Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

DISTRIBUTION OF A LOW DOSE COMPOUND WITHIN PHARMACEUTICAL TABLET BY USING MULTIVARIATE CURVE RESOLUTION ON RAMAN HYPERSPECTRAL IMAGES

Mathieu Boiret[a*], Anna de Juan[b], Nathalie Gorretta[c], Yves-Michel Ginot[a], Jean-Michel Roger[c]

[a]Technologie Servier, Orléans, France

[b]Grup de Quimiometria, Dept. Química Analítica, Universitat de Barcelona, Spain

[c]IRSTEA, UMR ITAP 361, Montpellier, France

*Corresponding author:

Tel.: +33 238 238 175

E-mail address: mathieu.boiret@fr.netgrs.com

Abstract

In this work, Raman hyperspectral images and Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) are used to study the distribution of actives and excipients within a pharmaceutical drug product. This article is mainly focused on the distribution of a low dose constituent. Different approaches are compared, using initially filtered or non-filtered data, or using a column-wise augmented dataset before starting the MCR-ALS iterative process including appended information on the low dose component. In the studied formulation, magnesium stearate is used as a lubricant to improve powder flowability. With a theoretical concentration of 0.5% w/w in the drug product, the spectral variance contained in the data is weak. By using a Principal Component Analysis (PCA) filtered dataset as a first step of the MCR-ALS approach, the lubricant information is lost in the non-explained variance and its associated distribution in the tablet cannot be highlighted. A sufficient number of components to generate the PCA noise-filtered matrix has to be used in order to keep the lubricant variability within the data set analyzed or, otherwise, work with the raw non-filtered data.

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

Different models are built using an increasing number of components to perform the PCA reduction. It is shown that the magnesium stearate information can be extracted from a PCA model using a minimum of 20 components. In the last part, a column-wise augmented matrix, including a reference spectrum of the lubricant, is used before starting MCR-ALS process. PCA reduction is performed on the augmented matrix, so the magnesium stearate contribution is included within the MCR-ALS calculations. By using an appropriate PCA reduction, with a sufficient number of components, or by using an augmented dataset including appended information on the low dose component, the distribution of the two actives, the two main excipients and the low dose lubricant are correctly recovered.

Keywords: Multivariate Curve Resolution, Alternating Least Squares, Raman hyperspectral imaging, Spectroscopy, Active and excipient distributions, Low dose compound

1. Introduction

In the last decade, the use of imaging coupled with vibrational spectroscopies (near infrared, mid infrared, fluorescence and Raman) has grown quickly in research and development environments. The spatial and spectral information contained in hyperspectral images can be associated with the distribution of the different constituents within the sample. Different areas such as polymer research [1], biomedical analysis [2], environment field [3] and pharmaceutical development [5] are using these new analytical tools based on vibrational hyperspectral imaging. During the analytical lifecycle of a pharmaceutical drug product, hyperspectral imaging became a very powerful technique to explore the compound distributions on the tablet surface or within a powder mixture [6]. This technology appeared as innovative and promising to ensure the final quality of the drug product [8] from the development to the production.

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

Because of the huge amount of data contained in hyperspectral images, a direct interpretation of the acquired images is often not possible. Therefore, several chemometric tools have previously been applied [10, 11]. Qualitative analyses such as Principal Component Analysis (PCA) have already been used with near infrared [12] and Raman [13] chemical imaging in order to study the compound distribution in a sample. Since PCA is mainly linked to the dataset variability and as calculated loadings do not have chemical meaning, this approach is used as a descriptive method. To extract quantitative information at a global and pixel level, principal component regression (PCR) and partial least squares regression (PLS-R) have already demonstrated through several studies that they were powerful chemometric techniques [15, 16]. However, these methods can be time consuming and difficult to implement as they usually require a calibration step to develop predictive models. To overcome this problem, resolution methods seem to be a good alternative.

The aim of resolution methods is to provide the distribution maps and pure spectra related to the image constituents of a sample from the information contained in the raw image [17]. Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) is one of the most famous tools applied on hyperspectral images [18, 19]. MCR-ALS decomposes the initial data in a bilinear model, assuming that the observed spectra (i.e. each pixel of the image) are a linear combination of the spectra of the pure components in the system. In order to ensure an accurate resolution, constraints have to be used during the optimization process. Indeed, due to rotational or intensity ambiguities, resolution of a multicomponent hyperspectral image might not be unique [21]. Different constraints were established and tested [23, 24]. In image resolution, non-negativity, spectral normalization and local rank analysis are generally the most successful tools. Local rank analysis describes the spatial complexity of an image by identifying the rank of a pixel neighbourhood area. Combined with reference spectra of the image constituents, the absence of one or more specific constituent in a pixel can be highlighted. Some constraints used for the resolution of a chemical process, such as unimodality, closure or hard-modelling should not be used to analyse hyperspectral images because

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

concentration profiles in the pixels of an image do not present the global continuous evolution that process profiles have [25].

Raman chemical imaging, because of its advantages such as negligible sample preparation, high chemical specificity and high spatial resolution, emerges as a new analytical tool in the quality control process of a solid drug product [26]. Final drug products are usually manufactured by using at least one active pharmaceutical ingredient (API) and several excipients. To improve powder flowability, most of the pharmaceutical manufacturing process includes a lubricant in the final drug formulation [27]. This compound is commonly present in a very low concentration in the powder blend and a spectroscopic bulk analysis will not be able to extract its contribution. Indeed, the corresponding variance of this constituent is very weak comparing with the other compounds of the sample. PCA, which aims at describing the directions of maximum global variance in the data, may have difficulties in retrieving information linked to a low dose constituent when the variance allocated to this component is similar in level to noise, which is often large in hyperspectral images. By offering the possibility to acquire images with a high spatial resolution, Raman chemical imaging coupled with appropriate chemometric methods appears as a promising technique to detect a low dose compound within a solid drug formulation.

In this work, MCR-ALS was applied on Raman chemical imaging data in order to provide the distribution of actives and excipients in a commercialised tablet. MCR-ALS was challenged by trying to identify the low dose lubricant in the hyperspectral image. The effect of using algorithms driven by finding directions of maximum variance explained is studied. In this sense, the effect linked to the first step of noise-filtering based on PCA, which is often used in MCR-ALS to remove noise and non-useful spectral information, is studied. By applying MCR-ALS on a noise-filtered PCA matrix, it is shown that the information of the low dose constituent may be lost during data reduction. The comparison between the MCR-ALS decomposition on a filtered and a non-filtered PCA matrix is

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

presented. Moreover, to keep the low dose constituent information during the PCA reduction, calculations are performed on an augmented matrix including the low dose constituent spectrum. The necessity of using appropriate pre-processing methods and constraints to find out the correct information linked to these low dose constituents is emphasized. This article shows the strategies to be followed in MCR-ALS analysis to retrieve correct information for low dose image constituents, from pre-processing, conditions to drive the iterative optimization to proper inclusion of constraints.

2. Materials and Methods

2.1. Samples

A commercial coated tablet of Bipreterax®, prescribed for arterial hypertension treatment and commercialised by "Les Laboratoires Servier", was used for the study. It is also known as Perindopril/Indapamide association. Final drug product contains respectively 4 mg of Perindopril (API1) and 1.25 mg of Indapamide (API2). Actives are known to have several solid state forms, but only one of them is present in this formulation. Major core excipients are lactose monohydrate, microcrystalline cellulose (Avicel). Magnesium stearate (MgSt), which is used as a lubricant, was added to the blend before compression with a theoretical mass concentration corresponding to 0.5% w/w. In order to analyse the tablet core, the coating was removed by eroding the sample with a Leica EM Rapid system (Leica, Wetzlar, Germany). A visual examination of the tablet did not provide any information concerning the distribution of the different compounds within the tablet.

2.2. Raman imaging system

The image was collected using a RM300 PerkinElmer system (PerkinElmer, Waltham, MA) and the Spectrum Image version 6.1 software. The microscope was coupled to the spectrometer and spectra

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

were acquired through it with a spatial resolution of 10 µm in a Raman diffuse reflection mode. Wavenumber range was 3200–100 cm^{-1} with a resolution of 2 cm^{-1}. Spectra were acquired at a single point on the sample, then the sample was moved and another spectrum was taken. This process was repeated until spectra of points covering the region of interest were obtained. A 785 nm laser with a power of 400 mW was used. Two scans of 2 s were accumulated for each spectrum. An image of 70 pixels per 70 pixels corresponding to 4900 spectra was acquired for a surface of 700 µm by 700 µm.

2.3. Pre-processing

Data were preprocessed in order to remove non-chemical biases from the spectra (scattering effect due to non-homogeneity of the surface, interference from external light source, spikes due to cosmic rays, random noise). First of all, data were spike-corrected in order to reduce the effect of cosmic rays [28]. The spectral range was reduced in order to focus only on the region of interest, corresponding to a Raman shift from 1800 cm^{-1} to 200 cm^{-1}. Reduced spectra were preprocessed by asymmetric least squares (AsLS) to correct baseline variations due to fluorescence contributions [29]. Finally, to enhance slight spectral variations, a Savitzky-Golay first derivative with a 2^{nd} order polynomial smoothing on a 9 points window [30] was applied.

2.4. Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)

A brief description of the MCR-ALS algorithm is given here. The algorithm was previously described in detail in Refs. [23, 24]. As any resolution methods, the main goal of MCR-ALS is decomposing the original matrix $\mathbf{D}_{(\mathbf{n},\mathbf{p})}$ (n samples or rows and p variables or columns) of a multi-component system into the underlying bilinear model which assumes that the observed spectra are a linear combination of the spectra of the pure components in the system:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \qquad (1)$$

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

where **C** is the matrix of concentration profiles, $\mathbf{S}^T$ the matrix of pure responses (i.e. spectra) and **E** contains the experimental error. In resolution of spectroscopic images, $\mathbf{D}_{(\mathbf{n},\mathbf{p})}$ is the matrix of the unfolded image, **C** contains the concentration profiles that, conveniently refolded, show the distribution maps of each image constituent and $\mathbf{S}^T$ contains the associated pure spectra [31].

In order to provide chemically meaningful profiles (i.e., pure spectra and distribution maps) and to reduce intensity and rotational ambiguities in the MCR solutions, constraints must be properly chosen during the iterative MCR-ALS process. Since concentrations of the constituents should not be negative, a non-negativity constraint was applied. Moreover, the calculated spectral profiles in matrix $\mathbf{S}^T$ were normalized at each iteration. To identify where the constituents of the drug product are present or absent in the image, the Fixe Size Moving Window Evolving Factor Analysis (FSMW-EFA) method was applied to the data [32]. This method provides the local complexity of a sample by performing singular value decomposition by moving a window of pixels across the full image. A window contains a specified number of spectra (at least 4, corresponding to a specific pixel and its neighbours). By calculating singular value maps of the sample, the presence of overlapped compounds in a pixel area can be displayed. By selecting a specific threshold, a corresponding local rank map can be provided by plotting the number of significant singular values above the threshold. This approach, due to its local character, is particularly well adapted to identify a compound with a low signal or with a low concentration within the sample because small local areas are analyzed one at a time. By comparing the local rank information with reference spectral information, missing constituents on particular pixels can be known [25].

Figures of merit of the optimization procedure are the lack of fit (lof) and the explained variance (R^2). The lack of fit is used to check if the experimental data were well fitted by the MCR-ALS procedure. These two criteria are calculated as follow:

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

$$\mathbf{lof}(\%) = \mathbf{100}\sqrt{\frac{\sum_{i,j}\mathbf{e}_{i,j}^2}{\sum_{i,j}\mathbf{D}_{i,j}^2}} \quad (2)$$

$$\mathbf{R}^2 = \frac{\sum_{i,j}\mathbf{D}_{i,j}^2 - \sum_{i,j}\mathbf{e}_{i,j}^2}{\sum_{i,j}\mathbf{D}_{i,j}^2} \quad (3)$$

where $\mathbf{D_{i,j}}$ is the input element of the original matrix $\mathbf{D_{(n,p)}}$ and $\mathbf{e_{i,j}}$ the related residual element after using the MCR-ALS model (see equaltion 1). Input element can be the original element from $\mathbf{D_{(n,p)}}$ or the element of a noise filtered PCA matrix $\mathbf{D_{PCA_{(n,p)}}}$ using the same number of components as in the MCR-ALS. A noise filtered PCA matrix $\mathbf{D_{PCA_{(n,p)}}}$ can be obtained as follows:

$$\mathbf{D_{PCA_{(n,p)}}} = \mathbf{U_{(n,k)}S_{(k,k)}V^T_{(k,p)}} \quad (4)$$

where **U**, **S** and $\mathbf{V^T}$ are calculated by singular value decomposition of the original $\mathbf{D_{(n,p)}}$ matrix and *k* is the number of the known constituents in the drug product. The PCA reduced matrix corresponds to a filtered matrix in a reduced space. This matrix should contain the major part of the spectral variance without noise.

MCR-ALS must be initialised by a first estimate of **C** or $\mathbf{S^T}$ matrix. Initial estimates are generally obtained by purest variable selection methods, such as SIMPLISMA (Simple-to-use Interactive Self-Modelling Mixture Analysis) [33]. This method identifies the most dissimilar spectra (or sample) in the dataset. However, due to the homogeneity of a pharmaceutical sample, it could be difficult to identify a pure pixel corresponding to a single constituent. Most of the time, the theoretical formulation of the sample is known during the development process. So pure reference spectra acquired with the same spectrometer and the same acquisition parameters can be selected as initial estimates to start the optimisation process.

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

In this article, three approaches will be tested and discussed in order to display the distribution of actives and excipients, including the low dose constituent. The first approach starts with the noise filtered PCA matrix $\mathbf{D}_{\mathbf{PCA}(\mathbf{n},\mathbf{p})}$ calculated from (4) using a component number *k* equal to the theoretical number of constituents in the formulation. The second approach consists of increasing the number of components to generate the noise filtered PCA matrix, from *k* to the maximum number of variables, the latter meaning working with the raw non-filtered data set. The third approach consists of using an augmented matrix, where the information of the low dose constituent is added, ensuring the extraction of its contribution during the noise filtering step.

3. Results and discussion

3.1. Exploratory analysis

Because of the spectral variability, applying multivariate data analysis on raw data would not lead to accurate results. Spectra were preprocessed in order to remove baseline variations and cosmic rays. A spike correction algorithm and asymmetric least squares were applied. In order to enhance low variations, a Savitzky-Golay first derivative with a window size of 9 points and a 2^{nd} polynomial order was calculated (Figure 2).

By observing the mean intensity plot of the image (mean intensity in each pixel), no useful information about compound distributions was extracted (results not shown). Therefore, chemometric tools have to be used in order to extract meaningful distributions of the different compounds. As a descriptive method, PCA was applied on the preprocessed data. By calculating appropriate principal components, that describe the maximum variance of the data set and are orthogonal to each other, PCA decomposes the preprocessed matrix in scores (related to distribution maps) and loadings (related to spectra) matrices [34, 35]. Figure 3 shows the image scores results of

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

the five first principal components. Different distributions and agglomerates were highlighted in the images. In this particular example, by knowing the studied formulation and by observing the calculated loading vectors, the distribution maps of PC1 and PC5 were linked to the lactose variability, while distribution maps of PC2, PC3 and PC4 were respectively linked to the distributions of API1, avicel and API2.

Even if PCA analysis provides a first approximation of the component distribution within the sample, the contribution of magnesium stearate was not extracted with this approach. Cumulative variance explained by the PCA model was shown in Figure 4. With 5 components, 98.5% of the variance was explained which means that 1.5% of the spectral variability was not captured by the model. Theoretical spectral variance $\mathbf{Var_i}$ of the magnesium stearate was estimated to 0.5% of the total variance and was calculated by using the following equation:

$$\mathbf{Var_i} = \mathbf{100} \times \frac{\sum_{i,j}\left(\mathbf{c_i s_i^T}\right)^2}{\sum_{i,j}\left(\mathbf{CS^T}\right)^2} \qquad (5)$$

where **C** and $\mathbf{S^T}$ are respectively the theoretical concentrations and the pure reference spectra of each constituent i. Due to the low concentration of magnesium stearate within the drug product, and because of the homogeneity aspect of the powder mixture before compression, the spectral variance of the lubricant might be lower or higher than 0.5%, depending on the studied area of the tablet.

From PC6, the variance contained in the principal components was lower than 0.2% of the total variance and reached a plateau of 0.02% of variance explained per component, which could be associated with a non-structured noise contained in the spectral matrix.

Several hypotheses could explain the non-identification of magnesium stearate within the spectral matrix. Due to its low concentration, the lubricant could either be present on a limited number of

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

pixels or could either be missing in the studied area. The associated spectral information could have led to overlapped features with other components or could have been spread into noise contributions.

PCA is mainly linked to the variability contained within the hyperspectral dataset, expressed as a combination of orthogonal components. Even if it provides a first approximation of the four major constituent distributions, the low spectral variability linked to the lubricant was not displayed on the five first components. Moreover, due to their unclear chemical meaning, loadings are difficult to interpret. To overcome this issue, MCR-ALS algorithm and appropriate constraints were used to enhance the chemical information of the decomposition.

3.2. MCR-ALS

3.2.1. Non-negativity and local rank constraints

MCR-ALS was initialized by using reference spectra of the five different constituents. Spectra were acquired with the same system and with the same parameters as the image. Image pre-processing tools were applied on the reference spectra (see section 2.3). To reduce rotational and intensity ambiguities, non-negativity and equality constraints were applied on the calculated concentrations. Lof and R^2 values were calculated according to equations (2) and (3).

By analysing the image locally, FSMW-EFA provides an estimation of the local complexity of the image [32]. Local rank map was obtained by calculating singular value decomposition on a 4 pixel window moving across the whole data. In general, the number of pixels has to be equal or higher than the total number of the image constituents but in this case, due to the high spatial resolution, the hypothesis was advanced that the five compounds could not be present in the same pixel. Four

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

eigenvalues were calculated for each pixel group. Each component singular values were sorted in increasing order (Figure 5). By choosing an appropriate threshold which separates significant singular values from noise, the local rank map was displayed (Figure 6). (Note that the threshold is selected visually, based on the fact that singular values associated with noise are very small and similar among them and lay at the bottom of plot in Figure 5). The number of missing components for a specific pixel was calculated by removing the local rank value of the pixel to the total rank of the matrix (chosen as the number of theoretical constituents). By calculating correlation coefficients between the raw pixel spectrum and each of the reference spectra, the constituent with the lowest correlation was identified as absent. The absence of a particular component in a pixel was not confirmed unless the correlation coefficient between the pixel spectrum and the reference spectrum of that component is equal or smaller than the largest element in the correlation matrix for that particular component. Results were afterwards encoded in an absence matrix $\mathbf{C}_{sel}$ (Figure 7) containing null values in the concentration elements of the missing components and "not-a-number" (NaN) values in other pixels (unconstrained pixels) [25].

3.2.2. Effect of PCA filtering on MCR-ALS results

In all cases, MCR-ALS was applied on the preprocessed data by using the constraints previously described (see section 3.2.1). The initial preprocessed matrix was reduced (noise-filtering) by using the five first vectors of the PCA decomposition of $\mathbf{D}_{(\mathbf{n},\mathbf{p})}$. MCR-ALS on the filtered PCA matrix provides an optimum value after 9 iterations. 97.9% of the variance was explained with a lack of fit calculated on the initial $\mathbf{D}_{(\mathbf{n},\mathbf{p})}$ and the reduced $\mathbf{D}_{\mathbf{PCA}_{(\mathbf{n},\mathbf{p})}}$ matrices respectively equal to 14.7 and 7.9. Correlation coefficients between calculated spectra and reference spectra were displayed in Table 1. The four first calculated spectra were highly correlated to the two actives and the two major excipients whereas the fifth component was not correlated to the magnesium stearate or to other constituents.

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

By starting MCR-ALS after a PCA reduction of the data, the magnesium stearate contribution was associated with the non-explained variance. In our example, the theoretical number of components in the drug product is equal to 5. The matrix $\mathbf{D}_{\mathbf{PCA}_{(\mathbf{n},\mathbf{p})}}$ calculated by (4), is then calculated by using the five first components of the PCA decomposition. With 5 components, 98.5% of the total variance was explained, which means that 1.5% of the variance was not included in the iterative MCR-ALS process. This part of the non-explained variance contains essentially noise but, due to the low concentration of magnesium stearate, could also contain the spectral contribution of this constituent.

In order to improve the MCR-ALS results and to extract magnesium stearate contribution, MCR-ALS analysis on a PCA-filtered matrix including progressively a larger number of principal components was tested. MCR-ALS decomposition was performed by using a PCA-filtered $\mathbf{D}_{\mathbf{PCA}_{(\mathbf{n},\mathbf{p})}}$ matrix using an increasing number of components, from 5 to the total number of variables. For the first iteration, the $\mathbf{D}_{\mathbf{PCA}_{(\mathbf{n},\mathbf{p})}}$ matrix was built by using the five first vector of the PCA reduction. The following MCR-ALS calculation was performed by adding an additional principal component to calculate the $\mathbf{D}_{\mathbf{PCA}_{(\mathbf{n},\mathbf{p})}}$ matrix. This process was repeated until the number of principal components was equal to the number of variables, corresponding to the use of the preprocessed non-filtered initial $\mathbf{D}_{(\mathbf{n},\mathbf{p})}$ matrix. For each MCR-ALS decomposition from 5 to 100 components, the highest correlation coefficient between the resolved spectra and the pure reference spectrum of magnesium stearate is displayed (Figure 9).

By using less than 20 principal components to reproduce the $\mathbf{D}_{\mathbf{PCA}_{(\mathbf{n},\mathbf{p})}}$ matrix, the contribution of magnesium stearate was not extracted. Using 20, the correlation between the calculated spectrum and the reference magnesium stearate spectrum was equal to 0.87 and reached 0.90 after a using 50 principal components to reproduce the matrix. As it is shown in Table 2, where MCR-ALS was applied on a $\mathbf{D}_{\mathbf{PCA}_{(\mathbf{n},\mathbf{p})}}$ built with k = 5, 10, 15, 20, 50, the results of the two active principal ingredients and

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

297 the two major excipients were not modified. In this case, using less than 20 components to build the
298 matrix $\mathbf{D}_{\mathbf{PCA}(\mathbf{n},\mathbf{p})}$ lose the magnesium stearate contribution.

299 In order to keep the maximum information, the initial preprocessed $\mathbf{D}_{(\mathbf{n},\mathbf{p})}$ matrix (i.e. the PCA non-
300 filtered dataset) was used to start the iterative MCR-ALS process. Non-negativity and local rank
301 constraints on concentrations were applied. Distribution maps were shown in Figure 10. The
302 optimum was reached after 3 iterations, with a lack of fit equal to 14.7 and a percentage of variance
303 explained equal to 97.8.

304 Correlations between calculated spectra and API1, AP12, lactose and avicel were respectively equal
305 to 0.98, 0.97, 0.99 and 0.95. Distributions and contributions of the different constituents were then
306 displayed in Figure 10. Major excipients (lactose and cellulose) are identified across the whole image
307 in distribution maps 3 and 4. Agglomerates of API 1 and API 2 were highlighted in the top left and
308 right distribution maps. The correlation between the calculated spectrum and the magnesium
309 stearate reference was equal to 0.90 (Figure 11). By using a non-filtered PCA matrix with appropriate
310 constraints, the information linked to the low dose constituent was extracted. The non-filtering
311 option can be the choice when there are no references that can indicate in an objective manner the
312 number of PCs necessary to include a minor constituent. As shown in Figure 10, only few pixels of the
313 image contained the lubricant ($\mathbf{C}_{\mathbf{opt.5}}$), which could be explained by its low concentration within the
314 drug product.

315

316 *3.2.3. Pure spectrum augmented matrix*

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

The preprocessed data matrix was column-wise augmented to form a multiset structure including the magnesium stearate preprocessed pure spectrum [36]. For this type of matrix augmentation, the bilinear model can be written as:

$$\begin{pmatrix}\mathbf{D_1}\\ \mathbf{D_2}\end{pmatrix} = \begin{pmatrix}\mathbf{C_1}\\ \mathbf{C_2}\end{pmatrix}.\mathbf{S^T} + \begin{pmatrix}\mathbf{E_1}\\ \mathbf{E_2}\end{pmatrix} = \mathbf{C_{augm}}.\mathbf{S^T} + \mathbf{E_{augm}} \quad (6)$$

where $\mathbf{S^T}$ is the pure spectral matrix of the different compounds present in the considered preprocessed $\mathbf{D_1}$ data matrix and the augmented $\mathbf{D_2}$ pure spectrum matrix. In these two matrices, the chemical compounds have to be the same, but their concentration profiles can be different. Non negativity of concentration and local rank constraints were applied on the data as it was described in section 3.2.1. In multiset analysis, a new constraint based on correspondence among species can be used. This constraint fixes the presence or absence of components in concentration matrix, always taking into account the sequence of components in the initial estimates to encode the information on presence/absence correctly. This presence or absence information is coded in binary format and introduced into the MCR algorithm. For $\mathbf{D_1}$, the correspondence among species vector was fixed to [1, 1, 1, 1, 1] as each constituent was supposed to be in the drug product whereas, for $\mathbf{D_2}$, only one value corresponding to the lubricant was fixed to 1, corresponding to the vector [0, 0, 0, 0, 1] (Note that this code is valid as long as MgSt is the fifth profile in the spectral estimates used in the MCR analysis). When a particular component is not present in a concentration matrix, the elements in the related profile are set to zero. This type of constraint contributes significantly to the elimination of rotational ambiguities.

By adding information of the low dose constituent in the matrix, the PCA reduction of the multiset provides a different model, which ensures the extraction of the lubricant information. The MCR-ALS can then be performed as usual, by using a first step of PCA reduction with 5 components. The

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

optimum was reached after 6 iterations, with a lack of fit equal to 8.3 (with respect to $\mathbf{D}_{\mathbf{PCA}_{(\mathbf{n},\mathbf{p})}}$) and 16.1 (with respect to $\mathbf{D}_{(\mathbf{n},\mathbf{p})}$) and a percentage of variance explained equal to 97.4%.

Correlations between calculated MCR-ALS $\mathbf{S}_{\mathbf{opt}}$ spectra and the five reference spectra were respectively equal to 0.98, 0.96, 0.99, 0.95 and 0.99 (Table 3) which ensure an appropriate resolution of the studied system. In Figure 12, distributions of API1, API2 and the two main excipients were in accordance with the previous results obtained from MCR-ALS on a PCA filtered or non-filtered dataset. However, because of the high correlation between the calculated $\mathbf{S}_{\mathbf{opt.5}}$ spectrum and the magnesium stearate reference spectrum, the distribution of the lubricant can be easily observed in the $\mathbf{C}_{\mathbf{opt.5}}$ distribution map. As for the PCA non-filtered approach, only few pixels were highlighted with the lubricant contribution, which could be explain by its low concentration within the drug product.

4. Conclusion

MCR-ALS was applied on Raman Chemical images in order to study the distribution of actives and excipients within a pharmaceutical drug product. This article was focused on the identification of a low dose constituent within a formulation. Three different approaches were tested. First, MCR-ALS was performed on a PCA reduced dataset built by using a number of components equal to the number of constituents within the formulation. Due to the low spectral variability of the lubricant, the PCA reduction did not extract the corresponding information and the MCR-ALS process was not able to find out this product. However, distribution of actives and major excipients were in accordance with the known formulation. In order to ensure the conservation of the low dose constituent contribution within the dataset, a sequential PCA reduction process was tested. For each iteration, a new PCA reduced dataset was generated (from 5 to 100 components) and used for MCR-

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

ALS calculations. It was shown that the lubricant information was not present in the iterative MCR-ALS process unless 20 components were used. From a PCA non-filtered dataset, the magnesium stearate distribution was detected by using appropriate non-negativity and local rank constraint. Results showed the distribution of the five constituents with high correlations between the calculated signals and the pure reference spectra. Finally, the initial preprocessed dataset was column-wise augmented with magnesium stearate preprocessed pure spectrum. By using a correspondence among species constraint properly defined, the PCA reduction of the matrix kept the lubricant information and then, the decomposition of the Raman chemical image provided high correlated calculated spectra with reference and well-defined actives and excipients distribution map.

This study demonstrates the ability of MCR-ALS to extract the contribution of a low constituent of a solid drug product from Raman hyperspectral images. The choice of appropriate pre-processing methods, constraints, data structures used and modus operandi was important to reach the objective. Raman Chemical images, known as a useful tool to study the distribution of compounds in a solid drug product, might be used to study the distribution of low dose constituents as a lubricant, an impurity or a crystalline form transformation.

References

[1] S. Kazarian, J. Higgins, A closer look at polymers, Chem. Ind. 10 (2002) 21-23.

[2] S. Koljenović, T.C. Bakker Schut, R. Wolthuis, B. de Jong, L. Santos, P.J. Caspers, J.M. Kros, G.J. Puppels, Tissue characterization using high wave number Raman spectroscopy, J. Biomed. Opt. 10 (2005) 031116.

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

[3] X. Zhang, R Tauler, Application of Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) to remote sensing hyperspectral imaging, Anal. Chim. Acta 762 (2013) 25-38.

[5] M. Bautista, J. Cruz, M. Blanco, Study of component distribution in pharmaceutical binary powder mixtures by near infrared chemical imaging, J. Spectral Imaging 3 (2012) 1-9.

[6] S.A. Schönbichler, L.K.H. Bittner, A.K.H. Weiss, U.J. Griesser, J.D. Pallua, C.W. Huck, Comparison of NIR chemical imaging with conventional NIR, Raman and ATR-IR spectroscopy for quantification of furosemide crystal polymorphs in ternary powder mixtures, Eur. J. Pharm. Biopharm. 84 (2013) 616-625.

[8] K. Kwok, L.S. Taylor, Analysis of Cialis® tablets using Raman microscopy and multivariate curve resolution, J. Pharm. Biomed. Anal. 66 (2012) 126-135.

[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] B. Vajna, G. Patyi, Z. Nagy, A. Bodis, A. Farkas, G. Marosi, Comparison of chemometric methods in the analysis of pharmaceuticals with hyperspectral Raman imaging, J. Raman Spectrosc. 42 (2011) 1977-1986.

[12] F. Clarke, Extracting process-related information from pharmaceutical dosage forms using near infrared microscopy, Vib. Spectrosc. 34 (2004) 25–35.

411 [13] S. Šašić, D.A. Clark, Defining a strategy for chemical imaging of industrial pharmaceutical samples
412 on Raman line-mapping and global illumination instruments, Appl. Spectrosc. 60 (2006) 494-502.
413
414
415 [15] J. Burger, P. Geladi, Hyperspectral NIR image regression part II: dataset preprocessing
416 diagnostics, J. Chemom. 20 (2006) 106-119.
417
418 [16] T. Furukawa, H. Sato, H. Shinzawa, I. Noda, S. Ochiai, Evaluation of homogeneity of binary blends
419 of poly(3-hydroxybutyrate) and poly(L-lactic acid) studied by near infrared chemical imaging (NIRCI),
420 Anal. Sci. 23 (2007) 871–876.
421
422 [17] S. Piqueras, L. Duponchel, R. Tauler, A. de Juan, Resolution and segmentation of hyperspectral
423 biomedical images by Multivariate Curve Resolution-Alternating Least Squares, Anal. Chim. Acta 705
424 (2011) 182-192.
425
426 [18] A. de Juan, R. Tauler, R. Dyson, C. Marcolli, M. Rault, M. Maeder, Spectroscopic imaging and
427 chemometrics: a powerful combination for global and local sample analysis, TrAc Trends Anal. Chem.
428 23 (2004) 70-79.
429
430 [19] C. Gendrin, Y. Roggo, C. Collet, Content uniformity of pharmaceutical solid dosage forms by near
431 infrared hyperspectral imaging: a feasibility study, Talanta 73 (2007) 733–741.
432
433
434 [21] H. Abdollahi, R. Tauler, Uniqueness and rotation ambiguities in Multivariate Curve Resolution
435 methods, Chemom. Intell. Lab. Syst. 108 (2011) 100-111.
436

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

[23] A. de Juan, R. Tauler, Chemometrics applied to unravel multicomponent processes and mixtures revisiting latest trends in multivariate resolution, Anal. Chim. Acta 500 (2003) 195-210.

[24] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB, Chemom. Intell. Lab. Syst. 76 (2005) 101-110.

[25] A. de Juan, M. Maeder, T. Hancewicz, L. Duponchel, R. Tauler, Chemometric tools for image analysis, in: R. Salzer and H.W. Siesler (Eds.), Infrared and Raman spectroscopic imaging, Wiley, 2009, pp. 65-109.

[26] E. Lee, Raman spectral imaging on pharmaceutical products, in: R. Salzer and H.W. Siesler (Eds.), Infrared and Raman spectroscopic imaging, Wiley, 2009, pp. 377-402.

[27] J. Wang, H. Wen, D. Desai, Lubrication in tablet formulations, Eur. J. Pharm. Biopharm. 75 (2010) 1–15.

[28] G. Post Sabin, A.M. de Souza, M.C. Breitkreitz, R.J. Poppi, Development of an algorithm for identification and correction of spikes in Raman imaging spectroscopy, Quim. Nova 35 (2012) 612-615.

[29] P. Eilers, Parametric Time Warping, Anal. Chem. 76 (2004) 404-411.

[30] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639.

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

[31] A. de Juan, M. Maeder, T. Hancewicz, R. Tauler, Use of local rank-based spatial information for resolution of spectroscopic images, J. Chemom. 22 (2008) 291-298.

[32] A. de Juan, M. Maeder, T. Hancewicz, R. Tauler, Local rank analysis for exploratory spectroscopic image analysis. Fixed Size Image Window-Evolving Factor Analysis, Chemom. Intell. Lab. Syst. 77 (2005) 64-74.

[33] W. Winding, J. Guilment, Interactive Self-Modeling Mixture Analysis, Anal. Chem. 63 (1991) 1425-1432.

[34] P. Geladi, H. Grahn, Multivariate image analysis in chemistry and related areas: chemometrics image analysis, John Wiley & Sons, Chichester, 1996.

[35] J.M. Amigo, J. Cruz, M. Bautista, S. Maspoch, J. Coello,M. Blanco, Study of pharmaceutical samples by NIR chemical-image and multivariate analysis, TrAc Trends Anal. Chem. 27 (2008) 696-713.

[36] R. Tauler, M. Maeder, A. de Juan, Multiset data analysis: Extended Multivariate Curve resolution, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive chemometrics, Elsevier, 2009, pp. 473-505.

Figure captions

Figure 2 : Preprocessed Raman spectra (AsLS and first derivative)

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

Figure 3: PCA scores: five first components associated with their explained variances. Different distributions and agglomerates were highlighted. PC1 and PC5 were linked to the lactose variability, while PC2, PC3 and PC4 were respectively linked to the distributions of API1, avicel and API2.

Figure 4: Cumulative variance explained of the PCA decomposition. From PC6, the variance contained in the principal components was lower than 0.2% of the total variance and reached a plateau of 0.02% of variance explained per component.

Figure 5: Singular values plot (top: non-sorted singular values, bottom: sorted singular values)

Figure 6: Local rank map obtained by choosing an appropriate threshold which separates significant singular values from noise.

Figure 7: $\mathbf{C_{sel}}$ matrix (Orange: absence of the constituent, White: presence of the constituent)

Figure 9: Highest correlation between the calculated spectra ($\mathbf{S_{opt}}$) and the reference spectrum of magnesium stearate (for each iteration of a PCA filtered matrix built from 5 to 100 components)

Figure 10: Distribution maps of drug substance constituents (PCA non-filtered dataset)

Figure 11: $\mathbf{S_{opt}}$ versus reference spectrum of magnesium stearate

Figure 12: Distribution maps of drug substance constituents (augmented matrix approach)

Table 1: Correlation between MCR-ALS calculated $\mathbf{S_{opt}}$ and the reference spectra (PCA filtered dataset)

Author-produced version of the article published in Journal of pharmaceutical and biomedical analysis, 2015, N°103, p. 35-43
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.jpba.2014.10.024

516 **Table 2:** MCR-ALS results according to the number of components used to build the PCA reduced
517 $\mathbf{D}_{\mathbf{PCA}_{(\mathbf{n},\mathbf{p})}}$ matrix

518

519 **Table 3:** Correlation between MCR-ALS $\mathbf{S}_{\mathbf{opt}}$ and the reference spectra (column-wise augmented
520 dataset)

521