



HAL
open science

Model-based co-clustering for hyperspectral images

Julien Jacques, Cyril Ruckebusch

► **To cite this version:**

Julien Jacques, Cyril Ruckebusch. Model-based co-clustering for hyperspectral images. Journal of Spectral Imaging, 2016, 5, pp.a3-1 - a3-6. 10.1255/jsi.2016.a3 . hal-01367941

HAL Id: hal-01367941

<https://hal.science/hal-01367941>

Submitted on 17 Sep 2016

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.

MODEL-BASED CO-CLUSTERING FOR HYPERSPECTRAL IMAGES

Julien Jacques¹ & Cyril Ruckebusch²

¹ *Univ Lyon, Lumière Lyon 2, ERIC, Lyon, France, julien.jacques@univ-lyon2.fr*

² *Université de Lille, Sciences et technologies, LASIR CNRS, Lille, France, cyril.ruckebusch@univ-lille1.fr*

September 17, 2016

Abstract. A model-based co-clustering algorithm for hyperspectral images is presented. This algorithm, which relies on the probabilistic latent block model for continuous data, aims to cluster both the pixels and the spectral features of the images. This approach has been applied to a benchmark Raman imaging dataset and revealed relevant information for spatial-spectral exploratory investigation of the data.

Corresponding author: julien.jacques@univ-lyon2.fr

Keywords: co-clustering, latent block model, hyperspectral images

1 Introduction

Co-clustering methods aim to design simultaneously a clustering of the rows and of the columns of a large array of data. The combination of row-clusters and column-clusters leads to the construction of a small number of homogeneous data blocks (see Figure 1), which are useful to summarize and interpret large data sets. In the era of big data and knowledge harvesting, co-clustering techniques have recently spotted a lot of interest from statisticians leading to the development of many algorithms and models. An extensive list of co-clustering methods is provided in reference,¹ distinguishing deterministic² and model-based techniques.³ Model-based approaches are attractive since they provide parsimonious data modeling in a probabilistic framework. This allows choosing the number of relevant co-clusters and opens possibilities for interpretation. In particular, the Latent Block Model (LBM⁴) is a powerful approach to find relevant block clustering and useful representations of the data.

In chemometrics, co-clustering can be viewed as an exploratory approach which allows to simultaneously group observations and variables of a data set, or, in other terms, to simultaneously group rows and columns of a two-way data matrix.⁵ This results in homogeneous blocks, which do not only share information about variable similarity as in classical one way clustering with e.g. k-means⁶ but also information about sample similarity. These blocks (co-clusters) can be used to summarize the data, to facilitate their interpretation or for feature selection. Being applied to hyperspectral images (HSI), co-clustering generates blocks that share simultaneously spatial (pixel) and spectral (variable) information to improve both the reliability and interpretability of the analysis. A statistical approach is sounded in HSI considering the huge number of pixels (individuals).

The potential of co-clustering in HSI analysis has only been pointed out very recently.^{7,8} In this work we propose using LBM to apply co-clustering to the investigation of Raman HSIs. LBM is

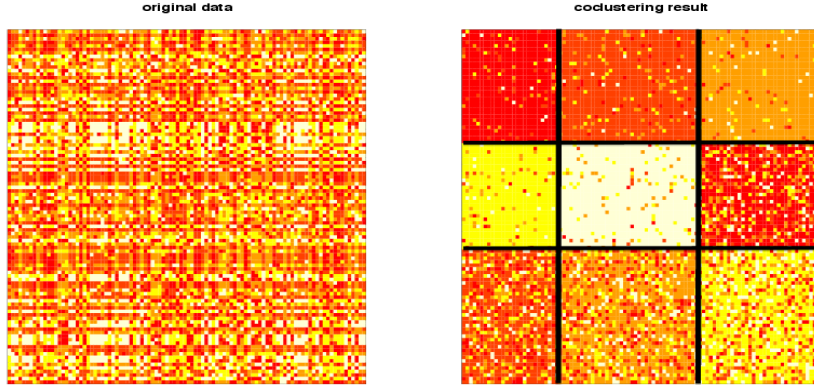


Figure 1: Example of data matrix \mathbf{x} of dimension $n \times d$ (left image) and of the co-clustering results for a partition of $K = 3$ row-clusters and $L = 3$ column-clusters leading to the representation of the data by 9 blocks (right image).

a probabilistic model which assumes that the data from the same block share the same probability distribution. Consequently, the whole dataset is assumed to be issued from a mixture model in which each component corresponds to the probability distribution of a block. These blocks are the most informative ones with the best spectral-spatial information.

The paper is organized as follows: Section 2 introduces the LBM model, its inference and model selection criteria for choosing the number of co-clusters. In Section 3, the proposed approach is applied to a HSI benchmark dataset consisting of Raman imaging of an oil-in-water emulsion. The results obtained are discussed in Section 4.

2 Latent block model for continuous data co-clustering

2.1 Latent block model

The latent block model assumes local independence i.e., the $n \times d$ random variables \mathbf{x} are assumed to be independent once the row partition $\mathbf{v} = (v_{ik})_{1 \leq i \leq n, 1 \leq k \leq K}$ and the column partition $\mathbf{w} = (w_{h\ell})_{1 \leq h \leq d, 1 \leq \ell \leq L}$ are fixed (note that a standard binary partition is used for \mathbf{v} and \mathbf{w}):

$$p(\mathbf{x}; \theta) = \sum_{\mathbf{v} \in V} \sum_{\mathbf{w} \in W} p(\mathbf{v}; \theta) p(\mathbf{w}; \theta) p(\mathbf{x} | \mathbf{v}, \mathbf{w}; \theta) \quad (1)$$

with (below the straightforward range for i, h, k and ℓ are omitted):

- V the set of all possible partitions of rows into K groups, W the set of partitions of the columns into L groups,
- $p(\mathbf{v}; \theta) = \prod_{ik} \alpha_k^{v_{ik}}$ and $p(\mathbf{w}; \theta) = \prod_{h\ell} \beta_\ell^{w_{h\ell}}$ where α_k and β_ℓ are the row and column mixing proportions, belonging to $[0, 1]$ and summing to 1,
- $p(\mathbf{x} | \mathbf{v}, \mathbf{w}; \theta) = \prod_{ihk\ell} p(x_{ih}; \mu_{k\ell}, \sigma_{k\ell}^2)^{v_{ik} w_{h\ell}}$ where $p(\cdot; \mu_{k\ell}, \sigma_{k\ell}^2)$ is the Gaussian density with mean $\mu_{k\ell}$ and variance $\sigma_{k\ell}^2$,

- $\theta = (\mu_{k\ell}, \sigma_{k\ell}^2, \alpha_k, \beta_\ell)$ is the whole mixture parameter.

This local independence assumption (independence of the $n \times d$ random variables \mathbf{x} once the row and column partitions are known), which is usual in mixture model for categorical data,⁹ allows to write the probability of the observed data matrix $p(\mathbf{x}|\mathbf{v}, \mathbf{w}; \theta)$ as a product of the probabilities of all the cells x_{ih} of the matrix. Without this assumption, the joint distribution of \mathbf{x} would be intractable.

Finally, since the row and column partitions are unknown in practice, we have to marginalize this product of probabilities over all the possible row partitions $\mathbf{v} \in V$ and column partitions $\mathbf{w} \in W$.

2.2 Model inference

The aim is to estimate θ by maximizing the observed log-likelihood

$$\ell(\theta; \check{\mathbf{x}}) = \sum_{\mathbf{v}, \mathbf{w}} \ln p(\mathbf{x}; \theta). \quad (2)$$

Due to the presence of missing data (\mathbf{v}, \mathbf{w}), this observed log-likelihood cannot be maximized analytically. The Expectation Maximization algorithm (EM¹⁰) is generally used to maximise log-likelihood in presence of missing data. This algorithm consists in maximizing the complete data log-likelihood:

$$\log p(\mathbf{x}, \mathbf{v}, \mathbf{w}; \theta) = \sum_{ik} v_{ik} \log \alpha_k + \sum_{j\ell} w_{j\ell} \log \beta_\ell + \sum_{ijkl} v_{ik} w_{j\ell} \log p(x_{ij}; \mu_{k\ell}, \sigma_{k\ell}^2) \quad (3)$$

by an iterative procedure which alternates an E step, in which the expectation of (3) conditionally on the observed data and the current value of θ is computed, and an M step in which this conditional expectation is maximized according to θ . But in case of co-clustering, the E step is computationally untractable, and some derivative of this algorithm should be used.³ In this work we opt for the Block EM algorithm.¹¹

Choice of the number of clusters In order to select the numbers of clusters, K in rows and L in columns, we propose to use the ICL-BIC criterion¹² for LBM with continuous data:

$$\text{ICL-BIC}(K, L) = \log p(\mathbf{x}, \hat{\mathbf{v}}, \hat{\mathbf{w}}; \hat{\theta}) - \frac{K-1}{2} \log n - \frac{L-1}{2} \log d - \frac{2KL}{2} \log(nd)$$

where $\hat{\mathbf{v}}$, $\hat{\mathbf{w}}$ and $\hat{\theta}$ are the respective estimations of the row partition, column partition and model parameters obtained at the end of the estimation algorithm and $\log p(\mathbf{x}, \hat{\mathbf{v}}, \hat{\mathbf{w}}; \hat{\theta})$ is the complete data log-likelihood given by:

$$\log p(\mathbf{x}, \hat{\mathbf{v}}, \hat{\mathbf{w}}; \hat{\theta}) = \sum_{ik} \hat{v}_{ik} \log \hat{\alpha}_k + \sum_{j\ell} \hat{w}_{j\ell} \log \hat{\beta}_\ell + \sum_{ijkl} \hat{v}_{ik} \hat{w}_{j\ell} \log p(x_{ij}; \hat{\mu}_{k\ell}, \hat{\sigma}_{k\ell}^2).$$

This criterion is expected to select the true numbers of clusters under given theoretical conditions. In practice, it can be used as a starting point for data exploration, the practitioner may change these numbers according to its interpretation.

Exploring all the possible combinations for (K, L) with $1 \leq K \leq K_{max}$ and $1 \leq L \leq L_{max}$ requires to estimate and to compute the ICL-BIC criterion $K_{max} \times L_{max}$ times. It can rapidly become computationally expensive. We propose here a greedy search in order to minimise the number of explored values for (K, L) :

1. start with $(K, L) = (1, 1)$
2. compute $\text{ICL-BIC}(K + 1, L)$ and $\text{ICL-BIC}(K, L + 1)$.
 - if both are lower than $\text{ICL-BIC}(K, L)$, then stop the search and return $(\hat{K}, \hat{L}) = (K, L)$.
 - else, if $\text{ICL-BIC}(K + 1, L) > \text{ICL-BIC}(K, L + 1)$, then $K \leftarrow K + 1$, else $L \leftarrow L + 1$.
3. return in 2.

Software The Block EM algorithm for estimating the Gaussian LBM model is implemented in the *blockcluster*¹¹ package for the R software.

3 Application

The Raman HSI data provided in Figure 2 corresponds to an oil-in-water emulsion sample analyzed in the fingerprint region of the spectral range. This data has been already used in previous works to benchmark data analysis methods. Full details about the dataset and the experimental setup can be found in earlier work.¹³ The dataset is available at: <https://mcra1s.wordpress.com/download/example-data-sets/examples-of-analytical-methods> (last accessed September 13th 2016).

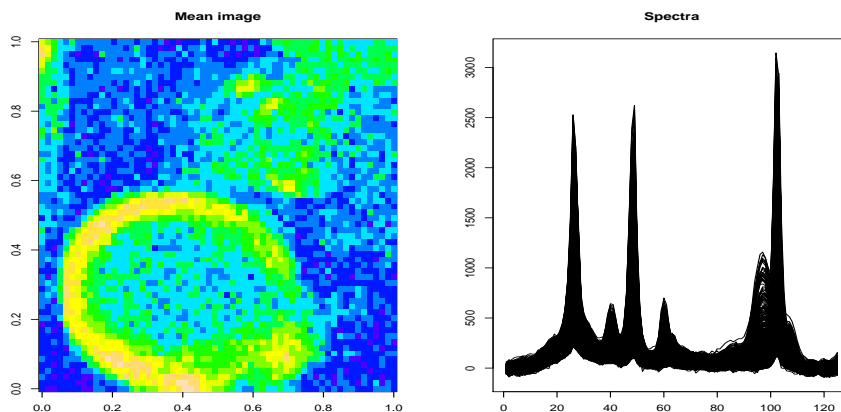


Figure 2: Raman HSI data of an oil-in-water emulsion. Mean image ($60 \times 60 \mu\text{m}$) and overlaid spectra in the range 900 to 1400 cm^{-1} .

The image in Figure 2 (left) is formed by 60×60 pixels and the spectra consists of 125 points in the range from 900 to 1400 cm^{-1} . The interest in analyzing this data is that the composition of the spectroscopic mixture is known.¹³ In the sampled area, according to previous work,^{13,14,15} four

different components can be extracted using multivariate curve resolution approaches corresponding to a structural phase, two droplet phases, and an aqueous phase. However, it should be noted that the structural and aqueous phases are mixtures of multiple chemical components. Thus, the chemical complexity of these emulsion should not be underestimated (10 to 12 different chemical compounds with relative concentration from 0.1% to 80%).¹³

The co-clustering results are provided in Figures 3 and 4. In Figure 3 (top), the results obtained for pixel clustering (left) with $K = 9$ pixel-clusters and $L = 6$ spectral-clusters, as well as the mean spectra for each of the 9 pixel-clusters, are shown. These numbers of pixel- and spectral-clusters were obtained using the ICL-BIC criterion described in Section 2.2. Considering the chemical complexity of the investigated sample, these results do not seem out of range. However, as already mentioned, this automatic procedure can be refined. Indeed, inspection of the mean spectra obtained for the different pixel-clusters reveals that some of them have a pair-wise correlation coefficient greater than 0.99. Adding to this the a priori information available about the investigated data, it was thus decided to merge these pixel clusters which leads to the 4 pixel-clusters representation shown in Figure 3 (bottom). The corresponding mean spectra for each cluster are also given.

One can observe that the different structures expected in this sample are revealed. The two droplet oil phases are clearly seen (blue and red pixels). These clusters were already identified individually before cluster averaging and they consist of different oil mixtures. After cluster averaging, the structural (purple) and aqueous phases (orange) which consist of around 80% water (plus two chemicals) are more clearly depicted.

In Figure 4 the results obtained for variable-clusters are shown. In the way chosen to represent the results, one plot is provided for each cluster and each plot contains the mean spectral value for each of the 4 pixel-clusters for the features that were retained for the construction of the block. The spectral-clusters which are the most relevant for the distinction of the spatial information are clusters 2 and 3 for which a clear distinction of the different mean cluster spectra is observed.

These spectral-clusters highlight spectral band where essentially difference intensities is observed for the main spectral features peaking at variable number 25, 50 and 105 (spectral cluster 2) and on the side of these bands (spectral cluster 3). On the other hand, the spatial-clusters 4 and 6 can be interpreted as the ones for which specific information regarding the droplet oil phase (coded in red) is obtained. Information around spectral variables 60 and 90 are of particular relevance, the latter corresponding to a shoulder observed on the original spectra.

4 Discussion

Whereas clustering groups objects according to the similarity of their features (variables), co-clustering groups both objects and variables simultaneously, it looks for which samples are related to which variables at the same time. In contrast to other approaches that have been proposed in the chemometric literature based on constrained outer product decomposition with sparsity on the latent factor of the bilinear model¹⁶ that are not necessary well-suited for spectral data (collinearities of the variables and the lack of sparsity), the proposed LBM provided reliable and useful (interpretable) results for Raman HSI data. More in general, co-clustering approaches provides results on the line of recent advances in spectral-spatial analysis of HSI images.¹⁷

Comparison of the results obtained for co-clustering with $K = 4$ pixel clusters (see Figure

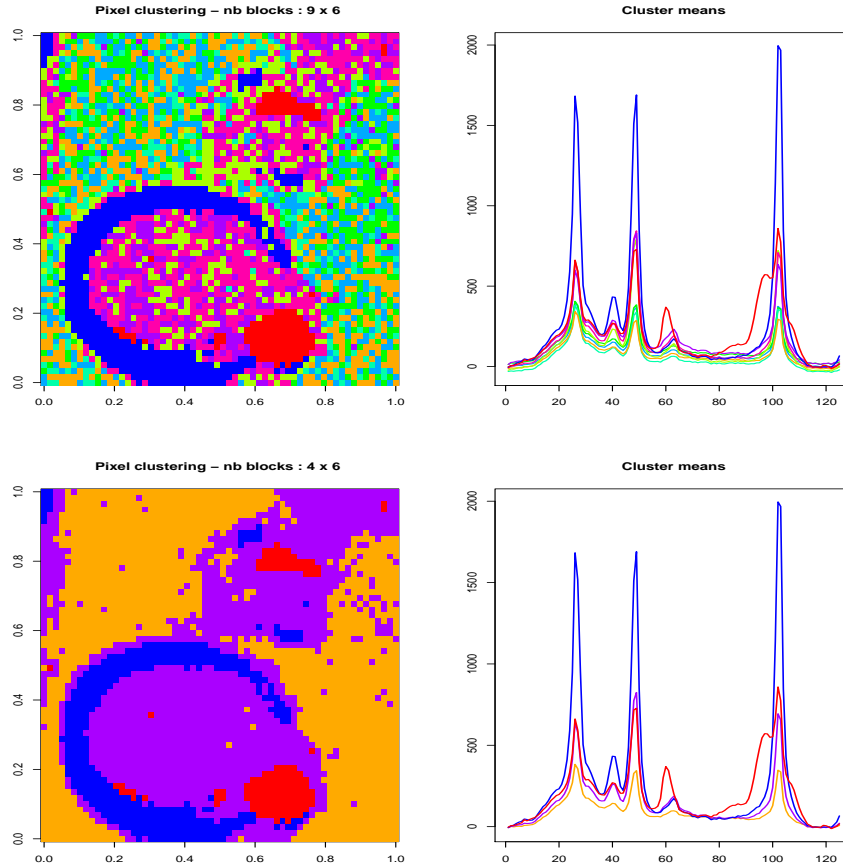


Figure 3: Pixel-clusters and corresponding mean spectra obtained for $K = 9$ (top) and $K = 4$ (bottom). One color per cluster.

3) can be made with the ones applying k-means. For this purpose, Figure 5 provides the results obtained for k-means pixel clustering considering 4 clusters. Overall it can be noticed that the images obtained are quite similar. However a closer inspection reveals that the two droplet oil phases are quite different. The structure of the smaller droplet is less clearly identified on the image obtained Figure 5. This is confirmed by the inspection of the spectra provided for the k-means clusters. They do not allow to clearly point the spectral differences in the region around the spectral variable 90, which role in the discrimination of this compound has been clearly identified from the interpretation of the spectral-cluster provide Figure 4.

The aim of this work was to demonstrate the applicability of the LBM co-clustering approach for HSI analysis and to illustrate the potential of the co-clustering for data investigation and spatial-spectral interpretation. It should be clear that co-clustering results cannot directly compare to the ones obtained applying multivariate resolution and unmixing methods which provide factorial data decomposition into pure component distribution maps and corresponding spectral features. However, in the example shown co-clustering has revealed interesting information that can be extremely useful for preliminary spatial-spectral investigation and that could potentially be used for further analysis.

Summing up, we believe that co-clustering can provide an efficient alternative for the investiga-

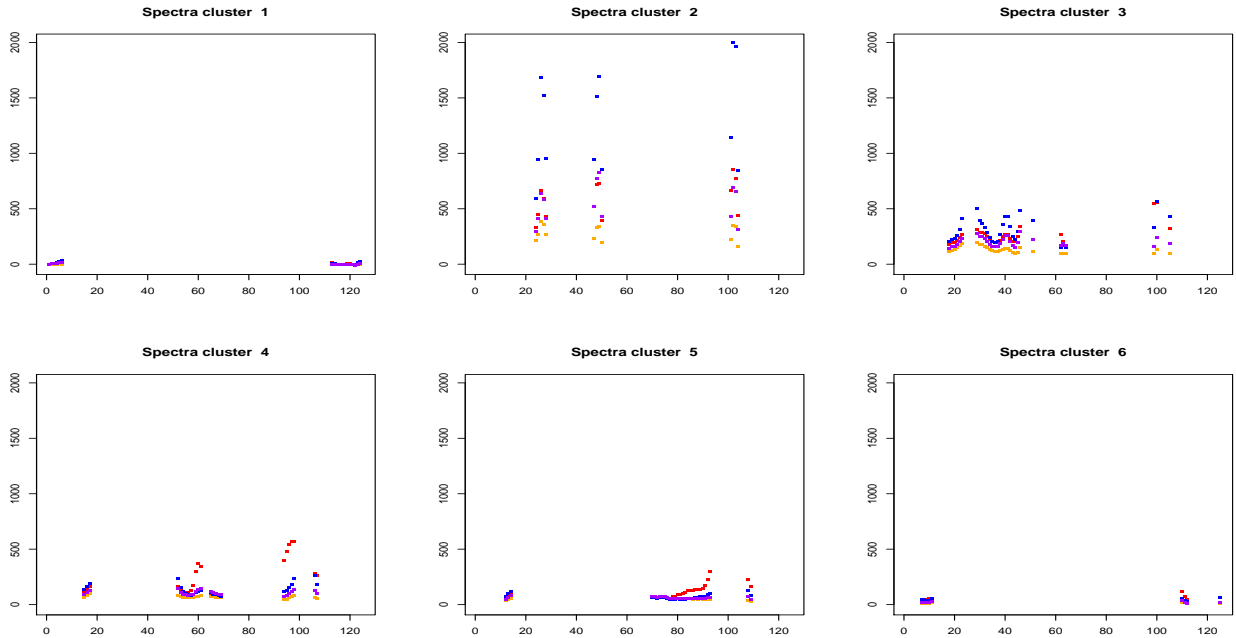


Figure 4: Spectral-clusters for $L = 6$. Mean cluster spectra are plotted to highlight spectral features.

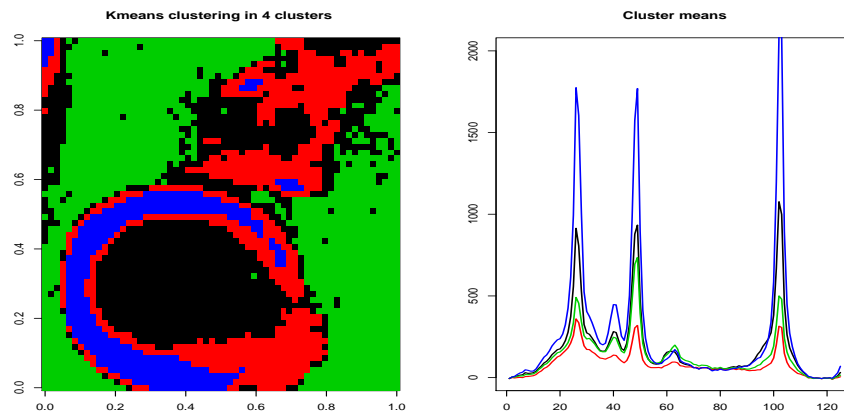


Figure 5: Pixel-clusters and corresponding mean spectra obtained with k-means for $K = 4$ clusters. One color per cluster.

tion of HSI data. The LBM framework could in future works be modified to allow the consideration of pixel neighborhood (sets of pixels) and account for spatial or morphological information.

References

- ¹ S. C. Madeira and A. L. Oliveira. "Biclustering algorithms for biological data analysis: A survey". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1:24–45, (2004).

- ² A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha. "A generalized maximum entropy approach to bregman co-clustering and matrix approximation". *Journal of Machine Learning Research*, 8:1919–1986, (2007).
- ³ G. Govaert and M. Nadif. *Co-Clustering*. Wiley-ISTE, (2013).
- ⁴ G. Govaert and M. Nadif. "Block clustering with bernoulli mixture models: Comparison of different approaches". *Computational Statistics and Data Analysis*, 52:3233–3245, (2008).
- ⁵ R. Bro, E.E. Papalexakis, E. Acar, and N.D. Sidiropoulos. "Coclustering-a useful tool for chemometrics". *Journal of Chemometrics*, 26(6):256–263, (2012).
- ⁶ J.A. Hartigan and M.A. Wong. "Algorithm as 1326 : A k-means clustering algorithm". *Applied Statistics*, 28:100–108, (1978).
- ⁷ W. Liu, S. Li, X. Lin, Y. Wu, and R. Ji. "Spectral–spatial co-clustering of hyperspectral image data based on bipartite graph". *Multimedia Systems*, 22(3):355–, (2016).
- ⁸ E. Izquierdo-Verdiguier. "Introducing co-clustering for hyperspectral image analysis". pages 465–468, (2015).
- ⁹ G. Celeux and G. Govaert. "Clustering criteria for discrete data and latent class models". *Journal of Classification*, 8:157–176, (1991).
- ¹⁰ A.P. Dempster, N.M. Laird, and D.B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, (1977).
- ¹¹ P.S. Bhatia, S. Iovleff, G. Goavert, V. Brault, C. Biernacki, and G. Celeux. "*blockcluster: Co-clustering package for binary, contingency, continuous and categorical data-sets*", (2015).
- ¹² A. Lomet. "*Sélection de modèle pour la classification croisée de données continues*". PhD thesis, Université de Technologie de Compiègne, Compiègne, France, (2012).
- ¹³ J.J. Andrew, M.A. Browne, I.E. Clark, T.M. Hancewicz, and A.J. Millichope. "Raman imaging of emulsion systems". *Journal of Applied Spectroscopy*, 52:790–796, (1998).
- ¹⁴ S. Hugelier, O. Devos, and C. Ruckebusch. "Constraining shape smoothness in multivariate curve resolution-alternating least squares". *Journal of Chemometrics*, 29:448–456, (2015).
- ¹⁵ A. De Juan, M. Maederb, T. Hancewicz, and R. Tauler. "Use of local rank-based spatial information for resolution of spectroscopic images". *Journal of Chemometrics*, 5:291–298, (2008).
- ¹⁶ E.E. Papalexakis, N.D. Sidiropoulos, and R. Bro. "From k-means to higher-way co-clustering: Multilinear decomposition with sparse latent factors". *IEEE Transactions on Signal Processing*, 61(2):493–506, (2013).
- ¹⁷ M. Fauvel, Y. Tarabalka, J.A. Benediktsson, J. Chanussot, and J.C. Tilton. "Advances in spectral-spatial classification of hyperspectral images". volume 101, pages 652–675, (2013).