset 2, illustrated on Fig. 2F–J). The last two datasets were also measured on the five paracetamol-based formulations (Dataset 3, illustrated on Figs. 3A-2E) and the five ibuprofen-based formulations (Dataset 4, illustrated on Fig. 3F–J), but with the second piece of device with higher resolution (Device 2, Table 1). It is worth noting that, many of the datasets were highly dispersed and heterogeneous, mostly because of blister effect, measurement noise and a possible inter-batch variability (e.g. of Fig. 2A and B). They were chosen as limiting cases to evaluate the performances of the prediction band. More homogeneous datasets may include process analytical technology (PAT) datasets measured with benchtop device with higher resolution and less noise.

Table 1. Description of the two near infrared spectrophotometers used to measure spectra, the features of the output signal and applied chemometric preprocessing.

Device Number	Device 1	Device 2
Equipment features		
Model name and manufacturer	MicroPhazir® of ThermoFisher Inc.	NIR-S-G1® of InnoSpectra Corp.
Spectral range analyzed	6266.8–4173.1 cm^{-1}	900–1700 nm
Number of spectral variables	100	256
Preprocessing		
Selected spectral range	6133.1–4238.2 cm^{-1}	9143–6399 cm^{-1}
Number selected of spectral variables	91 (over 100)	128 (over 256)
Smoothing	Savitzky-Golay: • Window size = 3 • Polynomial order = 2 • Derivative order = 1	Savitzky-Golay: • Window size = 5 • Polynomial order = 2 • Derivative order = 1
Normalization	Standard Normal Variate (SNV)	Standard Normal Variate (SNV)

Regarding the preprocessing of the signal, all spectra were smoothed using the Savitzky-Golay smoother [34] parametrized for each piece of device as reported in Table 1, followed by the standard normal variate (SNV) normalization.

The methodologies of the two evaluation studies are described in Section 3.2 Study 1: Simulations to evaluate the statistical properties of the $\beta\%$-prediction band with datasets 1 and 2, 3.3 Study 2: applications to two verification problems with datasets 3 and 4. The first study involves Monte Carlo validations with Datasets 1 and 2 to study some statistical properties of the prediction band, in comparison with the benchmark SIMCA model. The second study involves external validations and application cases with Datasets 3 and 4.

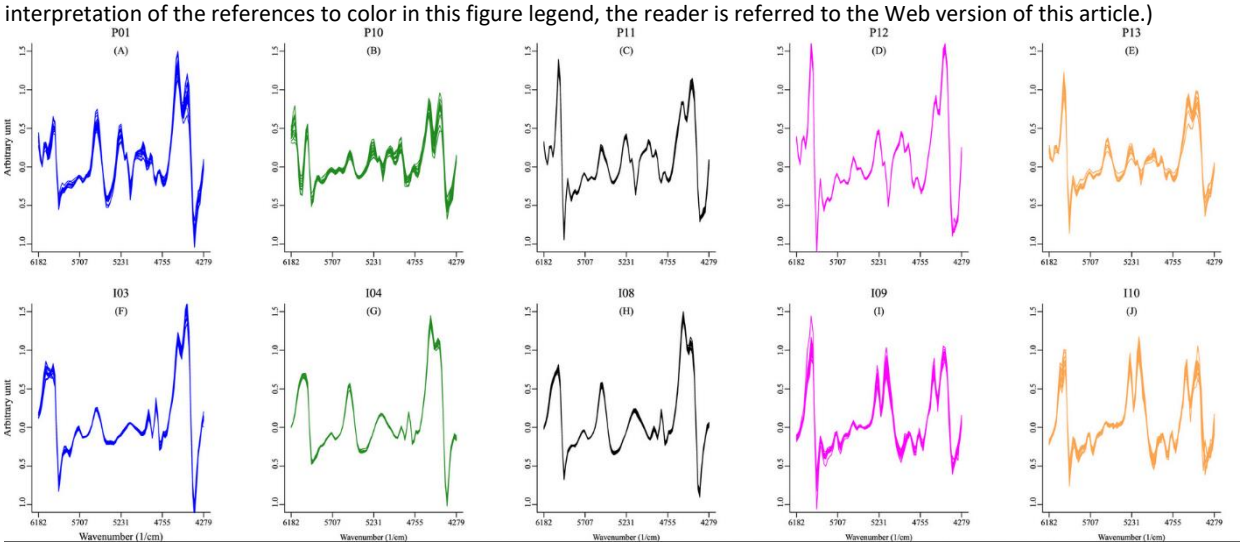
Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

Table 2. Description of the four near-infrared datasets used in Monte Carlo or external validation studies.

Descriptors					
Dataset 1 (Device 1)					
Drug formulations codes	P01	P10	P11	P12	P13
API (dosage in mg in brackets) and commercial name	Paracetamol (1000) Dafalgan™	Paracetamol (250), acetylsalicylic acid (250), caffeine (65) Excedryn™	Paracetamol (500), caffeine (65) Panadol plus™	Paracetamol (500) Mylan	Paracetamol (325), tramadol (37.5) EuroGenerics
Number of batches (spectra per batch in brackets)	04 (17–20)	04 (20)	03 (20)	01 (20)	02 (20)
Dataset 2 (Device 1)					
Drug formulations codes	I03	I04	I08	I09	I10
API (dosage in mg in brackets) and commercial name	Ibuprofen (400) EuroGenerics	Ibuprofen (600) EuroGenerics	Ibuprofen (600) TEVA	Ibuprofen (200) Nurofen	Ibuprofen (400) Nurofen
Number of batches (spectra per batch in brackets)	04 (20)	04 (20)	01 (20)	04 (20)	03 (20)
Dataset 3 (Device 2)					
Drug formulations codes	P01	P10	P11	P12	P13
API (dosage in mg in brackets) and commercial name	Paracetamol (1000) Dafalgan™	Paracetamol (250), acetylsalicylic acid (250), caffeine (65) Excedryn™	Paracetamol (500), caffeine (65) Panadol plus™	Paracetamol (500) Mylan	Paracetamol (325), tramadol (37.5) EuroGenerics
Number of batches (spectra per batch in brackets)	06 (20)	04 (10)	03 (10)	01 (10)	02 (10)
Dataset 4 (Device 2)					
Drug formulations codes	I03	I04	I08	I09	I10
API (dosage in mg in brackets) and commercial name	Ibuprofen (400) EuroGenerics	Ibuprofen (600) EuroGenerics	Ibuprofen (600) TEVA	Ibuprofen (200) Nurofen	Ibuprofen (400) Nurofen
Number of batches (spectra per batch in brackets)	06 (20)	04 (10)	01 (10)	04 (10)	03 (10)

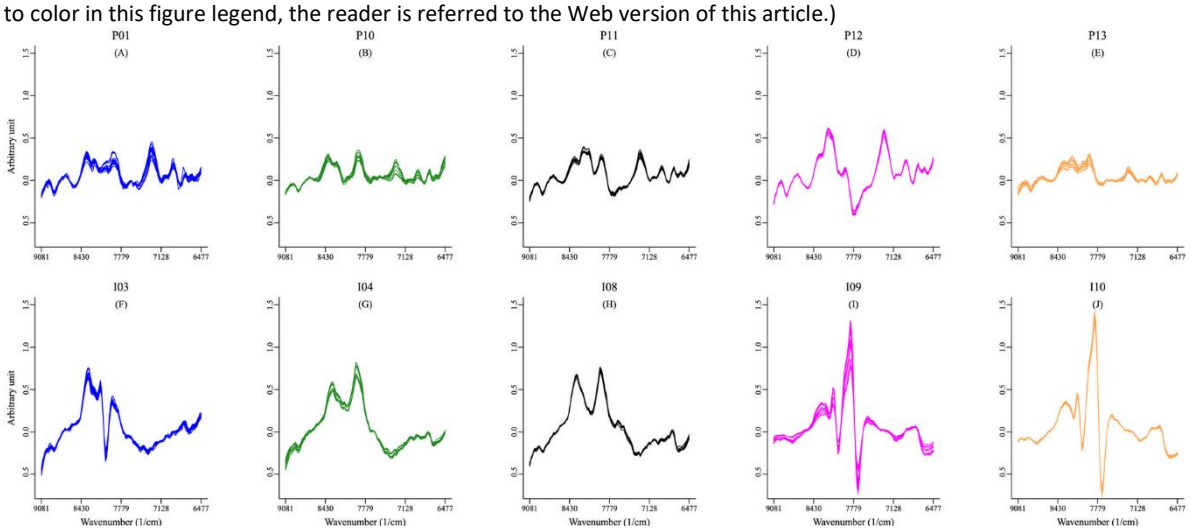
Note: (a) API: Active pharmaceutical ingredient.

Fig. 2. Samples of NIR (MicroPhazir® spectrophotometer) spectra of five paracetamol formulations (Dataset 1, panels A–E) and five Ibuprofen formulations (Dataset 2, panels F–J). Blue spectra represent the target classes; raw spectra were preprocessed by the Savitzky-Golay smoothing and SNV-normalization; see Table 1 for the detailed descriptions of the spectrophotometer and processing of the signal, and Table 2 for the drug formulations and the sample sizes. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)



Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

Fig. 3. Samples of NIR (NIR-S-G1® spectrophotometer) spectra of five paracetamol formulations (panels A–E, Dataset 3) and five Ibuprofen formulations (panels F–J, Dataset 4); blue spectra represent the target classes; raw spectra were preprocessed by the Savitzky-Golay smoothing and SNV-normalization; see Table 1 for the detailed descriptions of the spectrophotometer and processing of the signal, and Table 2 for the drug formulations and the sample sizes. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)



3.2. STUDY 1: SIMULATIONS TO EVALUATE THE STATISTICAL PROPERTIES OF THE β%-PREDICTION BAND WITH DATASETS 1 AND 2

The first study consists in Monte Carlo validations to evaluate five key classification properties of the proposed β%-prediction band and to compare them to the benchmark class-modelling model namely the SIMCA model. Datasets 1 and 2 (Table 2, Fig. 2) involving the two categories of formulations measured with Device 1 were used. Formulations coded as P01 (Dataset 1, Fig. 2A) and I03 (Dataset 2, Fig. 2F) were used as target classes for the paracetamol-based and the ibuprofen-based formulations respectively. The corresponding non-target classes were respectively formulations P10, P11, P12 and P13 for the paracetamols (Fig. 2B–E) and I04, I08, I09 and I10 for the ibuprofens (Fig. 2G–J, Table 2). The evaluated properties were as follows:

- Property 1 involves the coverage rate (i.e. average true positive rate) and the type 2 error rates (i.e. the average false positive rates) of the prediction band for each of the three testing procedures for new spectra described in Section 2.10, compared with those of the benchmark SIMCA model; it is important to mention that, for any probabilistic one-class classifier to be optimal, the coverage rate should agree with the *a priori* defined theoretical β%-confidence of band, while the type 2 error rates should be as close to zero as possible [2,7,8];
- Property 2 involves the sensitivity of the coverage and type 2 error rates of the prediction band to the type of PC decomposition, i.e. whether the smoothed PC decomposition enhances the classification performances of the prediction band;
- Property 3 involves the sensitivity of the coverage and type 2 error rates of the prediction band to the choice of the prior probability distribution on the singular values or eigenvalues for the Bayesian zero-mean model on the PCs' scores; this sensitivity analysis is recommended for any Bayesian prediction model [20]; ideally the performances of the prediction band should remain stable under different types of prior models on the singular values or eigenvalues [20];

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

- Property 4 involves the interpretability of the classification results of the prediction band, compared to the benchmark SIMCA model; and
- Property 5 involves the average computation time required for practical implementation.

To evaluate these properties, the study proceeds as follows for each of the two datasets or categories of formulations. In a first step, four 95%-prediction bands were built for the target formulation (either P01 or I03) using standard PC decomposition on a given calibration set. Each band is built with one of the four prior models described in Section 2.8 to evaluate the sensitivity of the performances of the prediction band to the choice of the prior model. These include the independent non-informative Jeffreys' priors on the reciprocal of the eigenvalues, and the independent non-informative Uniform(0,+∞), vague Normal$_+$(0,100) and weakly informative Normal$_+$(0,1) priors on the square-root of the eigenvalues. For each of the four resulting models, $B = 30$ bootstrap resamples of the calibration set were used to account for uncertainties in the PC decomposition and $M = 2000$ valid MC or MCMC samples of the scores' vector were generated at each bootstrap iteration, not including the burn-in or warm-up phase of 1000 samples when the MCMC (HMC) sampling is used. Hence, a total of 60,000 MC or MCMC valid predictions were obtained for each of the four models.

Regarding the calibration and test sets, they were defined as follows. For each target formulation (P01 or I03), half of the spectra was randomly sampled without replacement from the pooled batches and used as calibration set for the four models described above. Pooling the batches before sampling enabled to account for inter-batch variability. Although this added variability would be better accounted for through the use of mixed-effect models, implementing such models is non-trivial, hence it is left out of the scope of this manuscript. The remaining half of the data was used as test set to evaluate the true positive rates. All the spectra of batches of the non-target formulations (P10, P11, P12, P13 for the paracetamols and I04, I08, I09, I10 for the ibuprofens) were used as test sets to evaluate the false positive rates per non-target formulation. To accept or reject any single test spectrum, the three testing procedures defined in Section 2.10 were considered:

- Procedure 1 accepts a spectrum if at least $\kappa_1 x\ 100\%$ of its reconstructions in the $B = 30$ bootstrap PC-bases falls entirely inside the band limits; for the present study, $\kappa_1 = 0.90$ was used as threshold to decide acceptance of future spectra. Values of 0.50, 0.75 and 1.0 were also used to investigate the effect of lower or higher thresholds on model performances;
- Procedure 2 accepts a spectrum if at each spectral point, at least $\kappa_2 x\ 100\%$ of its $B = 30$ reconstructions in the bootstrap PC-bases falls inside the band limits; for the present study, $\kappa_2 = 0.90$ was used as threshold to decide acceptance of future test spectra. Values of 0.50, 0.75 and 1.0 were also used to investigate the effect of lower or higher thresholds on model performances;
- Procedure 3 accepts a spectrum if its reconstructions in the $B = 30$ bootstrap PC-bases have at most κ_3 points outside the band limits on average; for the present study, $\kappa_3 = 1$ was used as threshold to decide acceptance of future test spectra. Values of 0, 2 and 3 were also used to investigate the effect of the effect of lower or higher thresholds on model performances.

In a second step, the four bands were built with the same calibration datasets using the functional or smoothed PC decomposition with the same prior models as for the standard PC decomposition to evaluate whether smoothing the eigenfunctions impacts the classification performances.

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

In a third step, to provide comparison with the SIMCA model, four rigorous SIMCA methods differing in the method of estimation of critical limits for the T^2 and Q statistics were built for each dataset. These four variants reflect the most recent and important improvements of the original SIMCA [2,12]. They were all built with an *a priori* confidence level fixed at 95%, that is a content $\beta\% = 95\%$ [7,36]. The only parameter affecting their performances was the number of PCs which was optimized by maximizing the sensitivity in leave-one-out cross-validation (CV). The first method, herein termed the Jackson-Mudholkar method, uses the Hotelling's T-squared distribution and the Jackson-Mudholkar approximation to define the critical limits for the T^2 and Q statistics respectively [35,37]. The second method herein termed the Chi-square method uses the Hotelling's T-squared and the scaled Chi-square distributions respectively on the T^2 and Q statistics to define their critical limits [6,37]. The third and fourth methods are the so-called data-driven methods which use two different scaled Chi-square distributions on the T^2 and Q statistics to define their critical limits [4,6,37]. The third method uses moment-based estimators to estimate the parameters (degree of freedom and scaling factors) of each distribution from the calibration data, whereas the fourth method uses robust estimators to estimate these parameters [6,36]. These four SIMCA methods were all implemented with the function *simca* of the package *mdatools* [37] of R software [29].

In a fourth step, all the previous steps were repeated 100 times to evaluate the bias and variance of the classification performances. The average and standard deviations of the true and false positive rates were reported and discussed.

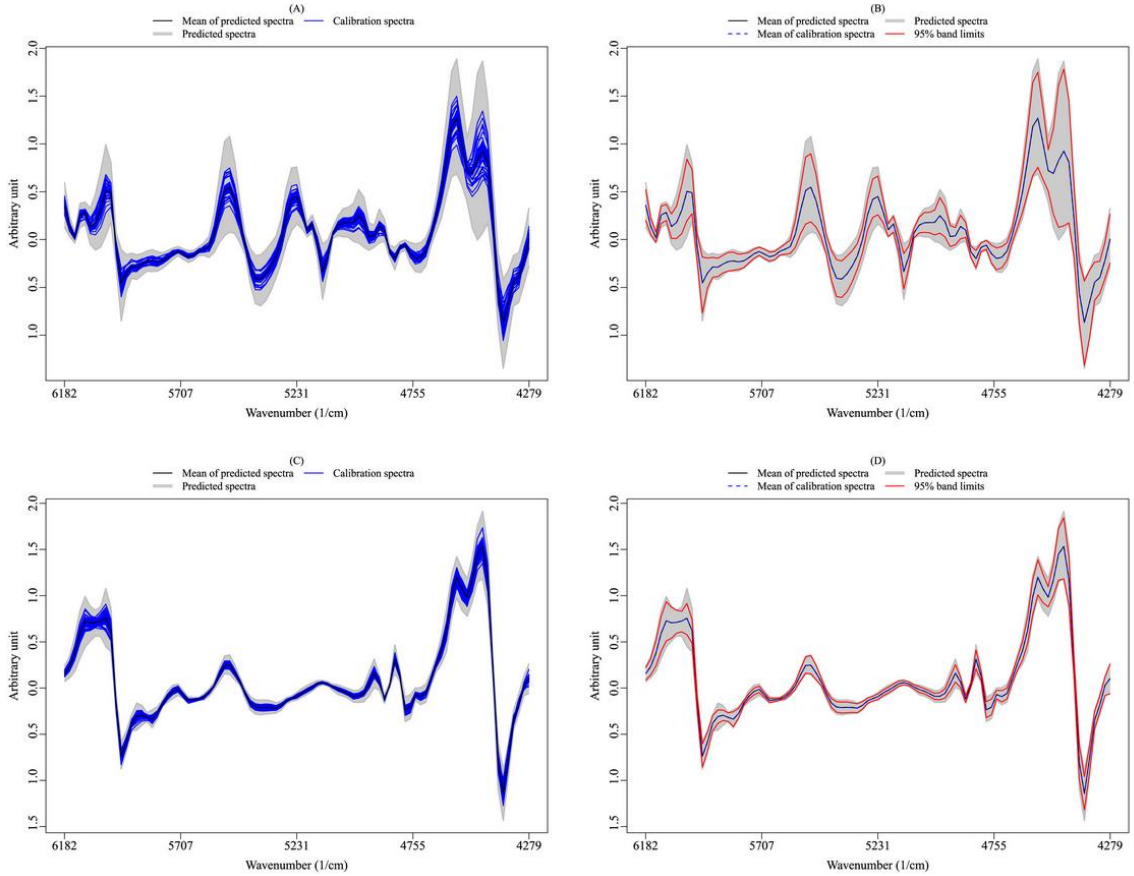
Apart from evaluating the performances, the practical application of the prediction band was illustrated, including the criticism and goodness-of-fit checks of the statistical models, the visualization of the band and outlyingness maps showing the outlyingness patterns of a target test spectrum against a non-target test spectrum.

3.3. STUDY 2: APPLICATIONS TO TWO VERIFICATION PROBLEMS WITH DATASETS 3 AND 4

The second study to evaluate the classification performances consists in external validations using the same strategy and datasets (Dataset 3 and 4) as Ciza and coworkers [33]. As in Ciza and coworkers [33], P01 and I03 were chosen as target formulations for paracetamol and ibuprofen respectively. For each target formulation, three batches were selected as "genuine" or "reference" and their spectra used to calibrate eight 95%-prediction bands resulting from the combination of the two models with normal likelihood on either standard or smoothed PCs' scores and the four prior models on either the eigenvalues or the singular values. Spectra of the three remaining batches were used as external validation data to evaluate the true positive rates. All the spectra of batches of non-target formulations (P10, P11, P12 and P13 for paracetamols and I04, I08, I09 and I10 for ibuprofens) were used as validation data to evaluate the false positive rates. A comparison of the performances of the prediction bands with the four benchmark SIMCA methods described in Study 1 in Section 3.2 was also done.

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

Fig. 4. Study 1: Adequacy of the prediction models and band limits for future spectra of the target paracetamol (P01, Dataset 1) and target ibuprofen (I03, Dataset 2) formulations (A) agreement between the trajectories of the predicted spectra and those of the calibration spectra of the target paracetamol formulation P01 (Dataset 1); (B) agreement between the mean of the predicted spectra and the mean of the calibration spectra of P01; (C) agreement between the trajectories of the predicted spectra and those of the calibration spectra of the target ibuprofen target formulation I03 (Dataset 2); (D) agreement between the mean of the predicted spectra and the mean of the calibration spectra of I03. Predictions were obtained using a Bayesian zero-mean model on smoothed PCs' scores with independent Jeffreys' priors on the reciprocal of the eigenvalues, repeated on naïve $B = 30$ bootstrap resamples of the calibration dataset to include uncertainties in the PC decomposition.



4. Results and discussion

4.1. STUDY 1: MONTE CARLO VALIDATIONS WITH DATASETS 1 AND 2 TO EVALUATE THE CLASSIFICATION PROPERTIES OF THE PREDICTION BAND

4.1.1. MODEL ADEQUACY

Basic statistical tools to evaluate the adequacy of each prediction model before its use for classification tasks are presented on Fig. 4 for both the paracetamol (P01) and the ibuprofen (I03) target formulations. These tools enable to visually check the agreement between the measured and the predicted spectra. It can be seen that the calibration spectra were overall within the range of predicted spectra for both target of formulations (Fig. 4A and C). Moreover, the means of the predicted and

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

observed spectra overlap almost perfectly for both P01 (Fig. 4B) and I03 (Fig. 4D) suggesting the unbiasedness of the predictions. These two figures also illustrates for each target formulation the limits of the band defined by the range of the 95% most central predicted spectra (Fig. 4B and D). Additional graphical tools that may be used to better understand how the model development proceeds and to practically evaluate the model adequacy at its various steps are provided and explained in Appendix A.1 and A.2.

4.1.2. CLASSIFICATION PERFORMANCES WITH STANDARD PC AND UNDER INDEPENDENT JEFFREYS' PRIORS

Table 3 shows for each of the paracetamol (P01) and ibuprofen (I03) target formulations, the average and standard deviation of the true positive rates of classification by two 95%-prediction bands based on the models with normal likelihood on either standard or smoothed PCs' scores and independent non-informative Jeffreys' priors on the reciprocal of the eigenvalues.

For testing procedure 1, the coverage rates were about the nominal 95% at $\kappa_1 = 90\%$ for both P01 and I03, i.e. by accepting a new tested spectrum if at least 90% of its reconstructions fall entirely inside the band limits. False positives rates were all 0.0%. Decreasing κ_1 to 50% enables to increase the true positive rates, while still discriminating the non-target spectra for both types of formulations. Increasing κ_1 to 90% decreases the true positive rate to about 91%. This procedure is the most rigorous and compliant with the concept of a simultaneous hypothesis testing.

Table 3. Average (standard deviation in brackets) of true positive rates of classification (in percentage, %) of the target paracetamol formulation (P01 in Dataset 1) and the target ibuprofen formulation (I03 in Dataset 2), by four 95%-prediction bands. Models with normal likelihood on either standard or smoothed PCs' scores and independent non-informative Jeffreys' priors on the reciprocal of the eigenvalues are used. The three testing procedures for a new spectrum defined in Section 2.10 are considered.

Testing procedure	Acceptance criteria	Paracetamol target formulation (P01)		Ibuprofen target formulation (I03)	
		Band based on standard PCs	Band based on smoothed PCs	Band based on standard PCs	Band based on smoothed PCs
1	κ_1				
	0.50	97.9 (3.3)	98.3 (2.3)	97.3 (3.5)	97.8 (2.8)
	0.75	96.7 (4.4)	97.1 (3.4)	96.2 (4.1)	96.6 (3.7)
	0.90	95.1 (5.2)	95.2 (4.3)	94.9 (4.8)	95.4 (4.6)
	1.00	90.6 (7.6)	90.0 (6.1)	91.4 (6.6)	91.9 (6.6)
2	κ_2				
	0.50	98.7 (2.5)	98.7 (1.9)	98.1 (2.9)	98.5 (2.5)
	0.75	97.2 (3.8)	97.6 (3.1)	96.8 (3.8)	97.3 (3.2)
	0.90	95.7 (4.9)	96.2 (3.9)	95.5 (4.6)	96.0 (4.0)
	1.00	90.6 (7.6)	90.0 (6.1)	91.4 (6.6)	91.9 (6.6)
3	κ_3				
	3	99.8 (0.9)	99.8 (0.6)	99.2 (1.5)	99.2 (1.4)
	2	99.6 (1.1)	99.7 (1.0)	98.5 (2.1)	98.8 (1.8)
	1	98.7 (2.5)	99.1 (1.7)	97.6 (3.1)	98.0 (2.5)
	0	90.6 (7.6)	90.0 (6.1)	91.4 (6.6)	91.9 (6.6)

Notes: (a) κ_1: minimum proportion of the reconstructions of a spectrum that falls entirely inside the band; κ_2: minimum proportion of the reconstructions of a spectrum that falls inside the band at each wavenumber; κ_3: average number of tolerated points outside the band for all reconstructions of a spectrum; (b) False positive rates are 0.0%(0.0%) for all cases of non-target formulations, except two cases of P13 at $\kappa_2 = 0.50$ where the false positive rates are 0.2% (0.8%) for the standard PC model and 0.4% (1.2%) for the smoothed PC model.

Classification performances with procedure 2 were similar to those of procedure 1, with almost unbiased true positive rates (95.7% for P01 and 95.5% for I03) at $\kappa_2 = 90\%$, i.e. by accepting an unknown tested spectrum if at each spectral point (wavenumber), at least 90% of its reconstructions fall inside the band limits. It is noted that with $\kappa_2 = 50\%$, a minor increase in the false positive rate of paracetamol formulation P13 was observed (0.2%). This is not surprising because this procedure, which proceeds by a pointwise testing, does not consider the whole trajectory of each reconstruction

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

of the tested spectrum. Hence, it is less rigorous than procedure 1. In addition, setting κ_2 to a relatively low value might have increased the risk of false positives for this procedure.

For testing procedure 3, tolerating on average 1 to 3 points per reconstruction outside the band increased the true positive rates to 98.7%–99.8% for P01 and 97.6–99.2% for I03, without affecting the false positive rates. Moreover, this procedure showed the lowest variability and seems to be more stable for both types of formulations. However, setting the average number of spectral points to tolerate outside the band (κ_3) requires some prior knowledge of the outlyingness pattern of spectra of the target class. A first guess might be obtained during a possible model validation step from the proposed outlyingness map (see explanations and examples in Sections 4.1.6 Interpretability of the classification results, 4.2 Study 2: application to two verification problems with datasets 3 and 4).

4.1.3. SENSITIVITY OF CLASSIFICATION PERFORMANCES TO THE TYPE OF PC DECOMPOSITION

Regarding the type of PC decomposition, the results show that smoothing the eigenfunctions did not substantially impact the classification performances of the prediction bands compared with the standard PC decomposition for both target P01 and I03 formulations (Table 3). This suggests that standard PC decomposition on these NIR data performs satisfactorily. In practice however, the authors would suggest to use the smoothed PCs as this is an additional and recommended step in regularizing the model, that is controlling potential excessive and noisy variations of the eigenfunctions, and hence preventing overfitting [21,28].

4.1.4. SENSITIVITY OF CLASSIFICATION PERFORMANCES TO THE PRIOR MODELS

Regarding, the sensitivity to the choice of the prior model, all the trends in the classification performances for both paracetamol-based and ibuprofen-based formulations under the Jeffreys' priors on the reciprocal of the eigenvalues varied little under the independent non-informative Uniform(0,+∞), vague $\text{Normal}_+(0,100)$ and weakly informative $\text{Normal}_+(0,1)$ priors on the singular values (see Tables in Appendix A Supplementary data, .3 for more details). This suggests that for these SNV-normalized NIR data, the performances of the proposed method are little sensitive to the choice of the non-informative or weakly informative prior models. However, in practice the authors would suggest to use the weakly informative $\text{Normal}_+(0,1)$ priors for such pharmaceutical SNV-normalized data. Especially, when the sample size is low and insufficient to precisely estimate the eigenvalues, such priors can further regularized the estimations [20].

4.1.5. COMPARISON OF PERFORMANCES WITH SIMCA MODEL

Compared to the benchmark SIMCA model with 95%-confidence level, the 95%-prediction band outperformed the four rigorous SIMCA methods in terms of average and variability of the true positive rates for both P01 and I03 formulations, the average and variability of the false positive rates being similar (Table 4). Indeed, all the four SIMCA methods showed substantial undercoverage for both target formulations, meaning that their true positive rates were far below the theoretical *a priori* fixed confidence level of 95%. The best true positive rates, 88.4% for P01 and 88.0% for I03, were obtained with the moment-based data-driven SIMCA (Table 4). This undercoverage behaviour of the SIMCA model is well-known and often occurs in practice [2,6], despite the method has undergone several

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

improvements reflected in the four variants herein used [4,6,37]. The authors argue that one major cause for this undercoverage behaviour is the fact that the SIMCA model does not integrate uncertainties about model parameters in the definition of its acceptance limits. For example, in the data-driven moment-based SIMCA [4,6,7], point estimates of degrees of freedom and scaling parameters of the two scaled Chi-square distributions on the Q and T^2 statistics are derived from the data, but neither uncertainties in these parameters nor the uncertainties in the PC decomposition are integrated in the decision rules or acceptance limits. Contrary to the SIMCA model, the proposed prediction band integrates all relevant modelling uncertainties in a statistically convenient manner, on one hand through the concept of predictive distribution of future PCs' scores and future spectra, and on the other hand through uncertainty in the reconstructions of tested unknown spectra. Especially, integrating uncertainties about reconstructions of tested spectra is a unique feature of risk-management by the proposed method. To the best of the authors' knowledge, this type of uncertainty has never been integrated in a CM method before. By tuning the proposed intuitive and interpretable acceptance criteria κ_1, κ_2 and κ_3, the classification performances can be further optimized. Because of all these uncertainty management features, the proposed method is fully predictive, risk-oriented and more correctly interpretable as probabilistic prediction space for a single future spectrum.

To conclude, these Monte Carlo validation results demonstrate the reliability of the proposed concept of β%-prediction band as CM method.

Table 4. Averages (standard deviations in brackets) of true positive rates of classification (in percentage, %) of the target paracetamol formulation (P01 in Dataset 1) and the target ibuprofen formulation (I03 in Dataset 2), by four optimized SIMCA methods (Jackson-Mudholkar method, chi-square method, data-driven moment-based method and data-driven robust method), with *a priori* fixed 95%-confidence level.

SIMCA method	Paracetamol target formulation (P01)			Ibuprofen target formulation (I03)		
	Optimal nPC	Sensitivity in LOOCV (%)	True positive rate (%)	Optimal nPC	Sensitivity in LOOCV (%)	True positive rate (%)
Jackson-Mudholkar method	1.3 (0.8)	87.4 (2.6)	83.6 (7.8)	1.5 (0.6)	89.5 (2.6)	86.2 (7.6)
Chi-square method	2.3 (1.5)	85.9 (2.8)	82.2 (7.3)	2.3 (0.8)	86.5 (3.1)	83.7 (7.8)
Data-driven moment-based method	1.6 (1.0)	89.5 (2.8)	88.4 (7.7)	1.3 (0.7)	88.7 (2.7)	88.0 (7.8)
Data-driven robust method	1.3 (0.6)	86.9 (5.3)	83.7 (9.6)	1.4 (0.5)	84.7 (3.9)	83.4 (9.4)

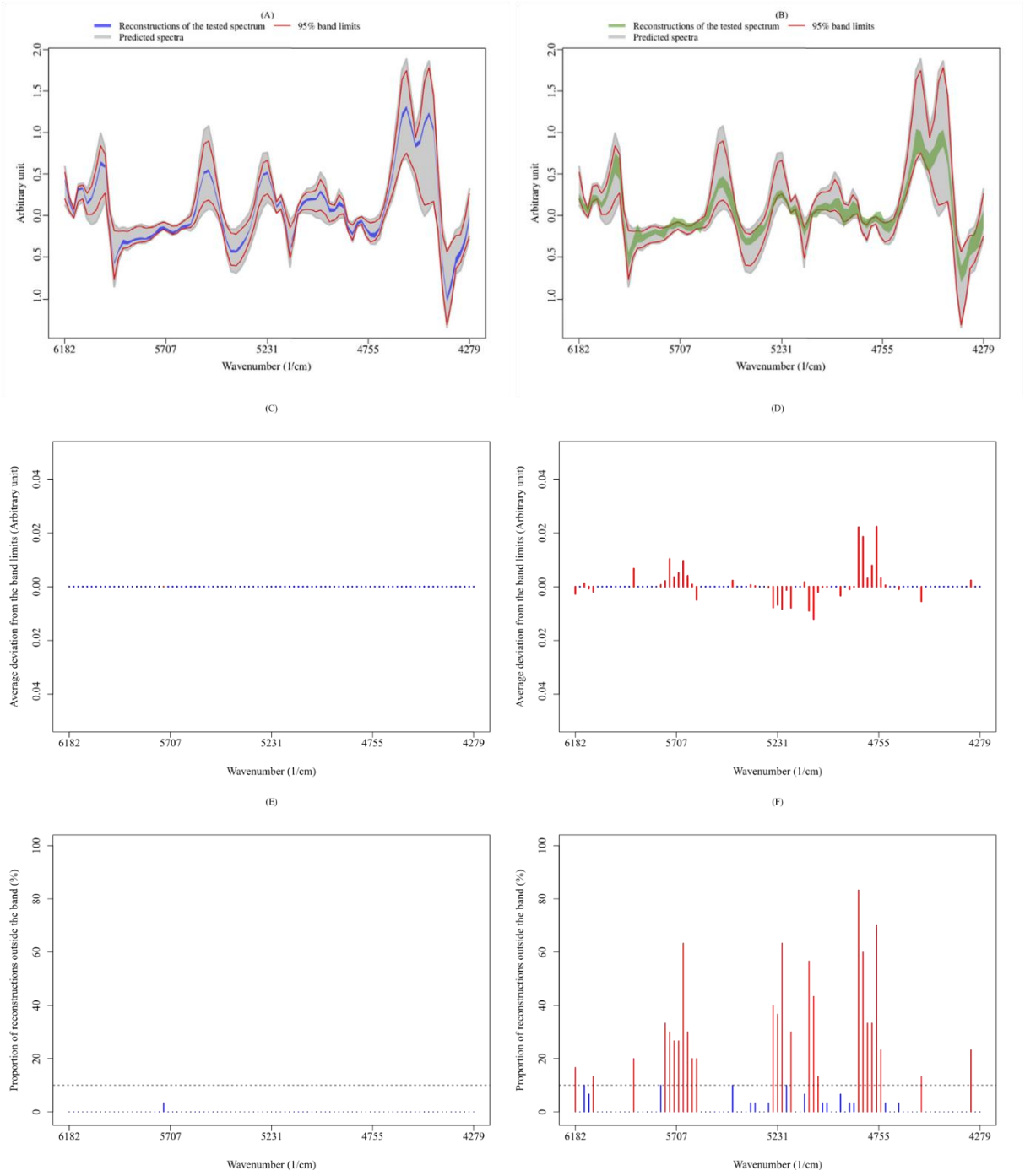
Notes: (a) nPC: number of principal components; (b) LOOCV: leave-one-out cross-validation; (c) False positive rates are 0.0%(0.0%) for all cases of non-target formulations.

4.1.6. INTERPRETABILITY OF THE CLASSIFICATION RESULTS

Besides a consistent control of coverage rate, another important feature of the proposed prediction band resides in the fact that it enables to easily visualize and characterize outlyingness patterns of tested spectra. Indeed, the band limits can be used as references defining critical or acceptable spectral trajectories for the target class. Hence, the range of spectral variables where a tested spectrum deviates from these references as well as the magnitude of these deviations can be derived and represented as an outlyingness map. This new concept of outlyingness pattern or map which enables a deeper understanding of how the band classifies spectra is illustrated on Fig. 5 for two paracetamol formulations. The acceptance criteria of a single spectrum for this illustration are those defined in the Experimental Section 3.2, that is $\kappa_1 = 90\%$, $\kappa_2 = 90\%$ and $\kappa_3 = 1$ respectively for procedures 1, 2 and 3. The top panels contrast the overlapping patterns of the reconstructions of one target test spectrum (P01, Fig. 5A) and one non-target test spectrum (P10, Fig. 5B) with the band. Clearly, the reconstructions of the non-target test spectrum (P10, Fig. 5B) were more spread than those of the target test spectrum (P01, Fig. 5A).

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

Fig. 5. Study 1: Comparison of the outlyingness patterns of one target spectrum (P01, Dataset 1) versus one non-target spectrum (P10, Dataset 1) w.r.t. a 95%-prediction band for formulation P01 (Dataset 1). (A): Overlap of the $B = 30$ reconstructions of the target spectrum with the band; (B): Overlap of the $B = 30$ reconstructions of the non-target spectrum with the band; (C): Outlyingness map showing the average deviation of the $B = 30$ reconstructions of the target spectrum from the band limits (blue bars have zero values; red bars have non-zero values); (D): Outlyingness map showing the average deviation of the $B = 30$ reconstructions of the non-target spectrum from the band limits (blue bars have zero values; red bars are non-zero values); (E): Proportion of the $B = 30$ reconstructions of the target spectrum outside the band at each wavenumber (the horizontal dashed line is a threshold of 10%, i.e. , $\kappa_2 = 90\%$ and red bars are above the threshold); (F): Proportion of the $B = 30$ reconstructions of the non-target spectrum outside the band at each wavenumber (the horizontal dashed line is a threshold of 10%, i.e. , $\kappa_2 = 90\%$ and red bars are above the threshold). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)



Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

Wavenumbers where they were outside the band are highlighted in red on their outlyingness maps (Fig. 5C and D). These maps contrast the average deviation of the reconstructions of the two tested spectra from the upper and lower band limits. The target test spectrum (P01, Fig. 5C) deviated at only one wavenumber with substantially lower magnitude than the non-target spectrum which deviated at several wavenumbers (P10, Fig. 5D). Only 1 out of the $B = 30$ reconstructions of this target test spectrum deviated, i.e. the proportion of its reconstructions entirely inside the band was 96.7%. This proportion was above the predefined minimal acceptance threshold $\kappa_1 = 90\%$. Hence, using the testing procedure 1, this target test spectrum was classified as true positive. On the contrast, the proportion of the $B = 30$ reconstructions of the non-target test spectrum P10 that was entirely inside the band was 0%, i.e. they all deviated from the band limits each at 1 to 16 points. This proportion was far below the predefined threshold $\kappa_1 = 90\%$. Hence this spectrum was classified as true negative.

Fig. 5E and F contrast the proportion of the reconstructions of the two tested spectra outside the band limits at each wavenumber. This proportion was below the predefined threshold of 10%, i.e. $100\% - \kappa_2$, at all wavenumbers for the target test spectrum (Fig. 5E); hence the spectrum was accepted as true positive using procedure 2 with $\kappa_2 = 90\%$. On the contrast, more than 10% of the reconstructions of the non-target test spectrum fell outside the band limits at several wavenumbers (Fig. 5F).

The average number of points per reconstruction outside the band was 0.0 (0.2) with a range of 0–1 for the target test spectrum, against 10.4 (4.0) with a range of 1–16 for the non-target test spectrum. Hence, using procedure 3 with the predefined threshold $\kappa_3 = 1$, the target test spectrum was accepted as true positive, while the non-target test spectrum was rejected as true negative.

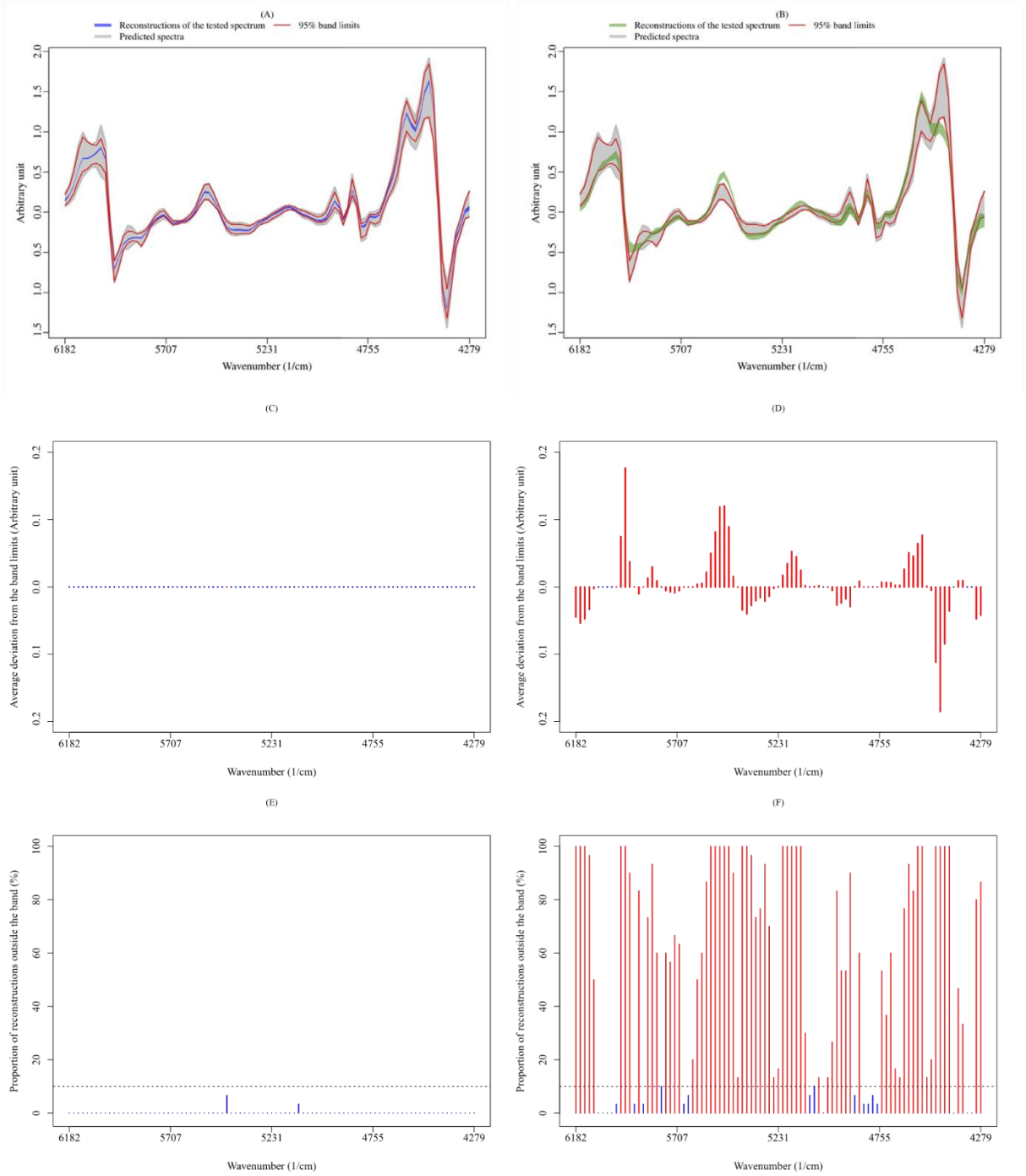
A similar detailed investigation of the outlyingness patterns of tested spectra was done for the ibuprofens on Fig. 6, contrasting the patterns for a target test spectrum (I03, Fig. 6A, C, 6E) with a non-target test spectrum (I04, Fig. 6B, D, 6F). It can be seen that the reconstructions of the target test spectrum (6A) were less spread than those of the non-target test spectrum (6B). They deviated at only two points with very low magnitudes (6C) compared with the non-target test spectrum whose reconstructions deviated at greater number of points with substantially higher magnitudes (6D). At most of the wavenumbers, more than 10% of the reconstructions of the non-target test spectrum (6E) were outside the band limits contrary to the target test spectrum (6F).

To the best of the authors' knowledge, no projection-based CM method including the SIMCA provides such rich information about outlyingness patterns and such easy physical understanding of the classification results.

It is important to note that the 95%-prediction band is different from a pointwise 95%-confidence band for the mean spectrum that is constructed by the mean-spectrum plus or minus 1.96 times its standard error at each spectral point, as provided by some chemometric software. This kind of band is a pointwise confidence band for the mean and not a prediction band for a next future spectrum. Furthermore, it is not a simultaneous band and hence, might not be effective in simultaneous testing of the whole trajectory of a future spectrum.

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

Fig. 6. Study 1: Comparison of the outlyingness patterns of one target spectrum (I03, Dataset 2) versus one non-target spectrum (I04, Dataset 2) w.r.t. a 95%-prediction band for formulation I03 (Dataset 2). (A): Overlap of the $B = 30$ reconstructions of the target spectrum with the band; (B): Overlap of the $B = 30$ reconstructions of the non-target spectrum with the band; (C): Outlyingness map showing the average deviation of the $B = 30$ reconstructions of the target spectrum from the band limits (blue bars have zero values; red bars have non-zero values); (D): Outlyingness map showing the average deviation of the $B = 30$ reconstructions of the non-target spectrum from the band limits (blue bars have zero values; red bars have non-zero values); (E): Proportion of the $B = 30$ reconstructions of the target spectrum outside the band at each wavenumber (the horizontal dashed line is a threshold of 10%, i.e. $\kappa_2 = 90\%$, and red bars are above the threshold); (F): Proportion of the $B = 30$ reconstructions of the non-target spectrum outside the band at each wavenumber (the horizontal dashed line is a threshold of 10%, i.e. $\kappa_2 = 90\%$, and red bars are above the threshold). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)



Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

4.1.7. COMPUTATION TIME FOR PRACTICAL IMPLEMENTATION

Regarding practical implementation, to generate the predictive distribution of the spectra for P01 on Device 1 using non-informative Jeffreys' priors on the reciprocal of the eigenvalues, the average computation time was 1.6 ± 0.3 s, that is less than 0.06 s per bootstrap iteration on a 2.2 GHz computer using just one core (no parallelization). With flat uniform priors or half-normal priors on the square-root of the eigenvalues using the HMC sampler of Stan probabilistic language [30] this computation time was about 10 times higher (17.9 ± 0.9 s, i.e. about 0.6 s per bootstrap iteration). Ranking the predicted spectra and setting the band limits took approximately 3.9 ± 0.3 s. Once the model is calibrated, the output for testing new spectra is a pair of vectors of band limits. Testing a single new spectrum involves projecting its $B = 30$ reconstructions by the bootstrap PC models onto that pair of vectors and computing the decision-making metrics for the three procedures, including the outlyingness map. This testing step took 0.25 ± 0.05 s without parallelization. These computation times remain almost the same with ibuprofens formulations with the same piece of equipment.

From these, it is clear that prediction bands over a complete spectrum can be built in a relatively short time, but careful model checking is mandatory. Moreover, testing new samples or batches against the obtained bands can be made nearly instantaneously, which allows envisaging the use of such methodology at a large scale.

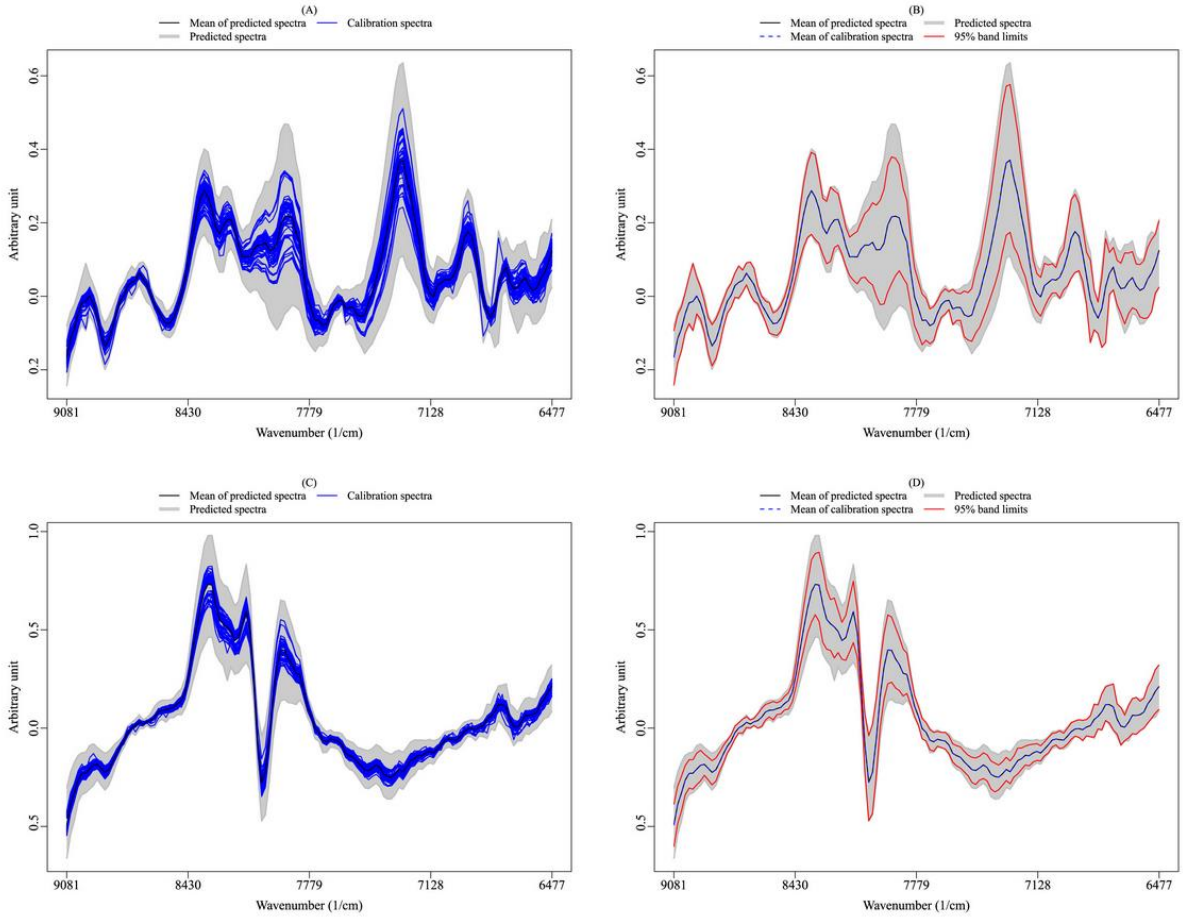
4.2. STUDY 2: APPLICATION TO TWO VERIFICATION PROBLEMS WITH DATASETS 3 AND 4

Graphics on Fig. 7 enable basic visual checks of the agreement between the predicted and the observed spectra for both formulations P01 and I03 before the use of their prediction bands for classification purposes. It can be seen that the measured spectra were overall within the range of predicted spectra for both formulations (Fig. 7A and B). The mean of the predicted and the measured spectra overlap almost perfectly suggesting the unbiasedness of the predictions for both formulations (Fig. 7B and D). The 95%-band limits used to classify spectra were also shown on Fig. 7B and D for both formulations.

Classification results for the model with independent Jeffreys' priors on the reciprocal of the eigenvalues of both standard and smoothed PCs are reported in Table 5 for both types of formulations. The results show that the 95%-prediction band yielded satisfactory classification performances with true positive rates about the nominal 95% at the threshold $\kappa_1 = 90\%$, $\kappa_2 = 90\%$ and $\kappa_3 = 1$ respectively for procedure 1, 2 and 3 for both formulations. Decreasing κ_1 and κ_2 to 50%, and increasing κ_3 to 3 enabled to increase the true positive rates without affecting the false positive rates, whatever the PC decomposition method (Table 5). The method also outperformed the four rigorous optimized SIMCA methods with *a priori* fixed 95% confidence levels in terms true positive rates, false positive rates being similar (Table 6). Amongst the SIMCA methods, the data-driven moment-based SIMCA yielded the closest true positive rates (80% for P01 and 85% for I03) to the nominal 95%. But it is noticed how the true positive rates varied substantially from one SIMCA model to another, especially for formulation P01 (Table 6).

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

Fig. 7. Study 2: Adequacy of the prediction models and band limits for future spectra of the target paracetamol (P01, Dataset 3) and target ibuprofen (I03, Dataset 4) formulations; (A) agreement between the trajectories of the predicted spectra and those of the calibration spectra of the target paracetamol formulation P01 (Dataset 3); (B) agreement between the mean of the predicted spectra and the mean of the calibration spectra of P01; (C) agreement between the trajectories of the predicted spectra and those of the calibration spectra of the target ibuprofen target formulation I03 (Dataset 4); (D) agreement between the mean of the predicted spectra and the mean of the calibration spectra of I03. Predictions were obtained using a Bayesian zero-mean model on smoothed PCs' scores with independent Jeffreys' priors on the reciprocal of the eigenvalues, repeated on $B = 30$ naïve bootstrap resamples of the calibration dataset to include uncertainties in the PC decomposition.



Outlyingness patterns of one target (P01) test spectrum and one non-target (P10) test spectrum with respect to the 95%-prediction band with acceptance thresholds $\kappa_1 = 90\%$, $\kappa_2 = 90\%$ and $\kappa_3 = 1$ respectively for procedure 1, 2 and 3, were investigated on Fig. 8. Clearly, the reconstructions of the target spectrum (Fig. 8A) were much less spread than those of the non-target spectrum (Fig. 8B). Wavenumbers where they deviated from the band limits and the average magnitude of deviations are depicted by the outlyingness maps (Fig. 8C and D). The target test spectrum (Fig. 8C) deviated at only two wavenumbers and with far lower magnitudes compared to the non-target test spectrum (Fig. 8D). 76.7% and 0.0% of the reconstructions of the target test spectrum and the non-target test spectrum respectively were entirely inside the band limits. Fig. 8E and F contrast the proportion of the reconstructions of both tested spectra outside the band limits at each wavenumber, and clearly the patterns were different. Clearly, despite both the target and non-target test spectra were classified as negatives with the predefined thresholds $\kappa_1 = 90\%$ and $\kappa_2 = 90\%$, the outlyingness map provides additional information to further discriminate between a false negative and true negative. The average number of points per reconstruction outside the band was 0.2 (0.4) with a range of 0–1 for the target

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

test spectrum against 27.7 (3.1) with a range of 22–34 for the non-target test spectrum. Hence, using $\kappa_3 = 1$, the target test spectrum is accepted as positive.

Table 5. True positive rates of classification (in percentage, %) of the target paracetamol formulation (P01 in Dataset 3) and the target ibuprofen formulation (I03 in Dataset 4), by four 95%-prediction bands. Models with normal likelihood on either standard or smoothed PCs' scores and independent non-informative Jeffreys' priors on the reciprocal of the eigenvalues are used. The three testing procedures for a new spectrum defined in Section 2.10 are considered.

Testing procedure	Acceptance criteria	Paracetamol target formulation (P01)		Ibuprofen target formulation (I03)	
		Band based on standard PCs	Band based on smoothed PCs	Band based on standard PCs	Band based on smoothed PCs
1	κ_1				
	0.50	95.0	96.6	98.3	98.3
	0.75	95.0	95.0	96.7	96.7
	0.90	95.0	95.0	95.0	96.7
	1.00	93.3	95.0	90.0	90.0
2	κ_2				
	0.50	96.6	98.3	98.3	98.3
	0.75	95.0	96.6	96.7	96.7
	0.90	95.0	95.0	95.0	96.7
	1.00	93.3	95.0	90.0	90.0
3	κ_3				
	3	98.3	98.3	98.3	98.3
	2	96.7	98.3	98.3	98.3
	1	96.7	96.7	98.3	98.3
	0	93.3	95.0	90.0	90.0

Notes: (a) κ_1: minimum proportion of the reconstructions of a spectrum that falls entirely inside the band; κ_2: minimum proportion of the reconstructions of a spectrum that falls inside the band at each wavenumber; κ_3: average number of tolerated points outside the band for all reconstructions of a spectrum; (b) False positive rates are 0.0% for all cases of non-target formulations.

Table 6. True positive rates of classification (in percentage, %) of the target paracetamol formulation (P01 in Dataset 3) and the target ibuprofen formulation (I03 in Dataset 4), by four optimized SIMCA methods (Jackson-Mudholkar method, chi-square method, data-driven moment-based method and data-driven robust method), with *a priori* fixed 95%-confidence level.

SIMCA method	Paracetamol target formulation (P01)			Ibuprofen target formulation (I03)		
	Optimal nPC	Sensitivity in LOOCV (%)	True positive rate (%)	Optimal nPC	Sensitivity in LOOCV (%)	True positive rate (%)
Jackson-Mudholkar method	2	88.0	68.3	1	93.0	80.0
Chi-square method	3	92.0	80.0	1	93.0	85.0
Data-driven moment-based method	2	92.0	80.0	2	93.0	85.0
Data-driven robust method	4	83.0	45.0	1	92.0	56.0

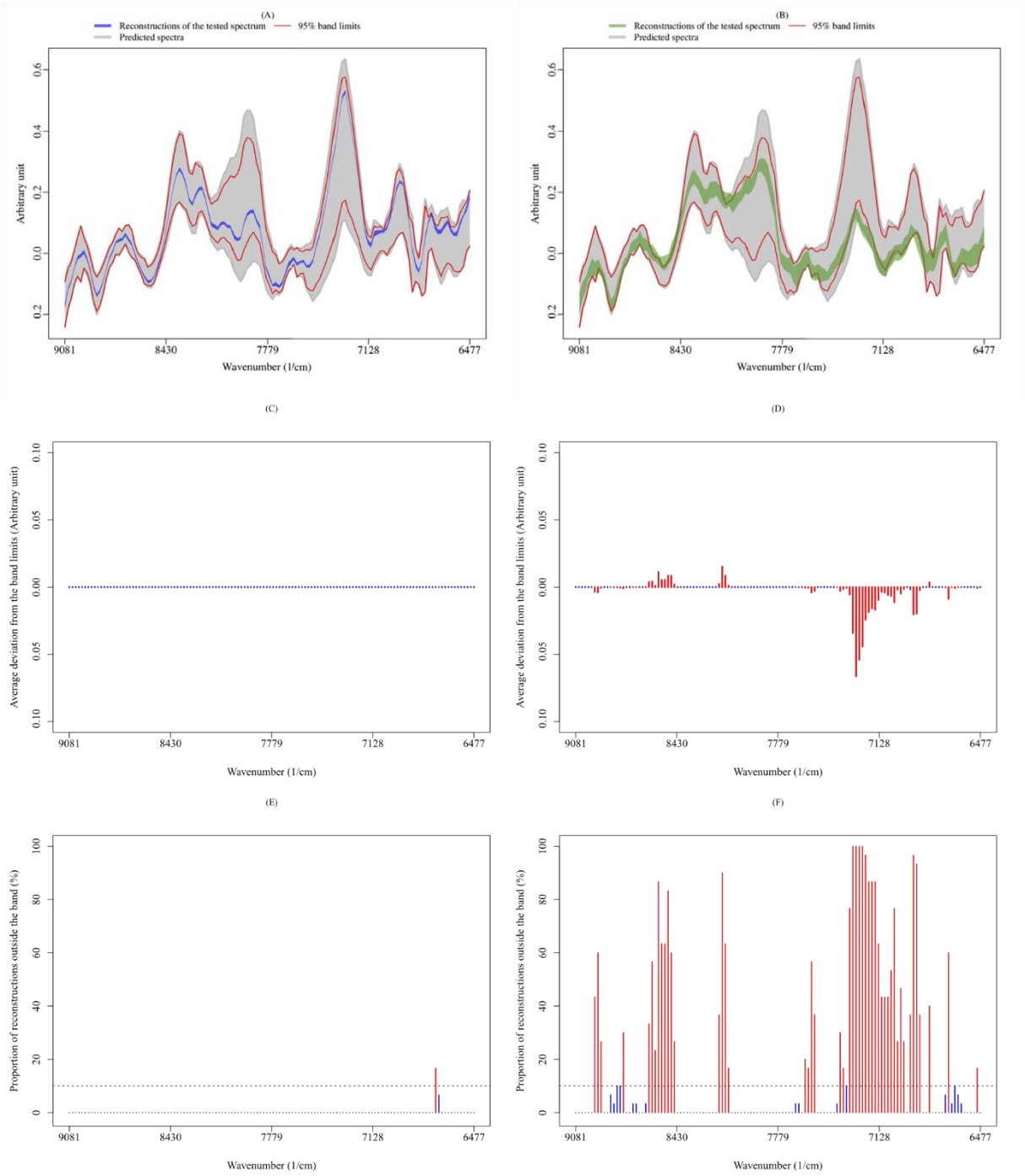
Notes: (a) nPC: number of principal components; (b) LOOCV: leave-one-out cross-validation; (c) False positive rates are 0.0% for all cases of non-target formulations.

A similar investigation of the outlyingness patterns of tested spectra can be done for the ibuprofens with Dataset 4 on Fig. 9. Clearly, the reconstructions of the target (I03, Fig. 9A, C, 9E) and non-target (I04, Fig. 9B, D, 9F) test spectra overlap the band with different patterns, in terms of spread of the reconstructions (9A versus 9B), the magnitude of the excursion outside the band limits (9C versus 9D), and the proportion of points outside the limits at each wavenumber (9E versus 9F).

It is worth noting that, the fact that some spectral ranges may be out of the band even for target class spectra demonstrates a possible spectral acquisition variability or inherent spectral variability due to blister composition. Therefore, acceptance criteria may be tuned to account for those issues, hence increasing the odds of true positive and while still correctly rejecting true negatives.

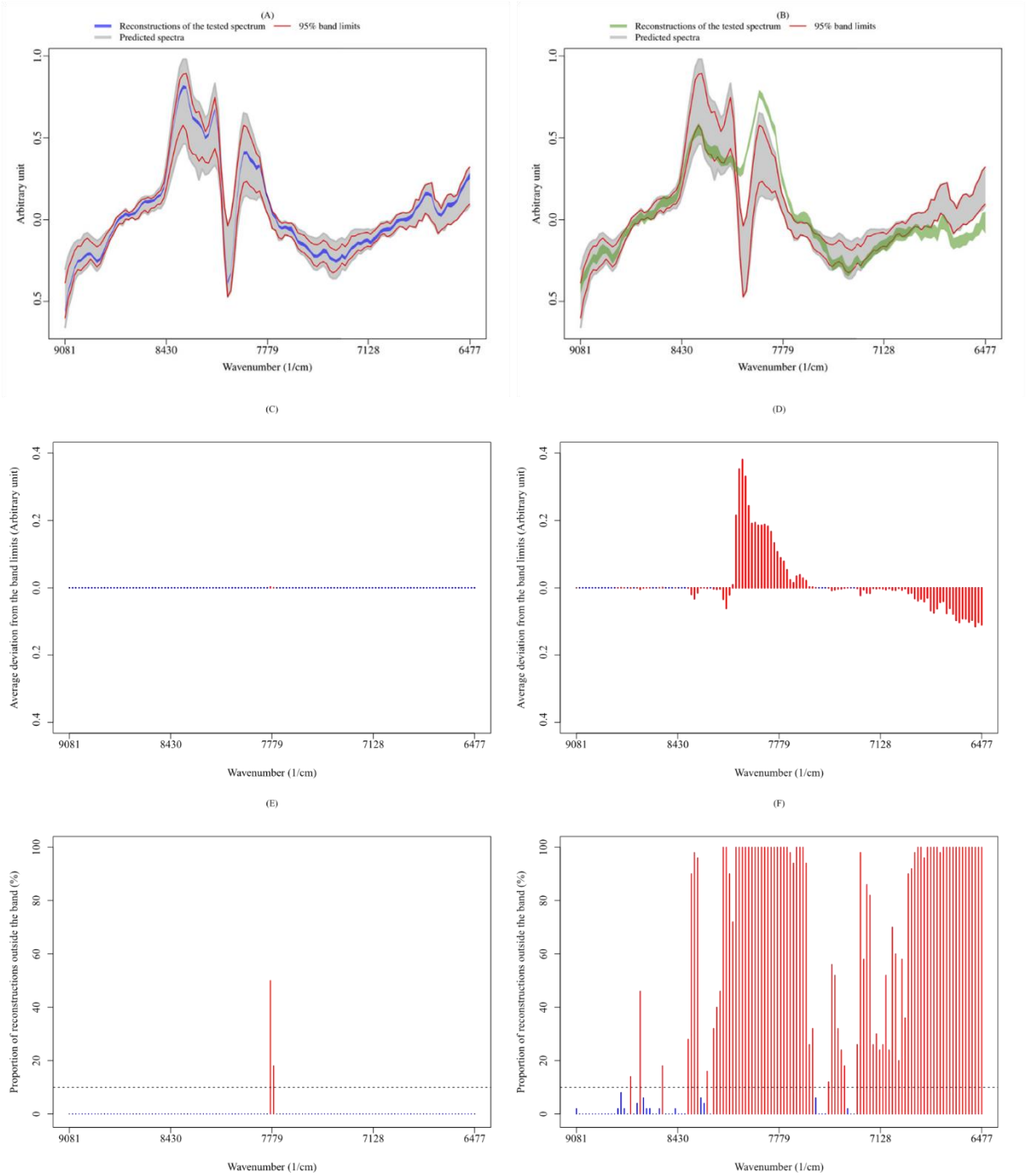
Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

Fig. 8. Study 2: Comparison of the outlyingness patterns of one target spectrum (P01, Dataset 3) versus one non-target spectrum (P10, Dataset 3) w.r.t. a 95%-prediction band for formulation P01 (Dataset 3). (A): Overlap of the $B = 30$ reconstructions of the target spectrum with the band; (B): Overlap of the $B = 30$ reconstructions of the non-target spectrum with the band; (C): Outlyingness map showing the average deviation of the $B = 30$ reconstructions of the target spectrum from the band limits (blue bars have zero values; red bars have non-zero values); (D): Outlyingness map showing the average deviation of the $B = 30$ reconstructions of the non-target spectrum from the band limits (blue bars have zero values; red bars have non-zero values); (E): Proportion of the $B = 30$ reconstructions of the target spectrum outside the band at each wavenumber (the horizontal dashed line is a threshold of 10%, i.e. $\kappa_2 = 90\%$, and red bars are above the threshold); (F): Proportion of the $B = 30$ reconstructions of the non-target spectrum outside the band at each wavenumber (the horizontal dashed line is a threshold of 10%, i.e. $\kappa_2 = 90\%$, and red bars are above the threshold). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)



Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

Fig. 9. Study 2: Comparison of the outlyingness patterns of one target spectrum (I03, Dataset 4) versus one non-target spectrum (I04, Dataset 4) w.r.t. a 95%-prediction band for spectra of formulation I03 (Dataset 4). (A): Overlap of the $B = 30$ reconstructions of the target spectrum with the band; (B): Overlap of the $B = 30$ reconstructions of the non-target spectrum with the band; (C): Outlyingness map showing the average deviation of the $B = 30$ reconstructions of the target spectrum from the band limits (blue bars have zero values; red bars have non-zero values); (D): Outlyingness map showing the average deviation of the $B = 30$ reconstructions of the non-target spectrum from the band limits (blue bars have zero values; red bars have non-zero values); (E): Proportion of the $B = 30$ reconstructions of the target spectrum outside the band at each wavenumber (the horizontal dashed line is a threshold of 10%, i.e. $\kappa_2 = 90\%$, and red bars are above the threshold); (F): Proportion of the $B = 30$ reconstructions of the non-target spectrum outside the band at each wavenumber (the horizontal dashed line is a threshold of 10%, i.e. $\kappa_2 = 90\%$, and red bars are above the threshold). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)



Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

5. Conclusion

Classification models should not only seek precision, but also integrate uncertainties management in decision-making [38]. This article introduces a fully predictive and probabilistic class-modelling method based on the concept of β%-prediction band for NIR spectra. The evaluation of the classification performances provides evidence that this proposed method for class-modelling possesses satisfactory predictive performances on NIR datasets, even better than the soft independent modelling of class analogy (SIMCA), whereas SIMCA is often considered as state-of-the-art of class-modelling. To the best of the authors' knowledge, this is the first class-modelling method based on bands, that is fully predictive and risk-oriented enabling a fully probabilistic decision making while providing easy physical interpretability. It uses whenever needed, appropriate techniques to account for all decision-making uncertainties including uncertainties about the band calibration and uncertainties about new spectra to be tested.

The proposed methodology has been tested for a drug identification task but it may be used for several other tasks such as PAT applications (end-point of blending process, raw materials attributes verification), food authentication, etc. It may of course be extended to other spectroscopy technologies such as Raman spectroscopy.

CRediT authorship contribution statement

Avohou T. Hermane: Conceptualization, Methodology, Software, Data curation, Formal analysis, Visualization, Writing - original draft, Writing - critical review (20%). **Sacré Pierre-Yves:** Data collection (100%), Data curation, Visualization, Writing - critical review (20%), Supervision, Project administration, Funding acquisition. **Lebrun Pierre:** Writing - critical review (20%), Supervision. **Hubert Philippe:** Visualization, Writing - critical review (20%), Supervision, Project administration, Funding acquisition, authorcontribution. **Ziemons Eric:** Visualization, Writing - critical review (20%), 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.

Acknowledgments

The authors are grateful to Corenthin Mees and Kris De Braekeleer from the Université Libre de Bruxelles who collected the Dataset 1 and 2 with the MicroPhazir® device of ThermoFisher Inc. The authors are also grateful to the whole team of the Vibra4Fake project.

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2020.11.039.

Funding

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

References

[1] M. Forina, P. Oliveri, S. Lanteri, M. Casale, Class-modeling techniques, classic and new, for old and new problems, Chemometr. Intell. Lab. Syst. 93 (2008) 132e148.

[2] P. Oliveri, Class-modelling in food analytical chemistry: development, sampling, optimization and validation issues e a tutorial, Anal. Chim. Acta 982 (2017) 9e19.

[3] M. Bevilacqua, R. Bucci, A.D. Magrì, A.L. Magrì, R. Nescatelli, F. Marini, Chapter 5 - classification and class-modelling, in: F. Marini (Ed.), Chemometrics in Food Chemistry, Data Handling in Science and Technology, vol. 28, Elsevier, Oxford, 2013, pp. 171e233.

[4] A.L. Pomerantseva, O.Ye Rodionova, Concept and role of extreme objects in PCA/SIMCA, J. Chemom. 28 (2014a) 429e438.

[5] O.Ye Rodionova, A.V. Titova, A.L. Pomerantsev, Discriminant analysis is an inappropriate method of authentication, Trac. Trends Anal. Chem. 78 (2016) 17e22.

[6] A.L. Pomerantsev, Acceptance areas for multivariate classification derived by projection methods, J. Chemom. 22 (2008) 601e609.

[7] A.L. Pomerantseva, O.Ye Rodionova, On the type II error in SIMCA method, J. Chemom. 28 (2014b) 518e522.

[8] O. Ye Rodionova, P. Oliveri, A. Pomerantsev, Rigorous and compliant approaches to one-class classification, Chemometr. Intell. Lab. Syst. 159 (2016) 89e96.

[9] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis, Theory and Practice, Springer-Verlag, New York, 2006.

[10] F. Ferraty, A. Goia, P. Vieu, Nonparametric functional methods, New tools for chemometric analysis, in: W. Ha€rdle, Y. Mori, P. Vieu (Eds.), Statistical Methods for Biostatistics and Related Fields, Springer-Verlag, Berlin, 2007, pp. 245e264.

[11] W. Saeys, B. de Ketelaerea, P. Dariusa, Potential applications of functional data analysis in chemometrics, J. Chemometr. 22 (2008) 335e344.

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

[12] S. Wold, M. Sjo€stro€m, 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. 243e282.

[13] K. Vanden Branden, M. Hubert, Robust classification in high dimensions based on the SIMCA method, Chemometr. Intell. Lab. Syst. 79 (2005) 10e21.

[14] M.P. Derde, D.L. Massart, UNEQ, A disjoint modelling technique for pattern recognition based on normal distribution, Anal. Chim. Acta 184 (1986) 33e51. [15] S. Lopez-Pintado, J. Romo, On the Concept of depth for functional data, J. Am. Stat. Assoc. 104 (2009) 718e734.

[16] G.J. Hahn, W.Q. Meeker, Statistical Intervals: A Guide for Practitioners, John Wiley & Sons, New-York, 1991, p. 392p.

[17] K. Krishnamoorthy, T. Mathew, Statistical Tolerance Regions: Theory, Applications, and Computation, John Wiley & Sons, Inc., Hoboken, 2009.

[18] E. Rozet, C. Hubert, A. Ceccato, D. Walthere, E. Ziemons, F. Moonen, K. Michail, R. Wintersteiger, B. Streel, B. Boulanger, P. Hubert, Using tolerance Intervals in pre-study validation of analytical methods to predict in-study results, the fitfor-future-purpose concept, J. Chromatogr. A 1158 (2007) 26e137.

[19] USP 41-NF 36, General Chapter, 1210, Statistical Tools for Analytical Procedure Validation, US Pharmacopeial Convention, Rockville, MD, 2017.

[20] 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. 675p.

[21] J.S. Morris, Functional regression, Annu, Rev. Stat. Appl. 2 (2015) 321e359.

[22] H. Chen, B.R. Bakshi, P.K. Goel, Towards Bayesian chemometrics e a tutorial on some recent advances, Anal. Chem. Acta. 602 (2007) 1e16.

[23] J. Goldsmith, S. Greven, C. Crainiceanu, Corrected confidence bands for functional data using principal components, Biometrics 69 (2013) 41e51.

[24] L. Xiao, V. Zipunnikov, D. Ruppert, C. Crainiceanu, Fast covariance estimation for high-dimensional functional data, Stat. Comput. 26 (2016) 409e421.

[25] J. Josse, F. Husson, Selecting the number of components in principal component analysis using cross-validation approximations, Comput. Stat. Data Anal. 56 (2012) 1869e1879.

[26] M. Gavish, D.L. Donoho, The optimal hard threshold for singular values is 4/3, IEEE Trans. Inf. Theor. 60 (2014) 5040e5053.

[27] D.L. Donoho, M. Gavish, Code Supplement to the Optimal Hard Threshold for Singular Values Is 4/3, 2014. http://purl.stanford.edu/vg705qn9070. (Accessed 27 September 2019). http://purl.stanford.edu/vg705qn9070.

Published in: Analytica Chimica Acta (2021), vol. 1144, pp. 130-149
DOI: 10.1016/j.aca.2020.11.039
Status: Postprint (Author's version)

[28] J. Goldsmith, F. Scheipl, L. Huang, J. Wrobel, J. Gellar, J. Harezlak, M.W. McLean, B. Swihart, L. Xiao, C. Crainiceanu, P.T. Reiss, Refund: Regression with Functional Data, R Package Version 0, 2018, pp. 1e17. https://CRAN.R-project.org/ package¼refund.

[29] R. R Core Team, A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2018. https://www.Rproject.org/.

[30] 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), https://doi.org/10.18637/jss.v076.i01.

[31] Y. Sun, M.G. Genton, D.W. Nychka, Exact and fast computation of band depth for large functional datasets: how quickly can one million of curves be ranked, Stat 1 (2012) 68e74.

[32] N. Tarabelloni, A. Arribas-Gil, F. Ieva, A.M. Paganoni, J. Romo, roahd, Robust Analysis of High Dimensional Data, R Package Version 1, 2018, .4.1. https:// CRAN.R-project.org/package¼roahd.

[33] 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) 469e478.

[34] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627e1639.

[35] J.E. Jackson, J.S. Mudholkar, Control procedures for residuals associated with principal component analysis, Technometrics 21 (1979) 341e349.

[36] Z. Malyjurek, R. Vitale, B. Walczak, Different strategies for class model optimization, A comparative study, Talanta 215 (2020).

[37] S. Kucheryavskiy, Mdatools e R package for chemometrics, Chemometr. Intell. Lab. Syst. 198 (2020) 1e10.

[38] Z. Ghahramani, Probabilistic machine learning and artificial intelligence, Nature 521 (2015) 452e459.