



HAL
open science

Distribution of a low dose compound within pharmaceutical tablet by using multivariate curve resolution on raman hyperspectral images

Mathieu Boiret, A. de Juan, N. Gorretta, Y.M. Ginot, Jean-Michel Roger

► To cite this version:

Mathieu Boiret, A. de Juan, N. Gorretta, Y.M. Ginot, Jean-Michel Roger. Distribution of a low dose compound within pharmaceutical tablet by using multivariate curve resolution on raman hyperspectral images. *Journal of Pharmaceutical and Biomedical Analysis*, 2015, 103, pp.35-43. 10.1016/j.jpba.2014.10.024 . hal-01465724

HAL Id: hal-01465724

<https://hal.science/hal-01465724>

Submitted on 13 Feb 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

3 Mathieu Boiret^{a*}, Anna de Juan^b, Nathalie Gorretta^c, Yves-Michel Ginot^a, Jean-Michel Roger^c

4
5 ^aTechnologie Servier, Orléans, France

6 ^bGrup de Quimiometria, Dept. Química Analítica, Universitat de Barcelona, Spain

7 ^cIRSTEA, UMR ITAP 361, Montpellier, France

8
9 *Corresponding author:

10 Tel.: +33 238 238 175

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

12
13 **Abstract**

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

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

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

37 **1. Introduction**

38 In the last decade, the use of imaging coupled with vibrational spectroscopies (near infrared, mid
39 infrared, fluorescence and Raman) has grown quickly in research and development environments.
40 The spatial and spectral information contained in hyperspectral images can be associated with the
41 distribution of the different constituents within the sample. Different areas such as polymer research
42 [1], biomedical analysis [2], environment field [3] and pharmaceutical development [5] are using
43 these new analytical tools based on vibrational hyperspectral imaging. During the analytical lifecycle
44 of a pharmaceutical drug product, hyperspectral imaging became a very powerful technique to
45 explore the compound distributions on the tablet surface or within a powder mixture [6]. This
46 technology appeared as innovative and promising to ensure the final quality of the drug product [8]
47 from the development to the production.

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

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

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

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

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

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

103 **2. Materials and Methods**

104

105 *2.1. Samples*

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

116 *2.2. Raman imaging system*

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

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

125 *2.3. Pre-processing*

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

134 *2.4. Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)*

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

$$140 \quad \mathbf{D} = \mathbf{C}\mathbf{S}^T + \mathbf{E} \quad (1)$$

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

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

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

164
$$\mathbf{Iof}(\%) = 100 \sqrt{\frac{\sum_{i,j} \mathbf{e}_{i,j}^2}{\sum_{i,j} \mathbf{D}_{i,j}^2}} \quad (2)$$

165
$$\mathbf{R}^2 = \frac{\sum_{i,j} \mathbf{D}_{i,j}^2 - \sum_{i,j} \mathbf{e}_{i,j}^2}{\sum_{i,j} \mathbf{D}_{i,j}^2} \quad (3)$$

166 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
167 after using the MCR-ALS model (see equation 1). Input element can be the original element from
168 $\mathbf{D}_{(n,p)}$ or the element of a noise filtered PCA matrix $\mathbf{D}_{\text{PCA}(n,p)}$ using the same number of components
169 as in the MCR-ALS. A noise filtered PCA matrix $\mathbf{D}_{\text{PCA}(n,p)}$ can be obtained as follows:

170
$$\mathbf{D}_{\text{PCA}(n,p)} = \mathbf{U}_{(n,k)} \mathbf{S}_{(k,k)} \mathbf{V}_{(k,p)}^T \quad (4)$$

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

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

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

191 **3. Results and discussion**

192

193 *3.1. Exploratory analysis*

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

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

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

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

$$217 \quad \mathbf{Var}_i = 100 \times \frac{\sum_{i,j} (c_i s_i^T)^2}{\sum_{i,j} (c s^T)^2} \quad (5)$$

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

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

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

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

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

236 3.2. MCR-ALS

237

238 3.2.1. *Non-negativity and local rank constraints*

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

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

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

263 3.2.2. Effect of PCA filtering on MCR-ALS results

264 In all cases, MCR-ALS was applied on the preprocessed data by using the constraints previously
265 described (see section 3.2.1). The initial preprocessed matrix was reduced (noise-filtering) by using
266 the five first vectors of the PCA decomposition of $\mathbf{D}_{(n,p)}$. MCR-ALS on the filtered PCA matrix
267 provides an optimum value after 9 iterations. 97.9% of the variance was explained with a lack of fit
268 calculated on the initial $\mathbf{D}_{(n,p)}$ and the reduced $\mathbf{D}_{PCA(n,p)}$ matrices respectively equal to 14.7 and
269 7.9. Correlation coefficients between calculated spectra and reference spectra were displayed in
270 Table 1. The four first calculated spectra were highly correlated to the two actives and the two major
271 excipients whereas the fifth component was not correlated to the magnesium stearate or to other
272 constituents.

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

281 In order to improve the MCR-ALS results and to extract magnesium stearate contribution, MCR-ALS
282 analysis on a PCA-filtered matrix including progressively a larger number of principal components
283 was tested. MCR-ALS decomposition was performed by using a PCA-filtered $\mathbf{D}_{\text{PCA}(n,p)}$ matrix using an
284 increasing number of components, from 5 to the total number of variables. For the first iteration, the
285 $\mathbf{D}_{\text{PCA}(n,p)}$ matrix was built by using the five first vector of the PCA reduction. The following MCR-ALS
286 calculation was performed by adding an additional principal component to calculate the $\mathbf{D}_{\text{PCA}(n,p)}$
287 matrix. This process was repeated until the number of principal components was equal to the
288 number of variables, corresponding to the use of the preprocessed non-filtered initial $\mathbf{D}_{(n,p)}$ matrix.
289 For each MCR-ALS decomposition from 5 to 100 components, the highest correlation coefficient
290 between the resolved spectra and the pure reference spectrum of magnesium stearate is displayed
291 (Figure 9).

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

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

299 In order to keep the maximum information, the initial preprocessed $\mathbf{D}_{(n,p)}$ 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 ($\mathbf{C}_{opt.5}$), which could be explained by its low concentration within the
314 drug product.

315

316 3.2.3. *Pure spectrum augmented matrix*

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

$$320 \quad \begin{pmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \end{pmatrix} \cdot \mathbf{S}^T + \begin{pmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \end{pmatrix} = \mathbf{C}_{\text{augm}} \cdot \mathbf{S}^T + \mathbf{E}_{\text{augm}} \quad (6)$$

321 where \mathbf{S}^T is the pure spectral matrix of the different compounds present in the considered
322 preprocessed \mathbf{D}_1 data matrix and the augmented \mathbf{D}_2 pure spectrum matrix. In these two matrices,
323 the chemical compounds have to be the same, but their concentration profiles can be different. Non
324 negativity of concentration and local rank constraints were applied on the data as it was described in
325 section 3.2.1. In multiset analysis, a new constraint based on correspondence among species can be
326 used. This constraint fixes the presence or absence of components in concentration matrix, always
327 taking into account the sequence of components in the initial estimates to encode the information
328 on presence/absence correctly. This presence or absence information is coded in binary format and
329 introduced into the MCR algorithm. For \mathbf{D}_1 , the correspondence among species vector was fixed to [
330 1, 1, 1, 1, 1] as each constituent was supposed to be in the drug product whereas, for \mathbf{D}_2 , only one
331 value corresponding to the lubricant was fixed to 1, corresponding to the vector [0, 0, 0, 0, 1] (Note
332 that this code is valid as long as MgSt is the fifth profile in the spectral estimates used in the MCR
333 analysis). When a particular component is not present in a concentration matrix, the elements in the
334 related profile are set to zero. This type of constraint contributes significantly to the elimination of
335 rotational ambiguities.

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

339 optimum was reached after 6 iterations, with a lack of fit equal to 8.3 (with respect to $\mathbf{D}_{\text{PCA}(n,p)}$) and
340 16.1 (with respect to $\mathbf{D}_{(n,p)}$) and a percentage of variance explained equal to 97.4%.

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

350

351 **4. Conclusion**

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

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

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

378 **References**

379 [1] S. Kazarian, J. Higgins, A closer look at polymers, *Chem. Ind.* 10 (2002) 21-23.

380

381 [2] S. Koljenović, T.C. Bakker Schut, R. Wolthuis, B. de Jong, L. Santos, P.J. Caspers, J.M. Kros, G.J.
382 Puppels, Tissue characterization using high wave number Raman spectroscopy, *J. Biomed. Opt.* 10
383 (2005) 031116.

384

385 [3] X. Zhang, R Tauler, Application of Multivariate Curve Resolution Alternating Least Squares (MCR-
386 ALS) to remote sensing hyperspectral imaging, *Anal. Chim. Acta* 762 (2013) 25-38.

387

388 [5] M. Bautista, J. Cruz, M. Blanco, Study of component distribution in pharmaceutical binary powder
389 mixtures by near infrared chemical imaging, *J. Spectral Imaging* 3 (2012) 1-9.

390

391 [6] S.A. Schönbichler, L.K.H. Bittner, A.K.H. Weiss, U.J. Griesser, J.D. Pallua, C.W. Huck, Comparison of
392 NIR chemical imaging with conventional NIR, Raman and ATR-IR spectroscopy for quantification of
393 furosemide crystal polymorphs in ternary powder mixtures, *Eur. J. Pharm. Biopharm.* 84 (2013) 616-
394 625.

395

396

397 [8] K. Kwok, L.S. Taylor, Analysis of Cialis® tablets using Raman microscopy and multivariate curve
398 resolution, *J. Pharm. Biomed. Anal.* 66 (2012) 126-135.

399

400

401 [10] C. Gendrin, Y. Roggo, C. Collet, Pharmaceutical applications of vibrational chemical imaging and
402 chemometrics: A review, *J. Pharm. Biomed. Anal.* 48 (2008) 533–553.

403

404 [11] B. Vajna, G. Patyi, Z. Nagy, A. Bodis, A. Farkas, G. Marosi, Comparison of chemometric methods
405 in the analysis of pharmaceuticals with hyperspectral Raman imaging, *J. Raman Spectrosc.* 42 (2011)
406 1977-1986.

407

408 [12] F. Clarke, Extracting process-related information from pharmaceutical dosage forms using near
409 infrared microscopy, *Vib. Spectrosc.* 34 (2004) 25–35.

410

- 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

- 437
- 438 [23] A. de Juan, R. Tauler, Chemometrics applied to unravel multicomponent processes and mixtures
439 revisiting latest trends in multivariate resolution, *Anal. Chim. Acta* 500 (2003) 195-210.
- 440
- 441 [24] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, A graphical user-friendly interface for MCR-ALS: a
442 new tool for multivariate curve resolution in MATLAB, *Chemom. Intell. Lab. Syst.* 76 (2005) 101-110.
- 443
- 444 [25] A. de Juan, M. Maeder, T. Hanczewicz, L. Duponchel, R. Tauler, Chemometric tools for image
445 analysis, in: R. Salzer and H.W. Siesler (Eds.), *Infrared and Raman spectroscopic imaging*, Wiley, 2009,
446 pp. 65-109.
- 447
- 448 [26] E. Lee, Raman spectral imaging on pharmaceutical products, in: R. Salzer and H.W. Siesler (Eds.),
449 *Infrared and Raman spectroscopic imaging*, Wiley, 2009, pp. 377-402.
- 450
- 451 [27] J. Wang, H. Wen, D. Desai, Lubrication in tablet formulations, *Eur. J. Pharm. Biopharm.* 75 (2010)
452 1–15.
- 453
- 454 [28] G. Post Sabin, A.M. de Souza, M.C. Breitzkreitz, R.J. Poppi, Development of an algorithm for
455 identification and correction of spikes in Raman imaging spectroscopy, *Quim. Nova* 35 (2012) 612-
456 615.
- 457
- 458 [29] P. Eilers, Parametric Time Warping, *Anal. Chem.* 76 (2004) 404-411.
- 459
- 460 [30] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares
461 procedures, *Anal. Chem.* 36 (1964) 1627–1639.
- 462

463 [31] A. de Juan, M. Maeder, T. Hanczewicz, R. Tauler, Use of local rank-based spatial information for
464 resolution of spectroscopic images, *J. Chemom.* 22 (2008) 291-298.

465
466 [32] A. de Juan, M. Maeder, T. Hanczewicz, R. Tauler, Local rank analysis for exploratory spectroscopic
467 image analysis. Fixed Size Image Window-Evolving Factor Analysis, *Chemom. Intell. Lab. Syst.* 77
468 (2005) 64-74.

469
470 [33] W. Winding, J. Guilment, Interactive Self-Modeling Mixture Analysis, *Anal. Chem.* 63 (1991)
471 1425-1432.

472
473 [34] P. Geladi, H. Grahn, Multivariate image analysis in chemistry and related areas: chemometrics
474 image analysis, John Wiley & Sons, Chichester, 1996.

475
476 [35] J.M. Amigo, J. Cruz, M. Bautista, S. Maspocho, J. Coello, M. Blanco, Study of pharmaceutical
477 samples by NIR chemical-image and multivariate analysis, *TrAc Trends Anal. Chem.* 27 (2008) 696-
478 713.

479
480 [36] R. Tauler, M. Maeder, A. de Juan, Multiset data analysis: Extended Multivariate Curve resolution,
481 in: S.D. Brown, R. Tauler, B. Walczak (Eds.), *Comprehensive chemometrics*, Elsevier, 2009, pp. 473-
482 505.

483
484 **Figure captions**

485
486
487 **Figure 2** : Preprocessed Raman spectra (AsLS and first derivative)

488

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

492

493 **Figure 4:** Cumulative variance explained of the PCA decomposition. From PC6, the variance contained
494 in the principal components was lower than 0.2% of the total variance and reached a plateau of
495 0.02% of variance explained per component.

496

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

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

500

501 **Figure 7:** C_{sel} matrix (Orange: absence of the constituent, White: presence of the constituent)

502

503

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

506

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

508

509 **Figure 11:** S_{opt} versus reference spectrum of magnesium stearate

510

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

512

513 **Table 1:** Correlation between MCR-ALS calculated S_{opt} and the reference spectra (PCA filtered
514 dataset)

515

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

518

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

521