Chapter 3

Multivariate Calibration for the Development of Vibrational Spectroscopic Methods

Ioan Tomuta, Alina Porfire, Tibor Casian and Alexandru Gavan

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.72598

Abstract

Vibrational spectroscopy, namely near infrared (NIR) and Raman spectroscopy, is based on the interaction between the electromagnetic radiation and matter. The technique is sensitive to chemical and physical properties and delivers a wide range of information about the analyzed sample, but in order to extract the information, multivariate calibration of the spectral data is required. The main goal of this work will be to present in detail the available multivariate calibration strategy for development of NIR and Raman spectroscopic methods, which was successfully applied in pharmaceutics.

Keywords: multivariate calibration, vibrational spectroscopy, NIR spectroscopy, Raman spectroscopy, design of experiments

1. Introduction

The development and implementation of vibrational spectroscopic methods such as near infrared (NIR) or Raman spectroscopy has increased significantly as the use of computer technology and chemometric methods has become more available. Considering the pharmaceutical domain, these methods have been extensively applied to quantify active pharmaceutical ingredients, excipients, or physical properties either as offline method for intermediate/final product characterization [1] or as real-time-monitoring methods implemented within blending [2], granulation [3], extrusion [4], tableting [5], coating [6], or freeze-drying processes [7].

The high-throughput analysis associated with vibrational spectroscopy favored its application to gain better process understanding, sustaining the pharmaceutical product development from a Quality by Design and Process Analytical Technology point of view [8], thus enhancing the opportunity to develop well-understood, well-controlled, and continuously optimized manufacturing processes and products [5]. The nondestructive nature of vibrational spectroscopic

methods is of great importance in the quality evaluation of production batches, as they allow the testing of a high number of samples or the entire process, depending on the type of method. Using classical methods, such as chromatography the quality of a 1–3 million tablet batch is certified on 20–30 tablets, and many functional excipients that directly influence product performance are not quantified. These limitations are exceeded by implementing process analytical instruments, such as NIR or Raman [9].

Near infrared spectra are generated by molecular vibrations that imply a change of the dipole moment (—CH, —NH, —OH, —SH) and are further complicated by overtones and combination bands that reduce the specificity of spectra. In case of Raman spectroscopy, the spectra are generated by inelastic scattering, caused by chemical groups that undergo a change in polarizability when excited with an incident light beam. These differences in molecular contribution to the generation of spectral data make the two methods complementary [10].

NIR and Raman spectra are considered a source of multivariate data, as they contain information related to physical and chemical properties of the analyzed sample. Thus, the application of chemometric methods for extracting predictive spectral variability and reducing orthogonal sources of variation is indispensable [11]. The sensitivity to both physical and chemical properties of the sample can be considered an advantage, if the analyst wants to predict several quality attributes of a drug product, such as content uniformity and crystalline structure. However, if only active content characterization is desired and polymorphism is not considered to be a critical attribute, but it is present, the calibration phase still has to include both aspects to ensure the accuracy of prediction for active content. The main disadvantage of vibrational spectroscopic methods relates to the need of an extensive calibration set that needs to include chemical, physical, instrumental, and environmental variability that is expected in future prediction sets and analysis conditions.

Vibrational spectroscopy is well suited to the means of multivariate calibration, as each observation is characterized by analytical signal/absorbance recorded at multiple wavelengths. Using multiple predictor variables instead of one wavelength overcomes some univariate calibration problems related to selectivity, precision, and diagnosis, resulting in a more robust calibration model [12].

2. Calibration set development strategies

The milestone in the development of a vibrational spectroscopic method is the chemometric model that is able to accurately predict the sample properties considered in calibration phase. Before building a model, there are several key steps that need to be considered, as they directly influence its quality and predictive performance. The first step would be the specification of responses along with variation ranges, followed by the selection of instrumental method and configuration, building a representative calibration set, recording of spectral data, data preprocessing, and developing the multivariate regression model that is further tested using external prediction sets. Each step plays an important role; however, a well-built calibration set is the best starting point to a well-performing model, as it is the source of spectral data that is used for further processing and model development.

In the calibration set development phase, the analyst has to incorporate the expected variability of future prediction sets, to ensure the representativeness of the samples. This expected variability is given first by the quality attributes that are to be predicted, for example, the concentration ranges of important formulation constituents. Frequently, this is not enough for a robust model, and other type of variability has to be included in the calibration process, such as process-induced variability or environmental variability. Production samples contain process-induced variability; however, constructing a calibration set solely on production samples is not appropriate as the factor ranges do not cover the required interval. A first option would be to prepare pilot-plant samples reproducing full-scale conditions. As the number of responses increases, the calibration set becomes larger and quickly becomes unfeasible due to the high costs of production. The second option would be to prepare laboratory samples in which the concentration ranges of desired components are varied simultaneously within appropriate ranges to avoid correlations [13].

The calibration set development strategy applied for the development of quantitative spectroscopic methods depends on the sample complexity (the number of responses and the number of interfering factors included in the calibration) and on the type of method that is developed, here considering off-line or real-time-monitoring methods. In the following section, a description of calibration opportunities will be provided starting from the simplest cases and heading toward more complex situations.

2.1. Different levels of the investigated property

The most simple calibration situations include a low number of responses, one or two, here considering a chemical and a physical property of a sample. In this case, the calibration set development strategy simply resumes to the preparation of a sample with different levels of the investigated property. Mbinze et al. developed quantitative NIR and Raman methods for the assay of antimalarial oral drops and prepared a calibration set by diluting a stock solution of quinine to obtain three concentration levels. For each level, three series with three replicates were prepared resulting in a calibration set with 27 samples [14]. Tomuta et al. used NIR to characterize meloxicam tablets by evaluating content uniformity, tablet hardness, disintegration, and friability. For content uniformity assay, the calibration set included active ingredient concentration range (five levels), days (three), and batches (three) as a source of variation, whereas in the case of physical properties assay the middle formulation was compressed on seven levels of compression force, ranging from 5 to 42 kN. Compressing the powder mixture with different forces yielded tablets with different hardness, disintegration, and friability. Different settings of a one-process factor were enough to induce variability in physical properties of the samples [15]. In a similar study, Tomuta et al. developed NIR method for physicochemical characterization of low active content indapamide tablets (2%, w/w) [1]. Virtanen et al. evaluated the crushing strength of theophylline tablets through Raman spectroscopy by considering both a process factor and a formulation factor to generate variability in tablet surface roughness. In this case, the tablets were prepared considering two particle sizes of theophylline, as raw material for the granulation phase, followed by mixing with lubricants and by compressing each granulate on five different compression forces [16].

The impact of polymorphism is a well-recognized phenomenon in the pharmaceutical industry, as the differences in crystalline structure of the same active ingredient generate different physical properties that get reflected in the quality of the final medicinal product. Croker et al. developed NIR and Raman methods to quantify FII and FIII of nootropic drug-piracetam from binary mixtures using a calibration set of 15 formulations with FII ranging from 0 to 100% [17].

Gómez et al. calibrated a Raman method for the content uniformity control of low-dosebreak-scored acenocumarol tablets by under and overdosing the powdered commercial medicinal product, by adding either lactose or the active pharmaceutical ingredient to the mixture. Two commercial products with different content uniformity were considered and the two calibration sets included 7 samples in the range of 1–3% (w/w) and 12 samples in the range of 0.35–1.50% [18]. Creating calibration sets by under-overdosing samples can result in correlated concentrations between API and excipients [19]. Collinearity between concentrations leads to spurious predictions by attributing changes to the correlated formulation component instead of the real contributor [20].

Changing the production scale generates samples that incorporate different types of variability from the primary conditions through which the calibration set was prepared. As laboratory-prepared samples lack manufacturing variability, the accuracy of prediction may be affected for production prediction sets. This limitation has been exceeded by extending the calibration set with production samples [13], adjusting the sampling strategy, pre-conditioning the calibration set to future expected environmental conditions [21], or by mathematically adding process variability to laboratory samples [20].

Blanco et al. developed NIR methods to control individual steps of paracetamol tablet manufacturing, resuming to an intermediate granulation step and tableting. Prior to building a calibration model, both laboratory-prepared samples and industrial production samples were taken into account to evaluate the eventual spectral differences. In case of the granule-active content assay, the calibration set was built solely on laboratory-prepared samples, whereas in the case of tablet assay the differences between laboratory and production samples made the calibration set include both, in order to ensure representativeness. For granule particle size characterization, samples collected over a period of 2 years ensured the presence of future expected variability in prediction set [22].

Blanco et al. used NIR to characterize mirtazapine tablets in terms of content uniformity and tablet hardness. For active ingredient content, the calibration set included production tablets from 20 batches and 34 laboratory-prepared samples, whereas for tablet hardness the laboratory samples were compacted in the range of 300–740 MPa. Including production samples for both responses reduced the systematic errors and gave better predictions [13]. By adding spectra from different manufacturing scales to the calibration set, the spectral variability becomes more representative, an important aspect for prediction accuracy. As the number of manufacturing samples is lower compared to the initial calibration set, proper weighting is necessary to avoid the dominating tendency of the larger dataset. To this regard, Farrel et al. applied Tikhonov regularization as a multi-criterion-based weighting selection method to augment the performance of NIR models regarding their ability to predict production scale products [23].

Blanco et al. proposed a method to incorporate physical variability that originates from production into the calibration. The concept relies on calculating a process spectrum, which added to the laboratory sample spectra incorporates process-related physical changes. The process

spectrum represents the difference between the laboratory sample spectrum and the intermediate/final product spectra of an identical composition prepared on a different scale. The variability given by the process spectra can be further increased by multiplying the data with different coefficients [20, 24].

In situations where solid-state transformations occur within the manufacturing process, it is frequently desired to construct the calibration set with components obtained through the same method to have more representative formulations. Netchacovitch et al. used Raman spectroscopy to determine crystalline itraconazole in amorphous solid dispersions prepared by hot-melt extrusion. Calibration set included three levels of concentration and was built by using crystalline API powder, six batches of grinded extrudates with amorphous API, and placebo-grinded extrudate [25].

Pan et al. calibrated NIR method for the quantification of low-level Irbesartan Form B from pharmaceutical tablets. Form B is known to have a limited solubility and is formed from Form A via a solution-mediated process. To incorporate physical variability into the calibration set, the sample preparation procedure supposed the use of specifications similar to the manufacturing process. The robustness of the method to process induced physical variability, the effect of tablet hardness, granule size, and atmospheric humidity was evaluated. It was demonstrated that the prediction accuracy was influenced only by relative humidity, generating a positive bias in samples stored at 50%RH. Therefore, the entire calibration and validation was reconsidered by pre-conditioning the samples at 25°C and 50%RH for 20 h, prior to recording the spectra and building the model. This way, the robustness of the method was increased to future expected variations in environmental conditions [21].

2.2. Design of experiment strategy

As the number of factors increases, the calibration set becomes more complex and different strategies have to be applied to avoid correlated responses. If two formulation components C1 and C2 are correlated, a change in the concentration of C1 can be spuriously predicted as a change in C2. In DoE, factors are varied simultaneously in a systematic manner, providing orthogonality, an essential condition for estimating regression coefficients [26]. There are several design types that can be used for calibration purposes, starting from the classic full factorials down to central composite, mixture, or D-optimal designs. Considering more complex formulations, NIR spectroscopy has been applied to determine the amount of amoxicillin in the presence of seven other excipients. By applying a three-factor (API, saccharose, and other excipients) experimental design, the concentration of factors was varied orthogonally [27]. Ferreira et al. used a calibration set prepared according to a DoE with three factors: hydrochlorothyazide, cellulose, and other excipients to train a NIR method for the quantification of the active ingredient in pharmaceutical samples [28].

Li et al. calibrated Raman method to quantify active ingredient content considering the presence of different sources of variability: degradation compound, relative humidity, change of scales, and compression force. Laboratory samples were prepared based on a 3^2 full-factorial design where the active ingredient ranged between 80 and 120%, from which a subset of samples were spiked with the degradation product, added in two molar ratios. Each powder mixture was compacted at 8 and 30 kN in laboratory scale and three design points were compacted at manufacturing scale [29]. Casian et al. developed NIR and Raman methods for

the quantification of two APIs found in significantly different concentrations from immediate release tablets. The calibration set was built on a full-factorial design with two factors and five levels with a total of 25 formulations [10]. The use of full-factorial designs is feasible with two factors if five levels of variation are used [52]. Adding one more factor will generate 125 experimental runs that are impractical [26, 53].

Netchacovitch et al. calibrated a Raman method to quantify low-level polymorphic impurities in a pharmaceutical formulation through a 12-run central-composite experimental design [25]. Central composite designs are extensions of the two-level full-factorial designs that are built by adding symmetrically axial points. Dependent on the position of axial points, factors can be varied on three levels (central composite face-centered design) or five levels (central composite circumscribed) [26].

Short et al. used NIR to evaluate relative density and crushing strength of four component tablets. Compared to other studies, where only the compaction pressure was considered as a factor to induce variability in the investigated response, in this case formulation composition was varied also. The calibration set consisted of 29 formulations (mixture design) with each formulation being compressed at different pressures [30]. Lyndgaard et al. developed a Raman method to quantify paracetamol content from tablets through blisters. The calibration set included 18 formulations, selected on the basis of a ternary mixture design (paracetamol, starch, and sucrose) with each factor being varied on six levels [31]. Igne et al. evaluated the effect of API physical form, excipient particle size, different manufacturer, and changes in environmental conditions on the performance of a NIR model. The calibration samples were prepared according to a 29-run quaternary mixture design with every formulation being compressed at two of five different forces. Only changes in the particle size of lactose produced biased predictions in both ambient and chamber conditions. The authors tested variable-selection methods to increase method robustness to raw material variability [32].

Griffen et al. used Raman spectroscopy to quantify all tablet constituents, three active ingredients and two excipients. In this case, the calibration set was built on a first-order (linear) five-level, five-factor mixture design that uniformly covered the concentration ranges of the components. The concentration of individual components ranged from 1 to 85% (w/w) [33]. Mixture designs are well suited for formulation application, where the sum of all ingredients adds up to 100% and where factors cannot be manipulated independently one from another. Porfire et al. used a D-optimal design with three variables and five levels to build a calibration set with 63 formulations with the purpose of quantifying encapsulated simvastatin and two functional excipients L-α-phosphatidylcholine and cholesterol from liposomes [34]. Saraguca et al. developed an NIR method to simultaneously quantify paracetamol and three other excipients from powder blends using a calibration set constructed on a 40-run D-optimal mixture design [19].

A D-optimal design is frequently applied for a high number of factors as it gives a lower number of runs compared to factorial designs. The D-letter originates from its criterion of selecting the best subset of factor combinations from a pool of theoretically possible combinations, which relies on maximizing the $X'X$ matrix Determinant [26]. In another study, Heinz et al. trained NIR and Raman to quantify ternary mixtures of alpha, gamma, and amorphous forms of indomethacin from ternary mixtures using a 13-sample calibration set built on a cubic model experimental design [35]. Lin et al. developed an at-line blend uniformity NIR method for simultaneous

quantification of four active ingredients with structural similarity, found in different concentrations. Calibration was built on six formulations, where five factors (four APIs and one diluent) were varied on six levels while avoiding correlations. The performance of the model was improved by adding a set of spectral data from a different production scale [43].

When DoE is used, correlations are significantly reduced dependent on the type of design, number of factors, and experimental runs. However, an increased number of factors will require a high number of experimental runs to avoid collinearity, which rapidly increases the costs. Several papers have addressed the question of how many samples are needed to ensure a robust calibration [19]. The fact that models with similar performance were developed on a reduced design compared to its full-factorial counterpart suggests the presence of redundant information in full-factorial designs [36].

Saraguca proposed a method that relies on building the model on a limited number of samples and uses the remaining formulations to test the predictive performance in terms of RMSECV and RMSEP. In the following steps, the calibration set was extended by transferring one formulation at a time from the test set until the calculated cross-validation and prediction errors stabilized. The sample selection procedure focused on maximizing the concentration variability of all components [19].

Alam et al. proposed a method for calibration set development in spectral space instead of concentration space. Orthogonality in spectral response will yield a better estimation of coefficients with a minimum number of samples, while orthogonality in concentration space will not necessarily translate into spectral orthogonality, as the contribution of each component to the sample spectrum is different. The method is based on decomposing the pure component spectra of a formulation into orthogonal directions (scores), which will be varied around a model tablet score through DoE. The model tablet score represents the score of the spectra recorded on a target formulation projected onto the orthonormal basis vector of the pure components spectra. After designing the spectral space calibration set, the composition of each spectra is retrieved by mathematical means [37].

2.3. Calibration strategy for calibration *in-line* monitoring methods

The application of vibrational spectroscopy for in-line monitoring implies the use of fiber optic probes mounted at the interface of the process itself to acquire spectral data with a defined rate. The simplest way to calibrate an *in-line* method is to acquire real-time spectra through the entire process length along with collecting samples at regular intervals. The response values obtained through reference methods are correlated with the spectral data, considering the process time as a link between the two [38–40]. More extensive calibrations also evaluate the effect of sample presentation, changing process, and formulation parameters, to challenge the robustness of the methods.

For coating application, the calibration strategy relies on the linear variation of spectral response as the contribution of the coating material increases and the tablet core contribution decreases [41]. Moes et al. developed quantitative NIR method using three batches of tablets by varying the tablet core weight (240–200–160 g) and the amount of coating suspension resulting in different coating thicknesses [42]. Möltgen et al. used five full-scale experimental

runs to develop a quantitative NIR method (one run) and to evaluate the effect of changing exhaust air temperature and spray rate (two runs) and the effect of tablet density and flow motion in the coater (two runs). For quantitative calibration, samples were collected through the entire process and analyzed using reference methods [6]. For the quantification of coating thickness by means of Raman spectroscopy, Kauffman et al. calibrated the method by considering film thickness and film composition variables. Tablets were coated on five levels ranging 0.5–6% weight gain by varying their residence time in the coater. As for film composition, three different TiO_2 levels were evaluated due to the strong Raman signal of this component offering the potential for an indirect measure [41]. In the case of thin coatings, the generation of a calibration set can become a difficult task and can become limited due to the lack of reference methods. In this situation, an alternative to classical regression methods would be the Science-Based Calibration (SBC) approach, which allows the calibration without a reference method by separating spectral variability into orthogonal (covariance matrix) and predictive parts (related to the coating). Möltgen et al. applied SBC to develop quantitative NIR method for in-line evaluation of thin hydroxypropyl methylcellulose (HPMC) coatings through four experimental runs. For calibration, the pure HPMC spectrum was used as the coating response spectrum and the covariance matrix included hardware, core, water, and process-related noise. The method developed without reference samples predicted accurately coating thickness values in the range of 8–28 μm demonstrating the value of SBC [43].

In order to predict granule moisture content in a six-segmented fluid bed dryer through NIR spectroscopy, a calibration set of 20 experiments was applied. Granules were prepared with five moisture levels by varying the drying air temperature and drying time. Each moisture level had four replicates prepared on two different days [3].

Clavaud et al. developed a global regression model for moisture content estimation from freeze-dried medicine. As expected, the calibration set was extensive, including three types of active ingredient with different concentrations, different vial diameters, and excipient amounts. To include intra- and inter-product variability, 5 batches and 100 samples were used for each product [44]. Martinez et al. calibrated NIR method for *in-line* quantification of two active ingredients in a batch-blending process by investigating the influence of sample presentation. With regard to this, the high-loading API was used either in the form of a cohesive powder or in a granular form prepared by melt-extrusion. The observed spectral differences were resumed to the polymer wavelength absorption band that coincided with the water region. The offline calibration of the method was built on 13 samples which included both forms of the high-loading API [2].

Wahl et al. evaluated *in-line* the content uniformity of ternary mixtures with an NIR mounted on the feed frame of a tablet press. For calibration, the active ingredient and two excipients concentrations were varied through eight experiments selected by means of a D-optimal design and two extra runs added to ensure equidistant steps in the content of each component. Spectral data were recorded in a dynamic acquisition mode, simulating real conditions [5].

Karande et al. developed NIR method for real-time monitoring of tableting based on a 105-sample calibration set generated through a simplex lattice design with four factors (chlorpheniramine maleate, lactose, microcrystalline cellulose, and magnesium stearate). Prior to building the calibration, the effect of sampling was evaluated by recording NIR spectra in both static and dynamic conditions. The differences between measurements revealed the importance of

ensuring similar sampling conditions for calibration as for actual real-time monitoring [9]. For another application, Karande et al. evaluated the effect of different spectral-sampling strategies on the performance of an NIR model, to accurately predict blend components in quaternary mixtures. Calibration samples (24 formulations-D-optimal mixture design) were recorded in three ways: laboratory mixing and static spectral acquisition; IBC (intermediate bulk container) mixing and static spectral acquisition; IBC mixing and dynamic spectral acquisition. Dynamic sampling yielded the best calibration model with highest accuracy, demonstrating the importance of selecting similar sampling conditions to the actual testing [45].

Based on the presented examples found in literature, the most frequently applied methods to design a calibration set were as follows:

- One chemical/physical property: formulations with three to five levels of variation for the response that span the desired range of concentration/physical property.
- One chemical and one physical property: formulations with three to five levels of variation for the chemical response and for the physical property calibration are considered only for target formulation (five levels).
- Two chemical/physical properties: any type of DoE (full-factorial, central composite, mixture design, D-optimal) to avoid collinearity and spurious predictions.
- Three chemical/physical properties: simple lattice mixture designs or D-optimal designs.
- In-line methods: models built by correlating sampled product properties with in-line collected spectra. Most rigorous studies also investigated the effect of process parameters on the NIR spectra.

3. Handling chemical, physical, and environmental interferences

The dependence of the NIR spectra on the sample's chemical and physical properties caused by absorption and scatter effects can be an advantage of this type of spectroscopy, but at the same time, the scatter effects caused by sample variations or even by environmental phenomena can create a series of analytical problems. In such cases, each type of interferences has to be considered in the calibration model development. In the following section, the importance of chemical, physical, and environmental interferences will be described, providing insights on specific spectral variations produced by each category and highlighting how to handle them in order to increase model robustness [1, 2].

Generally, a quality NIR analysis should provide a model that manages a correct interconnection of the spectral variables with the samples properties of interest. At the same time, an ideal calibration model will not react to instrument variation, environmental changes, background interferences, and will be mostly focused on the information of interest. Chemometrics is the science that enables the extraction of relevant information, as well as the reduction of unrelated information as well as interfering parameters.

Spectral interferences resulting from variable physico-chemical sample properties (e.g., particle size variation and moisture content) or instrumental effects (e.g., path-length variation,

light scattering, and random noise) can be reduced, eliminated, or standardized by using spectral pretreatments, prior to the multivariate data analysis [3]. Since the correct selection of spectral pretreatment can significantly improve the reliability of the model, this topic will be discussed in the following paragraphs. The most common pre-processing techniques can be divided into two groups: pretreatments for spectral normalization and for smoothing/differentiation. The first group achieves spectral normalization through scatter-correction methods. Scatter effects are common for all spectroscopic techniques and the phenomenon appears mostly because of the physical variabilities between samples or path-length variations. Two of those pre-processing concepts are standard normal variate (SNV) and multiplicative scatter correction (MSC) which also normalize the baseline shifts of different samples [4, 5]. The second set of pre-processing methods has the capacity to reduce or remove the noise by smoothing and differentiating the spectral values. The most common spectral derivatives are based on the Savitzky-Golay (SG) [6] and the Norris-Williams algorithms [7].

In most cases, in order to obtain best results, there is the need to apply both types of pretreatment techniques one after the other. Peeters et al. tested both types of pre-processings not only to reduce light scattering effects but also to minimize peak shifts of Raman and NIR spectra. They applied SNV, MSC, and first and second derivatives obtained by calculating 15-point quadratic Savitzky-Golay filters, in order to develop a method for the off-line prediction of tablet properties [8]. Sylvester et al. developed an in-line NIR-monitoring method for a freeze-drying process using the SNV pre-processing in order to remove multiplicative interferences caused by scatter and particle size variations and the first Savitzky-Golay derivative to reduce baseline shifts and to improve the spectral resolution [9]. The successful development of a real-time method for monitoring continuous powder flow from a tableting machine feeder was described by Alam et al. Savitzky-Golay derivatization was first applied for smoothing, followed by SNV for scatter correction [10]. Environmental interferences can be caused by sample, instrument, or even laboratory variations; this type of interferences causes misalignments or shifts of the spectra and is commonly overcome by applying alignment/warping techniques to the data [3]. Those methods stretch or compress the signal in order to match it in the best way possible with a given reference spectra [11, 12].

All pre-processing methods have the purpose to reduce the undesirable variability and interferences from the data, but there is always a risk of choosing an inappropriate type or applying a severe pre-processing that would also remove valuable information. Because of this, choosing the correct technique is one of the most important steps in data pre-processing and model development.

A last useful solution to deal with problems caused by interferences is wavelength selection method. The model development can be done based on the specific spectral domain that contains the information of interest. In order to select the domain of interest or to eliminate irrelevant wavelength domains, principal component analysis (PCA) can be performed. Prior to the PCA, the collected spectra should be pre-processed and column centered, then the analysis can be performed on the data matrix. Finally, the variables should be selected according to high peak loadings obtained for all relevant principal components (PCs), and the position of the resulting features should be compared with the original spectrum to validate the selection.

4. Data pre-processing

During the development of a multivariate calibration model, systematic variation such as baseline shifts and scatter effects, not relevant for the prediction of the response variables (Y), is present in predictor variables (X). Pre-processing methods are used in order to remove the systematic variation not related to the Y-matrix, which might impair the interpretation or predictive ability of the developed model.

The main goals of data-pre-processing are the following:

a. improvement of the robustness and accuracy of subsequent analyses;

b. improved interpretability: raw data are transformed into a format that will be better understandable by both humans and machines;

c. detection and removal of outliers and trends; and

d. reduction of the dimensionality of the data mining task and removal of irrelevant and redundant information [46].

The methods generally used for data pre-processing are divided into two categories. The first consists of classical pre-processing methods, used for normalization, smoothing, and differentiation. The second is represented by methods for variable selection and dimensionality reduction [47]. Among these methods, the most appropriate has to be chosen, such as to only remove unwanted variation, without excluding or altering chemically relevant information [48].

When used in an inappropriate way, pre-processing may introduce artifacts or cause loss of information. Thus, the purpose of the analysis is important for the selection of the most appropriate pre-processing method, because scattering is disruptive for compound identification and quantitation, but is useful to study the physical properties of the sample. As a consequence, the best pre-processing method, ensuring a correct data analysis and robust results, has to be chosen by testing and comparing the results of different methods [48].

4.1. Pre-processing methods

4.1.1. Spectral normalization

In many analytical methods, the variables measured for a given sample are increased or decreased from their true value by a multiplicative factor, which is called the scaling or gain effect. In spectroscopic methods, the scaling effect arises from path-length effects, scattering effects, source or detector variations, so the relative value of variables should be used during multivariate modeling rather than the absolute measured value. The sample normalization is one of the most important pre-processing methods, which is applied in an attempt to correct for multiplicative scaling effects, the shifts and the trends in baseline and curvilinearity, by identifying some aspect for each sample which should be essentially constant from one sample to the next, and correcting the scaling of all variables based on this characteristic [48].

Normalization methods can be subdivided into two main groups: simple normalization methods (min-max normalization, one-norm, vector normalization, standard normal variate),

requiring only the information from the spectrum to be normalized, and normalization methods requiring the presence of collective spectral data matrices or of reference spectra (multiplicative scatter correction and extended multiplicative signal correction (EMSC) [46]. Among these, the most used scattering correction algorithms include the SNV and MSC. The two pretreatments give similar results, being considered exchangeable, but the results obtained through both algorithms are compared usually, since they may be different [49]. SNV was proposed to reduce multiplicative effects of scattering, particle size, and multicollinearity changes over the NIR spectra. This approach starts with mean centering and consists of dividing mean-centered spectra by the standard deviation over the spectral intensities [50]. SNV normalizes each spectrum returning a mean of 0 and a variance of 1 spectra dataset [48]. The disadvantage is the assumption that multiplicative effects are uniform over the whole spectral range, so artifacts may be introduced by this transformation.

The de-trend method is another approach to correct for baseline shift, which removes the baseline curvature by expressing it as a quadratic function of the wavelengths. The modeled baseline is subtracted from the spectrum, so de-trend can be used after SNV to circumvent any curvilinear trend, where the baseline drift is a function of wavelength [50]. The MSC pretreatment performs a linear regression of each spectrum on a reference spectrum, which is usually the mean of all available spectra, for example, the average spectra of the calibration set, or a generic reference spectrum can also be applied [49].

4.1.2. Smoothing and differentiation

The smoothing algorithms are used in order to correct the spectral noise, while differentiation is used to enhance spectral resolution and to eliminate background absorption. The most common ways to achieve smoothing are the use of noise filters for de-noising and smoothing and Savitzky-Golay smoothing/derivative filters for smoothing/resolution enhancement. Noise filters are specific low-pass filters which can be used to reduce random noise. Their drawback is that the signal-to-noise ratio is increased at the expenses of distorting the signal. The most popular smoothing filters are the zeroth-order SG-smoothing/derivative filter, the binomial filter, and the moving average filter [46].

Derivatives are used for their capability to remove both additive and multiplicative effects in the spectra. The first derivative removes only the baseline; the second derivative removes both baseline and linear trend. The first derivative is estimated as the difference between two subsequent spectral measurement points, while the second-order derivative is estimated as the difference between two successive points of the first-order derivative spectra [51]. The most popular derivation method is SG algorithm, proposed by Golay and Savitzky in 1962 [52]. The method has the advantage that computation of the derivatives and smoothing are carried out in a single step. The algorithm used in this method is based on fitting a polynomial in a symmetric window on the raw data, in order to find the derivative at the center point. The parameters of the polynomial are calculated and the derivative of this function is found, this value being used as the derivative estimate for this center point. The same operation is subsequently applied to all points in the spectra. Two decisions are important to be made in this algorithm, i.e., the window width (width of the subset of the data) and the fitted

polynomial order. The highest derivative that can be determined depends on the degree of the polynomial used during the fitting [51].

4.1.3. Dimensionality-reduction methods

These methods rely on reducing the dimension of the predictor space spanned by a number of variables or wavelengths, in order to find the subspace mainly containing variations related to the response matrix. The orthogonal projection and the variable-selection methods are in this group. Orthogonal signal correction (OSC) and its modified version direct orthogonal signal correction ((D)OSC) are the most common among this group, developed to remove systematic variation in the descriptor matrix, that is not correlated to the response matrix. In other words, the pre-processing is performed in such a way that the removed parts are orthogonal (not linearly related) to the response matrix [53, 54]. The method has the advantage of correcting at once multiple artifacts.

An alternative OSC algorithm was developed by Trygg and Wold and is called orthogonal projection to latent structures (OPLS). The objective of OPLS is the same as of OSC, but the approach is different, i.e., the OPLS method analyzes the variation explained in each PLS component. The non-correlated systematic variation in descriptor matrix is removed, making interpretation of the resulting PLS model easier, and the non-correlated variation can be analyzed further [55].

Variable-selection techniques consist of selecting particular variables related to the response, instead of removing the interference modeled as a spectrum, the aim being to identify a subset of wavelengths that produces the smallest possible error [56]. Selecting the most correlated wavelengths may lead to better performance in PLS and PCR, but, on the other hand, selection of the most correlated wavelengths may eliminate those that correct for the influence of interfering compounds or factors [56].

4.2. Pre-processing strategy

In practical applications, combinations of pre-processing methods are usually employed in search for the best algorithm, involving more than one pre-processing step. According to Rinnan et al., several rules may serve as guidelines: scatter correction (except of normalization) should always be performed prior to differentiation; normalization can be used at both ends of the correction, but usually is easier to be assessed if it is done prior to any other strategy; MSC gives a smaller baseline correction than SNV with subsequent de-trending; it is not recommended to perform de-trending followed by SNV [51].

The ideal pre-processing strategy should only remove artifacts present in the data, without introducing any unwanted artifacts or variability in the data. When physical properties, that is, tablets' crushing strength, are evaluated through vibrational spectroscopy, typical pre-processing methods such as SNV, MSC, and the derivatives cannot be used, because they lose the baseline-shifting information, which is relevant for the physical properties. The data in this case should be modeled as such or after normalization [16]. Three approaches are described in the literature, for the selection of the most appropriate strategy: the trial-and-error approach; visual inspection and the use of data-quality parameters [57].

In the trial-and-error approach, all pre-processing methods are applied to the data and the pre-processed data are used as an input to a calibration model, which is further used to assess the quality of the pre-processing strategy by an internal measure, such as RMSEP or RMSECV [57]. For example, Karande et al. chose among various pre-processing methods through comparing the figures of merit (explained variance, R^2, RMSEC, and RMSECV) of the developed partial least-squares (PLS1) regression models, for the quantification of micronized drug and excipients in tablets by NIR spectroscopy. The raw calibration spectra were pretreated with SNV followed by first derivative and SNV followed by second derivative pre-processing. All models were developed using the entire spectral range or narrow spectral ranges. The best performance of the calibration method (highest explained variance, lowest RMSEC and RMSECV) was obtained using the whole spectral range, pretreated with SNV followed by first derivative spectral pre-processing [9] . The same approach has been used by Porfire et al. in the attempt to select the best pretreatment method in the development of calibration models for prediction of chemical composition and crushing strength of sustained-release tablets with indapamide. PLS regression was performed for non-processed spectra as well as for spectra treated by various pre-processing methods (i.e., FD, SD, SNV, MSC, FD + SNV, FD + MSC), and the most suitable pretreatment algorithm was chosen based on the results obtained for PLS model validation through cross-validation, i.e., based on its RMSECV and bias [58].

In visual inspection method, the effect of pre-processing is assessed before a model is constructed. Thus, because artifacts have been removed during pre-processing, samples should show more spectral overlap after pre-processing in visual inspection, and differences between groups of samples should be more pronounced. However, as visual inspection may be very difficult and not objective, the data are not usually inspected in "spectral mode" but in a lower dimensional space, obtained usually through principal component analysis [57]. PCA reduces the dimensionality of the problem by generating linear combinations of the original variables returning new "latent" variables. Each original variable is weighted by a loading representing the importance of the considered variable on the variance of the data. The variability of the data is expressed by new dimensions called principal components, and the projection of a pixel onto the PCs is called its score. The result of PCA is the decomposition of the pre-processed matrix in a score matrix and a loading matrix [48]. PCA is used for data overview, for example, for detecting outliers, groups, and trends among observations, for evaluating relationships among variables, and between observations and variables. In PCA, data in the matrix X are transferred into a new coordinate system defined by principal components. The direction in variable space occupied by the most varying data points will define the location of the first PC, and the second PC will be given by the largest variation orthogonal to the first component. PCs are extracted until only minor variation is left unexplained by the PC model, each component consisting of a score vector and a loading vector. Observations close to each other in a score plot have similar properties, and variables close to each other in a loading plot are correlated. Thus, the score plot is useful for the detection of strong outliers, clustering, and time trends [59] .

The detection of strong outliers through PCA is done by analyzing the score plot. The strong outliers are removed, as they may have a degrading impact on model quality. A statistic tool called Hotteling's T^2 may be used in conjunction with the score plot for the detection of strong outliers. This tool is a multivariate generalization of Student's t-test, defining the

normal area corresponding to 95 or 99% confidence. Subsequently, for a better understanding of the properties of grouped data, a splitting of data into smaller groups according to the nature of the clustering is done, and separate PCA models may be fitted. For the detection of weakly deviating observations (moderate outliers), which are not strong enough to show up as outliers in score plots, the residuals of each observation are used. The detection tool is called DmodX (a notation for distance to the model in *X*-space). A value of Dmodx is calculated for each observation, and the values are plotted in a control chart where the maximum tolerable distance (Dcrit) for the dataset is given. The plot of DmodX enables an overview of the unsystematic process variation, as moderate outliers have DmodX values higher than Dcrit [59].

Before PCA, scaling of data is usually performed, because variables have different numerical ranges so they will have different variance and they will weight differently in the data analysis. The most common approach is the unit variance (UV) scaling, consisting in dividing each variable by its standard deviation. The result is that each variable has equal variance, meaning that the "length" of each variable is identical, although the mean values still remain different [59].

Tôrres et al. used Hotelling's T^2 chart to analyze the NIR spectra of a training (calibration) set for the development of a monitoring method for the stability of captopril in tablets. Before being analyzed by PCA, NIR spectra were smoothed as described by Savitzky-Golay with a 21-point window and second-order polynomial and were processed by MSC for the correction of baseline variation due to non-homogeneity of particle's distribution [60]. The Hotelling's T^2 chart measures the distance from an observation to the center of the samples under normal operating conditions and evaluates whether a particular sample has a systematic deviation from the samples considered to be under statistical control [61]. As all samples from the training set were assumed to be normal, the training chart was not expected to identify systematic deviations in these samples in the training phase, so the number of PC retained in the model was selected to minimize the number of false alarms (false positives and false negatives) during the training phase of the control charts [60].

5. Regression methods

Regression analysis is a modeling technique used to investigate the relationship between dependent variables (responses or *y*'s variables) and independent variables (predictor, factors or *x*'s variables). According to the number of variable, three cases can be distinguished:

1. Simple linear regression—one *y* and one *x* variable.
2. Univariate linear regression—one *y* and several *x's* variables.
3. Multivariate linear regression—several *y's* and several *x's* variables [62].

The objective of a regression method can be achieved by means of a model where the observed result (dependent variables, response, *y*'s variables) is described as a function of independent variables (*x*'s variables) and the noise is left as residual.

In a regression analysis, the relationship between two data matrix *X* (BxK) and *Y* (NxM) are related to each other. A regression model can be written as in a matrix form as

$$\mathrm{Y} = \mathrm{XB} \tag{1}$$

where Y is the matrix of x's variables; X is the matrix of *y*'s variables; B is the matrix of regression coefficient, B(KxM).

A good estimate of regression coefficient (B matrix) provides a good fit to *Y* and good prediction of future unknown parameters y^{T}. More, the regression coefficient vector should be of mechanistic understanding and interpretable [59, 63, 64].

A large number of regression methods were developed, all with the goal of finding the best estimation of B. In the calibration of spectroscopic methods, only multivariate regression techniques can be applied, and the most used are (1) multiple linear regression (MLR), (2) principal component regression, (3) partial least-squares regression, and (4) orthogonal partial least-squares regression (O-PLS). In the last years, some advanced regression methods as (5) Bayesian ridge regression (Bayes-RR) (6) support vector regression (SVR) or (7) decision tree regression (DTR) have started to be used.

5.1. Multiple linear regression

Multiple linear regression is an extension of simple linear regression model. In the case of MLR determination, the relationship between *x*'s—variables and *y*'s—variables is achieved by means of a model where the responses (*y*'s—variables) are described as a function of analyzed factors (*x*'s—variables) and the noise is left in the residual (ε) (Eq. (1)) [65]

$$y = f(x_1, x_2, x_3 \ldots, x_n) + \varepsilon \tag{2}$$

The function *f* is approximated by a polynomial equation (Eq. (3)),

$$y = b_0 + b_1 x_0 + b_0 x_2 + \ldots + b_0 x_n + \varepsilon \tag{3}$$

where *bi* (*i* = 1, 2, 3, …,*n*) are the regression coefficients and describe the effect of each term on the response *y*.

The polynomial equation (Eq. (3)) can be written in matrix way as follows:

$$y = Xb + \varepsilon \tag{4}$$

where *X* are the matrix of *x*'s variables and *b* the vector, and the multiple linear regression is used to determinate vector *b*.

If there are orthogonalities between *x*'s variables, Eq. (4) can be written as

$$b = \left(X^T X\right)^{-1} X^T y \tag{5}$$

In this equation, matrix $X^T X$ become a diagonal matrix and *b* is easily calculated.

If not all the x's variables can be controlled, the number of x's variables extends the number of experimental runs or the number of experimental runs is larger than the number of x's variables, the co-linearity between x's variables arises and the orthogonality no longer exists, so the inverse of X^TX cannot be applied.

Except the cases when the calibration of spectroscopic methods is performed following the design of experiment strategy in the other multivariate calibration, the orthogonalities do not exist and the MLR cannot be applied. That is the reason why other regression methods based on latent variables as partial last squares are preferred and become popular. When the calibrations are performed based on latent variables, inside of using the original variables in the regression, a new set of orthogonal (latent) variables is calculated and leads to reduction of the original dimension of x's variables matrix and performs the least-square estimation.

5.2. Principal component regression

Principal component regression is a regression method based on principal component analysis and it is used when datasets are highly collinear. In a PCA regression, the original set of collinear variables is transformed to a new set of correlated variables. So, the principal component analysis is used to decompose the x's variables into a principal component (orthogonal basis) and a subset of components in order to predict y's variables. The basic idea of the principal component regression is to calculate the principal components and then use some of these components as predictors in a linear regression model fitted using the typical least-squares procedure [66, 67].

In the case of PCR determination, the relationship between x's variables and y's variables is achieved by means of a matrix of lower dimension (TP^T), called principal components, and a matrix of residuals (E).

$$X = 1\overline{X} + TP^T + E \tag{6}$$

where $\overline{X}$ contains X average; T is a matrix of scores that summarizes the X variables; P is a matrix of loadings showing the influence of the X variables; E is a matrix of residuals (the deviations between the original and the predicted values) [66].

The main idea of principal regression is to replace X matrix of row date to a smaller orthogonal score—loading matrix (TP^T matrix) that summarized the original X matrix, and then to relate the T-scores to y's variables.

The core of PCR is that a small number of principal components is enough to explain the variability into the data. In most of the cases, it might be found out that four to six principal components are enough to explain more that 90% of the variance into the data.

5.3. Partial least squares

The partial least-squares regression is the most popular method for the creation of models used in the development of NIR and Raman spectrometric methods and is used to develop a linear link between two matrices, the NIR/Raman spectral data and the reference values.

The PLS approach was first proposed by Herman Wold around 1975 for the modeling of complicated datasets in terms of chains of matrices, the so-called path models. PLS regression is preferable to develop calibration models because unlike MLR, it can analyze data with strongly collinear, noisy, and numerous *X*-variables, and also simultaneously model several response variables [68]. PLS was developed for situation in which the data have more independent variables than observations (the "small n, large p") or/and where collinearity is present among dataset [69].

The PLS finds a multivariate model (linear or polynomial) that describes the relationship between *Y* matrix (dependent variables) and *X* matrix (predictor variables) expressed as

$$\mathrm{Y} = \mathrm{f(X)} + \mathrm{E} \tag{7}$$

PLS may be easily understood geometrically if we imagine the matrices *X* and *Y* as *N* points in two spaces. The *X*-space with *K* axes, and the *Y*-space with *M* axes, where *K* is the number of columns in *X* matrix and *M* the number of columns in *Y* matrix. The objectives of PLS is to find a latent variable so that the best approximate *X*-space, the best approximate *Y*–space, and the greatest possible correlation between *X*-space and *Y* space.

A PLS model can be written as

$$X = 1\overline{X} + TP^T + E \tag{8}$$

$$Y = 1\overline{Y} + UC^T + F \tag{9}$$

$$T = U + H \tag{10}$$

where $\overline{X}$ contains the X average; $\overline{Y}$ contains the Y average; *T* is a matrix of scores that summarizes the *X* variables; *U* is a matrix of scores that summarizes the *Y* variables; E, F, H is a matrix of residuals (the deviations between the original and the predicted values) [12].

In a PLS algorithm, there are additional loading called weight. *P* is the matrix of weigh expressing the correlation between *X* and *U* and is used to calculate *T*. *C* is the matrix of weigh expressing the correlation between *Y* and *T* and is used to calculate *U* [12, 70].

5.4. Orthogonal partial least squares

OPLS has been developed in order to separate information in the *X* matrix that is correlated with Y matrix form Y-uncorrelated information. The idea of O-PLS algorithm was to remove systematic variation uncorrelated with the response with the goal and to reduce the number of components in order to increase interpretability of the model [55, 69, 71].

The main idea of O-PLS is to separate the systematic variation in *X* into two parts, one which is related to both *X* and *Y* (co-varying noise) and one which is orthogonal to *Y* (structured noise). Two O-PLS algorithms were developed, the first (O1-PLS) is unidirectional $\mathrm{X} \Rightarrow \mathrm{Y}$ and the second (O2-PLS) is bi-directional $\mathrm{X} \Leftrightarrow \mathrm{Y}$ and is able to separate these different types of variations in both *X* and *Y* matrices [63, 64]. The practical result of using O-PLS algorithms inside of PLS is cleaner models that are easier to display and interpret.

An O2-PLS model can be written as

$$X = 1\overline{X'} + TP^T + T_o P_o^T + E \tag{11}$$

$$Y = 1\overline{Y'} + UC^T + U_o C_O^T + F \tag{12}$$

$$T = UB_u + H_{TU} \tag{13}$$

$$U = TB_T + H_{UT} \tag{14}$$

where $\overline{X}$ is a contain de X average; $\overline{Y}$ is a contain de Y average; T is a matrix of scores that summarizes the X variables; U is a matrix of scores that summarizes the Y variables; P is a matrix of weigh that express the correlation between X and U; C is a matrix of weigh that express the correlation between Y and T; E, F, H_{TU}, H_{UT} are the matrixes of residuals.

The matrixes TP^{T} and UC^{T} hold the joint X/Y information overlap [12, 63, 64].

In the last years, O2-PLS has become the preferred regression technique for the development of calibration models in NIR and Raman spectroscopy.

5.5. Bayesian ridge regression

Another regression method recently proposed for multivariate calibration of spectroscopic methods is Bayesian ridge regression. The method presents similarities with least squares, and the estimated coefficients tend toward zero in order to avoid collinearity [44].

In a Bayes-RR regression model, higher-level prior Gaussian distributions can be introduced over α^2 and α, and the prediction can be performed by integrating over α^2, α, and the regression parameters w. Since this prior distribution is conjugate to the likelihood function, the predictive distribution is also Gaussian [72]

$$p(y|\alpha, \alpha^2) = \int p(y|w, \alpha^2) p(w|\alpha) dw \tag{15}$$

The Bayes-RR is a widely used regression technique in machine learning based on the ridge regression [73], and in the last years its performance for the development of excellent models for spectroscopic calibration has been proved [72, 74, 75].

5.6. Support vector regression (SVR)

The support vector machines (SVMs) are a set of learning methods mostly used for classification that can be used as a regression technique which is called the support vector regression. In the last years, SVM started to be used in chemometrics for NIR spectra classification and multivariate calibration. The SVR uses the same principles as the SVM and is based on finding the hyperplane maximizing the margin between classes. The hyperplane maximizing the margin is justified by statistical learning theory endowed with a probabilistic test error bound that is minimized when the margin is maximized. The regression is performed using kernel functions that transform the data into a higher dimensional feature space to make a linear

separation possible. The models obtained by SVR are more complex and difficult to interpret in comparison with those obtained by other regression techniques [44, 76, 77].

5.7. Decision tree regression

Decision tree regression is a type of decision tree algorithm that can be applied to solve regression problems. Decision trees represent one of the main techniques used for discriminant analysis, classification, and prediction in knowledge discovery. It is widely used because it closely resembles human reasoning and is easy to understand. The principle is to compute a regression in a tree structure from breaking down a dataset into smaller and smaller subsets. Recently, some applications in multivariate calibration of spectroscopic methods have been proposed [44, 77–79].

Acknowledgements

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS - UEFISCDI, project number PN-III-P2-2.1-BG-2016-0201.

Author details

Ioan Tomuta*, Alina Porfire, Tibor Casian and Alexandru Gavan

*Address all correspondence to: tomutaioan@umfcluj.ro

Department of Pharmaceutical Technology and Biopharmacy, University of Medicine and Pharmacy "Iuliu Hatieganu", Cluj-Napoca, Romania

References

[1] Tomuta I, Rus L, Iovanov R, Rus LL. High-throughput NIR-chemometric methods for determination of drug content and pharmaceutical properties of indapamide tablets. Journal of Pharmaceutical and Biomedical Analysis. 2013;**84**:285-292

[2] Martínez L, Peinado A, Liesum L. In-line quantification of two active ingredients in a batch blending process by near-infrared spectroscopy : Influence of physical presentation of the sample. International Journal of Pharmaceutics. 2013;**451**(1–2):67-75

[3] Fonteyne M, Arruabarrena J, De Beer J, Hellings M, van den Kerkhof T, Burggraeve A, et al. NIR spectroscopic method for the in-line moisture assessment during drying in a six-segmented fluid bed dryer of a continuous tablet production line: Validation of quantifying abilities and uncertainty assessment. Journal of Pharmaceutical and Biomedical Analysis. 2014;**100**:21-27

[4] Van Renterghem J, Kumar A, Vervaet C, Paul J, Nopens I, Vander Y, et al. Elucidation and visualization of solid-state transformation and mixing in a pharmaceutical mini hot melt

extrusion process using in-line Raman spectroscopy. International Journal of Pharmaceutics. 2017;**517**(1–2):119-127

[5] Wahl PR, Fruhmann G, Sacher S, Straka G, Sowinski S, Khinast JG. PAT for tableting: Inline monitoring of API and excipients via NIR spectroscopy. European Journal of Pharmaceutics and Biopharmaceutics. 2014;**87**(2):271-278

[6] Möltgen C, Puchert T, Menezes JC, Lochmann D, Reich GA. Novel in-line NIR spectroscopy application for the monitoring of tablet film coating in an industrial scale process. Talanta. 2012;**92**:26-37

[7] Kauppinen A, Toiviainen M, Lehtonen M, Järvinen K, Paaso J, Juuti M, et al. Validation of a multipoint near-infrared spectroscopy method for in-line moisture content analysis during freeze-drying. Journal of Pharmaceutical and Biomedical Analysis. 2014;**95**:229-237

[8] Liu R, Li L, Yin W, Xu D, Zang H. Near-infrared spectroscopy monitoring and control of the fluidized bed granulation and coating processes—A review. International Journal of Pharmaceutics. 2017;**530**(1–2):308-315

[9] Karande AD, Wan P, Heng S, Liew CV. In-line quantification of micronized drug and excipients in tablets by near infrared (NIR) spectroscopy: Real time monitoring of tableting process. International Journal of Pharmaceutics. 2010;**396**(1–2):63-74

[10] Casian T, Reznek A, Vonica-gligor AL, Van Renterghem J, De Beer T, Tomu I. Development, validation and comparison of near infrared and Raman spectroscopic methods for fast characterization of tablets with amlodipine and valsartan. Talanta. 2017;**167**:333-343

[11] Taylor P, Gabrielsson J, Trygg J, Gabrielsson J, Trygg J. Recent developments in multivariate calibration. Journal of Chemometrics. 2006;**36**:243-255

[12] Eriksson L, Byrne T, Johansson E, Trygg J, Vikstrom C. Multi- and Megavariate Data Analysis: Basic Principles and Applications. 3rd ed. Malmo: MKS Umetrics AB; 2013. 1-505 pp

[13] Blanco M, Alcal M. Content uniformity and tablet hardness testing of intact pharmaceutical tablets by near infrared spectroscopy A contribution to process analytical technologies. Analytica Chimica Acta. 2006;**557**:353-359

[14] Mbinze JK, Sacré PY, Yemoa A, Mavar Tayey Mbay J, Habyalimana V, Kalenda N, et al. Development, validation and comparison of NIR and Raman methods for the identification and assay of poor-quality oral quinine drops. Journal of Pharmaceutical and Biomedical Analysis. 2015;**111**:21-27

[15] Tomuta I, Iovanov R, Bodoki E, Vonica L. Development and validation of NIR-chemometric methods for chemical and pharmaceutical characterization of meloxicam tablets. Drug Development and Industrial Pharmacy. 2014;**40**(4):549-559

[16] Virtanen S, Antikainen O, Yliruusi J. Determination of the crushing strength of intact tablets using Raman spectroscopy. International Journal of Pharmaceutics. 2008;**360**(1–2):40-46

[17] Croker DM, Hennigan MC, Maher A, Hu Y, Ryder AG, Hodnett BKA. Comparative study of the use of powder X-ray diffraction, Raman and near infrared spectroscopy for

quantification of binary polymorphic mixtures of piracetam. Journal of Pharmaceutical and Biomedical Analysis. 2012;**63**:80-86

[18] Gómez DA, Coello J, Maspoch S. Raman spectroscopy for the analytical quality control of low-dose break-scored tablets. Journal of Pharmaceutical and Biomedical Analysis. 2016;**124**: 207-215

[19] Sarraguca MC, Lopes JA. Quality control of pharmaceuticals with NIR: From lab to process line. Vibrational Spectroscopy. 2009;**49**:204-210

[20] Blanco M, Peguero A. Analysis of pharmaceuticals by NIR spectroscopy without a reference method. Trends in Analytical Chemistry. 2010;**29**(10):1127-1136

[21] Pan D, Crull G, Yin S, Grosso J. Low level drug product API form analysis—Avalide tablet NIR quantitative method development and robustness challenges. Journal of Pharmaceutical and Biomedical Analysis. 2014;**89**:268-275

[22] Blanco M, Peguero A. Controlling individual steps in the production process of paracetamol tablets by use of NIR spectroscopy. Journal of Pharmaceutical and Biomedical Analysis. 2010;**51**:797-804

[23] Farrell JA, Higgins K, Kalivas JH. Updating a near-infrared multivariate calibration model formed with lab-prepared pharmaceutical tablet types to new tablet types in full production. Journal of Pharmaceutical and Biomedical Analysis. 2012;**61**:114-121

[24] Blanco M, Peguero ANIR. Analysis of pharmaceutical samples without reference data: Improving the calibration. Talanta. 2011;**85**(4):2218-2225

[25] Netchacovitch L, Dumont E, Cailletaud J, Thiry J, De Bleye C, Sacré P, et al. Development of an analytical method for crystalline content determination in amorphous solid dispersions produced by hot-melt extrusion using transmission Raman spectroscopy: A feasibility study. International Journal of Pharmaceutics. 2017;**530**(1–2):249-255

[26] Eriksson L, Johansson E, Kettaneh-Wold N, Wikstrom C, Wold S. Design of Experiments—Principles and Applications. 3rd ed. MKS Umetrics AB: Umea; 2008

[27] Silva MAM, Ferreira MH, Braga JWB, Sena MM. Development and analytical validation of a multivariate calibration method for determination of amoxicillin in suspension formulations by near infrared spectroscopy. Talanta. 2012;**89**:342-351

[28] Ferreira MH, Braga JWB, Sena MM. Development and validation of a chemometric method for direct determination of hydrochlorothiazide in pharmaceutical samples by diffuse reflectance near infrared spectroscopy. Microchemical Journal. 2013;**109**:158-164

[29] Li Y, Igne B, Drennen JK, Anderson CA. Method development and validation for pharmaceutical tablets analysis using transmission Raman spectroscopy. International Journal of Pharmaceutics. 2016;**498**:318-325

[30] Short SM, Cogdill RP, Wildfong PLD, Iii JKD, Anderson CA. A near-infrared spectroscopic investigation of relative density and crushing strength in four-component compacts. Journal of Pharmaceutical Sciences. 2009;**98**(3):3-8

[31] Lyndgaard LB, Van Den Berg F, De Juan A. Quantification of paracetamol through tablet blister packages by Raman spectroscopy and multivariate curve resolution-alternating least squares. Chemometrics and Intelligent Laboratory Systems. 2013;**125**:58-66

[32] Igne B, Shi Z, Iii JKD, Anderson CA. Effects and detection of raw material variability on the performance of near-infrared calibration models for pharmaceutical products. Journal of Pharmaceutical Sciences. 2014;**103**:545-556

[33] Griffen J, Owen A, Matousek P. Comprehensive quantification of tablets with multiple active pharmaceutical ingredients using transmission Raman spectroscopy—A proof of concept study. Journal of Pharmaceutical and Biomedical Analysis. 2015;**115**:277-282

[34] Porfire A, Muntean D, Achim M, Vlase L, Tomuta I. Simultaneous quantification of simvastatin and excipients in liposomes using near infrared spectroscopy and chemometry. Journal of Pharmaceutical and Biomedical Analysis. 2015;**107**:40-49

[35] Heinz A, Savolainen M, Rades T. Strachan CJ, Quantifying ternary mixtures of different solid-state forms of indomethacin by Raman and near-infrared spectroscopy. European Journal of Pharmaceutical Sciences. 2007;**2**:182-192

[36] Bondi RW, Igne B, Drennen JK, Anderson CA. Effect of experimental design on the prediction performance of calibration models based on near-infrared spectroscopy for pharmaceutical applications. Applied Spectroscopy. 2012;**66**(12):1442-1453

[37] Alam A, Drennen J, Anderson C. Designing a calibration set in spectral space for efficient development of an NIR method for tablet analysis. Journal of Pharmaceutical and Biomedical Analysis. 2017;**145**:230-239

[38] Lee M, Seo D, Lee H, Wang I, Kim W, Jeong M, et al. In line NIR quantification of film thickness on pharmaceutical pellets during a fluid bed coating process. International Journal of Pharmaceutics. 2011;**403**(1–2):66-72

[39] Gendre C, Genty M, Boiret M, Julien M, Meunier L, Lecoq O, et al. Development of a process analytical technology (PAT) for in-line monitoring of film thickness and mass of coating materials during a pan coating operation. European Journal of Pharmaceutical Sciences. 2011;**43**(4):244-250

[40] Gendre C, Boiret M, Genty M, Chaminade P, Manuel J. Real-time predictions of drug release and end point detection of a coating operation by in-line near infrared measurements. International Journal of Pharmaceutics. 2011;**421**(2):237-243

[41] Kauffman JF, Dellibovi M, Cunningham CR. Raman spectroscopy of coated pharmaceutical tablets and physical models for multivariate calibration to tablet coating thickness. Journal of Pharmaceutical and Biomedical Analysis. 2007;**43**:39-48

[42] Moes JJ, Ruijken MM, Gout E, Frijlink HW. Ugwoke MI, Application of process analytical technology in tablet process development using NIR spectroscopy: Blend uniformity, content uniformity and coating thickness measurements. International Journal of Pharmaceutics. 2008;**357**:108-118

[43] Möltgen C, Herdling T, Reich G. A novel multivariate approach using science-based calibration for direct coating thickness determination in real-time NIR process monitoring. European Journal of Pharmaceutics and Biopharmaceutics. 2013;**85**(3):1056-1063

[44] Clavaud M, Roggo Y, Dégardin K, Sacré P, Hubert P, Ziemons E. Global regression model for moisture content determination using near-infrared spectroscopy. European Journal of Pharmaceutics and Biopharmaceutics. 2017;**119**:343-352

[45] Karande AD, Liew CV, Heng PWS. Calibration sampling paradox in near infrared spectroscopy: A case study of multi-component powder blend. International Journal of Pharmaceutics. 2010;**395**:91-97

[46] Lasch P. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. Chemometrics and Intelligent Laboratory Systems [Internet]. 2012;**117**: 100-114. DOI: 10.1016/j.chemolab.2012.03.011

[47] Gabrielsson J, Trygg J. Recent developments in multivariate calibration. Critical Reviews in Analytical Chemistry. 2006;**36**(3–4):243-255

[48] Gabrielsson J, Jonsson H, Airiau C, Schmidt B, Escott R, Trygg J. OPLS methodology for analysis of pre-processing effects on spectroscopic data. Chemometrics and Intelligent Laboratory Systems. 2006;**84**(1–2 SPEC. ISS):153-158

[49] Fearn T, Riccioli C, Garrido-Varo A, Guerrero-Ginel JE. On the geometry of SNV and MSC. Chemometrics and Intelligent Laboratory Systems. 2009;**96**(1):22-26

[50] Barnes RJ, Dhanoa MS, Lister SJ. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Applied Spectroscopy. 1989;**43**(5):772-777

[51] Rinnan Å, van den Berg F, Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. TrAC—Trends in Analytical Chemistry. 2009;**28**(10):1201-1222

[52] Savitzky A, Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry. 1964;**36**(8):1627-1639

[53] Wold S, Antii H, Lindgren F, Ohman J. Orthogonal signal correction of near infrared spectra. Chemometrics and Intelligent Laboratory Systems. 1998;**44**:144-185

[54] Westerhuis JA, De Jong S, Smilde AK. Direct orthogonal signal correction. Chemometrics and Intelligent Laboratory Systems. 2001;**56**(1):13-25

[55] Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). Journal of Chemometrics. 2002;**16**(3):119-128

[56] Zeaiter M, Roger JM, Bellon-Maurel V. Robustness of models developed by multivariate calibration. Part II: The influence of pre-processing methods. TrAC—Trends in Analytical Chemistry. 2005;**24**(5):437-445

[57] Engel J, Gerretzen J, Szymanska E, Jansen JJ, Downey G, Blanchet L, et al. Breaking with trends in pre-processing? TrAC—Trends in Analytical Chemistry. 2013;**50**:96-106

[58] Porfire A, Filip C, Tomuta I. High-throughput NIR-chemometric methods for chemical and pharmaceutical characterization of sustained release tablets. Journal of Pharmaceutical and Biomedical Analysis. 2017;**138**:1-13

[59] Eriksson L, Byrne T, Johansson E, Trygg J, Vikström C. Multi- and Megavariate Data Analysis: Basic Principles and Applications. 2nd ed. MKS Umetrics AB: Malmo; 2006. 500 p

[60] Tôrres AR, Grangeiro S, Fragoso WD. Vibrational spectroscopy and multivariate control charts: A new strategy for monitoring the stability of captopril in the pharmaceutical industry. Microchemical Journal. 2017;**133**:279-285

[61] Zhu L, Brereton RG, Thompson DR, Hopkins PL, Escott REA. On-line HPLC combined with multivariate statistical process control for the monitoring of reactions. Analytica Chimica Acta. 2007;**584**(2):370-378

[62] Rencher AC. Methods of Multivariate Analysis. IIE Transactions. Vol. 37. 2nd ed. 2012. 727 p.

[63] Trygg J, Wold S. O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. Journal of Chemometrics. 2003;**17**(1):53-64

[64] Trygg J. Prediction and spectral profile estimation in multivariate calibration. Journal of Chemometrics. 2004;**18**(34):166-172 Available from: http://doi.wiley.com/10.1002/cem.860

[65] Rajalahti T, Kvalheim OM. Multivariate data analysis in pharmaceutics: A tutorial review. International Journal of Pharmaceutics. 2011;**417**(1–2):280-290. DOI: 10.1016/j.ijpharm.2011.02.019

[66] Jackson JE. A User's Guide to Principal Components. New York - Chichester - Brisbane - Toronto - Singapore: John Wiley & Sons, Inc; 2003. 592 p

[67] Zumel N, Mount J. Practical Data Science with R [Internet]. Manning Publications Co; 2014. 416 p. Available from: https://livebook.manning.com/#!/book/practical-data-science-with-r/table-of-contents/

[68] Wold S, Sjostrom M. PLS-regression—A basic tool of chemometrics.pdf. 2001;109-130

[69] Biagioni DJ, Astling DP, Graf P, Davis MF. Orthogonal projection to latent structures solution properties for chemometrics and systems biology data. Journal of Chemometrics. 2011;**25**(9):514-525

[70] Dunn K. Process improvement using data. 2016;(October):378

[71] Otto M. Chemometrics. Statistics and Computer Application in Analytical Chemistry. 3rd ed. Wiley-VCH Verlag GmbH & Co.: Weinheim; 2017. 385 p

[72] Chen T, Martin E. Bayesian linear regression and variable selection for spectroscopic calibration. Analytica Chimica Acta. 2009;**631**(1):13-21

[73] Taylor P, Frank E. Friedman JH, A statistical view of some chemometrics regression tools a statistical view of some chemometrics regression tools. Technometrics. April 2013;**2012**:37-41

[74] Hoerl AE, Kennard RW. Ridge regression: Biased estimation for problems nonorthogonal. Technometrics. 2000;**42**(1):80-86

[75] Brown PJ, Fearn T, Vannucci M. Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. Journal of the American Statistical Association. 2001;**96** (454):398-408 Available from: http://www.tandfonline.com/doi/abs/10.1198/016214501753168118

[76] Smola A, Schölkopf B. A tutorial on support vector regression. NeuroCOLT2 Technical Report Series. 1998;73. Available from: http://www.springerlink.com/index/KM7KRM46802R2114.pdf

[77] Mutihac L, Mutihac R. Mining in chemometrics. Analytica Chimica Acta. 2008;**612**(1):1-18

[78] Kotsiantis SB. Decision trees: A recent overview. Artificial Intelligence Review. 2013; **39**(4):261-283

[79] Czajkowski M, Kretowski M. The role of decision tree representation in regression problems—An evolutionary perspective. Applied Soft Computing. 2016;**48**:458-475. Available from:. DOI: 10.1016/j.asoc.2016.07.007