Interpretable one-class classification of Raman spectra using prediction bands estimated by wavelet regression.

T. Hermane Avohou*, Pierre-Yves Sacré, Philippe Hubert, Eric Ziemons.

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

ABSTRACT: Previously, we introduced a novel one-class classification (OCC) concept for spectra. It uses as acceptance space for genuine spectra of the target chemical, a prediction band in the wavelengths' space. As decision rule, test spectra falling substantially outside this band are rejected as non-complying with the target, and their deviations documented in the wavelengths space. This band-based OCC concept was applied to smooth signals like Near-Infrared (NIR) spectra. A regression model based on a smoothed principal component (PC) representation of the training spectra was used to predict unseen trajectories of future spectra. The boundaries of the most central predicted trajectories were chosen as critical trajectories. We now propose a methodology to construct such a band-based one-class classifier for Raman spectra which are sharper and noisier than NIR spectra. The spectra are transformed by a composition of wavelet and principal component (wPC) expansions instead of just a PC expansion in the previous methodology for NIR spectra. Wavelets can capture sharp features of Raman signals and provide a framework to efficiently denoise them. A Bayesian multinormal model is then used to derive predictions of future wPC scores of unseen spectra. These predicted wPC scores are then back-transformed to obtain predictions of future trajectories of unseen spectra in the wavelengths' space, whose most central region defines the acceptance band or space. This band-based one-class classifier successfully classified first derivatives of real pharmaceutical Raman spectra, while enjoying the advantage of documenting deviations from critical trajectories in the wavelengths' space, and hence being more interpretable.

Raman spectra provide specific structural fingerprints of molecules, and thus enable identification of chemicals.[1,2] Their peaks are more chemically interpretable, which is particularly desired in many applications such as identification of chemicals, detection of undesirable or novel compounds, etc.[2] The recent technological advances in instrumentation, including miniaturized measurement devices and the possibilities of cloud-based computing systems have made Raman spectroscopy practically accessible to many fields and applications.[1-3] Especially, in the quality by design context in the pharmaceutical industry, Raman spectroscopy is increasingly used to address qualitative analytical compliance testing problems, including the identification of raw materials and the verification of the compliance of the quality of intermediate drug products and bioprocesses with a reference.[1-3] In such problems, only representative spectra of the targeted identity or quality are available, while non-targeted or undesired identities or quality profiles are theoretically unlimited or cannot be representatively sampled. Hence, such qualitative testing problems are mathematically addressed by predicting the conformity of the Raman spectral features of unknown test samples to those of the representative reference set, based on classification rules defined using the reference set exclusively.[4,5] The type of mathematical techniques involved in such compliance verification tasks are known as rigorous one-class classification (OCC) methods.[4-7] Often there also exist samples of some non-target quality or products. Even if these samples are not representative of all the scenarios of non-target quality or product profiles one might encounter in the future, they can also be used together with the reference set to define the classification rule. In the latter case, the resulting type of models are termed the compliant OCC methods.[4-7]

Several OCC methods exist in chemometrics and may be applied to Raman data (see ref 4-8 for an overview). They are all based on distances of the spectra in a projection space, i.e. the spectra are projected onto a univariate or bivariate metric space where a threshold metric value is determined and used to build an enclosing envelop or acceptance space intended to contain a high proportion, say 100β% of metric values of future spectra of the reference product ($\beta \in]0,1]$).[4,6] The most prominent of such OCC methods is undoubtedly the soft independent modeling of class analogy (SIMCA).[5,7,8]

Recently, ref 9 introduced a band-based OCC paradigm for spectral data which uses as classification rule, the statistical concept of prediction band for future spectra in the wavelengths' space. This prediction band leverages the functional nature of the data, i.e. the fact that a spectrum is a discretized and noisy observation of a random smooth function of the wavelengths.[9,10] It extends the well-known concept of a prediction interval for random variables[11] to random curves. It is delimited by an upper limiting spectrum and a lower limiting spectrum, defined so that the band contains on average a prespecified proportion, say 100β% of future regular spectra of the target product ($\beta \in]0,1]$). Hence, it defines the most typical and acceptable

trajectories for unseen spectra of the target product in a way that, if a new spectrum overlaps entirely or mostly with this band, it is accepted as complying with the target product. Otherwise, it is considered as an extreme or outlier spectrum. This acceptance rule can be tuned by adding decision rules like those used in control charts, say tolerating a small number denoted κ of random points outside the band. This new OCC paradigm has two distinguishing features compared to classical distance-based OCC methods like the SIMCA:

- Firstly, it enables to analyze the deviations patterns of test spectra alongside the whole continuum of the wavelengths. Hence, it provides richer information about the patterns of deviations of false and true negative spectra, including the localization and magnitude of excursions outside the band in the wavelengths' space.

- Secondly, it leverages the concept of statistical prediction region[11] to conveniently accommodate uncertainties about the prediction of the compliance of a single unseen spectrum, which results in a more realistic decision-making framework.

To construct this prediction band, a function-on-scalar regression approach based on a standard or functional principal component (PC) decomposition was used.[9,12] Briefly, the training spectra are projected onto a PC basis. This basis is then optimally truncated and its retained eigenfunctions are regularized by smoothing splines to prevent overfitting.[9,12] Then, a Bayesian multinormal model is used to derive the so-called predictive distribution of the scores in the truncated PC subspace. This represents the probability distribution of scores of future spectra in that subspace.[9,13] This predictive distribution is then back-transformed to obtain the predictive distribution of trajectories of future spectra of the target product given the information present in the training dataset. Ultimately, a band is defined as the boundaries of the $100\beta\%$ most central (deepest) region of that predictive distribution, using the concept of depth statistics which generalizes the univariate quantiles to multivariate and functional observations.[9,14] The method successfully classified near-infrared (NIR) spectra. However, these spectra are known to be far smoother than Raman spectra which exhibit remarkably sharper bands, are noisier and harder to denoise.[2] These two peculiarities of Raman signals namely their complex shape and high noise level, are known to increase the risk of overfitting in modeling.[12,15] For such signals, truncated PC basis may not effectively prevent overfitting. Specifically designed basis expansions with more efficient denoising methods are needed as demonstrated in this work.[12,15]

This article proposes a novel regression framework for estimating the predictive distribution of future spectra - the building block of the band-based OCC. This regression framework is fit for noisy signals with sharp peaks, sampled on a high dimensional grid, like Raman or FT-IR spectra. The framework depicted on Figure 1, involves in a first step, expanding each (pre-processed) spectrum into a series of uncorrelated coefficients in an orthonormal wavelet basis. Wavelet bases (Figure 2) are known to be versatile because of their so-called time-scale localization property that enables them to flexibly model a wide class of smooth or piecewise-smooth signals having non-stationary features like sharp peaks and spikes.[12,15-18] In a second step, a thresholding function is applied to the coefficients of each spectrum to denoise and regularize them (Figure 1). This enables to select a subset of truly informative and noise-free coefficients, reducing the dimensionality and smoothing each spectrum in the wavelengths' space.[17-19] These adaptivity and denoising features make the wavelet regression framework particularly well-suited for modelling Raman spectra. In a third step, the regularized coefficients are jointly modelled after projecting them onto an uncentered PC basis for further dimensionality reduction, resulting in the so-called wavelet principal component (wPC) transform.[12,15,20] In a fourth step, a Bayesian multinormal model is used to derive the predictive distribution of the wPC scores.[9,13] This distribution is back-transformed by an inverse wPC transform to obtain the predictive distribution of unseen trajectories of future spectra (step 5, Figure 1). Once the predictive distribution is obtained, the concept of statistical depth is used to select its $100\beta\%$ most central region whose boundaries delimit the acceptance band in the wavelengths' space.[9,14]

The method successfully classified three real Raman pharmaceutical datasets described in ref 2,21. Its performances are compared to those of the previously developed PC-based band[9] and two advanced SIMCA methods.[8]

2. METHODS

2.1. Model formulation.

Let $x(t)$ be a Raman spectrum fingerprint of a given target product, t being a wavelength variable (i.e., the Raman shift in cm^{-1}). We aim at constructing a prediction band delimited by an upper limiting trajectory and a lower limiting trajectory as acceptance region for a whole future measurement of $x(t)$. To achieve this, N discretized and noisy spectra indexed by $i = 1, \cdots, N$ are measured on independent samples of the target product on a regular grid of K values t_k of t $(k = 1, \cdots, K)$. In the wavelet analysis context, K must be a power of 2, i.e., $K = 2^J$ with $J > 0$ being an integer.

Let $\mathbf{x}_i = [x_i(t_1), \cdots, x_i(t_K)] = [x_{i1}, \cdots, x_{iK}]$ be the $1 \times K$ vector of intensities of the ith spectrum, and $\mathbf{X}$ the $N \times K$ spectral matrix whose ith row is $\mathbf{x}_i$.

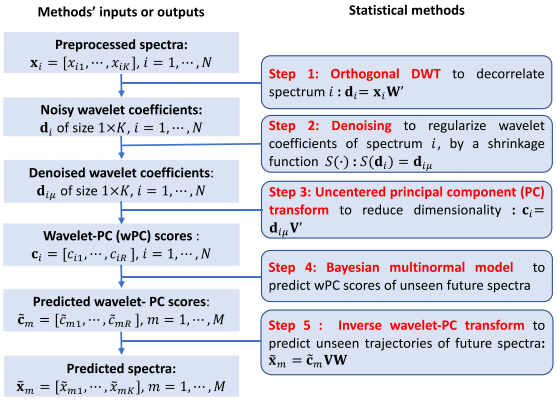

Figure 1. Regression workflow to predict the unseen trajectories of future Raman spectra. Notes: N, K, R, and M are respectively the numbers of training spectra, wavelengths, uncentered PCs, and predicted spectra; $\mathbf{W}'$ and $\mathbf{V}'$ are respectively the wavelet and uncentered PC transforms' matrices.

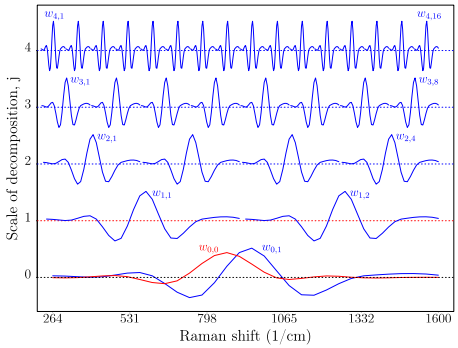

Figure 2. Example of wavelet basis for a signal of length $K=2^5$: Daubechies' least asymmetric wavelets (with 10 vanishing moments) with $J = 5$ scales of decomposition labeled as $j = 0, \cdots, 4$. There are 1 smoothing function at $j = 0$ (red curve, w_{00}) and 2^j detail functions at every j (blue curves, w_{jr}).

The model describing the observed variation of spectra is a regression model where the response to be predicted is the whole trajectory of the spectrum. It is termed "function-on-scalar" regression model,[9,12] and is expressed at the kth wavelength t_k as

$$x_i(t_k) = \mu_i(t_k) + e_i(t_k), \quad (1)$$

where $\mu_i(t_k)$ is the correlated spectrum-specific mean at t_k, and $e_i(t_k)$ is the white noise at t_k with the assumption $e_i(t_k) \sim \text{Normal}(0, \sigma_i^2)$, $\sigma_i > 0$.

Note that $\mu_i(t_k)$ can be further decomposed as $\mu_i(t_k) = \mu(t_k) + u_i(t_k)$ where $\mu(t_k)$ is the correlated class or overall mean signal at t_k and $u_i(t_k)$ is the zero-mean and correlated Gaussian deviation of $x_i(t_k)$ from the class mean $\mu(t_k)$ at t_k. Hence, the ith spectrum may be decomposed into the class mean plus its Gaussian deviation (error) from the class mean, plus an additive white noise. The vectorized version of (1) is written

$$\mathbf{x}_i = \boldsymbol{\mu}_i + \mathbf{e}_i \quad (2)$$

where $\boldsymbol{\mu}_i$ is the $1 \times K$ vector of spectrum-specific mean i.e. the noise-free spectrum for $\mathbf{x}_i$; $\mathbf{e}_i$ is $1 \times K$ vector of additive white noise for $\mathbf{x}_i$, with the assumption that $\mathbf{e}_i \sim \text{Multinormal}(\mathbf{0}, \sigma_i^2 \mathbf{I}_K)$, $\mathbf{I}_K$ being the $K \times K$ identity matrix.

2.2. Orthogonal discrete wavelet transform (DWT, step 1).

Each spectrum $\mathbf{x}_i$ is projected onto an orthogonal discrete wavelet basis depicted on Figure 2 (see also Supporting Information S1.1).

Briefly, for a signal of length K, the orthogonal wavelet basis is a set of K localized basis functions structured in $J = \log_2 K$ orthogonal frequencies labeled from the coarsest to the finest as $j = 0, \cdots, J-1$. Each frequency scale j includes 2^j similar detail functions denoted $w_{jr}(t_k)$ with $r = 1, \cdots, 2^j$. They are intended to capture a specific high frequency oscillation at different locations of the analyzed spectrum. The scale $j = 0$ includes in addition one smoothing or low frequency function denoted $w_{00}(t_k)$ (Figure 2).[16,18]

Similarly to the eigenfunctions in the well-known PC transform, each of the wavelet basis functions i.e. $w_{jr}(t_k)$ is correlated with any spectrum $\mathbf{x}_i$ to derive its scores or coefficients herein denoted d_{ijr} where d_{i00} always refers to the smoothing coefficient (see Supporting Information S1.2 for illustrations of the decomposition of real Raman spectra into wavelet coefficients and their levelwise reconstruction). Hence, $\mathbf{x}_i$ can be reconstructed at any wavelength t_k as linear combination of the wavelet functions as

$$x_i(t_k) = d_{i00} w_{00}(t_k) + \sum_{j=0}^{J-1} \sum_{r=1}^{2^j} d_{ijr} w_{jr}(t_k). \quad (3)$$

The decomposition in (3) can simply be written in matrix form. If $\mathbf{W}'$ denotes the $K \times K$ orthogonal discrete wavelet transform (DWT) matrix and $\mathbf{W}$ its inverse or transpose, then, the wavelet transform of $\mathbf{x}_i$ and hence of Model (2) is written

$$\mathbf{x}_i \mathbf{W}' = \boldsymbol{\mu}_i \mathbf{W}' + \mathbf{e}_i \mathbf{W}' \quad (4)$$

resulting in the wavelet domain model

$$\mathbf{d}_i = \mathbf{d}_{i\mu} + \mathbf{d}_{ie} \quad (5)$$

implying for the coefficient indexed by j and r that

$$d_{ijr} = d_{i\mu jr} + d_{iejr} \quad (6)$$

In Equations (4)-(6), $\mathbf{d}_i$ is the $1 \times K$ vector of uncorrelated noisy wavelet coefficients for $\mathbf{x}_i$; $\mathbf{d}_{ie}$ is the $1 \times K$ vector of uncorrelated wavelet coefficients for the white noise $\mathbf{e}_i$; $\mathbf{d}_{i\mu}$ is the $1 \times K$ vector of true and noise-free coefficients to be recovered for $\mathbf{x}_i$; d_{ijr}, $d_{i\mu jr}$, and d_{iejr} are the uncorrelated elements respectively of $\mathbf{d}_i$, $\mathbf{d}_{i\mu}$, and $\mathbf{d}_{ie}$, where $j = 0$ and $r = 0$ label the smoothing coefficient, while $j = 0, \cdots, J-1$ and $r = 1, \cdots, 2^j$ label the detail coefficients.

The denoising of $\mathbf{d}_i$ or recovering of the true coefficients $\mathbf{d}_{i\mu}$ is achieved *via* the wavelet thresholding.[17-19] It is worthwhile to recall that, because of the orthogonality of $\mathbf{W}'$, $\mathbf{d}_{ie} \sim \text{Multinormal}(\mathbf{0}, \sigma_i^2 \mathbf{I}_K)$, $\mathbf{I}_K$ being the identity matrix.

2.3. Wavelet denoising (step 2).

The Raman spectrum $x(t)$ is typically piecewise smooth. Hence, the true detail coefficients $d_{i\mu jr}$ $(j = 0, \cdots, J-1, r = 1, \cdots, 2^j)$ of $\mathbf{x}_i$ in (6) typically form a sparse sequence by scale j, i.e. many of them are zero.[18] This means that at a given scale j, the noisy detail coefficients d_{ijr} are a mixture of small magnitude coefficients which are likely exclusively noisy (i.e. the true value $d_{i\mu jr} = 0$ and the observed value d_{ijr} is noise only), and large magnitude coefficients which concentrate true signal contaminated with white noise (i.e. the true value $d_{i\mu jr} \neq 0$ and the observed value d_{ijr} is a mixture of noise and signal, see Supporting information S1.2 for the illustration of the structure and sparsity of wavelet coefficients of real Raman signals). Hence, small magnitude coefficients are ideally zeroed without loss of information whereas high magnitude coefficients might be shrunk towards 0, but not zeroed. This wavelet shrinkage procedure is a critical step. It possesses the properties of denoising, reducing the dimensionality, and regularizing (smoothing) each spectrum in the wavelengths' space.[18,19]

There exists a wide variety of well-established statistical procedures to perform shrinkage of sparse wavelet coefficients of a single signal vector under the Gaussian white noise assumption in (1)-(6) (see ref 17,18 for an overview). The core idea of these procedures is to estimate some threshold below which coefficients are zeroed, and above which they are either kept (hard-thresholding) or ideally shrunken by the threshold (soft-thresholding) to remove their noise component.[17,18] The soft-thresholding approach is used in this work.

Whatever the method used to estimate the threshold, it is proportional to the noise scale σ_i in (2) and (6), and hence an initial estimate $\hat{\sigma}_i$ of σ_i is required. It is commonly and robustly computed in the wavelet space from the median absolute deviation (MAD) of coefficients of the finest scale $J-1$, which are likely exclusively noisy,[18,19] as

$$\hat{\sigma}_i = 1.4826 \times \text{MAD}\{d_{i(J-1)r}, r = 1, \cdots, 2^{(J-1)}\}. \quad (7)$$

Here, we describe a simple threshold estimation method yet well-adapted to our noisy Raman data, called the "levelwise universal threshold".[17,19] Given the noise scale in (7), this method estimates the universal threshold for coefficients at each scale of decomposition.[17] It is based on the statistical principle that the largest of any sequence of n Gaussian random variables (with mean 0 and variance 1) is roughly $\sqrt{2 \cdot \log\ n}$. Estimating this threshold levelwise from coefficients with noise variance $\hat{\sigma}_i^2$ results in J thresholds, each denoted $\hat{\rho}_{ij}$ $(j = 0, \cdots, J-1)$ and computed as

$$\hat{\rho}_{ij} = \hat{\sigma}_i \sqrt{2 \cdot \log\,(2^j)}. \quad (8)$$

Then, denoising using the thresholds in (8) and the soft-thresholding function enables to remove the noise and recover the true coefficients as

$$d_{i\mu jr} = 0, \qquad \text{if } |d_{ijr}| < \hat{\rho}_{ij} \quad (9)$$

or

$$d_{i\mu jr} = \text{sign}(d_{ijr}) \cdot |d_{ijr} - \hat{\rho}_{ij}|, \quad \text{if } |d_{ijr}| > \hat{\rho}_{ij} \quad (10)$$

where d_{ijr} and $d_{i\mu jr}$ are respectively the noisy and denoised coefficients as defined in (6) and $\hat{\rho}_{ij}$ is the estimated maximal noise magnitude at scale j for spectrum $\mathbf{x}_i$.

Simply explained, the thresholding procedure in (9)-(10) zeroes coefficients smaller than the maximal noise magnitude, while coefficients greater than this threshold are shrunken to remove their noise component. This procedure is intended to ensure noise-free estimates of the detail coefficients. It is applied to each noisy coefficients' vector $\mathbf{d}_i$. This results in the $1 \times K$ vector of recovered wavelet coefficients i.e., $\mathbf{d}_{i\mu}$ with elements $d_{i\mu jr}$ for $\mathbf{x}_i$ as defined in (5)-(6), and in the $N \times K$ matrix of denoised coefficients of all spectra denoted $\mathbf{D}_\mu$.

The smoothed spectrum i.e., $\boldsymbol{\mu}_i$ in (2) can then be recovered from the denoised coefficients' vector $\mathbf{d}_{i\mu}$ as

$$\boldsymbol{\mu}_i = \mathbf{d}_{i\mu}\mathbf{W}. \quad (11)$$

2.4. Principal components transform and prediction of future scores and spectra (step 3 -5).

The recovered wavelet coefficients using the procedure in (7)-(10) at step 2 are modelled using a Bayesian normal model,[13] to obtain predictions of their future values for unseen spectra. Back transforming these predictions of future wavelet coefficients enables to obtain predictions of unseen trajectories of future spectra.

Firstly, the recovered coefficients are further transformed by an uncentered PC expansion,[22] enabling to jointly model them (i.e., their linear combinations), to further reduce the dimensionality and to speed-up computations. This composition of wavelet and PC expansions is termed the wavelet principal component (wPC) transform.[15,20]

Let

$$\mathbf{C} = \mathbf{D}_\mu \mathbf{V}' \quad (12)$$

be the uncentered PC expansion of $\mathbf{D}_\mu$ derived by its singular value decomposition (SVD),[22] where $\mathbf{D}_\mu$ is the $N \times K$ matrix of recovered wavelet coefficients for all spectra; $\mathbf{V}'$ is the $K \times R$ reduced matrix of right singular vectors of $\mathbf{D}_\mu$, $R \leq N$ denoting the number of right singular vectors or wPCs retained to preserve the bulk of the information, say 99.5% for a near-lossless transform; $\mathbf{C}$ is the $N \times R$ matrix of wPC scores whose ith row is $\mathbf{c}_i = [c_{i1}, \cdots, c_{iR}]$ of size $1 \times R$.

Secondly, under normality assumptions of the columns of $\mathbf{C}$, predictions of future values of each of the R wPC scores' variables can be obtained independently by a non-standardized Student-t distribution with $N-1$ degrees of freedom,[13] i.e.,

$$(\tilde{c}_r|\mathbf{C}) \sim \text{Student}_{(N-1)}[\bar{c}_r, (1+N^{-1}) \cdot s_r^2], \quad (13)$$

where $\tilde{c}_r$ is a future value of the rth wPC scores' variable (column of $\mathbf{C}$), "|" means "given", and $\mathbf{C}$ is the scores' data matrix; $\bar{c}_r = N^{-1}\sum_{i=1}^{N} c_{ir}$ and $s_r^2 = (N-1)^{-1}\sum_{i=1}^{N}(c_{ir} - \bar{c}_r)^2$ are respectively the sample mean and sample variance of c_{ir}; $\bar{c}_r$ and $(1+N^{-1}) \cdot s_r^2$ are respectively the location and squared-scale parameters of the Student-t distribution.[13] The details of the derivation of the prediction model in (13) can be found in regression textbooks, for example in ref 13.

Thirdly, sampling independently M times from (13) for each of the R wPC scores' variables (columns of $\mathbf{C}$), yields M predictions of the wPC scores' vectors for future unseen spectra, each denoted $\tilde{\mathbf{c}}_m = [\tilde{c}_{1m}, \cdots, \tilde{c}_{Rm}]$ with $m = 1, \cdots, M$. Stacking the M predicted wPC scores' vectors, results in a $M \times R$ matrix denoted $\tilde{\mathbf{C}}$ approximating the predictive distribution of the wPC scores' vectors.

Fourthly, back transforming the predicted wPC scores i.e., the $M \times R$ matrix $\tilde{\mathbf{C}}$, to the wavelets' space and then to the wavelengths' space by a double inverse matrix multiplication as

$$\tilde{\mathbf{X}} = \tilde{\mathbf{C}}\mathbf{V}\mathbf{W}', \quad (14)$$

enables to obtain a M-sample of predictions of unseen trajectories of future spectra denoted $\tilde{\mathbf{X}}$ of dimension $M \times K$, whose mth row $\tilde{\mathbf{x}}_m$ $(m = 1, \cdots, M)$ is a random predicted trajectory for $x(t)$. In (14) $\tilde{\mathbf{C}}$ is the $M \times R$ matrix whose row $\tilde{\mathbf{c}}_m$ $(m = 1, \cdots, M)$ are the predicted wPC scores' vectors for future unseen spectra, $\mathbf{W}$ and $\mathbf{V}$ are respectively the $K \times K$ transpose of the wavelet transform matrix defined in (4) and the $R \times K$ right singular vectors' matrix defined in (12). Hence, this M-sample $\tilde{\mathbf{X}}$ simulates possible trajectories of single future spectra one might expect given the spectral data at hand, $\mathbf{X}$.

2.5. Ranking predicted spectra to set band limits (step 6).

Using the predicted spectra $\tilde{\mathbf{X}}$ in (14), a simultaneous prediction band is defined for future spectra by setting upper and lower limits at every t_k. To achieve this, a center-outwards ordering of the predicted spectra is performed to select the $100\beta\%$ most central ones.[14] This ordering can be achieved by multivariate quantiles methods such as the Mahalanobis depth, or the modified band depth (mBD). The latter leverages the functional nature of the spectra and is preferred in this work.[14]

The mBD of a predicted spectrum, $\tilde{\mathbf{x}}_m$, w.r.t. all predicted spectra, $\tilde{\mathbf{X}}$, is the average fraction of spectral points where $\tilde{\mathbf{x}}_m$ falls inside bands delimited by all pairs of spectra of $\tilde{\mathbf{X}}$.[14] It is calculated as follows:

- Firstly, all pairs of spectra (rows) of $\tilde{\mathbf{X}}$ are identified.

- Secondly, $\tilde{\mathbf{x}}_m$ is projected onto the band formed by each pair and the proportion of points where it is inside that pair is calculated.

- Thirdly, this proportion is averaged for all pairs of spectra of $\tilde{\mathbf{X}}$.

Of two spectra, the one having the highest mBD value is said to be deeper or more central w.r.t $\tilde{\mathbf{X}}$.[14] This procedure is used to select the 100β% most central predicted spectra whose upper and lower boundaries define the limiting trajectories or acceptance space for the class.

2.6. Model tuning, decision rules, and spectrum analysis.

The estimated band is evaluated either by a one-step internal validation or by a cross-validation step, where decision rules like those used in control charts are defined and tuned to optimize its performances, for example by tolerating a few random excursions outside the band.

If $z(t)$ denotes an out-of-training spectrum and $\mathbf{z}$ of size $1 \times K$ its measured value, then the test of compliance of $\mathbf{z}$ with the training set proceeds in two steps. Firstly, $\mathbf{z}$ is projected onto the wavelet basis $\mathbf{W}'$ as in (4). Its coefficients are denoised using the same shrinkage procedures of (7)-(10), and its reconstruction is obtained by an inverse DWT as in (11). Secondly, the obtained reconstruction is projected onto the band and wavelengths where it falls outside are localized, and the magnitude of the deviations are quantified. The number of excursions outside the band denoted κ is used as a tuning parameter of the classification rule.

In the rigorous optimization strategy, the acceptance rule for a future unseen spectrum may be defined as the smallest κ, producing a sensitivity greater than 100β% ($\beta \in]0,1]$). Alternatively, in the compliant optimization strategy where samples of non-target products are available during the tuning step, these samples may also be used to choose κ to have a good compromise between prediction sensitivity and specificity.

2.7. Software.

The proposed methodology is implemented in R statistical software.[23] Orthogonal DWT, coefficients' thresholding, and inverse DWT are performed respectively with the *wd()*, *threshold() and wr()* routines of the *WaveThresh* package.[18] The uncentered PC decomposition of the denoised wavelet coefficients and the Monte Carlo sampling of the predicted wPC scores are done respectively with the *svd() and rt()* routines of base R. The ordering of predicted spectra by the mBD algorithm is done with the *MBD()* routine of the package *roahd*.[24]

3. EXPERIMENTAL

3.1. Datasets.

Three real Raman spectroscopy datasets described in ref. 2 and 21 were used to evaluate the classification performances. They were measured on tablets of three groups of model-drug formulations differing in the ratio of active pharmaceutical ingredient (API) and excipients' contents (see Supporting Information S2 for detailed descriptions of the formulations). Measurements were performed through the blisters with a portable Truscan RM® (Thermo Scientific, USA) covering a spectral range of 250-2875 cm^{-1}. The Truscan may be particularly noisy in the measurement mode.[2] To remove baseline effect due to fluorescence, the first derivative of all spectra was estimated using the Savitzky-Golay (SG) algorithm[25] with a polynomial degree of 2 and a window size specific to each group of drugs. The estimated derivatives were normalized to unit area and truncated to the spectral range of 264.5-1599.4 cm^{-1} with K = 1024 wavelengths.

Paracetamol dataset (Figure 3A). The first dataset comprises spectra of tablets of five paracetamol-based formulations having high API/excipients ratio.[21] Twenty tablets were measured per analysed batch. They were coded as P01 (reference 1, 4 batches, API: paracetamol 1000 mg), P10 (reference 2, 4 batches, APIs: paracetamol 250 mg, acetylsalicylic acid 250 mg and caffeine 65 mg), P11 (3 batches, APIs: paracetamol 500 mg and caffeine 65 mg), P12 (1 batch, API: paracetamol 500 mg) and P13 (2 batches APIs: paracetamol 325 mg and tramadol 37.5 mg). Derivatives were estimated with a SG window size of 15.

Ibuprofen dataset (Figure 3B). The second dataset comprises spectra of tablets of five ibuprofen-based formulations having different API/excipient ratios.[2,21] These formulations were challenging for Raman spectroscopy because of the high fluorescence background caused by a wide variety of coating nature and colors. They were coded as I03 (reference 1, 4 batches with 20 tablets each, API: ibuprofen 400 mg), I04 (reference 2, 4 batches with 20 tablets each, API: ibuprofen 600 mg), I08 (1 batch with 10 tablets, API: ibuprofen 600 mg), I09 (4 batches with 10 tablets each, API: ibuprofen 200 mg) and I10 (3 batches with 10 tablets each, API: ibuprofen 400 mg). Derivatives were computed with a SG window size of 35.

Simvastatin dataset (Figure 3C). The third dataset is described in ref 2 and comprises spectra of tablets of four simvastatin-based formulations having low ratios of API to excipients' contents. Twenty tablets were measured per analysed batch. They were coded as S01 (1 batch, API: simvastatin 20 mg), S02 (2 batches, API: simvastatin 20 mg), S03 (reference 1, 2 batches, API: simvastatin 20 mg) and S04 (reference 2, 2 batches, API: simvastatin 20 mg). Derivatives were computed with a SG window size of 15.

3.2. Evaluation method.

Two validation strategies were used to evaluate the classification performances of the wPC-based band. Firstly, a Monte Carlo validation was used to estimate the expectation and standard deviation of its true positive rate (TPR) and false positive rate (FPR). These criteria were compared to those of the PC-based band[9] and two advanced SIMCA methods.[8] Secondly, an external validation approach was used to evaluate the reliability of its predictions and demonstrate its applicability in real pharmaceutical settings and its advantages over the two benchmark SIMCA.

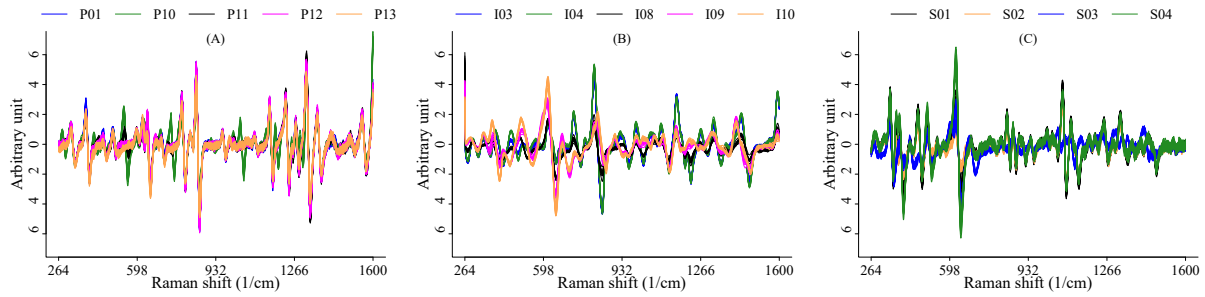

Figure 3. Preprocessed Raman datasets: (A) Paracetamol formulations; (B) Ibuprofen formulations; (C) Simvastatin formulations (see Supporting Information S2 for detailed descriptions of the formulations and PC scores' plots).

Parametrization of the models. All models were calibrated with a nominal probability content of $100\beta\% = 95\%$.

The wPC-based band was built using Daubechies' least asymmetric wavelet basis with 10 vanishing moments[16]. The number of vanishing moments is chosen to maximize the smoothness of these functions.[16] Used with the denoising and approximation models in (7)-(10) and (11), this basis enabled a remarkably good reconstruction and denoising of the preprocessed spectra, preserving the peaks of the three groups of drugs as shown in Figure 4. The predictive distributions of future spectra were approximated with a Monte Carlo sample of size $M = 60{,}000$ spectra.

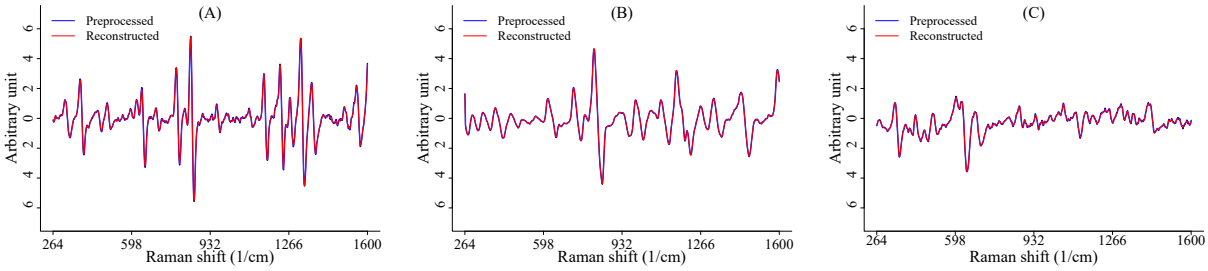

Figure 4. Quality of the reconstruction by the wavelet regression of one reference spectrum of: (A) the paracetamol formulation P01; (B) the ibuprofen formulation I03; (C) the simvastatin formulation S03.

The two benchmark SIMCA methods used are two versions of the recently introduced data-driven SIMCA (DD-SIMCA).[8] They use two different chi-square distributions to model the squared Mahalanobis scores distance (SD) and the squared Euclidean orthogonal distance (OD). The parameters (degrees of freedom and scaling factors) of the distributions are estimated from the data using either the moment method or the robust method.[8] The squared SD and OD are then scaled by their scaling factors, and summed resulting in a joint chi-square distributed total distance whose $100\beta\%$ critical limit defines a triangular acceptance area.[8] Both models are implemented with the function *simca()* of the package *mdatools*[26] of R.[23]

Validation methods. The Monte Carlo validation proceeded as follows. Reference formulations P10, I04 and S04 were used to calibrate the models respectively for the paracetamol, ibuprofen and simvastatin datasets. At each iteration, each reference set was randomly split into a training set to build models (40, 40 and 25 spectra respectively for P10, I04 and S04), an internal validation set to optimize sensitivity or tune acceptance rules (20, 20 and 5 spectra respectively for P10, I04 and S04), and a test set to evaluate the TPR (20, 20 and 10 spectra respectively for P10, I04 and S04). All spectra of non-references were used as test sets to evaluate the FPR per group of formulations. A total of 50 Monte Carlo iterations was used and, the average and standard deviation of TPR and FPR were computed for each OCC method.

The external validation proceeded as follows. Half of the batches of each of the reference formulations P01 (40 spectra), I03 (40 spectra) and S03 (20 spectra) was used to calibrate models, which were optimized by 10-fold cross-validation. The remaining batches of each reference and all batches of non-reference formulations were used as independent test sets to evaluate the reliability of the TPR and FPR.

4. RESULTS AND DISCUSSION

4.1. Predicted trajectories of future spectra and band limits.

Figure 5 illustrates the predicted trajectories of future spectra for reference formulations P01, I03 and S03 as examples. It provides qualitative visual evidence of the unbiased character of the prediction models as the mean predicted spectra almost perfectly overlap with the mean calibration spectra. The critical trajectories (red lines) are used as reference to classify the test spectra and evaluate their deviations patterns. A more comprehensive visual assessment proving the agreement between the distribution of the wPC scores predicted by (13) and the calibration wPC scores is provided in Supporting Information S3.

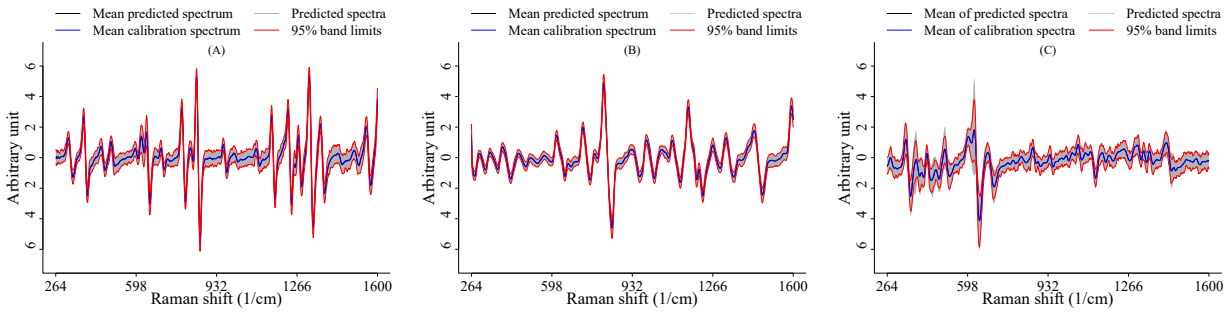

Figure 5. Predicted trajectories of future spectra and, upper and lower band limits for three reference formulations: (A) paracetamol P01; (B) ibuprofen I03; (C) simvastatin S03. Notes: A total of $M = 60000$ spectra were predicted.

4.2. Performances in Monte Carlo validation.

The wPC-based band was conservative recognizing on average more than 95% of spectra of the reference formulations (Table 1). On the contrary, the DD-SIMCA methods were overly liberal recognizing on average far less than the desired 95% of the reference spectra. Both the wPC-based band and the SIMCA effectively discriminated the non-reference formulations in each group (Table 1).

Contrary to the wPC-based band, the PC-based band performed poorly, failing to properly discriminate non-reference ibuprofen formulations (Table 1). This is due to the high level of noise of the Raman spectra causing overfitting that simple truncation of the PC basis and regularization of the retained eigenfunctions by smoothing splines cannot efficiently address.[12,15] For NIR spectra which are smoother with higher signal-to-noise ratio than Raman spectra, this overfitting was not observed, and the PC-based band performed accurately.[9] In the current method, the wavelet shrinkage provided a framework to efficiently denoise the Raman spectra and control overfitting, while preserving meaningful Raman peaks.[15,20]

4.3. Performances (reliability) in external validation.

The reliability of the prediction of the wPC-based band was confirmed by the external validation using independent test sets. It correctly recognized the bulk of the reference spectra inside each group of formulations, even in instances of low training sample sizes of simvastatin (Tables 2). On the contrary, both DD-SIMCA methods showed substantial undercoverage with TPRs far lower than the desired theoretical proportion of 95% (Table 2). All FPRs were zero for both the wPC-based band and the SIMCA methods. Despite the fact that the PC-based band showed high TPRs, it failed to properly reject non-reference formulations in the three groups of drugs.

It must be emphasized that, despite our proposed methodology generally classifies better than the SIMCA, it might underperform if its parameters are not well-estimated.

Table 1. Monte Carlo validation - Average (standard deviation in brackets) true and false positive rates (TPR and FPR) of classification of paracetamol, ibuprofen and simvastatin formulations by the wavelet principal component (wPC)-based band, the principal component (PC)-based band, the moment-based and robust data-driven SIMCA (mDD-SIMCA and rDD-SIMCA).

Method	Paracetamol (P10 vs. P01, P11-P13)			Ibuprofen (I04 vs. I03, I08-I10)			Simvastatin (S04 vs. S01-S03)		
	Parameter	TPR (%)	FPR (%)	Parameter	TPR (%)	FPR (%)	Parameter	TPR (%)	FPR (%)
wPC-based band	$\kappa = 3$[a]	96.9 (5.4)	0.0 (0.0)	$\kappa = 1$	96.5 (5.3)	0.0 (0.0)	$\kappa = 2$	97.8 (4.2)	0.0 (0.0)
PC-based band	$\kappa = 0$	98.0 (2.7)	0.0 (0.0)	$\kappa = 0$	99.9 (0.3)	25.5 (4.5)	$\kappa = 0$	99.0 (1.3)	0.0 (0.0)
mDD-SIMCA	$R = 1$[b]	89.1 (7.4)	0.0 (0.0)	$R = 1$	84.4 (7.2)	0.0 (0.0)	$R = 1$	58.5 (15.6)	0.0 (0.0)
rDD-SIMCA	$R = 1$	83.1 (9.5)	0.0 (0.0)	$R = 1$	75.5 (13.9)	0.0 (0.0)	$R = 1$	28.0 (27.5)	0.0 (0.0)

[a]κ = optimized number of tolerated points outside the band (sensitivities in the one-step optimization phase are reported in Supporting Information S4). [b]R = optimized number of PCs for the SIMCA.

Table 2. External validation – Reliability of the true and false positive rates (TPR and FPR) of classification of paracetamol, ibuprofen and simvastatin formulations by the wavelet principal component (wPC)-based band, the principal component (PC)-based band, the moment-based and robust data-driven SIMCA (mDD-SIMCA and rDD-SIMCA).

Method	Paracetamol (P01 vs. P10-P13)			Ibuprofen (I03 vs. I04-I10)			Simvastatin (S03 vs. S01, S02, S04)		
	Parame-ter	TPR (%)	FPR (%)	Parameter	TPR (%)	FPR (%)	Parame-ter	TPR (%)	FPR (%)
wPC-based band	$\kappa = 1$[a]	97.5	0.0	$\kappa = 1$	97.5	0.0	$\kappa = 1$	95.0	0.0
PC-based band	$\kappa = 0$	100.0	47.5	$\kappa = 0$	100.0	15.0	$\kappa = 0$	100.0	5.0
mDD-SIMCA	$R = 1$[b]	65.0	0.0	$R = 1$	80.0	0.0	$R = 1$	75.0	0.0
rDD-SIMCA	$R = 1$	25.0	0.0	$R = 1$	60.0	0.0	$R = 1$	75.0	0.0

[a]κ = optimized number of tolerated points outside the band (sensitivities in the 10-fold cross-validation optimization phase are reported in Supporting Information S4). [b]R = optimized number of PCs for the SIMCA.

4.4. Analysis of deviations' patterns.

Besides its probabilistic character, the most distinguishing feature of the band-based OCC is that it enables to investigate and use the patterns (i.e. the localization and magnitude) of deviations from the normal acceptable trajectories in the wavelengths' space.[9] Three illustrations based on the external validation results are given (Figure 6).

As first illustration, Figure 6A contrasts at each wavelength the average deviations of test spectra of P01 and P10 w.r.t the band calibrated for P01 on Figure 5A. Out of the 40 test spectra of P01, 5 spectra deviated randomly at 1 point, and 1 spectrum at 2 points. These deviations were barely noticeable in magnitude (Figure 6A). On the contrary, Each spectrum of P10 deviated substantially in magnitude with an average of 312 ± 54 points per spectrum (see the localizations and magnitudes on Figure 6A). An analysis of these deviations' patterns enabled us to attribute most of the deviating peaks to the additional APIs i.e., acetylsalicylic acid and caffeine). The same deviations' profiles can be derived for P11, P12 and P13 whose spectra deviated at 55 ± 6, 17 ± 4 and 123 ± 8 points on average.

As second illustration, Figure 6B contrasts the deviations' patterns of test spectra of I03 and I04 w.r.t the band constructed for I03 on Figure 5B. Only two I03 spectra went outside the band at 1 point and 2 points, while all spectra of I04 deviated substantially, on average at 64 ± 4 points per spectrum. An analysis of this deviations pattern enabled us to attribute the most deviating peak (around 480cm^{-1}) to starch derivatives which is consistent with the composition differences between the two formulations. Indeed, I04 has a higher proportion of sodium starch glycolate compared to I03). Likewise, spectra of I08, I09, and I10 deviated substantially at 650 ± 6, 726 ± 18 and 747 ± 5 points on average.

As third illustration, Figure 6C contrasts the deviations' patterns of test spectra of S03 and S04 w.r.t the band constructed for S03 on Figure 5C. On average, 0.3 ± 0.6 and 459 ± 12 points of spectra of S03 and S04 respectively were outside the band (see localizations and magnitudes on Figure 6C). The analysis of this deviations' patterns did not enable us to attribute the deviating ranges. Looking back at the raw spectra, we observed that S03 exhibited a much higher fluorescence background. Consequently, several peaks of the formulation were masked, or their intensity was lowered compared to S04. Therefore, it was difficult to attribute each individual peak to a specific compound.

Besides the interpretation and assignment of the deviating bands, the deviations' pattern may be considered as a new outlyingness signature (in position and magnitude) that might be used in further investigations. It may enable the analyst to identify previously encountered deviations.

It is worthwhile to mention that the acceptance rule of the band-based classifier which is based on the number of tolerated points outside the upper and lower limits, could be further tuned using the so-called compliant optimization strategy, if representative samples of the non-targets were available during the model tuning phase in cross-validation. In this case, the number of tolerated points outside the limits should be chosen to have a good compromise between sensitivity and specificity, i.e. a good accuracy.

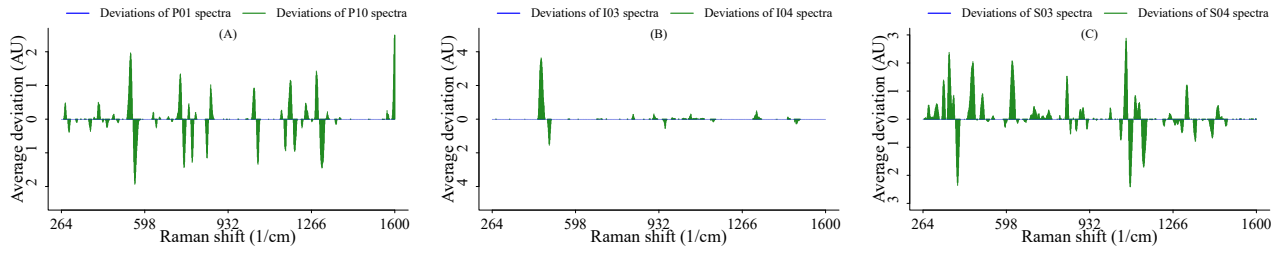

Figure 6. Comparison of average deviations from 95%-prediction limits of test spectra of a reference formulation (blue bars) vs a non-reference formulation (green bars) for: (A) paracetamol, P01 vs. P10; (B) ibuprofen, I03 vs. I04; (C) simvastatin, S03 vs. S04.

5. CONCLUSION

Raman spectroscopy is experiencing an increasing interest in many fields. The pharmaceutical industry, with the process analytical technology initiative, integrates several Raman sensors in the production lines to monitor and predict the quality of produced medicines. It also uses handheld devices for many quality-control tasks such as identification

of raw materials and post-marketing surveillance. Most of these applications (e.g., identification of raw materials, end-point detection of blending process, falsified medicines detection, etc.) use the Raman spectral information as fingerprint to check the conformity of the analyzed samples with known reference signatures.

This work presents a novel chemometric tool to achieve these conformity testing tasks. It uses wavelet regression to model sharp and noisy signals, and predict their whole trajectories. These predictions are then used to delimit an acceptance band for OCC. This band successfully classified real Raman spectra, and even showed sensitivities better than those of two recent versions of SIMCA.

Another value-added is the possibility to obtain and visualize the deviations' patterns of tested spectra w.r.t critical reference trajectories. These patterns may be used either to interpret the reason of the deviations (e.g. wrong compounds, different proportions, etc.) or to get a signature of deviations for a tested chemical.

Despite this work focused on Raman spectra, the proposed method can be extended to other spectroscopies exhibiting sharp features such as mid-infrared or Nuclear Magnetic Resonance spectra.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website: Additional details including (S1) Mathematical construction of wavelet basis construction and decomposition, (S2) Description of formulations and PC scores' plot, (S3) Assessment of agreement between predicted and calibration wPC scores, (S4) cross-validation results to optimize the number of tolerated points outside the band (Word file).

AUTHOR INFORMATION

Corresponding Author

*Email: thavohou@uliege.be.

Author Contributions

The conceptualization, software development, data analysis and drafting of the initial manuscript are done by T.H. Avohou. All authors contributed to the reviewing and approved the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

The authors acknowledge the support by the Wallonia Region of Belgium through the Grant N°7517. The authors also acknowledge Pierre Lebrun, PhD, for fruitful discussions during this work.

REFERENCES

(1) Gala, U.; Chauhan, H. Principles and applications of Raman spectroscopy in pharmaceutical drug discovery and development. *Expert Opin. Drug Discovery* **2015**, 10, 187-206.

(2) Deidda, R.; Sacré, P.-Y.; Clavaud, M.; Coïc, L.; Avohou, H.; Hubert, P.; Ziemons, E. Vibrational spectroscopy in analysis of pharmaceuticals: Critical review of innovative portable and handheld NIR and Raman spectrophotometers. *TrAC, Trends Analyt. Chem*. **2019**, 114, 251-259.

(3) Esmonde-White, K.A.; Cuellar, M.; Uerpmann, C.; Lenain, B.; Lewis, I.R. Raman spectroscopy as a process analytical technology for pharmaceutical manufacturing and bioprocessing. *Anal. Bioanal. Chem.* **2017**, 409, 637-649.

(4) De Luca, S.; Bucci, R.; Magrì, A.D.; Marini, F. *Class modelling techniques in chemometrics: Theory and applications*. In Encyclopedia of analytical chemistry: Applications, theory and instrumentation; Meyers, R.A., Ed.; John Wiley and Sons Ltd: Chichester, New York, 2018; pp 1-24.

(5) Pomerantsev, A.L.; Rodionova, O.Ye. New trends in qualitative analysis: Performance, optimization, and validation of multiclass and soft models. *TrAC, Trends Analyt. Chem*. **2021**, 143, 1-13.

(6) Brereton, R.G. One-class classifiers. *J. Chemom.* **2011**, 25, 225-246.

(7) Lemos, T.; Kalivas J.H. Self-optimized one-class classification using sum of ranking differences combined with a receiver operator characteristic curve. *Anal. Chem.* **2020**, *92*, 5354-536.

(8) Pomerantsev, A.L.; Rodionova, O.Ye. Popular decision rules in SIMCA: Critical review. *J. Chemom.* **2020**, 34, 429-438.

(9) Avohou, T.H.; Sacré, P.Y.; Lebrun, P.; Hubert, P.; Ziemons, E. A probabilistic class-modelling method based on prediction bands for functional spectral data: Methodological approach and application to near-infrared spectroscopy. *Anal. Chim. Acta* **2021**, 1144,130-149.

(10) Houhou, A.; Rösch, P.; Popp, J.; Bocklitz, T. Comparison of functional and discrete data analysis regimes for Raman spectra. *Anal. Bioanal. Chem.* **2021**, DOI:10.1007/s00216-021-03360-1.

(11) Meeker, W.Q.; Hahn, G.J.; Escobar, L.A. *Statistical intervals: A guide for practitioners and researchers*; John Wiley & Sons: New-York, 2017.

(12) Morris, J.S. Functional regression. *Annu. Rev. Stat. Appl.* **2015**, 2, 321-359.

(13) Box, G.P.E.; Tiao, G.C. *Bayesian Inference in Statistical Analysis*; John Wiley & Sons, Ltd: Boca Raton, 1992.

(14) López-Pintado, S.; Romo, J. On the concept of depth for functional data. *J. Am. Stat. Assoc.* **2009**, 104, 718-734.

(15) Johnstone, I.M.; Lu, A.Y. On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Stat. Assoc.* **2009**, 104, 682-693.

(16) Daubechies, I. *Ten lectures on wavelets.* SIAM: Philadelphia, 1992.

(17) Antoniadis, A.; Bigot, J.; Sapatinas, T. Wavelet in nonparametric regression: A comparative study. *J. Stat. Softw.* **2001**, 6, 1-81.

(18) Nason, G.P. *Wavelet methods in statistics with R*; Springer Science and Business Media: New York, 2008.

(19) Donoho, D.L.; Johnstone, I.M. Adapting to unknown smoothness via wavelets. *J. Am. Stat. Assoc.* **1995**, 12, 1200-1224.

(20) Roisilien, J.; Winje, B. Feature extraction across individual time series observations with spikes using wavelet principal component analysis. *Stat. Med.* **2012**, 32, 3662-3669.

(21) Ciza, P.H.; Sacré, P.-Y.; Waffo, C.; Coïc, L.; Avohou, T.H.; Mbinze, J.K.; Ngono, R.; Marini, R.D.; Hubert, P.; Ziemons, E. Comparing the qualitative performances of handheld NIR and Raman spectrophotometers for the detection of falsified pharmaceutical products. *Talanta* **2019**, 202, 469-478.

(22) Cadima, J.; Joliffe, I. On relationships between uncentred and column-centred principal component analysis. *Pak. J. Stat.* **2009**, 25, 473-503.

(23) R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, 2018; URL https://www.R-project.org/.

(24) Tarabelloni, N.; Arribas-Gil, A.; Ieva, F.; Paganoni, A.M.; Romo, J. roahd: Robust analysis of high dimensional data. R package version 1.4.1. 2018; https://CRAN.R-project.org/package=roahd.

(25) Savitzky, A.; Golay, M.J.E. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **1964**, 36, 1627-1639.

(26) Kucheryavskiy, S. mdatools - R package for chemometrics. *Chemom. Intell. Lab. Syst.* **2020**, 198, 1-10.