BUDAPEST UNIVERSITY OF TECHNOLOGY AND ECONOMICS
FACULTY OF CHEMICAL TECHNOLOGY AND BIOTECHNOLOGY

Multivariate curve resolution and regression methods in Raman chemical imaging

Theses of Ph.D. dissertation

Submitted by: Balázs Vajna

Supervisor: Prof. Dr. György Marosi

Department of Organic Chemistry and Technology

2012

Acronyms and abbreviations

API	active pharmaceutical ingredient
CLS	classical least squares, a way of modelling mixture spectra using the reference spectra of the pure components
GA	genetic algorithms (variable selection method)
iPLS	interval PLS (variable selection method, definition of PLS provided below)
MAF	maximum autocorrelation factors
MCR-ALS	multivariate curve resolution – alternating least squares
PCA	principal component analysis
PLS	partial least squares (regression)
PMF	positive matrix factorization, curve resolution method
$R^2_{(C,CV,P)}$	coefficient of determination (for calibration, cross-validation and prediction)
RMSE(C,CV,P)	root mean square error (of calibration, cross-validation and prediction)
SEM-EDS	scanning electron microscopy – energy-dispersive X-ray spectroscopy
SISAL	simplex identification via slplit augmented Lagrangian
SMCR	self-modelling curve resolution, (family of chemometric methods used for curve resolution)
SMMA	self-modelling mixture analysis
SRD	sum of ranking differences, mathematical method to compare objects (models, performance parameters, analytical methods etc.)
SS2D	sample-sample 2D correlation spectroscopy
SVM	support vector machines, nonlinear machine learning / regression algorithm
XRPD	X-ray powder diffraction

1. Introduction

Chemical imaging is one of the most rapidly developing areas of spectroscopy, combining a vibrational spectroscopic (mid- or near infrared, or Raman) technique with appropriate optics. This, in the case of *point mapping*, is a microscope, which allows the acquisition of spectra from a point with the diameter of 0.5-5 microns, depending on the applied magnification. Point mapping is carried out by the sequential acquisition of multiple spectra according to a pre-defined grid, moving the microscope stage after each spectrum collection step.

Each pixel of a chemical image contains the spectrum recorded from the corresponding spot on the sample surface. Since the number of pixels, as well as the number of wavenumber channels is huge, the vast amount of data contained in a chemical image cannot be properly processed using traditional evaluation methods. A comprehensive analysis of such images requires the use of *chemometric* (multivariate data analysis) methods. Such evaluation of chemical images is a rather unexplored area, hence this was selected as the primary topic of the present PhD dissertation. Investigations were focused on issues in pharmaceutical and polymer technologies, however, results are not constrained to these areas and the methods proposed here may be just as well applied in other fields.

The aim of this dissertation is to characterize various chemometric methods in the evaluation of Raman maps and thereby enhance the applicability of chemical imaging in the following areas:

(1) ***In-depth analysis of pharmaceutical formulations under development.*** The underlying motivation of such analyses is provided by the recently introduced *Quality by Design* and *Process Analytical Technology*, approaches recommended by the US Food and Drug Administration. Their aim is to enable accurate prediction of final product quality by

implementing appropriate analytical methods and data evaluation. In this framework, the focus of this study was to determine how chemical imaging, combined with chemometrics, can facilitate the exploration of the structure of samples and their correspondence with various physico-chemical and biological properties.

(2) ***Non-destructive analysis of illegal and counterfeit drugs.*** The main challenge in this case lies in the fact that usually no prior data is available about the samples to be characterized. Therefore the qualitative and quantitative composition, as well as the structural properties, have to be determined/estimated using the Raman map itself, without using any external source of information.

(3) ***Qualitative and quantitative characterization of polymer waste.*** The two main factors rendering the evaluation of Raman maps acquired from waste samples difficult are the immensely bad signal-to-noise ratio and the unknown chemical composition. Instead of the only applicable traditional approach of visually inspecting each pixel, which would take hours of labour for each Raman map, an automated method based on existing chemometric algorithms was to be developed.

2. State-of-the-art

The main advantage of chemical imaging is the possibility of local analysis and exploration of the spatial distribution of the components. The distribution and homogeneity of some components have a strong influence on certain physico-chemical attributes. Therefore extracting the concentration maps helps draw such conclusions which would otherwise remain unknown with the use of macroscopic (bulk) spectrometric methods[1,2].

The most straightforward method to visualize the spatial distribution of a component is plotting the integrated intensity of a characteristic vibrational band against the spatial coordinates of the pixels. The main drawback of this approach is that it can be only applied when a truly selective band exists for the component in question. Despite its disadvantages, this is the most widely used method.

Classical least squares (CLS) modelling can be applied when the vibrational bands of different components overlap with one another. In this case, the (mixture) spectra found in each pixel are modelled as the linear combination (the weighted average) of the pure component spectra. When the reference spectra are known, these weight factors can be easily computed[3] and can be used for the estimation of the concentrations. This method is also very widely used.

Numerous issues in the pharmaceutical technology can be resolved by the evaluation of chemical images using these two traditional methods. Applications range across the entire technology chain, from the initial powder blending, through process troubleshooting, up to the high throughput screening of the final products[1,2,4]. Several macroscopic properties can be assessed this way regarding the formulations and their manufacturing technology. No study has, however, systematically investigated and compared the spatial distribution of components within tablets manufactured by different technologies.

Raman mapping (among the chemical imaging methods) is particularly effective in the analysis of solid dispersions and polymorphic mixtures. Analyses of solid dispersions

[1] C. Gendrin, Y. Roggo, C. Collet, J. Pharm. Biomed. Anal. 48 (2008) 533-553.
[2] A.A. Gowen, C.P. O'Donnell, P.J. Cullen, S.E.J. Bell, Eur. J. Pharm. Biopharm. 69 (2008) 10-22.
[3] H. Mark, J. Workman: Chemometrics in Spectroscopy, first ed., Academic Press, 2007.
[4] K.C. Gordon, C.M. McGoverin, Int. J. Pharm. 417 (2011) 151-162.

manufactured by traditional preparation methods have already been reported, however, **there is no study dealing with such products prepared by supercritical extrusion and electrospinning**, two of the **most recently implemented technologies in the pharmaceutical technology**. Besides, even though the homogeneity of these dispersions can be theoretically related to the dissolution properties, an **actual comparison of the homogeneity, determined by Raman mapping, and the true dissolution properties cannot be found in the literature**.

Solid-state analysis of the active ingredients within formulations is also usually carried out by either univariate or CLS modelling, even in the most recent studies[5]. The detection and identification of trace amounts of unexpected (and possibly unknown) polymorphs, however, require the use of multivariate methods. Only one study applying multivariate techniques exists that addressed the issue of detecting polymorphic contaminations under the detection limit of bulk macroscopic methods, which used exploratory statistics and only dealt with the identification and *qualitative* characterization[6]. **The quantitative estimation of trace crystalline contaminations was therefore still unresolved** and was part of the agenda during the present work.

An increasing problem nowadays is the growing emergence of illegal and counterfeit products on the pharmaceutical and black markets[1]. **There is often no or limited prior information about these samples** – in many cases, even the components are completely unknown. In this case, both the pure component spectra and their concentration maps have to be estimated using the spectra found in the pixels of the chemical image[2]. This can mainly be carried out using exploratory statistics, especially its subset of *self-modelling curve resolution* (SMCR) methods. The most prominent methods used in chemical imaging are *principal component analysis* or PCA (among the basic exploratory techniques) and multivariate curve resolution – alternating least squares (among the SMCR methods). The latter applies the equations of CLS on the data matrix of the Raman map to iteratively compute the pure component spectra and the concentrations, enforcing physically meaningful constraints (such as *non-negativity* of concentrations and intensities, *closure* of all concentrations in a pixel to give a sum of 1, etc.) after each iteration step.

Other methods have also been applied, such as sample-sample 2D correlation spectroscopy (SS2D) or the SMCR methods of self-modelling mixture analysis (SMMA), positive matrix factorization (PMF) and simplex identification via split augmented Lagrangian (SISAL). Nevertheless, the investigation of unknown products is still a rather unexplored territory. ***Maximum autocorrelation factors* (MAF), a method well performing in image processing and mass spectroscopic imaging, has not been used in vibrational chemical imaging.** In addition, multiple methods have only been compared under optimal circumstances (i.e. without the presence of significant noise or disturbance factors)[7]. The current literature does not offer a way to determine **which method should be generally used for the evaluation of Raman maps of unknown products (and how) – in particular, how the noise level and the homogeneity of one or more components affect the curve resolution and which method is the least sensitive to these.** The influence of the noise level (or the lack of it) affects the measurement conditions to be applied, whereas the separation of spectra of homogeneously distributed components is generally a major challenge in curve resolution.

[5] S. Šašić, S. Mehrens, Anal. Chem. 84 (2012) 1019-1025.
[6] E. Widjaja, P. Kanaujia, G. Lau, W. Kiong Ng, M. Garland, C. Saal, et al., Eur. J. Pharm. Sci. 42 (2010) 45-54.
[7] C. Gendrin, Y. Roggo, C. Collet, J. Near Infrared Spectrosc. 16 (2008) 151-157.

Another possible application of Raman mapping aided by curve resolution is the characterization of polymer waste, as this combination makes the analysis of both major and minor components possible. Even though near infrared spectrometric imaging has already been applied to classify waste items according to their main polymeric ingredient (using *classification*, or, in other words, *supervised learning* algorithms)[8], the quantitative analysis of waste materials with chemical imaging is yet to be solved. On one hand, these studies employed global imaging with a camera and investigated intact waste items, not providing representative sampling. On the other hand, these supervised classification methods require a pre-defined *training set* for building the model, which inevitably results in the fact that (mainly) only the polymers present in this training set can be later identified in the waste. However, waste materials almost always contain **various unexpected and unknown constituents, the amount of which is also important to estimate**. An added difficulty is caused by the **immensely low quality of the mapping spectra**, due to the various components responding differently to the illuminating excitation laser. **The state-of-the-art literature does not offer a method to overcome these challenges and to provide the means for the quantitative analysis of polymer wastes by Raman mapping.**

Quantitative determination is just as important in the field of pharmaceuticals, especially for the excipients which, unlike the active ingredients, are usually not determined by specific analytical techniques. Although traditional CLS modelling is feasible for the rough estimation of component concentrations (with unknown error rates, though), it is often important to provide pixel-to-pixel concentrations and their visualized distribution images with low, and known, error rates.

There is rather little experience in chemical (especially in Raman) imaging in this respect. On one hand, only partial least squares regression has been used[9] (sometimes incorrectly[10], and for Raman mapping, in only one publication[11]) from the wide range of feasible and state-of-the art methods. PLS regression itself, for instance, can be enhanced by using **variable selection methods** (such as interval-PLS or genetic algorithms), **none of which has been previously tested in chemical imaging**. Besides, even though the intensity of Raman bands depends on the concentrations in a nonlinear manner under certain circumstances, **none of the available nonlinear regression methods have been applied or compared in this field**.

The comparison of multiple methods, an important aspect of the present dissertation, can be carried out using numerous traditional performance parameters. Additionally, a new non-parametric method has also been developed recently to carry out such comparative investigations. **The novel *sum of ranking differences* (SRD) method was also widely utilized in this study to select the most appropriate chemometric method for each practical issue to be solved with chemical imaging and appropriate evaluation.** In addition to solving existing technological issues, a further aim of this work was to deeper understand the chemometric methods applied here.

3. Experimental and evaluation methods

Raman mapping, the core analytical technique applied in this study, was carried out by a LabRam type micro-Raman spectrometer (Horiba Jobin Yvon). The excitation laser, the optical magnification, the spectrum acquistion time, the area of investigated surface and the

[8] A. Kulcke, C. Gurschler, G. Spock, R. Leitner, M. Kraft, J. Near Infrared Spectrosc. 11 (2003) 71-81.
[9] C. Ravn, E. Skibsted, R. Bro, J. Pharm. Biomed. Anal. 48 (2008) 554-561.
[10] Z. Rahman, A.S. Zidan, M.A. Khan, Int. J. Pharm. 400 (2010) 49-58.
[11] K.M. Balss, F.H. Long, V. Veselov, A. Orana, E. Akerman-Revis, et al., Anal. Chem. 80 (2008) 4853-4859.

number of spectra accumulated in each pixel was determined separately for each sample to provide as much and as representative information as possible, with minimal overall measurement time. Spectra were pre-treated with piece-wise linear base-line correction to remove the fluorescent background and were normalized to unit area to eliminate the intensity deviations arising from the surface roughness and therefore from the focusing error (Horiba LabSpec 5.41). Evaluation was then carried out by using the most appropriate spectrometric or chemometric method for each area of investigation, often comparing multiple methods wherever the best option was initially unknown or ambiguous.

For polymorphic stability studies, two batches of granules containing donepezil were used, manufactured differently in the framework of an industrial collaboration. For qualitative stability prediction studies, samples containing spironolactone were prepared by melt extrusion (HAAKE Minilab). To understand the correspondence between the spatial distribution of components and certain physico-chemical properties, tablets of identical composition were prepared with different manufacturing methods including direct compression (Fette EX-1), several variations of high-shear (Pro-C-ept 4M8) and fluidized bed (Glatt GCPG-1) granulation, supercritical extrusion (SCAMEX Rheoscam) and electro-spinning. The Raman maps of these samples were evaluated using univariate (integrated intensity under selected band) and CLS modelling. X-ray powder diffraction (XRPD, PANalytical X'Pert Pro MPD) measurements in-vitro dissolution tests (Erweka DT6) were carried out to validate Raman mapping results.

Interaction between active and inactive ingredients was studied using a poorly water soluble compound (the name of which is not disclosed here to comply with the protection of industrial rights) and various types of cyclodextrins, blended with 1:1 molar ratio and processed with lyophilization (freeze drying, Leybold Lyovac GT-3). Apart from Raman mapping, the homogeneity of these samples were analyzed using scanning electron microscopy – energy-dispersive X-ray spectroscopy (SEM-EDS, JEOL JSM-6380LA) and their crystallinity by XRPD.

Modelling the investigation of unknown tablets was carried out using three sets of samples. Model comparison at different noise levels was done on the Raman maps of a commercially available product (name not disclosed due to industrial rights). For testing curve resolution methods at different homogeneity levels of the active ingredient, tablets manufactured by different technologies (already introduced above) were used as analytes. The effect of the homogeneity of multiple components was assessed on the Raman maps of the commercially available products Isoptin SR-E (manufactured by melt extrusion) and Isoptin SR (manufactured by conventional wet granulation). Chemometric methods of PCA, MAF, SS2D, SMMA, SISAL, MCR-ALS and PMF were compared, optimizing the internal parameters and physical constraints for each algorithm in each case.

Quantitative characterization of polymer wastes was developed using shredded automotive waste, separated to different density fractions by magnetic density separation[12]. The assignment of pixels to polymers was carried out by an empirical approach based on SMCR methods. Quantitative determination of pharmaceutical solid dispersions was tested on spray-dried (Pro-C-Ept MicroSpray) samples containing caffeine model drug and poly(lactic-co-glycolic acid) with 11 different caffeine concentration levels. Various univariate and multivariate regression methods were tested by comparing their performance with cross-validation on the calibration set (6 concentration levels) and an independent validation set (5 concentration levels).

[12] B. Bodzay (témavez. Gy. Marosi): PhD thesis, BUTE Department of Organic Chemistry and Technology, 2011.

Multivariate calculations were carried out by Matlab 7.6 (Mathworks) and PLS_Toolbox 6.2. software. SISAL iterations were calculated by a Matlab toolbox provided by the developers. PMF calculations were performed by the PMF2 provided by the developer. Univariate statistical calculations were carried out by Matlab, Microsoft Excel and Statistica 8.0 (Statsoft) software.

4. Results and achievements

4.1. Polymorphic stability studies of pharmaceuticals using univariate methods

One of the most prominent challenges in the current pharmaceutical technology is the preparation and stabilization of active ingredients in amorphous form. Such experiments have to be aided by selective and sensitive analytical methods.

X-ray powder diffraction, albeit being the most straightforward option, is often not appropriate to detect the beginning crystallization of a low-dosage amorphous active ingredient, when it is surrounded by high amounts of crystalline excipients. Raman mapping was tested on such a sample set and it was found that the crystallization of an amorhous active ingredient, with approximately 5% concentration in mass fractions compared to the overall sample, can be well detected with Raman mapping. This can be done under the limit of detection of bulk analytical techniques, already with univariate evaluation of maps by using the integrated area under selected peaks.

The stability of amorphous form in solid dispersions is greatly influenced by the presence of trace crystalline seeds, which ideally should be identified right after manufacturing. Extrudates prepared under different conditions were successfully compared by Raman mapping revealing the differences in the residual crystallinity below the XRPD limit of detection.

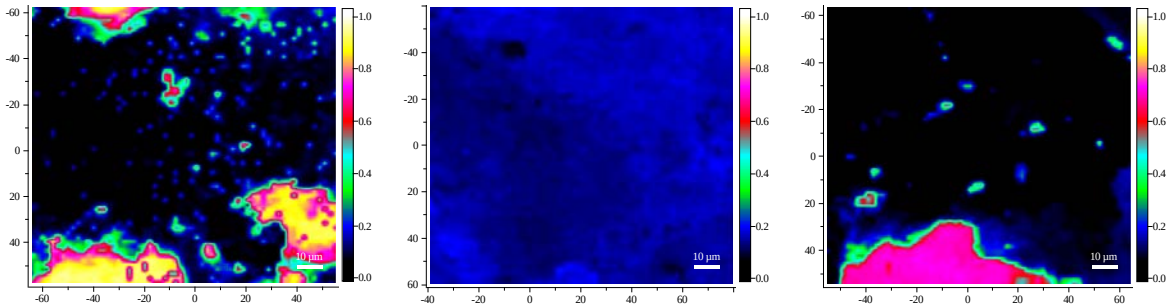
4.2. Technology evaluation using multivariate modelling with reference spectra

Real-life samples often come with the challenge that one or more of the components lack a truly selective vibrational band, hence its distribution unable to be revealed independently from other ingredients. However, it is important to extract the selective distribution maps for both the active and inactive ingredients in order to assess the dissolution and other physico-chemical properties.

The investigation of multiple differently manufactured tablets revealed distribution patterns characteristic to the preparation method (Figure 1). Since some of the pure component spectra significantly overlapped with each other, evaluation was carried out by CLS modelling using the entire spectral region. Descriptive statistics of the pixel-to-pixel concentrations support the differences found among the technologies using the visualized maps. Reproducibility studies proved the validity of the conclusions. High magnification studies revealed fine patterns in the distribution of the minor lubricant component Mg stearate, which showed unambiguous correspondence with the crushing strength of the tablets. The inspection of the characteristic bands of the active ingredient provides information regarding the polymorphic changes occurring during the manufacturing process. Furthermore, the overall intensity of the mapping spectra shows an unambiguous, nonlinear correspondence with the compaction force applied during tablet compression (and thus indirectly with the crushing strength). It was also shown that the semi-quantitative estimation of concentrations

can be improved by an empirical correction method that takes the structural properties of the pharmaceuticals into consideration.

In connection with these detailed analyses, the homogeneity of solid dispersions prepared by supercritical extrusion and electrospinning was also assessed, using the estimated concentration maps. The Raman maps enable sensitive detection of locally heterogeneous areas, which can be related to the beginning of crystallization within the sample (below the limit of detection of XRPD). The descriptive statistics were shown to be related to the dissolution rate of the active ingredient.



a: direct compression *b*: high-shear granulation *c*: fluidized bed granulation

Figure 1. Spatial disptribution of API in differently manufactured tablets (Raman mapping at 100× magnification)

4.3. Investigation of drug-excipient interaction using multivariate curve resolution

Locally appearing areas within solid dispersions somewhat enriched in the active ingredient are always suspected to contain crystalline seeds. Similarly, complexation between a drug and its carrier matrix may sometimes be incomplete to a certain extent. Although Raman mapping with appropriate magnification is theoretically feasible to detect if this happens locally within the sample, the traditional evaluation methods are often incapable of extracting the required information. The reason is that the reference spectra of either the locally appearing polymorphic form or the *pure* resulting complex is unknown. In these cases, these pure spectra are best estimated by self-modelling curve resolution methods.

An analysis of binary drug-cyclodextrin mixtures proved that pure component spectra resolved by MCR-ALS can directly prove the occurrence or lack of complexation. If some non-complexed drug remains present (or is formed in the timeframe between the preparation and analysis), one of the resolved spectra proves its existence by featuring its vibrational bands. The concentrations estimated by MCR-ALS can be directly used for the quantitative estimation of the crystalline, non-complexed content, well below the XRPD limit of detection, and even when the drug-excipient complex is itself crystalline.

4.4. Characterization of unknown products using exploratory statistics

When illegal or counterfeit products are examined, the components and the manufacturing methods are often unknown. Therefore, in this part of the study, all samples were treated as completely unknown. The analysis of a model tablet provided the means to compare subspace projection methods (PCA, MAF), covariance analysis (SS2D) and self-modelling curve resolution algorithms (SMMA, MCR-ALS, PMF). The noise level was

varied to see its effect on the resolved spectra and concentrations. Then, the model tablets introduced in Section 4.2. were investigated to determine the effect of the distribution and homogeneity of the active ingredient on the goodness of the curve resolution process. In this case, SMMA, MCR-ALS, PMF and SISAL were compared.

The results regarding these sample sets can be summarized in the following:

- Methods were developed to estimate the number of components present in a sample.
- It was determined, using traditional model performance parameters and SRD ranking method, that the pure spectra of highest quality and concentration maps closest to those obtained with CLS are yielded by MCR-ALS. This applies for high noise levels and practically completely homogeneous API distribution as well.
- When unknown products are analyzed by Raman mapping, chemometric evaluation proposed in this study enables the resolution of the highest possible number of relevant pure component spectra and concentration maps, enabling the identification of components using a spectral library, and performing in-depth technology evaluation using the concentration maps and our results described in Section 4.2.

The comparative study of extruded Isoptin SR-E and wet granulated Isoptin SR proved that the spectra of two homogeneously distributed can also be separated using MCR-ALS with the joint constraints of non-negativity, closure and Windig's angle (or contrast) constraint, thereby enabling the analysis of solid dispersions. This approach revealed misleading information about the composition of Isoptin SR-E in the original publication discussing the dissolution properties of Isoptin SR-E[13]. Then, using, the correct composition determined via the combined use of curve resolution and a spectral library, along with the estimated concentration maps, an improved interpretation was given to the results originally shown in the cited paper.

4.5. Quantitative characterization of polymer waste using clustering and curve resolution methods

The composition of polymer waste, such as that of the illegal drugs, is often partially or completely unknown. An added, major difficulty lies in the presence of contaminations and high variety of components, which results in low signal-to-noise ratio and strong contribution of various disturbing factors to the mapping spectra (such as strong fluorescent background, intensity cut-off due to detector saturation and presence of dyes with intensive Raman bands). These render the traditional integrated intensity and CLS methods completely inapplicable. Even though in this case every pixel usually corresponds to one polymer only, and visual inspection and library identification of each mapping spectrum would provide reliable results, this approach would take hours of human labour for each Raman map.

An empirical approach, based on SMCR algorithms, was developed to identify and quantify those constituents which can be detected via their Raman signals. The method assigns each pixel to the most prominent polymer contained within the sampled volume, based on the concentrations estimated by SMCR algorithms for each polymer in that particular pixel. The pixel will be assigned to a polymer if its resolved concentration exceeds a pre-defined threshold level – if multiple such polymers exist, then the pixel will be assigned to the polymer with the highest SMCR resolved concentration. By comparing multiple SMCR algorithms, MCR-ALS was selected as the most suitable method to provide the basis for pixel assignment. The resulting evaluation process is able to provide quantitative estimation for

[13] W. Roth, B. Setnik, M. Zietsch, A. Burst, J. Breitenbach, et al., Int. J. Pharm. 368 (2009) 72-75.

major as well as trace components, within the acceptable error margins for polymer waste materials.

4.6. Improving quantitative characterization via Raman mapping using regression and machine learning algorithms

The CLS method described in Section 4.2. is already feasible for giving a rough estimation on the mass fractions of the components present. However, the (in)accuracy of such estimation cannot be determined just based on the reference pure component spectra. The calculation of concentration maps with known accuracy require a calibration set and the use of regression methods. Since no comparative study can be found for regression with chemical imaging, the present study investigates numerous methods with various complexity, from univariate linear regression to nonlinear machine learning algorithms.

Normalization of mapping spectra usually cannot be avoided due to the surface roughness and the resulting inaccuracy in focusing, leading to intensity deviations. It was shown, however, that this preprocessing method results in nonlinear correspondence between the computed and the actual concentrations. This phenomenon is especially strong where the Raman-activity (integrated intensity at unit acquisition time, over the entire wavenumber range) of the components are very different. Although *partial least squares* (PLS), a method well known in chemometrics, already provided acceptable prediction as shown in Figure 2a, it was further improved by the use of variable selection methods. Variable selection enables filtering the uninformative wavelength channels out that have a negative influence on the model accuracy. In contrast to what is usual in spectrometry, bands with small or intermediate intensity tend to be selected via variable selection and not those with the highest intensity, particularly due to their nonlinear behaviour.

Using either interval PLS or genetic algorithms, the root mean square error (RMSE) of estimation was reduced to fall below 3.5% in mass fractions per pixel. It was shown that iPLS can also be applied to aid univariate analysis by selecting the vibrational band with the best predictive attributes in an automated manner. It was found that in the case of normalized mapping datasets the selected band is ***not*** among the most intensive ones, as the most intensive peaks show stronger nonlinearity and intensity deviation with respect to the concentration.

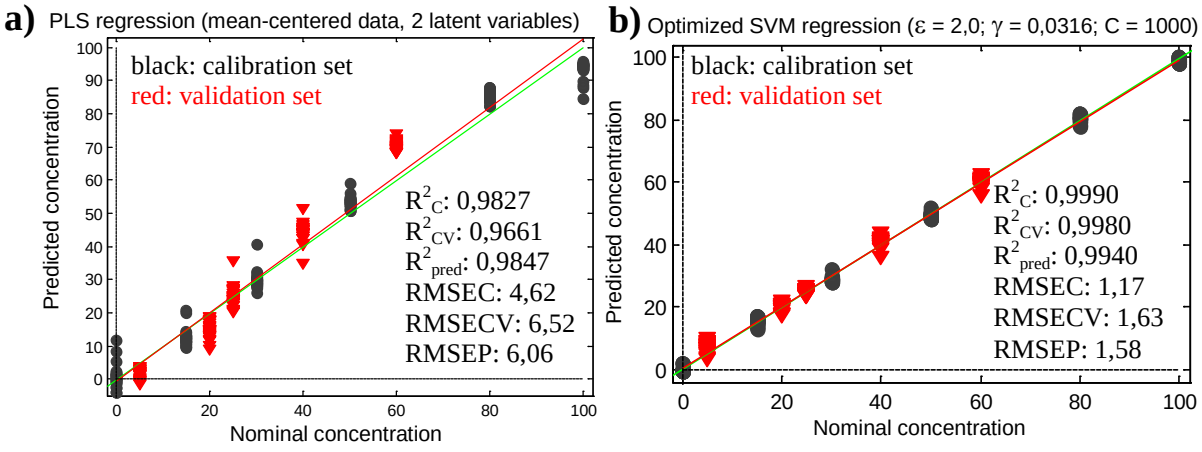


Figure 2. Predicted versus nominal concentrations with optimized **a)** PLS and **b)** SVM regression

This study was the first to compare nonlinear regression techniques to account for the effects of normalization (and other factors causing nonlinearity). It was shown that the highest predictive accuracy, in this case with an error below 2% in mass fractions, can be achieved using support vector machines (Figure 2b).

The comparison of regression methods was carried out using well-known performance parameters (RMSE, coefficient of determination, etc.) as well as with the novel SRD method. This technique was applicable in ranking the regression methods but just as well in the comparison of different performance parameters.

5. Theses of dissertation

1. It was proven that the concentration maps estimated by CLS modelling of Raman maps enable the **characterization of the heterogeneity** of the components within pharmaceutical products, providing the means to determine the manufacturing technology of the investigated product. The descriptive statistics and the visualized concentration maps provide information about the dissolution and mechanical properties of the products. [**I-V**,XVI-XVIII,XXIV]

2. It was shown that the occurrence of **complexation can be detected** by MCR-ALS decomposition of Raman maps. The proposed algorithm yields two spectra, one of which determines whether the complexation has taken place or not. The other estimated spectrum contains the peaks of the non-complexed, crystalline active ingredient if it is present. It was shown that the **quantitative estimation of the crystalline active ingredient can be carried out** using its concentration map estimated by MCR-ALS, regardless of whether the API-cyclodextrin complex itself is crystalline or amorphous. [**VI**,1,6,9,12]

3. The fact was recognized that **the number of components within a tablet of unknown composition can be determined** using SMMA method, if the noise level is low or moderate (0-15%) and if the occurrence of polarization effects is low. Supplementary application of PCA is also advised at high noise levels. When the polarization effects are very strong, 10-12 component spectra should be resolved in order to overestimate the number of components actually present, after which spectra corresponding to the same component should be averaged. [**VII**,**VIII**]

4. It was confirmed that the **spectra and concentration maps of the components can be best estimated by MCR-ALS and PMF** methods, **regardless of the level of noise and the degree of homogeneity** of the active ingredient. The accuracy of the spectrum resolution only slightly depends on the input spectra used to initialize the algorithm. MCR-ALS can resolve the spectrum of a homogeneously distributed component even when only non-negativity constraints are used, while the accuracy of the spectra reolved by PMF can be optimized by tuning the so-called rotational parameter. Among the spectrum resolution methods, MCR-ALS provides the concetration values closest to CLS using the actual reference spectra.
[**VII**,**VIII,**XVII,XVIII,2-6,10-12,14]

5. **Spectra of components forming a solid disperson** were resolved for the first time, applying MCR-ALS and the joint use of nonnegativity, closure and Windig's angle (contrast) constraints. The approach requires that the correlation between the true pure component

spectra should be small. This method enables accurate resolution of pure component spectra from solid dispersions such as those manufactured by extrusion. [**V**]

6. A method based on Raman mapping was developed for the first time to **identify and quantitatively characterize the components of polymer waste**. The empirical method developed to assign polymers to the different pixels provides similar accuracy to the manual pixel-by-pixel identification of spectra, despite the very low quality of the Raman spectra collected from the waste samples. The method is able to estimate the concentration of major as well as trace components within the tolerable error margin in the analysis of polymer waste. [**IX**,XIX]

7. Raman mapping was successfully applied for **quantitative characterization**. It was shown that normalization, which is used to remove the intensity deviation coused by focusing errors, results in nonlinear correspondence between the intensities and the concentrations. It was shown that the degree of nonlinearity and the error of prediction can be reduced by using variable selection methods, which remove the bands with the highest intensities (due to their nonlinear behaviour) from the dataset. **Univariate regression can be enhanced with a multivariate approach by applying interval PLS variable selection** with wide window size. The comparison of various multivariate regression and machine learning methods showed that the **concentration error in the distinct pixels can be best reduced using support vector regression, below 2% (in mass fractions)**.[**I**,XXIX,6,12,13,18]

6. Practical application of scientific results

All investigations shown in this dissertation were motivated and driven by actual issues in the pharmaceutical and/or polymer technology. The appropriate chemometric method was selected and optimized for each task. As a result, the analytical protocols proposed in this study can be directly used in both the industrial and the academic practice.

One of the most relevant area of application is expected to be the analysis of amorphous active ingredients, as Raman mapping enables the detection of locally appearing crystalline seeds. This technique can be applied not only in posterior stability tests after storing the sample for long time intervals, but also to predict the stability, in a qualitative manner, right after manufacturing. The estimations provided below the XRPD limit of detection have gained increased interest in the pharmaceutical industry.

Results regarding unknown pharmaceuticals are mostly expected to be applied in the analysis of illegal and counterfeit drugs by pharmaceutical and federal authorities. With the evaluation of Raman maps proposed in this study, the identification of the components present is only one of the outcomes. Besides, it becomes possible to draw conclusions about certain conditions of manufacturing. This information helps determine whether the source of two narcotic tablets is the same or not. Chemical imaging, unlike other methods, can distinguish “high-quality” counterfeits from the originals. (These can be just as dangerous for patients as producers of counterfeits do not comply with the strict pharmaceutical regulations.) It has to be also emphasized that Raman mapping enables completely non-invasive analysis, thus, tablets investigated in this manner can be further analyzed by other analytical methods, making Raman mapping a truly valuable tool in pursuing illegal products.

Polymer waste can be directly analyzed by following the steps outlined in this study and by applying the empirical method developed and published in the framework of this

dissertation. Applying this approach saves hours of human labour for each Raman map, while maintaining its accuracy at an acceptable level. Such analyses have been later performed on various waste samples from the automotive, electronic and construction industries. The practical applications are described in detail in another PhD study[12].

Regarding quantitative analysis of pharmaceutical solid dispersions, many aspects of the results shown here are not specific to the samples analyzed in this study but can be generally applied. The conclusions drawn here are also not restricted to imaging but can be often generalized to other Raman (and possibly other vibrational) spectrometric applications. The present dissertation offers a critical and reliable method to test and compare multiple regression algorithms. Furthermore, even when legal restrictions or industrial policy requires the use of straightforward univariate regression, interval PLS, according to our guidelines, can be easily used to automatically select the best vibrational peak to use.

7. List of related publications

7.1. Published in journals with impact factor

[**I**] **B. Vajna**, I. Farkas, A. Szabó, Zs. Zsigmond, Gy. Marosi: *Raman microscopic evaluation of technology dependent structural differences in tablets containing imipramine model drug*, Journal of Pharmaceutical and Biomedical Analysis, 51 (2010) 30-38. IF: **2,733** Cited by: 16 (10*)

[**II**] G. Patyi, A. Bódis, I. Antal, **B. Vajna**, Zs. Nagy, Gy. Marosi: *Thermal and spectroscopic analysis of inclusion complex of spironolactone prepared by evaporation and hot melt methods*, Journal of Thermal Analysis and Calorimetry, 102 (2010) 349-355. IF: **1,752** Cited by: 9 (4*)

[**III**] Zs. Nagy, M. Sauceau, E. Rodier, **B. Vajna**, K. Nyúl, Gy. Marosi, J. Fages: *Use of Supercritical CO_2 aided and Conventional Melt Extrusion for Enhancing the Dissolution Rate of an Active Pharmaceutical Ingredient*, Polymers for Advanced Technologies, 23 (2011) 909-918. IF: **1,776** Cited by: 1 (0*)

[**IV**] Zs. Nagy, A. Balogh, **B. Vajna**, A. Farkas, G. Patyi, Gy. Marosi: *Comparison of electrospun and extruded, Soluplus® based solid dosage forms of improved dissolution*, Journal of Pharmaceutical Sciences, 101 (2011) 322-332. IF: **3,031** Cited by: 6 (6*)

[**V**] **B. Vajna**, H. Pataki, Zs. Nagy, I. Farkas, Gy. Marosi: *Characterization of melt extruded and conventional Isoptin formulations using Raman chemical imaging and chemometrics*, International Journal of Pharmaceutics, 419 (2011) 107-113. IF: **3,607** I: 1 (1*)

[**VI**] **B. Vajna**, I. Farkas, A. Farkas, H. Pataki, Zs. Nagy, J. Madarász, Gy. Marosi: *Characterization of drug-cyclodextrin formulations using Raman mapping and multivariate curve resolution*, Journal of Pharmaceutical and Biomedical Analysis, 56 (2011) 38-44. IF: **2,733** Cited by: 3 (3*)

[**VII**] **B. Vajna**, G. Patyi, Zs. Nagy, A. Farkas, Gy. Marosi: *Comparison of chemometric methods in the analysis of pharmaceuticals with hyperspectral Raman imaging*, Journal of Raman Spectroscopy, 42 (2011) 1977-1986. IF: **3,137** Cited by: 6 (2*)

[**VIII**] **B. Vajna**, A. Farkas, H. Pataki, Zs. Zsigmond, T. Igricz, Gy. Marosi: *Testing the performance of pure spectrum resolution from Raman hyperspectral images of differently manufactured pharmaceutical tablets*, Analytica Chimica Acta, 712 (2012) 45-55. IF: **4,310** Cited by: 1 (1*)

[**IX**] **B. Vajna**, B. Bodzay, A. Toldy, I. Farkas, T. Igricz, Gy. Marosi: *Analysis of car shredder polymer waste with Raman mapping and chemometrics*, Express Polymer Letters, 6 (2012) 107-119. IF: **1,575** Cited by: 1 (0*)

[**X**] B.B. Marosfői, Gy. Marosi, A. Szabó, **B. Vajna**, A. Szép: *Laser pyrolysis micro-spectroscopy for modelling fire-induced degradation of ethylene-vinyl acetate systems***,** Polymer Degradation and Stability 92 (**2007**) 2231-2238. IF: **1,752** Cited by: 2 (0*)

[**XI**] M. Berkesi, K. Hidas, T. Guzmics, J. Dubessy, R.J. Bodnar, Cs. Szabó, **B. Vajna**, T. Tsunogae: *Detection of small amounts of H_2O in CO_2-rich fluid inclusions using Raman spectroscopy*, Journal of Raman Spectroscopy, 40 (**2009**) 1461-1463. IF: **3,147** Cited by: 13 (6*)

[**XII**] F. Billes, H. Pataki, O. Ünsalan, H. Mikosch, **B. Vajna**, Gy. Marosi: *Solvent effect on the vibrational spectra of Carvedilol*, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 95 (2012) 148-164. IF: **2,098**

[**XIII**] G. Szebényi, G. Romhány, **B. Vajna**, T. Czvikovszky: *EB treatment of carbon nanotube-reinforced polymer composites*, Radiation Physics and Chemistry, 81 (2012) 1383-1388. IF: **1,132**

[**XIV**] I.M. Szilágyi, B. Fórizs, O. Rosseler, Á.Szegedi, P. Németh, P. Király, G. Tárkányi, **B. Vajna**, K. Varga-Josepovits, K. László, A.L. Tóth, P. Baranyai, M. Leskelä: *WO_3 photocatalysts: Influence of structure and composition*, Journal of Catalysis, in press, **available online**, DOI: 10.1016/j.jcat.2012.07.013, IF: **6,002**
[**XV**] H. Pataki, I. Csontos, Z. K. Nagy, **B. Vajna**, M. Molnár, L. Katona, Gy. Marosi: *Implementation of Raman Signal Feedback to Perform Controlled Crystallization of Carvedilol*, Organic Process Research and Development, in press, **available online**, DOI: 10.1021/op300062t (**2012**) IF: **2.391**

* Citations by independent authors

7.2. Published in journals without impact factor

[XVI] **Vajna B.**, Nagy Zs.K., Patyi G., Zsigmond Zs., Antal I., Marosi Gy.: *Application of chemical imaging in pharmaceutical technology*, Acta Pharmaceutica Hungarica 79 (2009) 104-116.
[XVII] **B. Vajna**, P. Szepesváry, Gy. Keglevich, Gy. Marosi: *Acquisition methods of chemical images and their evaluation with chemometric methods*, Hungarian Journal of Chemistry, 116 (2010) 77-85.
[XVIII] **B. Vajna**, A. Bódis, Gy. Marosi: *Multivariate data analysis in chemical imaging*, Journal of the Hungarian Chemical Society, 65 (**2010**) 313-319.
[XIX] **B. Vajna**, K. Palásti, B. Bodzay, A. Toldy, S. Patachia, R. Buican, C. Catalin, M. Tierean, *Complex analysis of car shredder light fraction*, The Open Waste Management Journal, 3 (2010) 47-56.
[XX] Nagy Zs. K., Patyi G., Bodzay B., **Vajna B**., Marosi Gy.: *From composites to nanomedicines*, Plastic and Rubber, 12 (2009) 450-454.
[XXI] Z.K. Nagy, G. Patyi, B. Bodzay, **B. Vajna**, G. Marosi: *Prüfungen und Herstellungsverfahren von Composites bis zu Nanomedikamenten*, Gummi Fasern Kunststoffe 64 (2011) 100-104.
[XXII] Marosi Gy., Patyi G., Nagy Zs.K., **Vajna B.**, Szabó A., Anna P., *Some examples from the field of technology and analysis of pharmaceutical products*, Hungarian Journal of Chemistry, 114 (2008) 137-140.
[XXIII] **B. Vajna** (szerk: K. László): *Conference of MSc students - abstracts of the best contributions October 2007*, ezen belül*: Application of Raman-microspectrometry in pharmaceutical developments*, Periodica Polytechnica Chemical Engineering 52 (2008) 73-74.
[XXIV] Pataki H,, Palásti K., **Vajna B**., Csontos I., Marosi Gy.: *Gyógyszerhatóanyag-kristályosodás valós idejű vizsgálata és módosítása segédanyaggal*, Acta Pharmaceutica Hungarica, 81 (**2011**) 109-124.

7.3. International conference abstracts

[XXV] H. Pataki, **B. Vajna**, Zs. Nagy, Gy. Marosi: *Investigation of crystallization processes using in-line Raman spectroscopy*, in: 4th BBBB International Conference on Pharmaceutical Sciences, conference proceeding book, page 116; European Journal of Pharmaceutical Sciences, 44 (2011) 1-204. IF: **3,291**

7.4. Manuscipts submitted or under preparation

[XXVI] H. Pataki, I. Markovits, **B. Vajna**, Zs. K. Nagy, Gy. Marosi: In-line monitoring of carvedilol crystallization using Raman spectroscopy, *Crystal Growth and Design*, submitted (**2012**)
[XXVII] S. Lakio, **B. Vajna,** I. Farkas, H. Salokangas, Gy. Marosi, J. Yliruusi: *Challenges in detecting Magnesium stearate distribution in tablets*, The AAPS Journal, submitted, returned for major revision (**2012**)
[XXVIII] B. Gyarmati, **B. Vajna**, Á. Némethy, A. Szilágyi: *Redox- and pH-responsive cysteamine-modified poly(aspartic acid) showing reversible sol-gel transition*, Macromolecular Rapid Communications, submitted (**2012**)
[XXIX] **B. Vajna**, I. Farkas, P. Sóti, Zs. Nagy, F. van der Gucht, H. Pataki, A. Farkas, Gy. Marosi: *Quantitative characterization of solid dispersions with nonlinear regression and variable selection methods*, writing in progress

7.5. Selected oral presentations

[1] **B. Vajna**, G. Marosi: *Local analysis of polymorphism with micro-Raman spectrometry*, Symposium on Pharmaceutical Chemistry and Technology ’08, Zalakaros, 29-30 September 2008.

[2] **B. Vajna**: *Application of chemometric methods in Raman mapping*, Symposium on Molecular Spectroscopy, Eötvös University, Budapest, 11 December 2009.
[3] **B. Vajna**, K. Palásti, P. Szepesváry, G. Marosi: *Application of chemometric methods in Raman mapping*, Analytical Days, Hungarian Chemical Society, Budapest, 28-29 January 2010.
[4] **B. Vajna**, G. Marosi: *Application of chemometric methods in Raman mapping*, KeMoMo-QSAR Symposium, Hungarian Academy of Sciences, Szeged, 29-30 April 2010.
[5] **B. Vajna**, P. Szepesváry, G. Marosi: *Testing the performance of pure spectrum resolution from Raman images of differently manufactured pharmaceutical tablets*, 5th International Symposium on Computer Applications and Chemometrics in Analytical Chemistry, MTESZ székház, Budapest, 2010. június 21-25.
[6] **B. Vajna**, P. Szepesváry, G. Marosi: *Analysis of Pharmaceuticals Supported by Chemometric Methods*, Symposium on Pharmaceutical Chemistry and Technology '10, Velence, 4-5 October 2010.
[7] **B. Vajna**, G. Marosi: *Multivariate data analysis (chemometrics) in Raman spectroscopy and imaging*, PANNON Sciences Association for Public Benefit, Vörösberény, 27-28 June 2009.
[8] **B. Vajna**, G. Marosi: *Introduction to multivariate data analysis (chemometrics)*, PANNON Tudományok Közhasznú Egyesület, Erdőtarcsa, 2010. jún. 18-20.

7.6. Selected poster presentations

[9] **B. Vajna**, A. Szabó, G. Marosi: *Micro-Raman spectrometry for detection of local polymorphic forms in solid pharmaceuticals*, 7th Central European Symposium on Pharmaceutical Technology and Biodelivery Systems, Ljubljana, 18-20 September 2008.
[10] **B. Vajna**, H. Pataki, G. Marosi: *Characterization of unknown tablets with micro-Raman mapping and chemometrics*, 7th World Meeting on Pharmaceutics, Biopharmaceutics and Pharmaceutical Technology, Valletta, Malta, 8-11 March 2010.
[11] **B. Vajna**, A. Farkas, P. Szepesváry, G. Marosi: *Chemometric resolution of pure component spectra in Raman chemical imaging*, 12th Chemometrics in Analytical Chemistry, Antwerp, Belgium, 17-21 October 2010.
[12] **B. Vajna**, H. Pataki, G. Marosi: *Raman Mapping and Chemometrics for the Characterization of Unknown Pharmaceuticals*, 4th FIP Pharmaceutical Sciences World Congress & AAPS Annual Meeting and Exhibition, New Orleans, USA, 13-18 November 2010.
[13] I. Farkas, **B. Vajna**, A. Farkas, Zs. Nagy, Gy. Marosi: *Characterization of poly(lactic co-glycolic acid) with Raman mapping using different chemometric regression methods*, Conferentia Chemometrica 2011, Sümeg, Hungary, 18-21 September 2011.
[14] A. Farkas, A. Balogh, Zs. Nagy, **B. Vajna**, I. Farkas, Gy. Marosi: *Combined use of Raman mapping for the analysis of pharmaceutical products with narrow concentration distribution and highly correlated Raman spectra*, Conferentia Chemometrica 2011, Sümeg, Hungary, 18-21 September 2011.
[15] H. Pataki, **B. Vajna**, G. Marosi: *Investigation of crystallization processes using in-line Raman spectrometry*, 7th World Meeting on Pharmaceutics, Biopharmaceutics and Pharmaceutical Technology, Valletta, Malta, 8-11 March 2010.
[16] A. Krupa, R. Jachowicz, **B. Vajna**, G. Marosi, D. Majda: *Orodispersable tablets containing solid dispersions of ibuprofen with neusilin US2 to improve the the insoluble drug delivery*, 7th World Meeting on Pharmaceutics, Biopharmaceutics and Pharmaceutical Technology, Valletta, Malta, 8-11 March 2010.
[17] A. Farkas, **B. Vajna**, T. Firkala, I. Farkas, Gy. Marosi: *Investigation of small trace amounts of drug by surface enhanced Raman chemical imaging supported by MCR-ALS method*, 13th Chemometrics in Analytical Chemistry, Budapest, 25-29 June 2012.
[18] I. Farkas, **B. Vajna**, A. Farkas, Zs. Nagy, H. Pataki, Gy. Marosi: *Transmission Raman technique: an innovative instrument in chemometric regression*, 13th Chemometrics in Analytical Chemistry, Budapest, 25-29 June 2012.