Received: 26 July 2016 ■ *Revised: 16 November 2016* ■ *Accepted: 16 November 2016* ■ *Publication: 1 December 2016*

JOURNAL OF SPECTRAL IMAGING

Paper Presented at IASIM 2016, July 2016, Chamonix, France

openaccess

Weighted fuzzy clustering for (fuzzy) constraints in multivariate image analysis–alternating least square of hyperspectral images

Siewert Hugelier,[a,*] Patrizia Firmani,[b] Olivier Devos,[a] Myriam Moreau,[a] Christel Pierlot,[c] Federico Marini[b] and Cyril Ruckebusch[a]

[a]Université de Lille, Sciences et Technologies, LASIR, CNRS, Villeneuve d'Ascq Cedex F-59655, France.
E-mail: siewert.hugelier@ed.univ-lille1.fr

[b]Dipartimento di Chimica, Università di Roma "La Sapienza", I-00185 Rome, Italy

[c]Université de Lille, ENSCL, UMR 8181–UCCS–Unité de Catalyse et Chimie du Solide, Villeneuve d'Ascq Cedex F-59655, France

In order to investigate hyperspectral images, many techniques such as multivariate image analysis (MIA) or multivariate curve resolution–alternating least squares (MCR–ALS) can be applied. When focusing on the use of MCR–ALS, constraints are of the utmost importance for a correct resolution of the data into its individual contributions. In this article, a fuzzy clustering pattern recognition method (fuzzy C-means) is applied on experimental data in order to improve the results obtained within the MCR–ALS analysis. The big advantage of a fuzzy clustering technique over a hard clustering technique, such as k-means, is that the algorithm determines the probability of a pixel to be assigned to a component, indicating that a pixel can be part of multiple clusters (or components). This is, of course, an important property for dealing with data in which a lot of overlap between the components in the spatial direction occurs. This article deals briefly with the implementation of the constraint into the MCR–ALS algorithm and then shows the application of the constraint on an oil-in-water emulsion obtained by Raman spectroscopy, in which the different components can be decomposed in a clearer way and the interface between the oil and water bubbles becomes more visible.

Keywords: MCR–ALS, constraint, hyperspectral, fuzzy clustering, fuzzy C-means, oil-in-water emulsion, Raman spectroscopy

Introduction

To investigate the spatial distribution of individual components present in a complex sample, one usually falls back on using spectral or hyperspectral imaging (HSI) techniques, and analysing the data with multivariate image analysis (MIA)[1,2] or multivariate curve resolution–alternating least squares (MCR–ALS).[3,4] A key factor in the correct resolution of a mixture by using MCR–ALS is the application of constraints, which limit the number of possible solutions to the problem. For hyperspectral imaging data, many of the constraints available for traditional process data (e.g. unimodality, closure etc.) cannot be applied as there is an unfolding step of the hyperspectral image cube to the data matrix **D**. We, therefore, propose the implementation and application of a constraint based on a clustering technique enabling the improvement of the MCR–ALS resolution of the (hyperspectral) data. The purpose of the constraint is to classify the different data elements (further referred to as pixels for readability) to the correct components.

ISSN: 2040-4565
doi: 10.1255/jsi.2016.a7

Clustering techniques are pattern recognition methods and have been widely applied in chemistry for data analysis.[5,6] The method used here is the fuzzy C-means (FCM) clustering algorithm.[7,8] Contrary to a hard clustering technique, such as k-means,[9] where a pixel is assigned to one and only one cluster, a fuzzy clustering technique assigns membership levels of the different pixels to the clusters. Most obviously, within the MCR–ALS framework, the number of clusters is the same as the number of components. A parallel can be made between fuzzy clustering and the local rank technique.[10] In local rank, the rank of pixels is estimated beforehand and this information is then used to determine the presence or absence of the pixels in the final component distribution maps. In fuzzy clustering, one pixel can belong to multiple clusters, which can be seen as a different version of the local rank constraint.

Within this article, we focus on the explanation of the implementation and application of the fuzzy C-means clustering technique within the MCR–ALS analysis and use hyperspectral images of an oil-in-water sample to demonstrate the technique.

Theory: MCR–ALS with fuzzy clustering

Even though many detailed explanations can be found about MCR–ALS,[1,11,12] a brief overview is provided in order to explain the basic concept of the technique and understand some of its limitations. MCR–ALS can be applied to (spectroscopic) data that follow an approximation of the Beer–Lambert law, and thus can be decomposed into a bilinear model of pure signal contributions, as shown in Equation (1):

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \tag{1}$$

where $\mathbf{D}$ ($m \times n$) represents the data matrix of the multicomponent system, $\mathbf{C}$ ($m \times k$) contains the concentration profiles of the k components, $\mathbf{S}^{\mathbf{T}}$ ($k \times n$) the corresponding spectral profiles of the k components and $\mathbf{E}$ ($m \times n$) is the matrix that expresses the residuals of the model (i.e. mainly noise). For HSI data, the bilinear model still holds, but now for every data pixel. This means that the original data cube has to be unfolded into a data matrix, from the original spatial–spectral dimensions to a two-way data matrix.[2,13]

The decomposition problem described in Equation (1), in which we search for $\mathbf{C}$ and $\mathbf{S}^{\mathbf{T}}$, is an inverse problem, and thus no unique solutions can be found. Many possible solutions have the same quality of fit,[14] leading to an uncertainty in the solutions which we refer to as ambiguity.[3] In order to limit the extent of these ambiguities and obtain reliable and meaningful results, constraints are added to the optimisation process to reduce the possible solutions for $\mathbf{C}$ and $\mathbf{S}^{\mathbf{T}}$. Note that these constraints are often derived from the physical nature of the system. In this paper, a fuzzy clustering constraint is added to aid in the resolution of the data. The algorithm used for this purpose is the fuzzy C-means approach.[7,8] It is a form of soft clustering in which each data point (pixel) can belong to more than one cluster. Cluster membership probabilities indicate the degree to which data points (pixels) belong to each cluster. As for any constraint, the constraint is applied at each iteration of the optimisation process. This has the advantage, over using it as a post-processing technique, that the classification membership probabilities are updated at each step of the optimisation process. These membership probabilities ("weights") are calculated by a minimisation of the objective function, $\mathbf{J_m}$, shown in Equation (2):

$$\mathbf{J_m} = \sum_{i=1}^{N}\sum_{j=1}^{C}\left(u_{ij}\right)^{m}\left\|x_i - c_j^2\right\|^2 \tag{2}$$

With m being a positive number between 1 and infinity, u_{ij} is the membership of x_i in cluster j [Equation (3)], x_i is a data pixel of the two-dimensional matrix $\mathbf{C}$, c_j is the centre of the cluster j [Equation (4)] and ||...|| the norm expressing the similarity between the measured data and the centre.

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_k\right\|}\right)^{\frac{2}{m-1}}} \tag{3}$$

$$c_j = \frac{\sum_{i=1}^{N}\left(u_{ij}\right)^{m}.x_i}{\sum_{i=1}^{N}\left(u_{ij}\right)^{m}} \tag{4}$$

This iterative minimisation process is stopped when it converges or, otherwise said, reaches a local minimum or saddle point of $\mathbf{J_m}$. The value of m determines the level of cluster fuzziness (when m is equal to 1, the membership converges to 0 or 1). Without prior information, m is usually chosen to be equal to 2 during the fuzzy C-means approach. The method differs from k-means by the addition of the membership value u_{ij} and the fuzzifier m.

The membership probabilities obtained by the fuzzy clustering algorithm (the values u_{ij}, obtained from the minimised objective function $\mathbf{J_m}$) are then used as weights for the concentration profiles (by element-wise multiplication) to obtain the new, constrained, concentration profiles ($\hat{\mathbf{C}}$). The overall MCR–ALS process with the fuzzy clustering constraint can be summarised in Figure 1. Additionally, we want to bring attention to the fact that the constraint is implemented in the MCR–ALS loop after the non-negativity constraint and can be applied on one or several of the components of the analysis.

Experimental and data

In order to create the oil-in-water emulsion sample, a thickener, consisting of a paraffin oil (Sigma-Aldrich, France) mix and water, was prepared and mixed together with octane (Sigma-Aldrich, France) in a 1 : 99% ratio. To stabilise this emulsion, two surfactants (span 60, Sigma-Aldrich, France and tween 60, Sigma-Aldrich, France) were added in a 10–90%

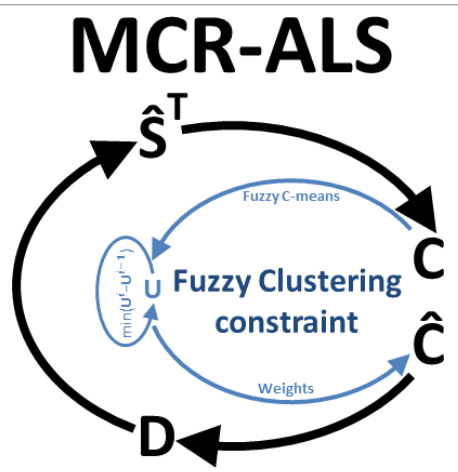

Figure 1. MCR–ALS process with a fuzzy clustering (fuzzy C-means) constraint. The weights, u_{ij}, obtained by the fuzzy clustering algorithm are used as weights on the concentration profiles, C, to obtain new constrained concentration profiles, $\hat{C}$.

ratio, respectively. The hyperspectral image of the sample was collected with a Horiba Scientific Labram HR Evolution Raman spectrometer over a range of 100–4000 cm^{-1}. To avoid degradation of the sample due to laser influence, a Linkam THM600 sample holder was used, which is a temperature-controlled microscope stage with an accuracy up to 0.01°C. The temperature was kept constant at room temperature (24°C). The spectra were recorded with a green laser (515 nm) for an acquisition time of 5 s and accumulation of 10 times with a spectral resolution of 1.9 cm^{-1}. The acquired image is 30 pixels × 30 pixels (spatial resolution of 0.9 µm, 100× objective). The mean image and the spectra can be seen in Figure 2. Note that the hyperspectral image has been limited to the fingerprint region, i.e. the spectral range between 675 cm^{-1} and 1550 cm^{-1} for data analysis.

All calculations were performed in MATLAB version R2014a (The MathWorks, Natick, MA, USA) using the MCR–ALS command line code (http://mcrals.wordpress.com/download/mcr-als-command-line/). The fuzzy clustering constraint has been implemented inside this MCR–ALS program.

Results and discussion

Initial exploratory analysis of the oil-in-water emulsion data reveals the presence of two main components, namely an aqueous phase and an oily phase. Additionally, we can also expect a third component, representing the interface between the two immiscible phases. This interface component can be important in understanding an underlying reaction taking place at the boundary between the two separate main components. Exchange of information can take place and involves the surfactants that were added to stabilise the emulsion. The constitution of this interface component and thus its spectral profile should be obtained free from other contributions so that this can be investigated. However, this phase, which is a minor component in the entire hyperspectral image, cannot be retrieved by MCR–ALS when using initial estimates obtained by SIMPLISMA.[15] It is only revealed when performing a two-component MCR–ALS analysis with non-negativity and inspecting the structure of the residuals (results not shown). Applying the fuzzy clustering constraint reveals that those interface pixels cannot be assigned to the aqueous phase or the oily phase. Using this information to our advantage, three component MCR–ALS analyses were performed, the results of which are presented in this manuscript. By using just non-negativity as a constraint during the analysis (on both **C** and $\mathbf{S}^T$), the results as shown in Figure 3a are obtained.

As can be seen, two main components, the oily phase and the aqueous phase, can be easily detected. The interface between these two phases can also be seen, but as it is a minor component, it is still mixed with contributions coming from the other phases. In order to improve the results obtained for the interface component, additional information should be given to the system. A possible solution would be to segment the data or to use a hard-clustering technique (e.g. k-means), but a drawback of the use of these methods is that a pixel is

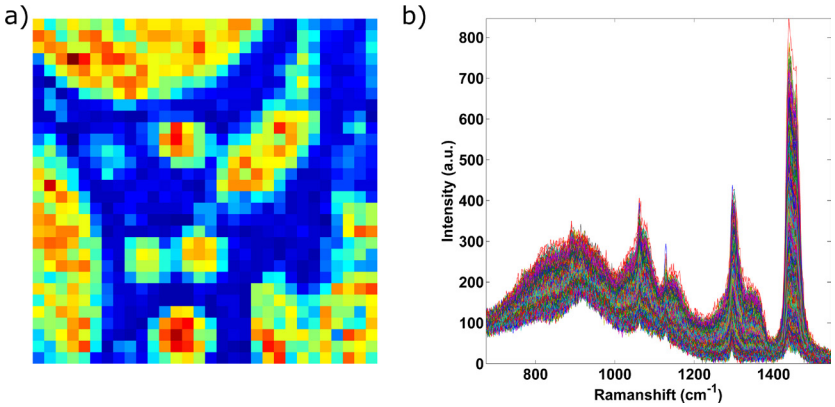

Figure 2. Raman hyperspectral image of an oil-in-water emulsion (30 pixels × 30 pixels × 500 spectral channels). (a) The mean image of the data; (b) the spectra obtained for every pixel for a range of 675–1550 cm^{-1}.

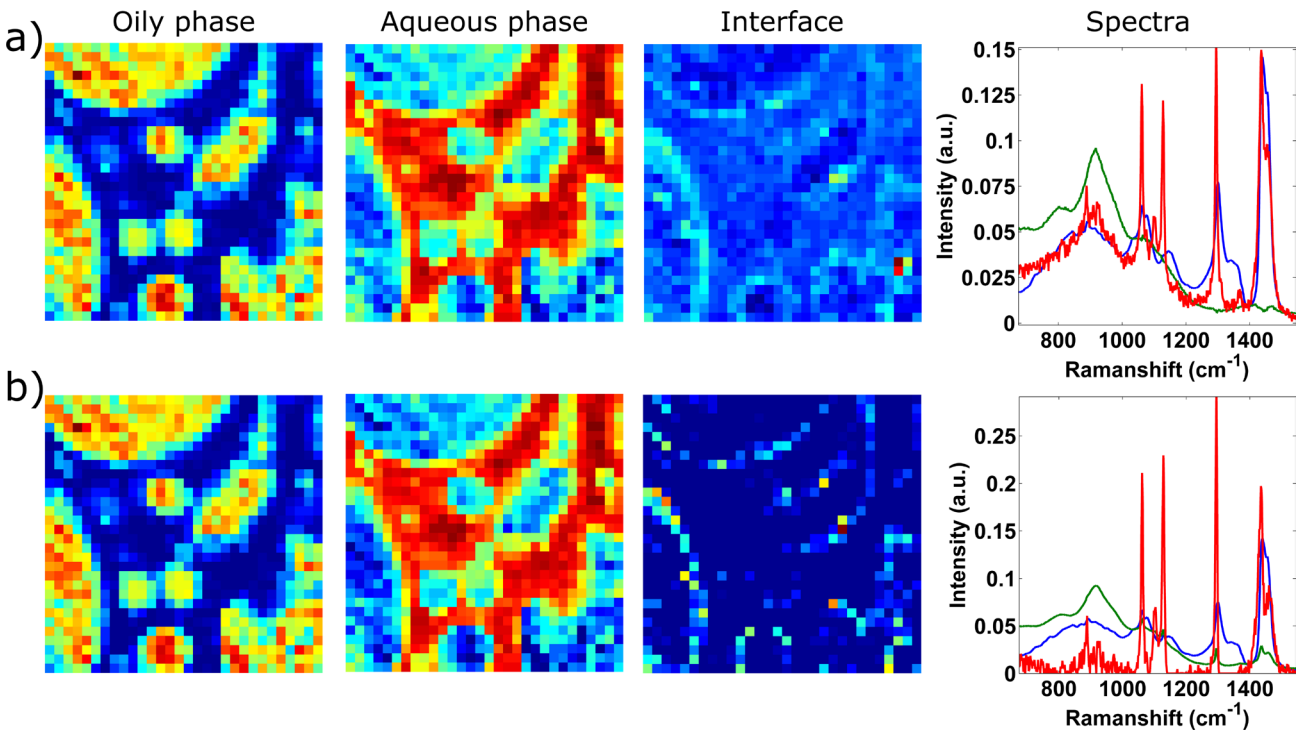

Figure 3. Distribution maps (intensity coded from blue to red) and spectral profiles for the components of the oil-in-water emulsion system (blue: oily phase, green: aqueous phase, red: interface). MCR–ALS results obtained using (a) non-negativity (LOF: 5.44%); (b) non-negativity and fuzzy clustering of component 3 (LOF: 5.56%).

assigned to one and only one component. It is thus impossible to have contributions of the three different components (with their respective spectra) in the same pixel. It can thus be said that these methods are too strict for a hyperspectral image, in which every pixel is a mixture of the different components. An alternative is thus to use a fuzzy clustering approach (i.e. fuzzy C-means), in which a pixel can have contributions from all the components. The fuzzy clustering constraint was applied onto the interface component during the MCR–ALS analysis. The results obtained are shown in Figure 3b. These results show not only a clear improvement in the selection of the interface component in the distribution map, but also that a better contrast in the component distribution maps for the other two components has been obtained (mainly visible for the aqueous component). The effect of the constraint is to enhance the contributions from pixels in which the interface component is strongly present and eliminate or reduce the ones from pixels in which the probability of presence of the interface component is small in a stepwise way (see Figure 4).

Looking at the spectral profiles, it should be clear that the interface component differs from the oily phase component in the spectral regions between 1100 cm^{-1} and 1150 cm^{-1} and

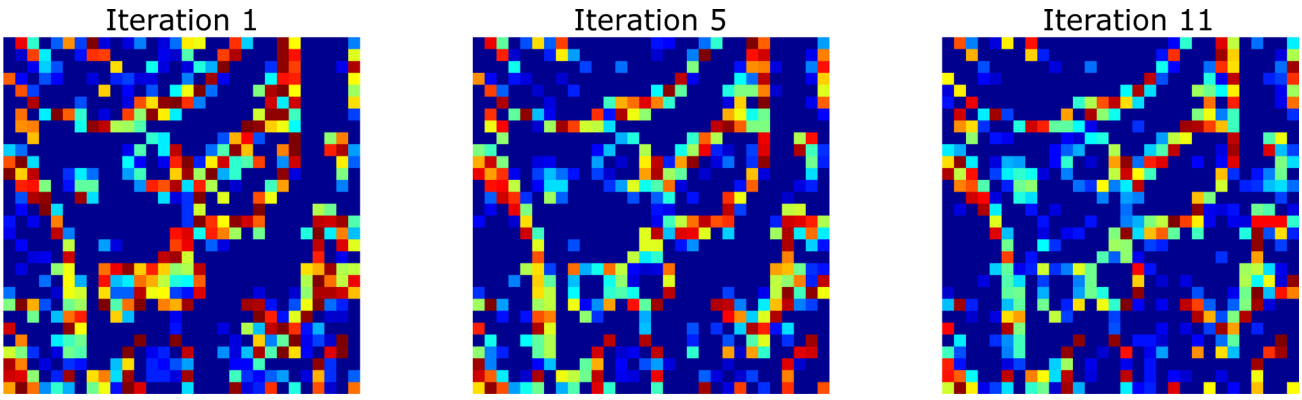

Figure 4. Evolution of the fuzzy clustering weights of the interface component during the MCR–ALS analysis (intensity coded from blue to red).

between 1300 cm^{-1} and 1400 cm^{-1}, and from the aqueous phase component in the peak at 1100 cm^{-1}, but mainly in intensity ratios, which explains why the interface is very hard to distinguish from the aqueous phase. Additionally, it should be clear from Figure 3a that when looking at the component distribution map of the interface component, several other pixels not belonging to the interface between the oily phase and aqueous phase are non-zero. That means that these pixels will be considered as "part of the interface" and thus influence the spectral profile associated with this component, leading to combination of the three different components. By using the proposed approach, we manage to gradually remove the pixels for which a low intensity was found (due to the clustering properties) as shown in Figure 3b, and obtain only the pixels that are truly part of the interface for this component. This will make the spectral profile for this component more pure and thus more correct (as there is no mixing anymore with the other components). This gradual optimisation of the component distribution map is what is shown in Figure 4. As can be seen here, due to the iterative process we apply in MCR–ALS, the weights obtained for this interface component change with each least squares step. At first, lots of other contributions are still present but it gradually optimises its result to finding only the interface component. The ALS procedure does the rest.

In conclusion, using a fuzzy clustering algorithm within the MCR–ALS analysis helped reveal this minor component of the system and shows its feasibility on this type of data. Additionally, supplementary material, containing the application of the fuzzy clustering approach on remote sensing data, is also available. It shows the feasibility of applying the approach not only on emulsion data, but also on other data where no interface components are present. For this data, the fuzzy approach was preferred over using k-means clustering to avoid the assignment of one pixel to one and only one cluster as lots of overlap between the different components in the spatial direction was present (see Supplementary material).

Conclusion

A continuous effort to find new constraints for the MCR–ALS analysis of hyperspectral images is necessary to facilitate the decomposition of them into their individual contributions. We have, therefore, implemented a fuzzy clustering constraint into the MCR–ALS algorithm that can be used, but not only, on hyperspectral images. The constraint enhances the pixels with a strong presence and removes contributions in which the probability of presence is low. We have then demonstrated the constraint on a hyperspectral image of an oil-in-water emulsion in which we successfully revealed a minor interface component, apart from two main contributions, which was not clear before. More generally, our experience is that the proposed alternative is useful at the interface between two components, i.e. where the specific spectral signature observed might results from their (non-linear) interaction.[16] This includes in remote sensing, for instance, borders between water and soil components.

Supplementary material

This paper contains Supplementary material.

References

1. K.H. Esbensen and P. Geladi, "Strategy of multivariate image analysis (MIA)", *Chemometr. Intell. Lab. Syst.* **7(1–2),** 67–86 (1989). doi: https://doi.org/10.1016/0169-7439(89)80112-1
2. J.M. Prats-Montalbán, A. de Juan and A. Ferrer, "Multivariate image analysis: a review with applications", *Chemometr. Intell. Lab. Syst.* **107,** 1–23 (2011). doi: https://doi.org/10.1016/j.chemolab.2011.03.002
3. R. Tauler, A. Smilde and B. Kowalski, "Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution", *J. Chemometr.* **9,** 31–58 (1995). doi: https://doi.org/10.1002/cem.1180090105
4. A. de Juan, R. Tauler, R. Dyson, C. Marcolli, M. Rault and M. Maeder, "Spectroscopic imaging and chemometrics: a powerful combination for global and local sample analysis", *Trends Anal. Chem.* **23(1),** 70–79 (2004). doi: https://doi.org/10.1016/S0165-9936(04)00101-3
5. R.G. Brereton, "Pattern recognition in chemometrics", *Chemometr. Intell. Lab. Syst.* **149,** part B, 90–96 (2015). doi: https://doi.org/10.1016/j.chemolab.2015.06.012
6. H. Parastar and A. Bazrafshan, "Fuzzy C-means clustering for chromatographic fingerprints analysis: A gas chromatography-mass spectrometry case study", *J. Chromatogr. A* **1438,** 236–243 (2016). doi: https://doi.org/10.1016/j.chroma.2016.02.049
7. J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA (1981).
8. J.C. Bezdek, R. Ehrlich and W. Full, "FCM: The fuzzy C-means clustering algorithm", *Comput. Geosci.* **10,** 191–203 (1984). doi: https://doi.org/10.1016/0098-3004(84)90020-7
9. J.B. MacQueen, "Some methods for classification and analysis of multivariate observations", *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.* Vol. 1 (1967). http://projecteuclid.org/euclid.bsmsp/1200512992
10. A. de Juan, M. Maeder, T. Hancewicz and R. Tauler, "Use of local rank-based spatial information for resolution of spectroscopic images", *J. Chemometr.* **22,** 291–298 (2008). doi: https://doi.org/10.1002/cem.1099
11. C. Ruckebusch and L. Blanchet, "Multivariate curve resolution: a review of advanced and tailored applications and challenges", *Anal. Chim. Acta* **765**, 28–36 (2007). doi: https://doi.org/10.1016/j.aca.2012.12.028

12. A. de Juan, J. Jaumot and R. Tauler, "Multivariate curve resolution (MCR). Solving the mixture analysis problem", *Anal. Meth.* **6,** 4964–4976 (2014). doi: https://doi.org/10.1039/C4AY00571F

13. S. Hugelier, O. Devos and C. Ruckebusch, "On the implementation of spatial constraints in multivariate curve resolution – alternating least squares for hyperspectral image analysis", *J. Chemometr.* **29,** 557–561 (2015). doi: https://doi.org/10.1002/cem.2742

14. R. Tauler and M. Maeder, "Two-way data analysis: multivariate curve resolution – error in curve resolution", in *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, Ed by S.D. Brown, R. Tauler and B. Walczak. Elsevier, Amsterdam, The Netherlands, p. 20 (2009). doi: https://doi.org/10.1016/b978-044452701-1.00051-x

15. W. Windig and J. Guilment, "Interactive self-modeling mixture analysis", *Anal. Chem.* **63,** 1425–1432 (1991). doi: https://doi.org/10.1021/ac00014a016

16. N. Dobigeon, Y. Altmann, N. Brun and S. Moussaoui, "Linear and nonlinear unmixing in hyperspectral imaging", in *Data Handling in Science and Technology: Resolving Spectral Mixtures*, Ed by C. Ruckebusch. Elsevier, Amsterdam, The Netherlands, p. 41 (2016). doi: https://doi.org/10.1016/b978-0-444-63638-6.00006-1

Supplementary material

Application on remote sensing data

The constraint has been applied on several other data sets than the one presented above. The data tested were mainly oil-in-water emulsions, in order to improve the resolution related to the interface component, and on remote sensing data. An additional example of the latter is given in Figure S1, which consists of airborne hyperspectral imaging (HSI) data, obtained over Moffet Field, CA, USA.[17] After an exploratory analysis, four different components could be distinguished, of which one was assigned to water and the three others are related to land. It is important to point out that the contributions of these different components can overlap, due to the presence of soil in shallow creeks or vegetation on land etc. That is why the application of the k-means (hard-) clustering method is not advised, as it leads to the assignment of a single pixel to one and only one cluster. We have, therefore, opted for a fuzzy clustering approach. As can be seen from Figure S1a, the MCR–ALS analysis was first carried out by only applying non-negativity on the component distribution maps and the spectral profiles [lack of fit (LOF): 4.96%]. We can see that there is still quite some overlap between the different components in places where it is not expected (confirmed by using the raw photo of the data).

We then carried out the MCR–ALS analysis by adding fuzzy C-means to the pool of constraints (on component 4), as shown in Figure S1b. We can notice that the LOF increased slightly, which is expected when an additional constraint is added. This result shows the removal of several water contributions in the distribution map of the fourth component, directly leading to an increase in intensity in the distribution map of component 2. Additionally, it can also be observed that by using the constraint on component 4, the distribution map of component 3 is influenced, as a decrease in the intensity of the water contributions is also observed.

This example thus shows that a fuzzy clustering approach can be useful when investigating data showing lots of overlap. It is flexible in its usage and also has a direct impact on the components to which it was not applied.

Supplementary reference

17. R.O. Green, M.L. Eastwood, C.M. Sarture, T.G. Chrien, M. Aronsson, B.J. Chippendale, J.A. Faust, B.E. Pavri, C.J. Chovit, M. Solis, M.R. Olah and O. Williams, "Imaging spectroscopy in the airborne visible/infrared imaging spectrometer (AVIRIS)", *Remote Sens. Environ.* **65,** 227–248 (1998). doi: https://doi.org/10.1016/S0034-4257(98)00064-9

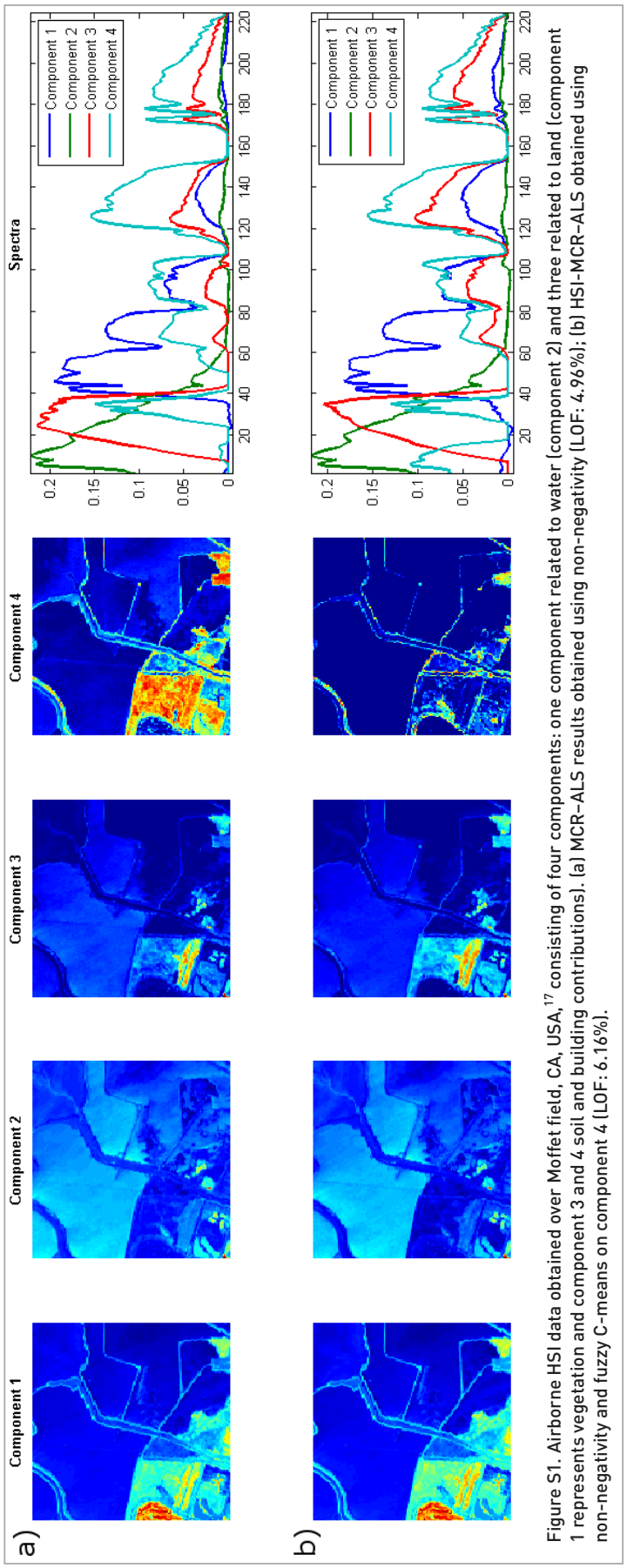

Figure S1. Airborne HSI data obtained over Moffet field, CA, USA,[17] consisting of four components: one component related to water (component 2) and three related to land (component 1 represents vegetation and component 3 and 4 soil and building contributions). (a) MCR-ALS results obtained using non-negativity (LOF: 4.96%); (b) HSI-MCR-ALS obtained using non-negativity and fuzzy C-means on component 4 (LOF: 6.16%).