Title: Quantification and handling of nonlinearity in Raman micro-spectrometry of pharmaceuticals

Author: Brigitta Nagy Attila Farkas Attila Balogh Hajnalka Pataki Balázs Vajna Zsombor K. Nagy György Marosi

PII: S0731-7085(16)30275-8
DOI: http://dx.doi.org/doi:10.1016/j.jpba.2016.05.036
Reference: PBA 10679

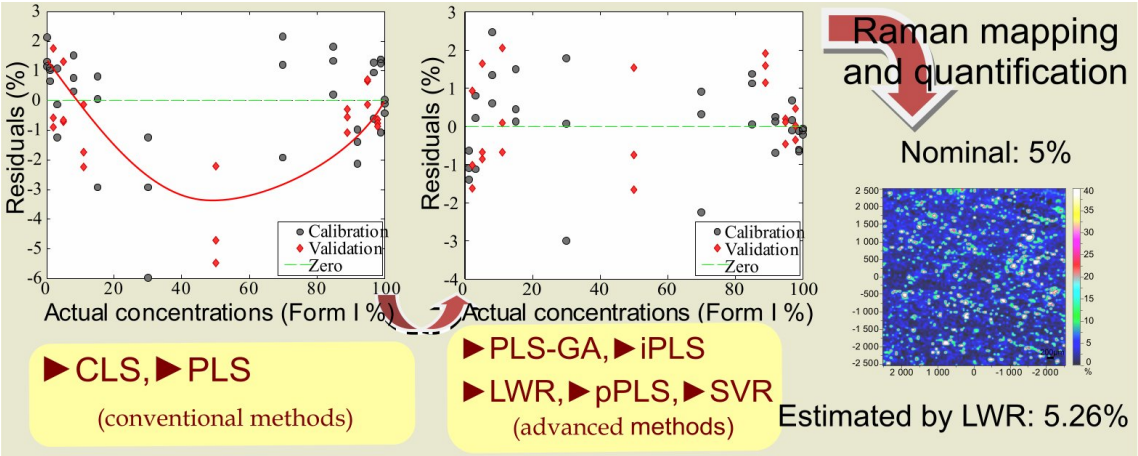
To appear in: *Journal of Pharmaceutical and Biomedical Analysis*

Received date: 2-3-2016
Revised date: 12-5-2016
Accepted date: 20-5-2016

Please cite this article as: Brigitta Nagy, Attila Farkas, Attila Balogh, Hajnalka Pataki, Balázs Vajna, Zsombor K.Nagy, György Marosi, Quantification and handling of nonlinearity in Raman micro-spectrometry of pharmaceuticals, Journal of Pharmaceutical and Biomedical Analysis http://dx.doi.org/10.1016/j.jpba.2016.05.036

Highlights

GAs and iPLS improved errors of determined concentrations with Raman spectroscopy.
The whole non-linearity of the dataset can be removed by LWR, pPLS or SVM.
High non-linearity appears if a component of high Raman-activity occurs in samples.
Raman map was created with real concentration applying proper quantitative regression.



Quantification and handling of nonlinearity in Raman micro-spectrometry of pharmaceuticals

Brigitta Nagy[a], Attila Farkas[a], Attila Balogh[a], Hajnalka Pataki[a], Balázs Vajna[a], Zsombor K. Nagy[a], György Marosi[a,]*

[a] Department of Organic Chemistry and Technology, Budapest University of Technology and Economics, H-1111 Budapest, Budafoki út 8., Hungary

* Corresponding author.

Tel.: +36 1 463 3654

E-mail: gmarosi@mail.bme.hu

Abstract

This work demonstrates how nonlinearity in Raman spectrometry of pharmaceuticals can be handled and accurate quantification can be achieved by applying certain chemometric methods including variable selection. Such approach proved to be successful even if the component spectra are very similar or spectral intensities of the constituents are strongly different. The relevant examples are: blends of two crystalline forms of carvedilol ("CRYST-PM" blend) and a three-component pharmaceutical model system ("PHARM-TM" blend). The widely used classical least squares regression (CLS) and partial least squares regression (PLS) quantification methods provided relatively poor root mean squared error of prediction (RMSEP) values: approximately 2-4% and 4-10% for CRYST-PM and PHARM-TM respectively. The residual plots of these models indicated the nonlinearity of the preprocessed data sets. More accurate quantitative results could be achieved with properly applied variable selection methods. It was observed that variable selection methods discarded the most intensive bands while less intensive ones were retained as the most informative spectral ranges. As a result not only the accuracy of concentration determination was enhanced, but the linearity of models was improved as well. This indicated that nonlinearity occurred especially at the intensive spectral bands. Other methods developed for handling nonlinearity were also capable of adapting to the spectral nature of both data sets. The RMSEP could be decreased this way to 1% in CRYST-PM and 3-6% in PHARM-TM. Raman maps with accurate real concentrations could be prepared this way. All quantitative models were compared by the non-parametric sum of ranking differences (SRD) method, which also proved that models based on variable selection or nonlinear methods provide better quantification.

Keywords: hyperspectral imaging; pharmaceutical analysis; chemometrics; multivariate regression; variable selection; nonlinear behavior

1. Introduction

Chemical imaging, including Raman mapping, has been a particularly rapidly emerging technique in analysis and characterization of various substances in the last decade. This method has several advantages over non-imaging spectroscopic techniques. Besides qualitative and quantitative characterization of ingredients in bulk materials, it can provide further information about spatial distribution of major and even minor (sometimes trace) components [1]. Many diverse issues have been already solved using chemical imaging in the field of life sciences and diagnostics [2,3], forensic sciences and counterfeiting [4,5], food analysis [6,7], plastics [8] and artworks [9]. In recent years, the application of this approach has sparked explosively growing interest particularly in the pharmaceutical industry [10,11]. Raman (or NIR) chemical imaging is greatly applicable for performing detailed analysis in various steps of manufacturing processes [12,13]. Researchers pay more and more attention to quantification, further highlighting the relevance of this topic.

There are many issues in which Raman spectroscopy and hyperspectral imaging has helped tackle serious challenges, such as identifying unexpected chemical substances or new polymorphs [14,15], investigating blend homogeneity [16], or testing polymorphic stability [17,18]. These qualitative studies reveal various types of information about the samples, allowing better understanding of pharmaceutical processes. However, the interest, regarding pharmaceutics, is mostly focusing on the complex view of the structure and the real quantification of component. Since FDA approved the guidance on Process Analytical Technology (PAT) [19], the significance of spectroscopic techniques has extremely grown thanks to their superior adaptability into continuous manufacturing processes [20-23]. Ongoing quality control can be achieved through accurate spectral evaluation, to which, however, the use of chemometrics is essential.

Since a vibrational (such as Raman, IR or NIR) spectrum contains hundreds or thousands of wavenumbers of interest, it is a multivariate entity, which progresses to a further level of complexity when combined with hyperspectral imaging (which requires processing of even thousands of spectra within the same dataset). Numerous chemometric methods may serve the quantification efforts [10,24-26]. In the simplest cases univariate approaches using one selective wavenumber can provide useful results. However, in many cases a sufficiently selective band does not exist [10], in which case multivariate methods has to be used encompassing the whole spectral range or parts of it [27]. One of the most easily interpretable methods is the classical least squares (CLS). It can be adapted fast for simple spectrum characterization when the spectra of all components are known and spectrum of each compound can be generated from pure spectra using spectral contributions (as estimated concentrations). However, some interfering effects, such as spectral similarity of components or material interactions or nonlinear behavior, may occur making it necessary to use more sophisticated approaches. Some authors have reported successful application of widespread chemometric methods such as partial least squares regression (PLS) or principal component regression (PCR) [28-30]. These methods were successfully used in polymorphic studies, where the component spectra are only slightly different. PLS is especially preferred for quantification of selected polymorphs in a mixture.

A large part of the spectral range is often non-informative in the quantitative evaluation. In such cases variable selection methods, such as interval partial least squares (iPLS) [31,32] or genetic algorithms (GA) [33-35], are promising candidates for treating the spectra and retaining only the sufficiently descriptive variables. In some cases, a certain degree of nonlinearity appears in the data, caused by interaction between components or due to spectral preprocessing. This phenomenon can be handled by polynomial partial least squares (pPLS) [36,37], locally weighted regression (LWR) [38,39] or support vector machine for regression

(SVR) [40,41]. pPLS works similarly to the conventional PLS regression, but it uses higher degree inner relations by determining polynomial functions between score values of dependent and independent variables called X-block scores and Y-block scores. LWR was applied based on PLS projection by fitting local PLS models to a specified range of adjacent observations. Applying SVR, the regression is carried out in a higher-dimensional feature space, in which the nonlinearity of the original input data can be handled properly. The construction of the suitable hyperplane is performed by a kernel function and regularized by several parameters (see Section 2.4.3.)

Although the conventional data analysis methods (CLS, PCR, PLS) proved to be successful in many applications, advances in chemical imaging and in pharmaceutical process-monitoring calls for novel chemometric methods [25,42]. Nevertheless up to now only few pharmaceutical related studies have demonstrated the advantageous use of aforementioned chemometric ways for variable selection and handling of nonlinearity. Support vector machine, for instance, has been proposed as a promising candidate [42] and used in Raman [43] and UV [44] quantitative spectroscopic studies, iPLS was applied in determination of Vitamin B12 in pharmaceutical tablets [45] and in the quantification of ibuprofen-nicotinamide co-crystals [46], while LWR has been used in a NIR quantitative analysis [47]. However, detailed comparative study demonstrating the relevance of the mentioned methods for quantitative determination of polymorph ratio and tablet composition based on Raman mapping has not published yet. Thus, the aim of this study is to evaluate the applicability of these tools in analysis of two model systems of pharmaceutical importance.

2. Materials and methods

2.1 Materials

Two-component mixtures of crystalline polymorphs of carvedilol model drug (referred to as CRYST-PM) were studied. The commercial carvedilol product (EGIS Pharmaceuticals Plc., Budapest, Hungary) consisted of pure Form II polymorph. Form I was obtained by a solvent mediated polymorphic transition process. First, 25 g Form II was dissolved entirely in 120 mL ethyl-acetate (Merck, Germany) heating the solution until 77 °C. The solution then was cooled while 2.5 g of Form I polymorph was added as seed crystals. The recrystallization occurred at 50 °C in three hours [48]. Crystals were removed by filtration. After drying the product purity was verified by Raman mapping.

The three component pharmaceutical model system (PHARM-TM) contained imipramine (EGIS Pharmaceuticals Plc., Budapest, Hungary) as model drug and microcrystalline cellulose (FMC BioPolymer, Princeton, USA) and maize starch (Colorcon, West Point, USA) as excipients. Each component was sieved to ensure the same particle size range (50-100 μm), to avoid segregation.

2.2 Preparation of mixtures and tablets

Uniform binary mixtures were achieved by grinding and mixing carvedilol Form I and Form II in a mortar with pestle, creating nineteen blends with different mass ratios (see Table 1). The total weight of each mixture was 2.00 g and the measurements of the components were carried out on analytical balance (precision of 0.1 mg). As the precision of the weighted quantity of the components were within 5 mg, there was no significant difference between the prepared actual and nominal concentrations. The mixtures were prepared right before the Raman measurements. Sufficient homogeneity was obtained by ten minutes of thorough homogenization, which was checked by collecting three Raman maps per mixtures (See Section 2.3.). As there were no differences in the spatial distribution of the Raman maps of the three samples, the homogeneity of the mixtures were deemed to be representative.

PHARM-TM was prepared and homogenized in the same manner and then it was compressed into a flat tablet (Camilla '95 OL57; Manfredi, Torino, Italy). Before analyzing the tablet form was broken in half and the fracture surface was mapped. Fig. 1 represents the nineteen nominal compositions. Due to long measurement time (see Section 2.3), the representative sampling were not intended to achieve by repeated Raman mapping instead, the homogeneity of the measured samples was checked by macropixel analysis [49] (see details in supplementary material).

2.3 Raman mapping experiments

Raman spectra were collected using a Horiba Jobin-Yvon LabRAM system coupled with an Olympus 97 BX-40 optical microscope (Olympus Corporation, Tokyo, Japan). Raman mapping of CRYST-PM blends were carried out with an external 532 nm frequency-doubled Nd-YAG laser source, while the samples of PHARM-TM were illuminated by a 785 nm diode laser (TEC 510 type, Sacher Lasertechnik, Marburg, Germany). An objective of 10× magnification was used for optical imaging and spectrum acquisition. The laser beam is focused through the objective, and backscattered radiation is collected with the same objective, as usual in most confocal spectroscopic systems. The collected radiation is directed through an edge filter that removes the Rayleigh photons, then through a confocal hole and the entrance slit onto a grating monochromator that disperses the light before it reaches the CCD detector.

The Raman maps of polymorphs were collected with a spectral range of 345-1790 cm^{-1}. The acquisition time for one spectrum was 1 s and 3 spectra were averaged per pixel.
Samples were investigated in 441 points by collecting a 21×21 pixel sized image with 50 µm step size. Each sample was mapped three times to ensure representativeness of the sampling.

Mixtures of PHARM-TM were investigated in a spectral range of 458-1678 cm^{-1}. Two spectra were accumulated using an acquisition time of 30 s in each spatial position. 13×13 pixels sized images (169 points) were collected with a step size of 50 μm.

2.4 Data analysis

Basic evaluation of the Raman maps (*i.e.* unfolding of the 3D data cube, basic preprocessing) were carried out using LabSpec 5.41 (Horiba Jobin Yvon, France). Chemometric analyses were performed using MATLAB 8.2. (MathWorks, USA) and PLS Toolbox 7.8.2. (Eigenvector Research, USA). Sum of ranking differences (SRD) method was carried out using a VBA macro in Excel 2010 (Microsoft, USA) made available by the developers (http://aki.ttk.mta.hu/srd/).

The spectrometric map was originally acquired in a 3-dimensional hypercube form (sized $n \times m \times \lambda$), which contained the spectral intensities. Two of the cube dimensions are the spatial coordinates, while the third corresponds to the wavelength channels. Before any chemometric analysis, the data had to be unfolded into a 2-dimensional matrix along to the coordinates in order to treat the data in the mentioned programs. Wavelength channels are in columns of the data matrix (in this case number of the channels is 1000) and each row contains a spectrum of a measured point.

Quantitative models were built by using the averaged spectra of the Raman maps. The sample sets were divided into calibration and validation sets as marked in Table 1 and Fig. 1. All

model building was conducted on calibration set, while the validation set was used to examine the predictive power of these models. On the calibration sets, contiguous blocks cross-validation was also performed by leaving out one concentration level at a time.

All Raman mapped spectra were preprocessed by baseline correction and normalization. Other preprocessing methods such as mean centering, autoscaling, multiplicative scatter correction (MSC) and standard normal variate (SNV) were also tested in various combinations. The combinations that provided the lowest RMSE value were applied for the final model building.

The goodness of fit, *i.e.* the performance of the model was evaluated by comparing the predicted and actual concentrations, using the two most widely used measures, namely the coefficients of determination (R^2) and the root mean square errors (RMSE). The indicators of goodness were calculated for calibration (R^2_{cal}, RMSEC), prediction (R^2_{pred}, RMSEP) as well as cross-validation (R^2_{cv}, RMSECV).

R^2 values come from the conventional analysis of variance in model fitting [50]. However, coefficient of determination was not in itself used to drive any decisions; it was used as a secondary metric to the RMSE values. The root mean square error (RMSE) was a simple measure to compare the calculated concentrations from the model and the "real" (actual or measured by other means) concentrations (see Equation 1).

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{\left(y_i^{(calc)} - y_i^{(real)}\right)^2}{N}} \qquad \text{(Eq. 1)}$$

N is the number of elements of the set, $y_i^{(real)}$ is the actual concentration in spectrum *i*. The $y_i^{(calc)}$ refers to the predicted concentration and may come from the calibration set (RMSEC), validation set (RMSEP) or the cross-validation iterations (RMSECV).

The comparison of prediction and cross-validation played an important role in the examination of overfitting. To test the predictive power of the models, in this study the RMSEP values of the validation set were compared with the RMSECV values. Inspecting the residual plots is an efficient way of revealing the nonlinearity of the models.

Sum of ranking differences [51,52], which is a non-parametric statistical process, was applied to compare the models. The objects (the quantitative models in this case) can be ordered and ranked based on a comparison to a reference (*e.g.* estimated concentrations are compared to actual concentrations in each pixel). The ranking differences are then summed. These SRDs are scaled to 0 to 100 percent, indicating the fully perfect match (0%) and the completely reversed ranking, respectively. Accordingly, the method having the lowest SRD is regarded as the best. By generating high number of random rankings, a Gaussian distribution is obtained to test whether the rankings of the examined methods differ significantly from the randomly generated rankings (see supplementary material).

2.4.1 Reference methods

Univariate regression and classical least squares (CLS) principal component regression (PCR) and partial least squares (PLS) models were used as reference models methods to estimate the concentrations in the mixtures. These methods are detailed in the literature [28-30].

2.4.2 Variable selection methods

Spectra with thousands of wavelength channels (variables) usually contain some non-informative variables. Although PCR and PLS project data to lower dimensions, aiming to retain only the substantial information, in fact each latent variable created still takes all

original variable into account, which reduces the predictive power of the model. To make models more accurate, non-informative regions should be completely removed before regression. For this purpose, two variable selection methods were tested.

Interval PLS [31,32]: Interval PLS is based on a systematic search of the best combination of variables to reach a better estimation against PLS on the whole spectral range. First, the spectra are divided into variable windows, with a window size specified by the user. Utilizing variable windows enables us to select meaningful bands of the spectra instead of standalone wavenumbers. However, the application of wider windows can obstruct the entirely elimination of non-informative variables, while narrow windows could lead to overfitting. Hence, determining the appropriate window size is a substantial matter of optimization, which was carried out by minimizing the RMSECV values. The searching algorithm can run in two different modes. Applying *forward selection* mode, the initial step is a PLS model built with one variable window and RMSECV is calculated. In the next steps, variable windows are individually added and PLS models are built. Variable windows are added until the best subset of variable windows (resulting model with the lowest RMSECV) is found. *Reverse* mode operates in the opposite direction. The first PLS model is built by eliminating only one window from the spectra, then variable windows are excluded successively until the RMSECV of the PLS models cannot be reduced further. In this way, iPLS determine a subset of variables with which the goodness of PLS model is enhanced. Nevertheless, it has to be noted that being a stepwise method, iPLS tend to stuck in local minimum, hence it does not necessarily results the best subset.

PLS regression aided by Genetic algorithms (PLS-GA) [33-35]: Compared to the iPLS method, PLS-GA is suitable for finding the variable subsets resulting in the global minimum of RMSECV and then the best subset can be applied in a PLS regression. The searching algorithm operates on the analogy of natural selection. Sets of variable windows (*genes*) are

used to create regression models (*individuals*) which are compared to one another based on the RMSECV (*the fitness function*) values. These models form a "population". The best half of the models is left in the population, while the rest are discarded. The remaining "survivor" genes are "crossed over" to fill up the population to its original size (while some mutation may occur as well), creating a new "generation" to next calculation. The algorithm stops when the required convergence criterion or the maximum number of generation is reached.

2.4.3 Regression methods for handling nonlinearity

Polynomial PLS (pPLS) [36,37]: PLS regression can be used for nonlinear curve fitting by using higher powers of scores. During the model building process the nonlinear/polynomial function has to be determined to describe the relationship in the inner model of PLS.

Locally weighted regression (LWR): [38,39] It is worth noting that LWR is a linear method locally. Nevertheless it can be used if data sets exhibit nonlinearity as it decomposes a single model in a series of local models [53]. Curve fitting with locally weighted regression method is carried out by considering only a defined number of adjacent points during the regression process. Adjacent points are also weighted LWR the distance from the given calibration point. The weight function and the number of adjacent points had to be determined by the user. The regression can be carried out either with the original variables or in the principal component space.

Support vector regression (SVR) [40,41]: Support vector regression is a machine learning method that is based on the transformation of the nonlinear data matrix into a higher dimension space where the data points fit to a linear curve. The transformation is carried out by a kernel function. Gaussian radial basis function kernel was applied, which is referred as a type of nonlinear SVR [43,54]. The effectiveness of SVR depends on parameter (γ) which influences the "curliness" of the kernel. In order to fit the regression function, two other

parameters also had to be set. The margin of tolerance (ε) affects the number of points involved in curve fitting, as only the calibration points outside or on the margin (called support vectors) are actually taking part in the regression. The cost (C) value represents the penalty associated with errors larger than ε.

3. Results and discussion

The two selected model systems allowed us to study the quantification of pharmaceutical samples in various aspects. Carvedilol polymorph mixtures (CRYST-PM) made it possible to study the accuracy of the quantification of components with high spectral similarity (Fig. 2). This examination can be essential to convince us, for instance, about purity or long-term stability of polymorphic form of active pharmaceutical ingredient (API). Detection and quantification limits at low concentration are required to recognize the appearing impurities. For these reasons, data analysis was mainly carried out at low concentration levels.

Effective quantification of blends of an API and excipients (PHARM-TM model system) is also a significant issue during drug formulation steps or in quality control. Although the model system consisted of three components only, data evaluation was difficult due to the fact that the spectra of the selected two typical excipients were similar to each other (see Fig. 3). In addition, imipramine has much larger Raman activity than the excipients. Consequently, applying the necessary spectral preprocessing (*e.g.* normalization), leaded to nonlinearity in the spectral data (meaning that the peak intensities have a nonlinear correspondence to the actual concentrations). Our approach aimed at the development of an effective model that is suitable to quantify all components in these blends.

3.1 Preprocessing

Before any chemometric analysis, it is required to preprocess the data. Baseline correction was needed to eliminate fluorescent background scattering. In CRYST-PM the same baseline was subtracted from all the measured spectra approaching it with a piecewise linear baseline attached to the assigned wavenumbers. However, in the PHARM-TM system, determining one piecewise linear function, which is appropriate for the correction of each spectrum is a cumbersome procedure due to the spectral and fluorescent differences of the components. For this reason the method of asymmetric least squares [55] was applied to fit a nonlinear baseline, which is defined by two parameters: the asymmetry parameter p, the value of which is ranging usually from 10^{-3} to 10^{-1}, and the smoothness parameter λ, generally set exponentially in the range of 10^2 and 10^9. In this study, the baseline were tuned empirically to p=10^{-2} and λ= 10^5.

In mapping studies, normalization is also commonly used to eliminate the intensity fluctuation caused by the error of focusing due to surface roughness. If this is not treated by some sort of normalization (bringing back each spectrum to the same scale) then all quantitative algorithms will be inherently affected by the goodness of the focusing in each pixel. However, spectral normalization, balancing the Raman-activity differences of the components, leads to loss of information and distorts quantification. Although it has some unwanted side effect of spreading out spectral differences on the entire spectrum this disadvantage is much smaller than the negative impact caused by the lack of normalization. As spectra of different polymorphs usually have similar Raman-activity, normalization constituted only slight interference in the CRYST-PM samples. In contrast, in the case of three component system

(PHARM-TM) normalization biased strongly the real contribution of imipramine (as the Raman intensities of the components were highly different).

In addition to baseline correction and normalization, mean centering, autoscaling, multiplicative scatter correction (MSC) and standard normal variate (SNV) were also tested in various combinations in the case of each calibration method. The best combination was applied for each model; these can be seen in Table 2 and Table 3.

3.2 Reference methods

In order to understand the significance of using variable selection and methods handling the nonlinearity, first, we carried out the quantification studies in comparison with reference methods described in Section 2.4. Preliminary studies with univariate regression showed that the selectivity of the visually chosen wavelength (725.9 cm^{-1} and 753.1 cm^{-1} in the case of CRYST-PM and 1594 cm^{-1}, 480 cm^{-1}, 1096 cm^{-1} for PHARM-TM) influence significantly the accuracy of the quantification. In the case of both model systems the method resulted in quantification errors higher than 3% (Table 2-3).

Classical least squares (CLS) regression method with CRYST-PM did not yield any better results than univariate regression: RMSEs exceeded 3% and it was not able to discriminate between the adjacent concentration levels. The residual plot (difference of predicted and actual concentration) shows nonlinearity in the data (Fig. 7.a). In the case of PHARM-TM the CLS model compared to univariate regression provided better determination for each component, however the RMSE values still remained about 10%. The highest error was observed in PHARM-TM (Fig. 4), which can be explained by the fact that this is the case when normalization caused the greatest bias in the spectrum of the API. The residual plot of imipramine shows nonlinearity too (Fig. 4.c).

Using principal component regression (PCR) and partial least squares (PLS) methods the main issue is the following: in order to avoid overfitting the number of principal components/latent variables (LVs) have to be determined cautiously. The optimization is carried out by the minimization of RMSECV, however, it is not sufficient to recognize overfitting. Our previous study [56] pointed out that overfitting is more effectively tackled when the search for the optimal LV numbers is not solely performed according to the minimum of RMSECV. It is also advised to keep in mind that the number of LVs should not greatly exceed the (known) number of components in the calibration samples. Considering all these factors, two latent variables were used in the PCR and PLS model of carvedilol system and this way the RMSE values were reduced to 2% (see Table 2). However, to analyze impurities in the samples (*i.e.* achieve better predictions for small concentration levels for each component), more accurate models are required. It also has to be mentioned that strong nonlinearity was observed related to the accuracy of estimations in the residual plots (Fig. 7.b).

PHARM-TM was quantified by using 3 LVs. The PLS model with mean centering was capable of decreasing the RMSEs of imipramine by partially diminishing the bias induced by normalization. Nevertheless, the errors of determination were still rather high (about 6-8%) due to a certain degree of nonlinearity remaining, which had to be reduced.

3.3 Variable selection methods

When applying variable selection methods, it should be kept in mind that a band in the spectrum includes more than one variable; hence groups of adjacent variables (wavenumbers) should be treated together. This was set by using a pre-defined size of variable windows instead of single variables. Different window sizes were tested and the optimum was determined based on the RMSECV values. Interval PLS and PLS-GA methods resulted in the

best predictions for CRYST-PM when the window size was set to contain 25 variables (it equals approximately 36 cm^{-1} in this case). After the model construction, the RMSEP values were calculated. When window sizes of 10 or 5 were used, RMSEP exceeded RMSECV, indicating that narrow window sizes cause overfitting (see Fig. 5). It was found that there are more combinations of wavelength intervals which characterize the model system equally well (RMSE values equal in their first decimal); in other words, iPLS algorithm found different local minimums of RMSECV.

Genetic algorithms aided PLS-GA method was performed applying RMSECV as the fitness function. The mutation rate, *i.e.* the probability of the mutation during the double cross-over was set to 0.5% and the population size of 64 was applied. The algorithm stopped when the convergence criterion of 50% or 200 generations were reached. These parameters were optimized empirically based on previous study [56], however it was found that they have much less impact on the RMSE values than the window size. The run of PLS-GA was monitored through diagrams (Fig. 6) showing the evolution of RMSECV values in the function of the number of applied variable windows in a particular generation (Fig. 6.a) as well as in the function of the generations (Fig. 6.b). Fig. 6.b depicts how the algorithm approaches a minimum value of RMSECV and Fig. 6.c presents that the number of variable windows is reduced by the end of the process. The final result of the variable selection is illustrated in Fig. 6.d, where the chosen variable windows of the last generation are depicted on the frequency of their occurrence. Consecutive runs and repeated analyses showed that certain variable windows were frequently selected. The PLS models built by these runs equally resulted in improved model goodness. No significant difference was identified between the repeated GA runs, thus only the results of one calculation is detailed in Table 2.

Additionally, it is important to note that the most intensive bands were discarded by all GA runs, which can be explained by the massively nonlinear or overlapping behavior of these

peaks. Hence this result is an important argument against the application of univariate methods, even if selective bands seem to be suitable to distinguish particular components. The iPLS and PLS-GA models of CRYST-PM provided error of determination of approximately 1% with moderate level of nonlinearity (Fig. 7.c).

The use of variable selection methods led to positive effects in the PHARM-TM results as well. The variable selection affected the RMSEs of each component to a different extent, which made the comparison of the models more complex. Most of the selected variable windows included the peaks of imipramine, while the wide bands of the excipients were selected less frequently. Accordingly, the greatest improvement in the goodness of prediction was reached in the quantification of imipramine. The best iPLS and PLS-GA models provided quantification of each component with an error around 5% (see Table 3). If the accuracy of quantification of a particular component was improved individually in PHARM-TM the concentration estimation of the other two components tended to deteriorate. Comparing iPLS and GA-based algorithm we found no significant differences in the best possible accuracy. It always depends on the problem at hand to decide which model is required to work with (*e.g.* if mapping one particular component was of interest, or the spatial distribution of all ingredients are of equal importance). In the present study, those models were considered as best, which served the optimum for all three components.

3.4 Methods for handling nonlinearity

In the residual plots of CRYST-PM, nonlinearity was still observed (Fig. 7.a-c), hence regression methods for handling nonlinearity were required to achieve even better fit. pPLS models were built by testing different degrees of a polynomial function of PLS scores. A model using the second power caused the lowest RMSECV, (Fig. 8) thus, this model was considered as optimum. The previously observed parabolic -nature of residuals (for linear

models) were completely eliminated (Fig. 7.d). When building a PLS model with higher powers of scores the RMSECV increased to such an extreme extent that suggested the model to be overfitted.

Locally weighted regression was applied in the principal component space (number of LVs set to two) with a tricubic weight function. In the optimization step the number of adjacent points had to be defined: *i.e.* how many consecutive observations should be considered in a local model? This number was appointed from 2 to 36 (the total number of calibration points) for LWR models to search the optimum. The lowest detected RMSECV in the regression model were found when approximately the half of the calibration points was set as adjacent as Fig. 9 shows. At 4 adjacent points another local minimum can be found. It is not unusual as three observations were applied per concentration level. The local models were improved significantly considering the fourth point. Fig. 9 also depicts that even if all points are included (as adjacent), lower error can be achieved by the LWR model than by the conventional linear PLS model. Comparing to the PLS-GA the performance of pPLS is similar in Table 2. In this case variable selection eliminated the nonlinearity almost perfectly, thus the application of nonlinear methods is not necessarily required. However, to find the best calibration model, it can be useful to test which method handle the nonlinearity better and provides better quantification.

Support vector regression (SVR) was carried out by transforming the data into a higher dimension space by a Gaussian kernel function. As SVR models can be easily overfitted, careful optimization was required. In the first step, the algorithm was allowed to optimize the three parameters (γ, C, ε) to find the minimum of RMSECV. This optimization determined 33 support vectors (from the total number of 36), which indicates a "very curly" regression function and is thus most likely overfitted. Changing the value of ε, which in fact only slightly influenced the RMSECV, allowed us to reduce the number of support vectors to 7. As a result

of this manual optimization, a γ-C-ε combination was found that it obtained a low RMSECV by keeping the number of support vectors low (and thus the regression function smooth and reliable, corresponding to the principle of parsimony) and could be used for prediction (RMSEP was 1.13%) at the same time.

In the course of quantification of PHARM-TM, different degree of nonlinearity was observed (Fig. 4) in the residual plots of the different components. This made the application of the methods for handling nonlinearity more complicated. The nonlinear nature was the most significant in the case of imipramine, which is the consequence of the main factor that the spectrum of the API was the most biased by the normalization (as previously mentioned). During the application of methods for reducing nonlinearity the task is to find a function that characterizes the nonlinearity of each component at the same time. As LWR and SVR cannot be interpreted with more than one dependent variable at the same time (meaning that a separate model would have to be built for each component) these methods could not be used in that manner as the reference methods. For these reasons in the case of PHARM-TM, only pPLS is discussed in this study and is compared to the other methods.

Second power found to be optimal when pPLS with 3 latent variables and different degrees of polynomial functions were tested. This resulted in reduction of the nonlinearity and the errors of API determination decreased by one percent (RMSEP 2.93). However, this model improved the quantification of excipients only slightly (keeping the errors of determination around 6 %) when nonlinearity was not significant.

3.5 Comparison of the models

The aforementioned comparisons after the model building process were based on the root mean square errors and on the coefficients of determination. A more advanced approach to compare the models is provided by the Sum of ranking differences (SRD) method [51,52]. In

this study the actual concentrations as well the means of concentrations estimated by the different models were selected as references. In this way, SRD proved to be able to indicate sensitively if a model was unable to discriminate correctly among the adjacent concentration levels due to inaccurate estimation. Hence the method was especially useful for the evaluation of the CRYST-PM's models. SRD was not applied on the PHARM-TM results as the locations of investigated concentrations were so different from each other, that the use of SRD for the concentration values made no sense.

Fig. 10 shows the comparison of the sums of rankings. The SRD method provided a ranking of the models that indicated the overall power of the models. Fig. 10 clearly shows that the reference methods, including PCR and PLS, performed the worst in this context (these methods were the least able to distinguish between adjacent concentration levels). SRD evaluated the iPLS and pPLS models as much better models having low SRD values (independently of the used reference). The univariate method led to moderately low SRD exceeding the efficacy of some multivariate methods in this context. Consequently, the intensity of the peak of 753 cm^{-1} increases by raising Form I concentration but in highly nonlinear way. The result of SRD was validated by comparing the SRD values with Gaussian random ranking and leave-many-out cross validation was performed as well. (see supplementary material)

3.6 Models applied on Raman-maps

The most accurate models according to the different comparisons were further tested on such Raman maps which were not used for either model building or validation. During the model building process optimization was performed by the minimization of RMSECV. However, having been aware of the fact that RMSECV alone is not suitable to identify overfitting, extreme values of parameters was not accepted even if it could have result in lower

RMSECV. After the model construction, the relation of RMSECV and RMSEP values was monitored to check if the overfitting was avoided successfully.

Quantifying new samples with known concentrations allowed us to investigate the true efficacy of a built model and confirm the successful elimination of overfitting. For this reason a Raman map with 101×101 pixels (5×5 mm area) from a sample with 5% carvedilol Form I and 95% Form II content was investigated. In the case of the PHARM-TM system, a sample containing 50% maize starch, 25% imipramine and 25% MCC were Raman mapped in 49×49 data points (2.4×2.4 mm area).

Prediction of the 5% overall carvedilol Form I content improved compared to the conventionally used CLS model by more advanced models. Concentration maps estimated by a reference and an advanced method were emphasized and visualized in Fig. 11. Fig. 11.a shows that CLS predicted 7.06 %, while LWR provided 5.26% according to Fig. 11.b. Largest improvement was observed in latter case. iPLS model provided 5.45%. Although PLS-GA model achieved accurate quantification according to the RMSEs it overestimated the overall Form I content more than PLS did on the large test map (5.87% with PLS-GA vs 5.76% with PLS) suggesting slight overfitting.

During the analysis of the PHARM-TM it was noticed that considerable differences appeared in the concentration maps of MCC and imipramine, while the prediction for maize starch was not affected strongly by the models. The models with variable selection predicted the actual concentration of imipramine outstandingly well but these were less accurate with the prediction of the excipients.

The Raman maps shown in Fig. 11 represent, owing to the applied quantitative chemometric methods, real concentrations at each point instead of the so called “spectral concentrations” generally used in publications.

4. Conclusions

Quantification of two-component polymorph mixtures and three-component blends of API and excipients was carried out by Raman chemical imaging combined with multivariate calibration methods, some of which applied the first time in this field. This work demonstrates how accurate quantitative determination can be achieved using regression for handling nonlinearity and variable selection procedures compared to the widely used CLS, PCR and PLS methods.

With the aid of the proposed method Raman mapping can be used for determining the real quantitative composition of pharmaceutical samples. Furthermore, it enables us to analyze extreme low concentrations providing a powerful tool for contaminant analysis of drug polymorphs.

It was found that a variable selection model or a model for eliminating nonlinearity improves significantly the analysis of a particular component having the most nonlinear features. Nevertheless, the quantification of other components can be improved with these models simultaneously.

The quantitative models shown in this study can become an integral part of a continuous pharmaceutical manufacturing process such as controlled crystallization or a content uniformity test after tableting, in line with the concept of PAT. For this purpose further advanced spectroscopic techniques are needed to be implemented performing truly real-time control. For example, transmission Raman spectroscopy is a promising way for shortening exposure times, although spatial information is lost in that case. Alternatively, novel techniques able to replace the lengthy scanning procedure by simultaneous measurement of all points can be developed further [57-59].

Acknowledgements

The work was financially supported by the Hungarian projects: OTKA PD 108975, OTKA K112644 and New Széchenyi Development Plan (TÁMOP-4.2.1/B-09/1/KMR-2010-0002).

References

[1] L. Zhang, M.J. Henson, S.S. Sekulic, Multivariate data analysis for Raman imaging of a model pharmaceutical tablet, Anal. Chim. Acta 545 (2005) 262-278.

[2] A. Whitley, F. Adar, Confocal spectral imaging in tissue with contrast provided by raman vibrational signatures, Cytometry Part A 69A (2006) 880-887.

[3] S.Y. Lin, M.J. Li, Cheng, W.-T., FT-IR and Raman vibrational microspectroscopies used for spectral biodiagnosis of human tissues, Spectroscopy 21 (2007), 1-30.

[4] P.Y. Sacré, E. Deconinck, L. Saerens, T. De Beer, P. Courselle, R. Vancauwenberghe, P. Chiap, J. Crommen, J.O. De Beer, Detection of counterfeit Viagra® by Raman microspectroscopy imaging and multivariate analysis, J. Pharm. Biomed. Anal. 56 (2011), 454-461.

[5] G.P. Sabin, V.A. Lozano, W.F.C. Rocha, W. Romão, R.S. Ortiz, R.J. Poppi, Characterization of sildenafil citrate tablets of different sources by near infrared chemical imaging and chemometric tools. J. Pharm. Biomed. Anal. 85 (2013) 207-212.

[6] A.A. Gowen, C.P. O'Donnell, M. Taghizadeh, P.J. Cullen, J.M. Frias, G. Downey, Hyperspectral imaging combined with principal component analysis for bruise damage detection on white mushrooms (Agaricus bisporus), J. Chemom. 22 (2008) 259-267.

[7] A.A. Gowen, Y. Feng, E. Gaston, V. Valdramidis, Recent applications of hyperspectral imaging in microbiology, Talanta 137 (2015) 43-54.

[8] B. Vajna, B. Bodzay, A. Toldy, I. Farkas, T. Igricz, G. Marosi, Analysis of car shredder polymer waste with Raman mapping and chemometrics, Express Polym. Lett. 6 (2012), 107-119.

[9] P. Ropret, C. Miliani, S.A. Centeno, Č. Tavzes, F. Rosi, Advances in Raman mapping of works of art, J. Raman Spectosc. 41 (2010) 1462-1467.

[10] C. Gendrin, Y. Roggo, C. Collet, Pharmaceutical applications of vibrational chemical imaging and chemometrics: A review, J. Pharm. Biomed. Anal. 48 (2008), 533-553.

[11] G.P.S. Smith, C.M. McGoverin, S.J. Fraser, K.C. Gordon, Raman imaging of drug delivery systems, Adv. Drug Deliv. Rev. 89 (2015) 21-41.

[12] J. Rantanen, Process analytical applications of Raman spectroscopy, J. Pharm. Pharmacol. 59 (2007), 171-177.

[13] M. Khorasani, J.M. Amigo, J. Sonnergaard, P. Olsen, P. Bertelsen, J. Rantanen, Visualization and prediction of porosity in roller compacted ribbons with near-infrared chemical imaging (NIR-CI), J. Pharm. Biomed. Anal. 109 (2015), 11-17.

[14] B. Vajna, H. Pataki, Z. Nagy, I. Farkas, G. Marosi, Characterization of melt extruded and conventional Isoptin formulations using Raman chemical imaging and chemometrics, Int. J. Pharm. 419 (2011) 107-113.

[15] S. Piqueras, L. Duponchel, R. Tauler, A. de Juan, Monitoring polymorphic transformations by using in situ Raman hyperspectral imaging and image multiset analysis, Anal. Chim. Acta 819 (2014) 15-25.

[16] P.F. Chavez, P. Lebrun, P.Y. Sacré, C. De Bleye, L. Netchacovitch, S. Cuypers, J. Mantanus, H. Motte, M. Schubert, B. Evrard, P. Hubert, E. Ziemons, Optimization of a

pharmaceutical tablet formulation based on a design space approach and using vibrational spectroscopy as PAT tool, Int. J. Pharm. 486 (2015) 13-20.

[17] E. Widjaja, P. Kanaujia, G. Lau, W.K. Ng, M. Garland, C. Saal, A. Hanefeld, M. Fischbach, M. Maio, R.B.H. Tan, Detection of trace crystallinity in an amorphous system using Raman microscopy and chemometric analysis, Eur. J. Pharm. Sci. 42 (2011) 45-54.

[18] H. Ueda, Y. Ida, K. Kadota, Y. Tozuka, Raman mapping for kinetic analysis of crystallization of amorphous drug based on distributional images, Int. J. Pharm. 462 (2014) 115-122.

[19] Food and Drug Administration, Guidance for Industry: PAT – A Framework for Innovative Pharmaceutical Development, Manufacturing and Quality Assurance, Washington, D.C., U.S.A, 2004.

[20] M. Fonteyne, J. Vercruysse, F. De Leersnyder, B. Van Snick, C. Vervaet, C., J.P. Remon, T. De Beer, Process Analytical Technology for continuous manufacturing of solid-dosage forms, TrAC-Trends Anal. Chem. 67 (2015) 159-166.

[21] J. Rantanen, J. Khinast, The future of pharmaceutical manufacturing sciences, J. Pharm. Sci. 104 (2015) 3612-3638.

[22] K. Knop, P. Kleinebudde, PAT-tools for process control in pharmaceutical film coating applications, Int. J. Pharm. 457 (2013) 527-536.

[23] K. Nikowitz, F. Folttmann, M. Wirges, K. Knop, K. Pintye-Hódi, G. Regdon, P. Kleinebudde, Development of a Raman method to follow the evolution of coating thickness of pellets, Drug Dev. Ind. Pharm. 40 (2014) 1005-1010.

[24] K.C. Gordon, C.M. McGoverin, Raman mapping of pharmaceuticals, Int. J. Pharm. 417 (2011) 151-162.

[25] P.Y. Sacré, C. De Bleye, P.F. Chavez, L. Netchacovitch, P. Hubert, E. Ziemons, Data processing of vibrational chemical imaging for pharmaceutical applications, J. Pharm. Biomed. Anal. 101 (2014) 123-140.

[26] N. Kumar, A. Bansal, G.S. Sarma, R.K. Rawal, Chemometrics tools used in analytical chemistry: An overview, Talanta 123 (2014) 186-199.

[27] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta 667 (2010) 14-32.

[28] R. Kramer, Chemometric techniques for quantitative analysis, Marcel Dekker, New York, 1998.

[29] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial. Anal. Chim. Acta 185 (1986) 1-17.

[30] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Int. Lab. Syst. 58 (2001) 109-130.

[31] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy, Appl. Spectrosc. 54 (2000) 413-419.

[32] R. Leardi, L. Nørgaard, Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, J. Chemom. 18 (2004) 486-497.

[33] C.B. Lucasius, M.L.M. Beckers, G. Kateman, Genetic algorithms in wavelength selection: a comparative study, Anal. Chim. Acta 286 (1994) 135-153.

[34] R. Leardi, A. Lupiáñez González, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemom. Int. Lab. Syst. 41 (1998) 195-207.

[35] R. Leardi, Genetic algorithms in chemometrics and chemistry: a review, J. Chemom. 15 (2001) 559-569.

[36] I.E. Frank, A nonlinear PLS model, Chemom. Int. Lab. Syst. 8 (1990) 109-119.

[37] R.M. Balabin, R.Z. Safieva, E.I. Lomakina, Comparison of linear and nonlinear calibration models based on near infrared (NIR) spectroscopy data for gasoline properties prediction, Chemom. Int. Lab. Syst. 88 (2007) 183-188.

[38] W.S. Cleveland, S.J. Devlin, Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting, J. Amer. Stat. Assoc. 83 (1988) 596-610.

[39] T. Naes, T. Isaksson, B. Kowalski, Locally weighted regression and scatter correction for near-infrared reflectance data, Anal. Chem. 1990, 62 (7), 664-673.

[40] V. Vapnik, Statistical Learning Theory. John Wiley and Sons, New York, 1988

[41] A. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (2004) 199-222.

[42] A.A. Gowen, C.P. O'Donnell, P.J. Cullen, S.E.J. Bell, Recent applications of Chemical Imaging to pharmaceutical process monitoring and quality control, Eur. J. Pharm. Biopharm. 69 (2008) 10-22.

[43] U. Thissen, M. Pepers, B. Üstün, W.J. Melssen, L.M.C. Buydens, Comparing support vector machines to PLS for spectral regression applications, Chemom. Int. Lab. Syst. 73 (2004) 169-179.

[44] I.A. Naguib, E.A. Abdelaleem, M.E. Draz, H.E. Zaazaa, Linear support vector regression and partial least squares chemometric models for determination of Hydrochlorothiazide and Benazepril hydrochloride in presence of related impurities: A comparative study, Spectrochim. Acta, Part A 130 (2014) 350-356.

[45] L. Sratthaphut, N. Ruangwises, Genetic Algorithms-Based Approach for Wavelength Selection in Spectrophotometric Determination of Vitamin B12 in Pharmaceutical Tablets by Partial Least-Squares, Procedia Eng. 32 (2012) 225-231.

[46] F.L.F. Soares, R.L. Carneiro, Evaluation of analytical tools and multivariate methods for quantification of co-former crystals in ibuprofen-nicotinamide co-crystals, J. Pharm. Biomed. Anal. 89 (2014) 166-175.

[47] D. Pérez-Marín, A. Garrido-Varo, J.E. Guerrero, Non-linear regression methods in NIRS quantitative analysis, Talanta 72 (2007) 28-42.

[48] H. Pataki, I. Markovits, B. Vajna, Z. K. Nagy, G. Marosi, In-Line Monitoring of Carvedilol Crystallization Using Raman Spectroscopy, Cryst. Growth Des. 12 (2012) 5621-5628.

[49] J. G. Rosas, Marcelo Blanco, A criterion for assessing homogeneity distribution in hyperspectral images. Part 2: Application of homogeneity indices to solid pharmaceutical dosage forms, J. Pharm. Biomed. Anal. 70 (2012) 691-699.

[50] J.O. Rawlings, S.G. Pantula, D.A. Dickey, Applied Regression Analysis: A Research Tool, second edn., Springer, New York, 1998

[51] K. Héberger, Sum of ranking differences compares methods or models fairly, TrAC-Trends Anal. Chem. 29 (2010) 101-109.

[52] K. Kollár-Hunek, K. Héberger, Method and model comparison by sum of ranking differences in cases of repeated observations (ties), Chemom. Int. Lab. Syst. 127 (2013) 139-146.

[53] F. Despagne, D.L. Massart, Development of a Robust Calibration Model for Nonlinear In-Line Process Data, Anal. Chem. 72 (2000) 1657-1665.

[54] J. Peng, L. Li, Support vector regression in sum space for multivariate calibration, Chemom. Int. Lab. Syst. 130 (2014) 14-19.

[55] P.H.C. Eilers, A Perfect Smoother, Anal. Chem. 75 (2003) 3631-3636.

[56] A. Farkas, B. Vajna, P.L. Sóti, Z.K. Nagy, H. Pataki, H., F. Van der Gucht, G. Marosi, Comparison of multivariate linear regression methods in micro-Raman spectrometric quantitative characterization, J. Raman Spectrosc. 46 (2015) 566-576.

[57] N.L. Schwindt, Hyper Spectral Imaging Using Fiber Optic Spatial Decomposition for Raman Spectroscopy, Master's Thesis, Colorado School of Mines, Division of Engineering, Golden, CO, USA, 2005.

[58] M. Okuno, H.O. Hamaguchi, Multifocus confocal Raman microspectroscopy for fast multimode vibrational imaging of living cells, Opt. Lett. 35 (2010) 4096-4098.

[59] E. Schmälzlin, B. Moralejo, M. Rutowska, A. Monreal-Ibero, C. Sandin, N. Tarcea, J. Popp, M. Roth, Raman Imaging with a Fiber-Coupled Multichannel Spectrograph, Sensors 14 (2014) 21968-21980.

Table 1 CRYST-PM mass ratios and composition of calibration and validation sets

Sample	Form I	Form II	type of set	Sample	Form I	Form II	type of set
CAR_1	0%	100%	cal.	**CAR_11**	70%	30%	cal.
CAR_2	1%	99%	cal.	**CAR_12**	85%	15%	cal.
CAR_3	2%	98%	val.	**CAR_13**	89%	11%	val.
CAR_4	3%	97%	cal.	**CAR_14**	92%	8%	cal.
CAR_5	5%	95%	val.	**CAR_15**	95%	5%	val.
CAR_6	8%	92%	cal.	**CAR_16**	97%	3%	cal.
CAR_7	11%	89%	val.	**CAR_17**	98%	2%	val.
CAR_8	15%	85%	cal.	**CAR_18**	99%	1%	cal.
CAR_9	30%	70%	cal.	**CAR_19**	100%	0%	cal.
CAR_10	50%	50%	val.				

Table 2: Performance characteristics of quantitative models of CRYST-PM (bc: baseline correction. nm: normalization. mncn: mean centering. SNV: standard normal variate)

	Univ.	**CLS**	**PCR**	**PLS**	**iPLS**	**PLS-GA**	**pPLS**	**LWR**	**SVR**
Preprocessing	bc. nm	bc. nm	bc. nm. mncn. SNV				bc. nm. mncn	bc. nm	bc.nm. mncn. SNV
RMSEC	3.20	3.48	1.82	1.68	1.14	1.01	1.06	0.89	1.16
RMSECV	-	-	4.00	2.16	1.30	1.18	1.67	1.74	1.54
RMSEP	3.16	3.63	1.88	1.91	1.20	1.19	1.24	1.38	1.13
$\mathbf{R^2_{cal}}$	0.9941	0.9931	0.9981	0.9984	0.9993	0.9994	0.9994	0.9996	0.9992
$\mathbf{R^2_{cv}}$	-	-	0.9912	0.9973	0.9990	0.9990	0.9985	0.9984	0.9987
$\mathbf{R^2_{pred}}$	0.994	0.9921	0.9984	0.9983	0.9994	0.9994	0.9993	0.9992	0.9994

Table 3: Performance characteristics of quantitative models of PHARM-TM

		Univ.	CLS	PCR	PLS	iPLS	PLS-GA	pPLS
RMSEC	**Starch**	9.36	6.82	6.03	6.09	4.87	5.62	6.00
	Cellulose	60.02	5.76	5.83	5.47	4.67	3.00	3.73
	API	11.68	10.04	6.33	5.67	3.46	3.95	5.97
RMSECV	**Starch**	-	8.50	7.77	8.10	6.95	7.69	8.60
	Cellulose	-	7.11	8.01	7.69	6.54	4.79	6.47
	API	-	12.60	8.71	8.37	5.61	7.39	7.67
RMSEP	**Starch**	7.38	6.17	6.48	6.53	5.28	5.37	6.01
	Cellulose	18.98	6.71	6.20	5.39	3.95	4.23	6.34
	API	15.10	12.51	4.28	3.90	4.37	2.50	2.93
R^2_{cal}	**Starch**	0.9275	0.9601	0.9674	0.9669	0.9788	0.9718	0.9670
	Cellulose	0.2871	0.9712	0.9696	0.9733	0.9805	0.9919	0.9875
	API	0.8908	0.9174	0.9641	0.9712	0.9893	0.9861	0.9681
R^2_{cv}	**Starch**	-	0.9406	0.9463	0.9417	0.9570	0.9473	0.933
	Cellulose	-	0.9573	0.9431	0.9477	0.9617	0.9795	0.962
	API	-	0.8753	0.9322	0.9374	0.9720	0.9512	0.9473
R^2_{pred}	**Starch**	0.8178	0.8934	0.8690	0.8648	0.9319	0.8764	0.882
	Cellulose	0.3646	0.6952	0.6898	0.7421	0.8601	0.8434	0.713
	API	0.9509	0.9000	0.9000	0.9438	0.9602	0.9654	0.9271

Figure captions

Fig. 1 Ternary diagram of the mixtures of PHARM-TM (filled circles indicate validation samples)

Fig. 2 Raman spectra of the two kinds of polymorphs of carvedilol

Fig. 3 Raman spectra of imipramine active ingredients and the two kinds of excipients (microcrystalline cellulose /MCC/ and maize starch)

Fig. 4 Residuals in CLS modelling of PHARM-TM for a) maize starch, b) microcrystalline cellulose and c) imipramine

Fig. 5 Model errors achieved with different variable window sizes (number of adjacent variables which is treated together) in a) genetic algorithm b) interval PLS for the data set of CRYST-PM

Fig. 6 Monitoring plots of a PLS-GA run (CRYST-PM): a) information about the fitness and the used window number in the single models, b) converging the values of the best (brownish line) and average fitness of the constructed models (blue line) along the generations, c) the average window numbers in models, d) the frequencies of each window in the final models

Fig. 7 Trends of residuals for calibration and validation points in CRYST-PM using a) CLS, b) PLS, c) iPLS and d) pPLS chemometric methods (arbitrarily fitted red lines demonstrate nonlinearity)

Fig. 8 Dependence of root mean square errors of calibration, cross-validation and prediction on the polynomial degree of scores in pPLS model of CRYST-PM

Fig. 9 Change of root mean square errors in LWR model depending on the number of local points considered

Fig. 10 Percentages of SRD for best models of the examined methods in the case of CRYST-PM applying actual concentrations as reference (0% indicates the same ranking as the reference ranking; 100% belongs to the fully opposite ranking

Fig. 11 Distribution maps of carvedilol mixture containing 5% of Form I quantified by a) CLS (7.06%) and b) LWR(5.26%).

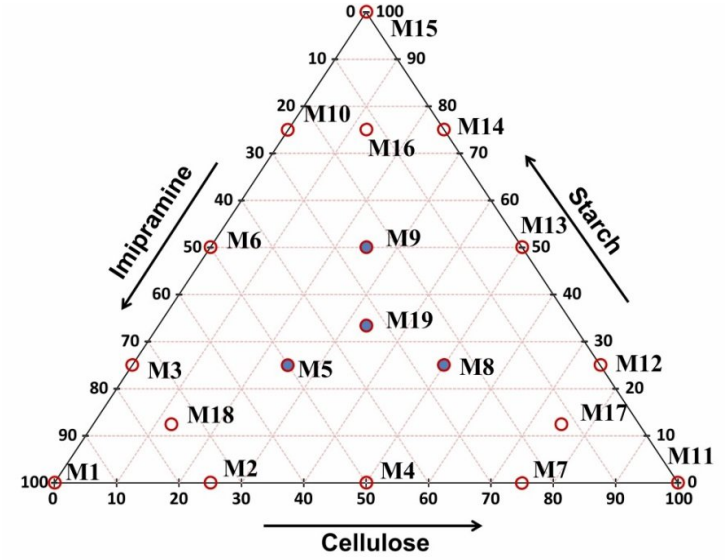


Figure 1

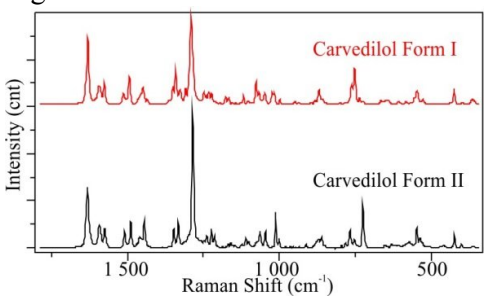


Figure 2

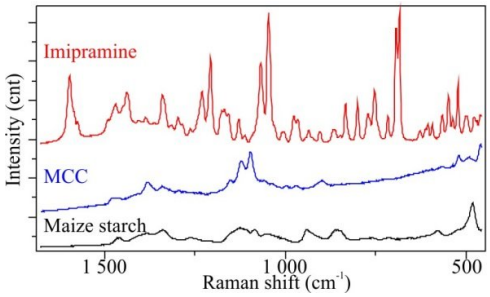


Figure 3

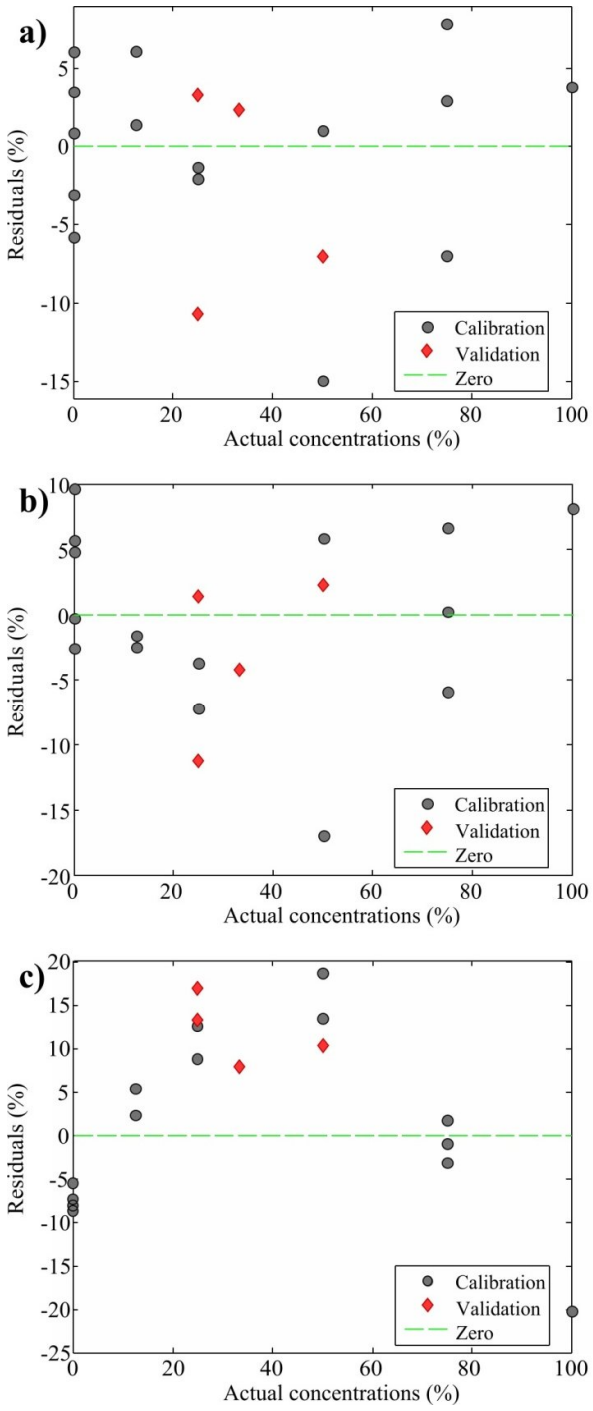


Figure 4

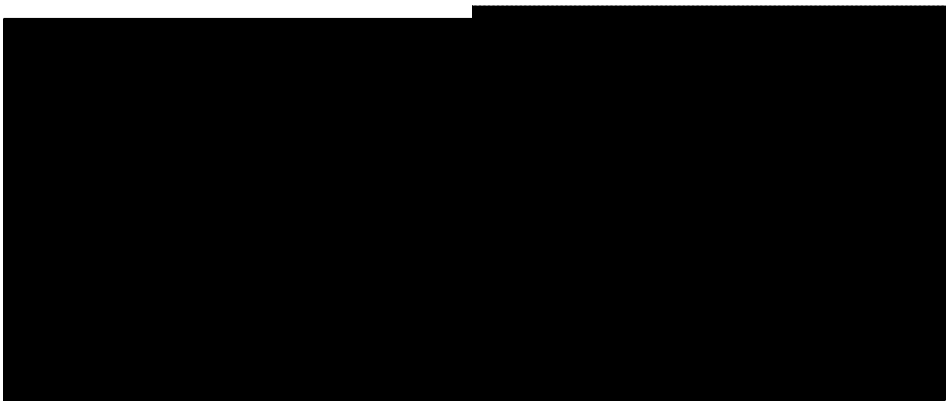


Figure 5

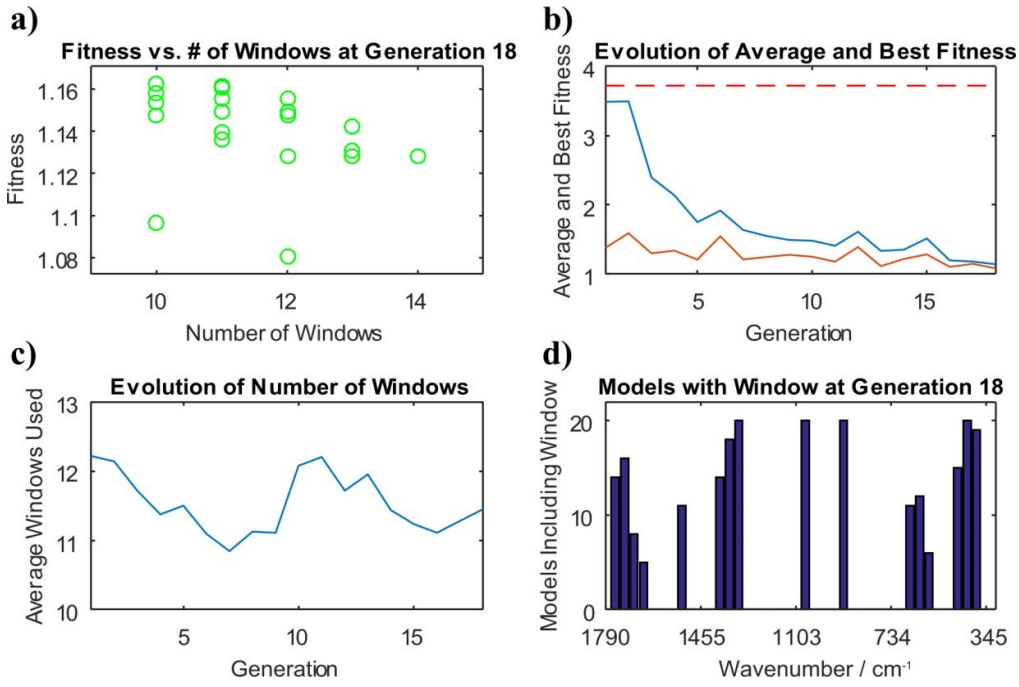


Figure 6

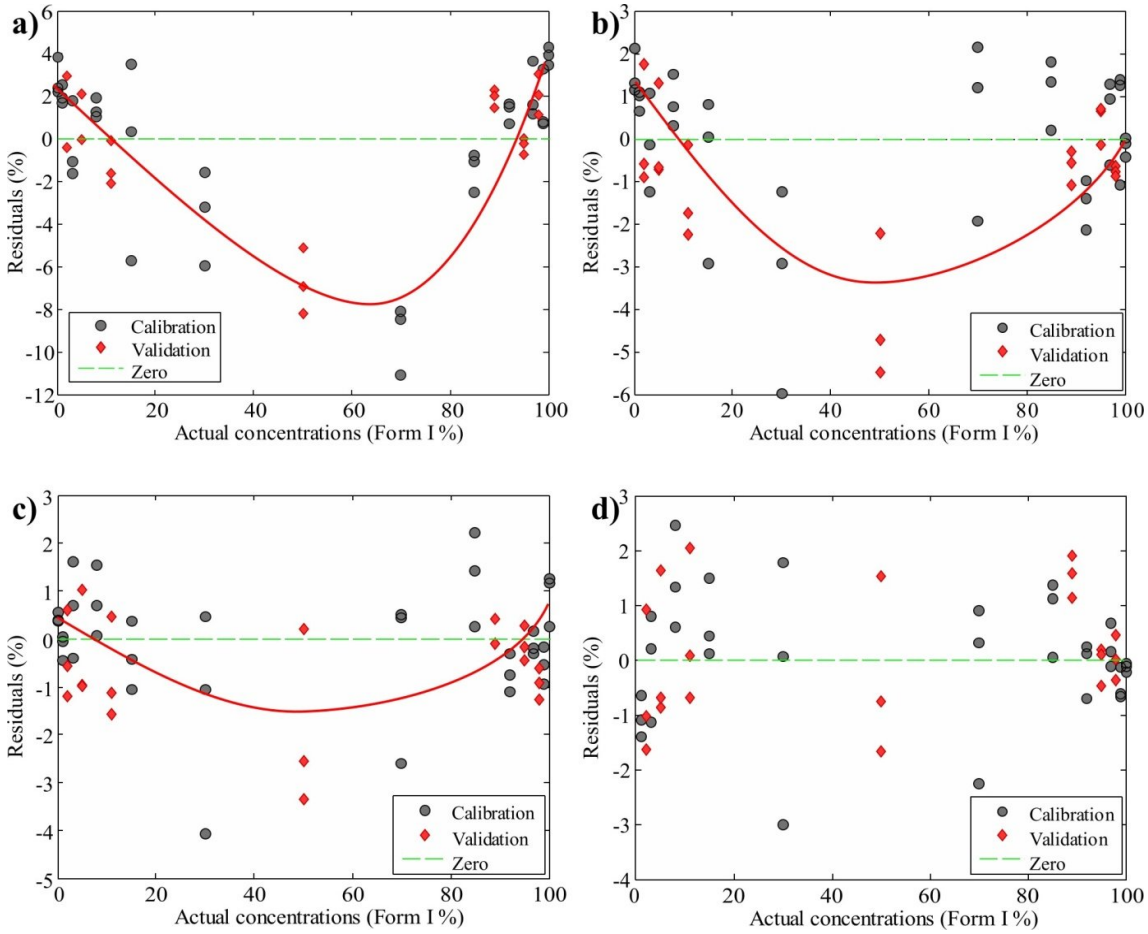


Figure 7

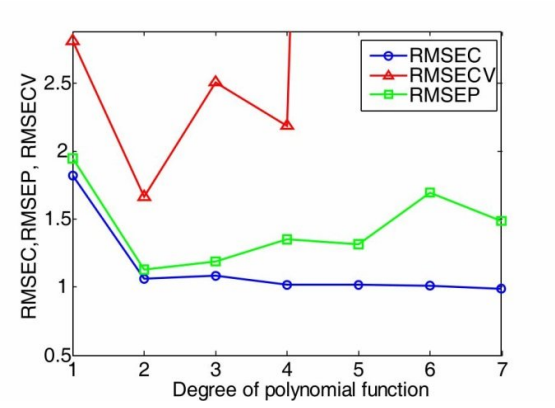


Figure 8

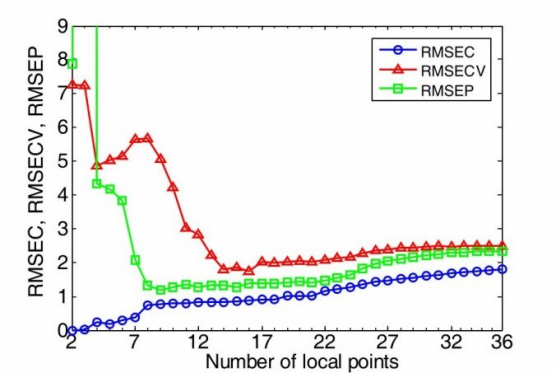


Figure 9

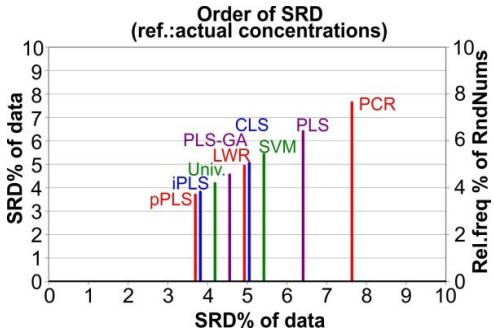


Figure 10

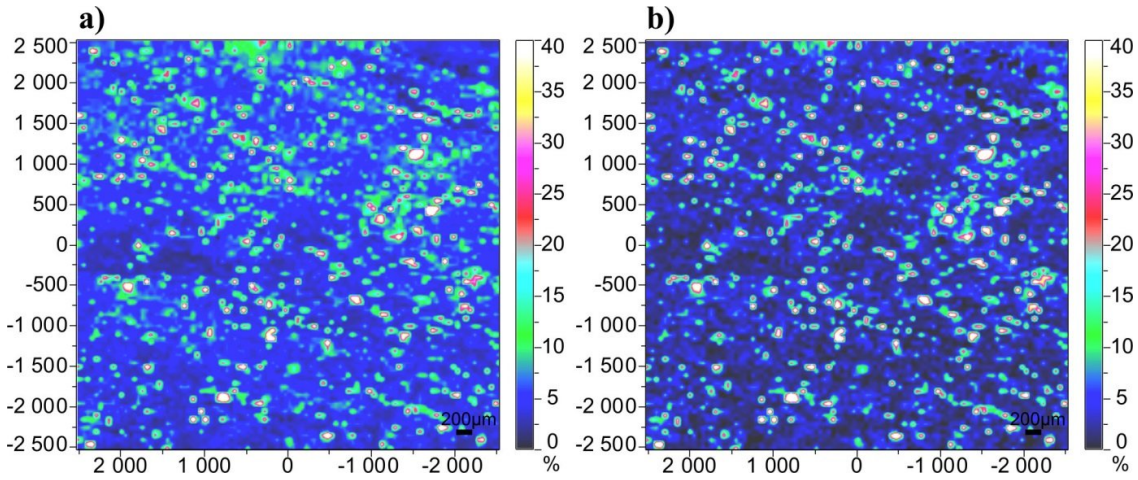


Figure 11