Application of Non-negative Matrix Factorization to LC/MS data

Jérémy Rapin, Antoine Souloumiac, Jérôme Bobin, Anthony Larue, Chistophe Junot, Minale Ouethrani, Jean-Luc Starck

Application of Non-negative Matrix Factorization to LC/MS data

Jérémy Rapin, Antoine Souloumiac, Jérôme Bobin, Anthony Larue, Chistophe Junot, Minale Ouethrani and Jean-Luc Starck (firstname.lastname@cea.fr)

***Abstract*—Liquid Chromatography-Mass Spectrometry (LC/MS) provides large datasets from which one needs to extract the relevant information. Since these data are made of non-negative mixtures of non-negative mass spectra, non-negative matrix factorization (NMF) is well suited for its processing, but it has barely been used in LC/MS. Also, these data are very difficult to deal with since they are usually contaminated with non-Gaussian noise and the intensities vary on several orders of magnitude. In this article, we show the feasibility of the NMF approach on these data. We also propose an adaptation of one of the algorithms aiming at specifically dealing with LC/MS data. We finally perform experiments and compare standard NMF algorithms on both simulated data and an annotated LC/MS dataset. This lets us evaluate the influence of the noise model and the data model on the recovery of the sources.**

***Index Terms*—BSS, NMF, sparsity, multiplicative noise, LC/MS**

I. Introduction

A. Liquid chromatography-mass spectrometry data

The aim of LC/MS is to detect, quantify and identify molecules from liquid samples. The liquid sample is first injected into a chromatographic column, through which the different compounds exhibit different kinds of physico-chemical interactions with the stationary phase. These compounds thus leave the column at different times, referred to as retention times. At each time t, the compounds leaving the column are sprayed, ionized in the source of the mass spectrometer, and then separated according to their mass to charge ratios in the analyzer (i.e., an orbitrap analyzer in the present study). Each ion having a specific mass-to-charge ratio, the LC/MS process provides a two dimension separation (although imperfect) in both mass and retention time domains. It yields 2D data such as the ones shown in Fig. 1. These data are coined $\mathbf{Y} \in \mathbb{R}^{m\times n}$ in the article, and each of the m lines of this matrix is a n-sample long spectrum at a given acquisition time.

A. Larue, J. Rapin and A. Souloumiac are with CEA, LIST, 91191 Gif-sur-Yvette Cedex, France.

J. Bobin, J. Rapin and J.-L. Starck are with CEA, IRFU, Service d'Astrophysique, 91191 Gif-sur-Yvette Cedex, France.

C. Junot and M. Ouethrani are with CEA, DSV/iBiTec-S, Service de Pharmacologie et d'Immunoanalyse, Laboratoire dEtude du Métabolisme des Médicaments, 91191 Gif-sur-Yvette Cedex, France.

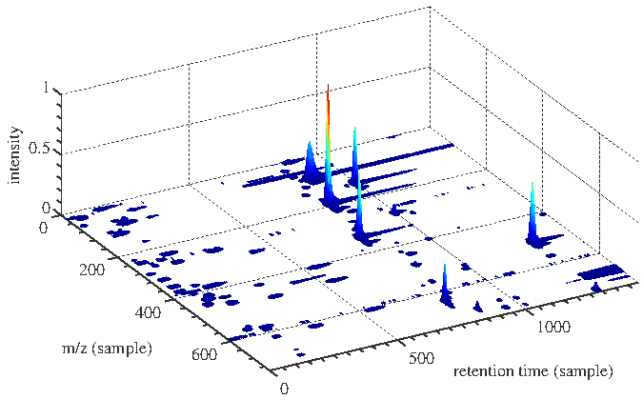


Fig. 1: LC/MS chromatogram of a sample (filtered for visualization purpose)

B. Non-negative matrix factorization

Non-negative matrix factorization (NMF) aims at decomposing the data as non-negative mixtures of non-negative signals, the sources. The first publications dealing with these particular settings come from Paatero & Tapper [1] and Lee & Seung [2]. The non-negative assumption arises naturally in many applications such as hyperspectral imaging [3], nuclear magnetic resonance [4], [5] or LC/MS [6], [7]. Indeed, in LC/MS, the mass spectra are non-negative, and the mixtures are related to the relative concentrations, which cannot be negative either. Under the instantaneous linear mixture model, each of the m observation $\mathbf{Y}_{i,\cdot} \in \mathbb{R}^{1\times n}$ is a linear mixture of r elementary non-negative spectra $\mathbf{S}_{j,\cdot} \in \mathbb{R}^{1\times n}$:

$$\mathbf{Y}_{i,\cdot} = \sum_{j=1}^{r} \mathbf{A}_{i,j}\mathbf{S}_{j,\cdot} + \mathbf{Z}_{i,\cdot}\ ,\ \forall i \in \{1,..,m\}, \qquad (1)$$

where $\mathbf{A}_{i,j}$ are the non-negative mixtures coefficients and $\mathbf{Z}_{i,\cdot}$ accounts for noise and model imperfections. Under matrix form, this can be recast as: $\mathbf{Y} = \mathbf{AS} + \mathbf{Z}$. From $\mathbf{Y}$, the aim is to recover both $\mathbf{A} \in \mathbb{R}^{m\times r}$ and $\mathbf{S} \in \mathbb{R}^{r\times n}$, which is usually done by solving a problem of type:

$$\underset{\mathbf{A}\geq\mathbf{0},\ \mathbf{S}\geq\mathbf{0}}{\operatorname{argmin}}\ \mathcal{D}(\mathbf{Y}||\mathbf{AS}) + \mathcal{J}(\mathbf{S}), \qquad (2)$$

where $\mathcal{D}$ is a divergence measuring the discrepancy between the data $\mathbf{Y}$ and the factorization $\mathbf{AS}$, and $\mathcal{J}$ is a regularization function providing prior information about the spectra. This problem is however non-convex and NP-Hard [8] and finding an optimal solution is therefore very difficult. Different NMF algorithms or even different initializations of a same algorithm therefore yield different factorizations.

C. Contribution

Although NMF is particularly well suited for processing LC/MS data, it has barely ever been used on this type of data. In this article, we test reference and state-of-the-art NMF algorithms on an annotated LC/MS dataset. We also propose an adaptation of the non-negative generalized morphological component analysis (nGMCA) aiming at specifically dealing with LC/MS data. The comparison highlights the behaviors of the algorithms in difficult settings, with large dynamics, multiplicative noise and potential non-linearities, and shows the efficiency of our adaptation. In the final section, based on the experiments, we discuss further improvements which could be brought to the existing algorithms in order to better handle LC/MS data.

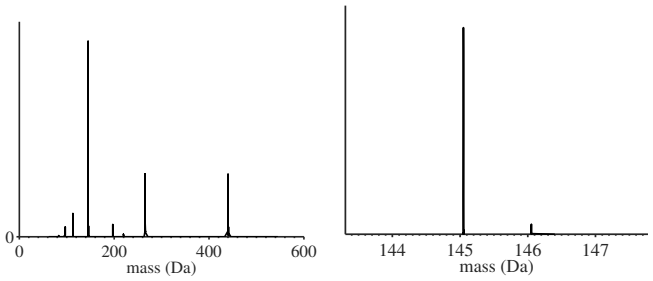
II. LC/MS DATASET

The data considered in this article were acquired from a mixture of eleven commercial chemical compounds for which the mass spectra and retention times are known (see Fig. 5 for the list of compounds). This sample was analyzed in an LC/MS pipeline using an orbitrap mass analyzer [9], [10]. We focus on a time range going from 2 to 18min and masses from 69 to 644 Dalton (Da) since these ranges concentrate most of the information of the sample. Since all ions related to the eleven molecules and their retention times were known, we can build a reference source matrix $\mathbf{S}^{\text{annot.}}$, in which each line is the mass spectrum of one of the molecules.

A. Mass and elution profiles

A typical mass profile $\mathbf{Y}_{t,\cdot} \in \mathbb{R}^{1\times n}$ at a time t is provided in Fig. 2a. This mass profile is a mixture of elementary spectra which are characteristic of specific compounds. Unmixing them can therefore help identify the chemical compounds of the liquid. In LC/MS, it is well known that ions other than molecular species are produced by atmospheric pressure ionization methods during the desolvation process, including natural isotopes, adduct ions, fragment ions formed by spontaneous in-source fragmentations of the precursor ion by release of small size neutrals and also multimers in the case of ESI mass spectra [11]. Thus, many peaks of the mass profile corresponds to the m/z ratios related to one of these ions. The figures are scaled in Da —1Da having a value very close to the mass of a nucleon— since all the ions have the same charge state here (-1e). Zooming on the main peak of the previous figure yields Fig. 2b, where a small peak is visible at $+1$Da, which is typical of the presence of carbon-13 ^{13}C. This figure also highlights the large range of intensities which must be extracted from the data. In practice, the main peak has a width of about ten samples, showing the extreme precision of the Orbitrap spectrometer which was used for the acquisition of these data.

Fig. 3 shows a temporal profile $\mathbf{Y}_{\cdot,\mu} \in \mathbb{R}^{m\times 1}$, or in other words the evolution of the intensity of a specific mass μ during the acquisition time. Such a profile is also called elution profile. Elution profiles are typically smooth, similar to a Gaussian with a width slightly smaller than a minute and a heavy tail after the maximum value.



(a) Full mass profile (b) Zoom around the main peak

Fig. 2: Examples of mass profile at a time t.

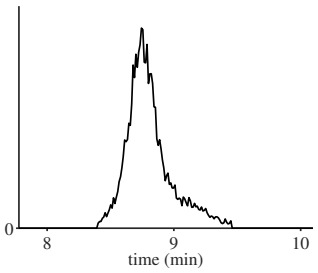


Fig. 3: Temporal profile of a mass in the dataset.

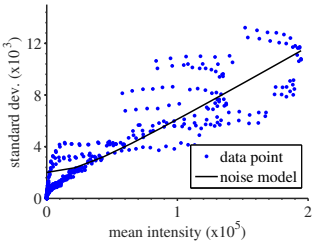


Fig. 4: Mean-standard deviation cloud for the main temporal profiles.

B. About the noise contamination

A feature of LC/MS data is clearly visible on the temporal profile of Fig. 3: the larger the values, the noisier they are. In order to corroborate this observation, we plot in Fig. 4 the point cloud of an estimation of the data standard deviation with respect to its mean amplitude, computed on non-null chunks of the most energetic elution profiles. A correlation between amplitude and standard deviation is then visible. We therefore model the noise contamination with multiplicative noise, which standard deviation is proportional to the value of eacg signal coefficient. Numerous studies have come to similar conclusions [12], [13]. Considering that the noise is a mixture of an additive component and a multiplicative component yields the following model for the standard deviation of the noise on each coefficient of the data:

$$\mathbf{\Sigma}_{i,j} = \sqrt{\sigma^2_{\text{add}} + \sigma^2_{\text{mult}} \mathbf{Y}^2_{i,j}}, \tag{3}$$

where σ_{add} is the additive noise standard deviation, and σ_{mult} is the coefficient of the multiplicative noise. The black line in the figure corresponds to the best fit model, with values $\sigma_{\text{mult}} = 0.074$ and σ_{add} =15,000. Biologists using these data tend to consider that peaks start to be significant at an order of magnitude of 10^4, which comforts our estimation.

C. Processing pipeline

The processing of LC/MS data is usually performed with software such as XCMS [14] or MZmine [15], which perform (i) automatic peak detection, (ii) alignment of features in the m/z and chromatographic retention time domains, and (iii) results are returned as a peak table containing variable identity (i.e., m/z and retention time) and signal abundances (i.e., peak intensities and/or area of extracted ion chromatographic

peaks) in the samples. As mentioned in [6], NMF is however particularly well-suited for the processing of LC/MS data and combines the extraction and gathering stages into a unique step. In this article, we evaluate the performances of several NMF approaches on an annotated dataset and show the feasibility of this approach, with the following processing pipeline:

1) Mass gridding: Each spectrum acquired by the mass spectrometer has a specific mass grid, different from the ones acquired at other times. We therefore construct a grid shared between all spectra so as to arrange the data into a matrix form.
2) Cleaning: masses which appear only a very limited number of consecutive times in the data, or at too low an intensity, are not considered significant and removed from the grid. Conversely, masses which are constant during all the analysis, such as the one corresponding to the solvents, are also eliminated from the data. The resulting data is coined $\mathbf{Y}$.
3) Subsampling: In order to further reduce the data size, $\mathbf{Y}$ is subsampled along the mass dimension, yielding a data matrix $\mathbf{Y}_{\text{sub}}$ with lower dimensionality. This reduces the precision but makes computation faster and diminishes the noise influence, now estimated at σ_{add} =12,000 and $\sigma_{\text{mult}} = 0.023$ (compared to the estimation in section II-B). Also, the initial precision can mostly be recovered thanks to the inversion step 5).
4) Factorization: the data is decomposed as $\mathbf{Y}_{\text{sub}} \approx \mathbf{A}\mathbf{S}_{\text{sub}}$ with $\mathbf{A}$ and $\mathbf{S}$ non-negative matrices, using an NMF algorithm.
5) Inversion: The full precision is recovered by solving the inverse problem 4 using the mixing matrix $\mathbf{A}$ found at the previous stage and the non-subsampled data $\mathbf{Y}$ explicited on the shared mass grid of stage 2).

$$\underset{\mathbf{S}\geq\mathbf{0}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{Y}-\mathbf{A}\mathbf{S}\|_2^2, \tag{4}$$

III. NMF Algorithms

Algorithms in NMF can aim at solving different formulations of the problem, with different divergences $\mathcal{D}$ and regularizations $\mathcal{J}$. Also, because of its non-convexity, even one algorithm with different initializations can produce different results. In this section, we therefore evaluate the performances of several algorithms on LC/MS data.

A. Standard non-regularized algorithms

The first converging iterative algorithm proposed in NMF is the multiplicative update of Lee & Seung [16]. Using the Kullback-Leibler divergence, it aims at solving Problem (2) with no regularization and with the Kullback-Leibler divergence (5) or the euclidian distance (6). Both versions are tested in the experiments, respectively under the names "mult. (KL)" and "mult. (L2)".

$$\mathcal{D}(\mathbf{Y}||\mathbf{A}\mathbf{S}) = \sum_{i,j} \mathbf{Y}_{i,j} \log \frac{\mathbf{Y}_{i,j}}{(\mathbf{A}\mathbf{S})_{i,j}} - (\mathbf{A}\mathbf{S})_{i,j} \tag{5}$$

$$\mathcal{D}(\mathbf{Y}||\mathbf{A}\mathbf{S}) = \|\mathbf{Y}-\mathbf{A}\mathbf{S}\|_2^2 \tag{6}$$

In order to take into account the multiplicative noise in LC/MS, Dubroca et al. proposed in [6] to modify the previous algorithm using a non-stationary model for the noise. At each sample, they provide an expected noise standard deviation $\mathbf{\Sigma}_{i,j}$ depending on the amplitude of the data (cf. equation (3)). The maximum likelihood is then given by:

$$\mathcal{D}(\mathbf{Y}||\mathbf{A}\mathbf{S}) = \|(\mathbf{Y}-\mathbf{A}\mathbf{S}) \oslash \mathbf{\Sigma}\|_2^2, \tag{7}$$

where $\oslash$ is the elementwise matrix division. This algorithm is coined "mult. (non-stat.)" in the following experiments. These multiplicative algorithms are widely used, easy to implement, but often deemed to be slow and not very efficient [17], [18].

B. Algorithms using sparse regularizations

In the wide sense, a sparse signal is a signal which concentrates its energy into only a few large non-zero coefficients, or can be well approximated in such a way. This is definitely the case in LC/MS since both spectra and mixing coefficients are mostly null with few large coefficients. In the NMF literature, the sparse assumption has been shown to help recovering relevant factorizations [19], [20].

Many iterative algorithms aim at solving Problem (2) with the euclidian divergence (6) and a sparsity inducing regularization $\mathcal{J}(\mathbf{S}) = \|\mathbf{\Lambda} \odot \mathbf{S}\|_1$, where $\mathbf{\Lambda}$ is a parameter matrix, and $\odot$ the pointwise multiplication. This is the case for HALS [21] and nGMCA [18][1], which are both tested in the experiments. In nGMCA, the $\mathbf{\Lambda}$ parameter must be set at the noise standard deviation. In a recent implementation of HALS [22], [23][2], the parameter $\mathbf{\Lambda}$ is automatically handled in order to obtain a user-defined sparsity rate[3]. In the experiments, the sparsity rate of the sources is set as the sparsity rate of the observations $\mathbf{Y}_{i,\cdot}$, since no better estimation is available in practice.

In [20], Kim & Park have proposed to use a regularization $\mathcal{J}(\mathbf{S}) = \lambda \sum_{t=1}^{n} \|\mathbf{S}_{\cdot,t}\|_1^2$. This term aims at favoring sparsity in a way which makes a single source dominate over the other ones. We use here the default values of the parameters from the implementation[4], since there is no straightforward way to tune them.

These algorithms may not be well-suited to the case of a multiplicative noise contamination because they use an euclidean divergence. Also, all the iterative algorithms are very sensitive to stationary points and can therefore converge to highly suboptimal solutions.

C. Adaptation of nGMCA for LC/MS data

Considering the drawbacks of the previous methods, we propose to adapt nGMCA to the case of LC/MS data by:

- using divergence (7) instead of divergence (6) so as to be more robust to multiplicative noise.
- add a stochastic term $\mathbf{D}$ in order to avoid the multiple stationary points, which arise from the specific structure of LC/MS data.

[1] http://www.cosmostat.org/GMCALab.html
[2] https://sites.google.com/site/nicolasgillis/code
[3] ratio of coefficients smaller than 10^{-6} times the largest one.
[4] http://www.cc.gatech.edu/~hpark/nmfsoftware.php

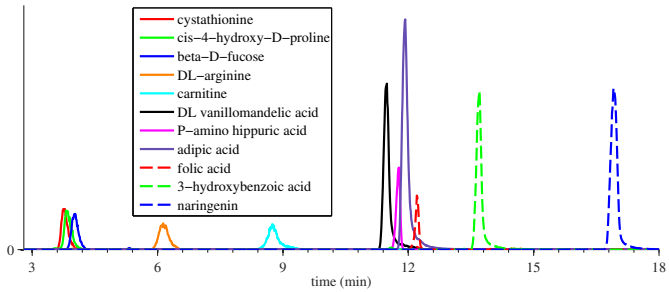


Fig. 5: Elution profiles $\mathbf{A}^{\text{annot.}}$ of the annotated sources, estimated by an inversion (infinite norm shared between $\mathbf{A}^{\text{annot.}}$ and $\mathbf{S}_{\text{sub}}^{\text{annot.}}$ for visualization purposes).

At each iteration k, both $\mathbf{A}$ and $\mathbf{S}$ are updated once —such as in standard nGMCA— so as to approach the solution of:

$$\underset{\mathbf{A}\geq\mathbf{0},\ \mathbf{S}\geq\mathbf{0}}{\text{argmin}}\ \frac{1}{2}\| \left(\mathbf{Y} + \sigma_D^{(k)}\mathbf{D}^{(k)} - \mathbf{AS}\right) \oslash \mathbf{\Sigma}\|_2^2 + \|\mathbf{\Lambda}^{(k)} \odot \mathbf{S}\|_1. \tag{8}$$

The coefficients of $\mathbf{D}^{(k)}$ are taken as independent centered and reduced Gaussian variables, and $\sigma^{(k)}$ decreases at each iteration and reaches 0 at the end of the algorithm so as to causing a bias. In the same way than in nGMCA, $\mathbf{\Lambda}^{(k)}$ is automatically chosen according to the estimated noise level. This version is coined "non-stat. nGMCA (stoch.)" in the experiments. In the following section, it is compared to all the other algorithms on simulated data.

IV. Experiments

A. Results on simulated data

So as estimate the actual elution profiles one should observe, we first undersample the reference source matrix on the subsampled grid so as to obtain $\mathbf{S}_{\text{sub}}^{\text{annot.}}$. We define the corresponding elution profiles by solving the inverse problem 9. These profiles are shown in Fig. 5. From there, one can synthesize a data matrix using Equation 10.

$$\mathbf{A}^{\text{annot.}} = \underset{\mathbf{A}\geq\mathbf{0}}{\text{argmin}}\ \frac{1}{2}\|\mathbf{Y}_{\text{sub}} - \mathbf{AS}_{\text{sub}}^{\text{annot.}}\|_2^2. \tag{9}$$

$$\mathbf{Y}_{\text{synth.}} = \mathbf{A}^{\text{annot.}}\mathbf{S}_{\text{sub}}^{\text{annot.}} + \mathbf{Z}. \tag{10}$$

We consider here that the coefficients $\mathbf{Z}_{i,j}$ are independent Gaussian variables with standard deviation fixed to $\mathbf{\Sigma}_{i,j}$, as computed in equation (3). The additive noise contamination is controlled by σ_{add}, which is fixed to 12,000, as estimated in the actual data. Still, the noise level was estimated only on non-null chunk and additive noise in the data is much scarcer in practice. The parameter σ_{mult} controls the multiplicative noise level. In Fig. 6, it varies from 0 to 0.1 in order to assess the sensitivity of the selected algorithms to multiplicative noise on the synthesized LC/MS data. The evaluation is carried out by comparing the obtained spectra with $\mathbf{S}_{\text{sub}}^{\text{annot.}}$ using the source distortion ratio (SDR) introduced in [24]. This criterion increases for higher quality reconstructions.

All the algorithms show significant sensitivity to the multiplicative noise contamination, as can be seen in Fig. 6. One can first notice the globally poor results of the multiplicative update algorithms, which do not use any sparse regularization, with a very slight advantage for the non-stationary version mult. (non-stat.). Concerning the other algorithms, which make use of sparse regularizations on $\mathbf{S}$, the stochastic non-stationary version of nGMCA obtains the best results. Kim & Park algorithm is globally 5dB under this version of nGMCA, but close to 5dB above all the other algorithms. One can recall that this algorithm was used with its standard parameters and was therefore not tuned at all. This highlights the fact that, as well as an adequate noise model, an accurate source model can be beneficial. Indeed, the regularization used in this algorithm favors sources which do not share any coefficient. This model is well verified on the dataset.

B. Qualitative study on the real data

The above observations were made on very realistic simulated data. Evaluation of the algorithms on the real dataset is however trickier. Indeed, the annotations only provide the main peaks time and mass localization for each source, but not exactly the spectra which should be recovered. In this section, we observe results obtained on the real dataset so as to develop quantitative criteria. We consider the search of 18 sources, among which we wish to recover the 11 annotated sources. nGMCA based algorithms are provided with estimations of the noise level: σ_{add} =12,000 and $\sigma_{\text{mult}} = 0.05$.

The elution profiles $\mathbf{A}^{\text{annot.}}$ in Fig. 7a can be compared with the ones obtained on the real dataset by nGMCA and non-stat. nGMCA (stoch.) (respectively Fig. 7b and 7c). A couple of observations can be made from it:

- the database is not perfect and may not contain all the compounds present in the mixture: nGMCA and non-stat. nGMCA (stoch.) agree on the existence of an prominent non-annotated source (purple line at 12.4min, source #12 and #14 respectively). Its mass spectrum presents a structure typical of the presence of adducts.
- source #6 obtained with nGMCA (dashed magenta line at 11.45min) interferes with DL-vanillomandelic acid

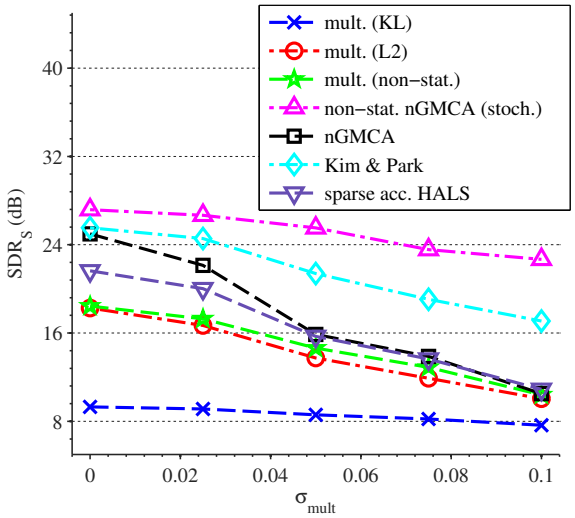
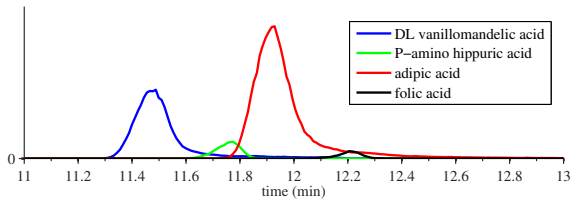
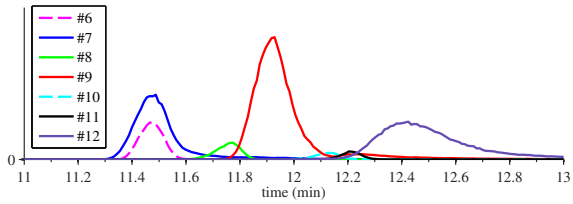


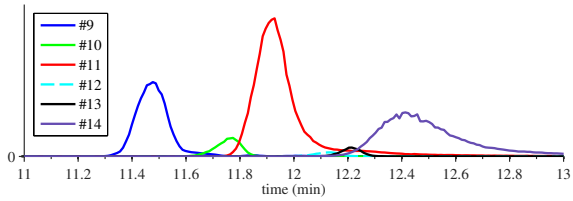
Fig. 6: SDR of the recovered sources with respect to the amount of multiplicative noise (average of 100 simulations).



(a) inversion of the annotated spectra.



(b) NMF with nGMCA.



(c) NMF with non-stat. nGMCA (stoch.).

Fig. 7: Zoom on the elution profiles between 11 and 13 minutes (real dataset).

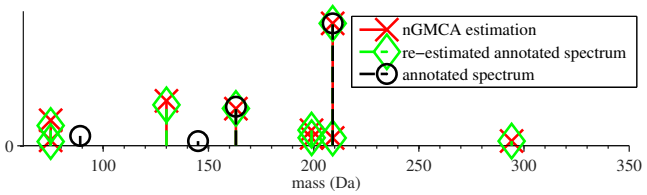
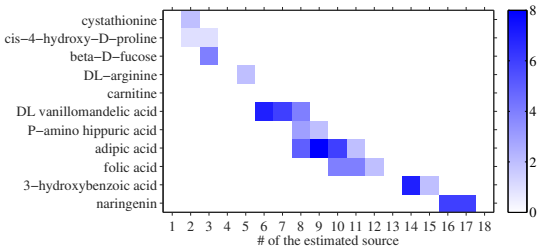


Fig. 8: Beta-D-fucose mass spectrum (annotations and reconstructions for the real dataset).

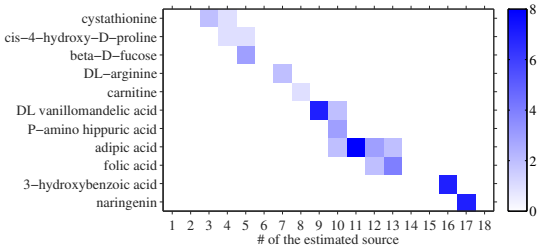
(source #7, blue line at 11.45min), and source #10 (dashed cyan line at 12.1min) with adipic acid (source #9, red line at 11.9min). The interferences with DL-vanillomandelic acid (source #9) do not appear with non-stat. nGMCA (stoch.), and the ones with the adipic acid are weaker, which tends to suggest a better reconstruction quality obtained with this algorithm.

Fig. 8 provides an example of mass spectrum reconstruction with nGMCA for beta-D-fucose, in red crosses. The green diamonds show the spectrum such as it is annotated, once set on the mass grid. One can notice the preeminence of two peaks, appearing on both spectra at about 163 and 209Da. Still, many peaks differ between these spectra. The black circles correspond to the same source estimated based on the inversion of $\mathbf{A}^{\text{annot.}}$, using Equation (4). This sums up to estimating the sources from the reference elution profile obtained from the annotations. The obtained spectra are then very similar to the ones yielded by nGMCA, which shows that, in practice, the data are more complex than what the annotations indicate.

Fig. 9a and Fig. 9b display the number of peaks respectively



(a) with nGMCA.



(b) with non-stat. nGMCA (stoch.).

Fig. 9: Number of peaks recovered peaks from the searched compounds (real dataset).

estimated by nGMCA and non-stat. nGMCA (stoch.) which are compatible with the annotations (both in mass and in time). It is difficult to give a quantitative evaluation of these results on the real dataset. Still, these figures provide some insights on the quality of the identifications of the searched sources:

- all the annotated sources are not necessarily recovered. On the top figure (nGMCA), the line associated to carnitine is empty and thus no peak from this source has been identified by the algorithm. Non-stat. nGMCA (stoch.) however recovers at least one peak from each of the sources on this example.
- some estimated sources contain peaks from several compounds (columns with more than one non-null coefficient). These additional peaks are interferences which must be avoided. A comparison between Fig. 9a and Fig. 9b, seems to indicate that the non-stat. nGMCA (stoch.) is less prone to interferences than nGMCA.
- some compounds are recognized in several estimated sources (rows with more than one non-null coefficients). This means that these compounds have been split into several parts. This is for instance what was observed for sources #6 and #7, and the sources #9 and #10 of the reconstructions with nGMCA on Fig. 7b. This splitting of the compounds into several parts seems once again less frequent with non-stat. nGMCA (stoch.).

C. Quantitative study on the real data

Considering the above qualitative example, one can design quantitative criteria. To this extent, the 200 largest recovered peaks are extracted from each reconstruction obtained with an NMF algorithm. This choice of a predefined number of peaks allows a more accurate comparison between the algorithms, since it does not privilege very selective thresholding strategies which could artificially reduce the amount of interferences.

Such as in the previous section, peaks are affected to a compound when they are compatible both in mass and in time with the annotations. Since we estimate 18 sources so as to recover the 11 annotated compounds, we define the matrix $\mathbf{E} \in \mathbb{R}^{11\times 18}$ such that each element $\mathbf{E}_{i,j}$ is the sum of the quadratic energies of the peaks of the j^{th} estimated source which are affected to the i^{th} compound. We also define the vector $e \in \mathbb{R}^{18}$ such that e_j is the sum of the quadratic energy of the peaks of the j^{th} estimated source which are not affected to any compound. We then affect each compound to a unique estimated source by choosing a permutation of the columns of $\mathbf{E}$ which maximizes $\sum_{i=1}^{11} \mathbf{E}_{ii}$. Thus the i^{th} compound is affected to the i^{th} estimated source. We are then able to define the following quantitative criteria:

- correct energy ratio: mean energy ratio of the estimated source which belongs to the compound it is affected to:

$$\text{corr.} = \frac{1}{11}\sum_{j=1}^{11} \frac{\mathbf{E}_{jj}}{e_j + \sum_{i=1}^{11} \mathbf{E}_{i,j}}. \tag{11}$$

- interference energy ratio: mean energy ratio of the estimated sources peaks belonging to an annotated compound which is different from the one to which the estimated source is affected, with respect to the total energy of the estimated source (energy ratio on the column of $\mathbf{E}$, taking into account the non-identified peaks):

$$\text{interf.} = \frac{1}{11}\sum_{j=1}^{11} \left(1 - \frac{\mathbf{E}_{jj} - e_j}{e_j + \sum_{i=1}^{11} \mathbf{E}_{i,j}}\right). \tag{12}$$

- splitting energy ratio: mean energy of the peaks of a compound which are not affected to it, with respect to the total energy of the identified peaks for this compound (energy ratio over the lines of $\mathbf{E}$). This criterion aims at evaluating the tendency of algorithms to split the compounds into several estimated source, while in the ideal case each of these compounds should be gathered in a unique estimated source. It is computed as follows:

$$\text{split.} = \frac{1}{11}\sum_{i=1}^{11} \left(1 - \frac{\mathbf{E}_{ii}}{\sum_{j=1}^{18} \mathbf{E}_{i,j}}\right). \tag{13}$$

- identification ratio: the number of compounds for which at least one peak was identified, with respect to the number of annotated compounds (i.e. 11).
- recovery: we will consider that all the sources have been recovered if at least one peak of each annotated compound has been identified. This does not mean that the factorization is perfect but it is nevertheless a necessary condition for it.

All this criteria are averaged over 200 realizations, corresponding to different initializations of the algorithms. The results are given in table I and the standard deviations are provided between parenthesis. This table is teaching in several aspects:

- evaluation of the algorithms: most of the criteria have similar behaviors. The most discriminative criteria is the recovery rate which goes from 29.5% to 97%. Most of the algorithms obtain between 30 and 40% on this criterion. Kim & Park's algorithms however still outperforms most of the other algorithms, with a recovery rate at 56%. Only non-stat. nGMCA (stoch.) reaches a better recovery rate, at 97%, while keeping one of the lowest interference ratio.
- robustness to the initialization: the variability of all the algorithms is significant. The initialization is therefore of the uttermost importance. This tends to confirm the presence of a large number of critical points to which the algorithms are very sensitive. The lower variability of non-stat. nGMCA (stoch.) —by more than 1 point— indicates that the stochastic term is helpful in order to be more robust to these critical points.
- data fidelity term: the non-stationary data fidelity term seems beneficial for nGMCA and for the multiplicative update algorithms, since mult. (non-stat.) performs significantly better than mult. (L2). The version with a Kullback-Leibler divergence obtains a recovery rate comparable to the one of mult. (non-stat.) (43.5%), although at the price of a deterioration on all the other criteria. The use of a particular data fidelity term is therefore far from being neutral. On these data, the full integration of the non-stationarity in the nGMCA framework is the most efficient approach.
- regularizations: The sparse regularization used in Kim & Park's algorithm seems efficient and robust. The standard $\ell 1$ regularization is also beneficial when used accurately in the non-stationary version of nGMCA.

In summary, the adaptation of nGMCA to non-stationary noise and the use of a stochastic term yields very good and robust results. Kim & Park's algorithm benefits here from the fact that the mixtures in this dataset are not very complex, with spectra sharing few common peaks and elution profiles which do not overlap too much. It would however not perform so well in more difficult settings where the dominant mixture model is not verified anymore, such as with a larger number of chemical compounds for instance.

V. Discussion about the modelization

The experiments above yield promising results for the use of NMF on LC/MS data. They however allow for many potential improvements.

A. Noise model

Noise is not yet fully understood in LC/MS. A more accurate noise model may be a contamination with a mixture of Poisson noise, at low intensities, and multiplicative noise at larger intensities, as suggested in [12]. In this paper, we have used several types of data fidelity term, including Kullback-Leibler divergence, and regular and weighted euclidean distances. The use of a non-stationary prior thanks to the utilisation of a weighted euclidean distance seemed better suited than the standard euclidean distance. Modifying $\mathcal{D}$ to take into account a more precie noise model could prove helpful.

algorithm \ criterion	corr. (%)	interf. (%)	split. (%)	iden. (%)	recov.
mult. (non-stat.)	74.6 (+/-6.0)	13.9 (+/-7.0)	14.7 (+/-5.3)	94.4 (+/-5.4)	44
mult. (KL)	69.2 (+/-5.6)	14.6 (+/-5.9)	15.2 (+/-4.8)	94.0 (+/-5.9)	43.5
mult. (L2)	74.7 (+/-5.6)	13.9 (+/-6.8)	13.9 (+/-4.8)	93.6 (+/-5.5)	37.5
Kim & Park	78.1 (+/-5.1)	9.6 (+/-7.4)	10.8 (+/-3.8)	95.9 (+/-4.5)	54.6
HALS	75.3 (+/-4.1)	12.4 (+/-6.2)	13.4 (+/-3.7)	93.5 (+/-4.3)	29.5
nGMCA	77.2 (+/-4.9)	11.2 (+/-5.2)	11.0 (+/-3.5)	93.8 (+/-5.0)	36
non-stat. nGMCA (stoch)	80.3 (+/-3.4)	4.5 (+/-2.4)	10.3 (+/-3.1)	99.7 (+/-1.6)	97

TABLE I: Mean results (and their standard deviations) over 200 different initializations.

B. Mixture model

The mixture model, which states that the observations are linear combinations of the sources (cf. equation (1)), is not perfect in practice. Several chemical compounds can indeed interact with each others. If one has access to a more accurate model, it may prove very helpful to modify the divergence in order to take it into account.

Another possibility consists in adding a robust term to the problem such as in [25]. nGMCA's framework could easily encompass the additional minimization of this robust term. This variable would gather features of the data which have high energy but cannot be modeled by low rank non-negative matrices, such as deviations from the linear model. By doing so, it would prevent these high energy features to contaminate the estimation of the spectra and elution profiles.

C. Data model for the spectra and the elution profiles

The use of regularizations for the spectra $\mathbf{S}$ such as the sparse ones introduced in section III can be beneficial. Among the ones tested in the experiments, the sparse regularization $\sum_{t=1}^{n} \|\mathbf{S}_{.,t}\|_1^2$ used in Kim & Park's algorithm was accurate on this dataset, although it could fail if the dominant mixture model is not verified anymore.

The algorithms introduced in this article use the sparsity of the spectra in $\mathbf{S}$ in order to help disambiguate the sources. $\mathbf{A}$ is however also sparse. This knowledge could thus be used in order to obtain more relevant factorizations. Also, while $\mathbf{A}$ is indeed sparse in the direct domain, it could be even sparser in a wavelet domain since the elution profiles are smooth. The use of sparsity in a transformed domain in NMF was tested with success in [26]. Still, its use on the mixtures along with a sparse regularization of the spectra was only used —with good results— in BSS [27], i.e. without the non-negative assumption of NMF.

REFERENCES

[1] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," Environmetrics, vol. 5, no. 2, pp. 111–126, 1994.

[2] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol. 401, no. 6755, pp. 788–791, 1999.

[3] S. Moussaoui, H. Hauksdóttir, F. Schmidt, C. Jutten, J. Chanussot, D. Brie, S. Douté, and J. A. Benediktsson, "On the decomposition of Mars hyperspectral data by ICA and Bayesian positive source separation," Neurocomputing, vol. 71, no. 10-12, pp. 2194–2208, 2008.

[4] D. A. Snyder, F. Zhang, S. L. Robinette, L. Bruschweiler-Li, and R. Brüschweiler, "Non-negative matrix factorization of two-dimensional NMR spectra: Application to complex mixture analysis," The Journal of Chemical Physics, vol. 128, no. 5, pp. 1–4, 2008.

[5] I. Toumi, B. Torrésani, and S. Caldarelli, "Effective Processing of Pulse Field Gradient NMR of Mixtures by Blind Source Separation," Analytical Chemistry, vol. 85, no. 23, pp. 11 344–11 351, 2013.

[6] R. Dubroca, C. Junot, and A. Souloumiac, "Weighted NMF for high-resolution mass spectrometry analysis," in Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European, Aug. 2012, pp. 1806–1810.

[7] W. S. B. Ouedraogo, "Méthode géométrique de séparation de sources non-négatives : Applications limagerie dynamique TEP et la spectrométrie de masse," Ph.D. dissertation, Université de Grenoble et Université de Tunis El Manar, 2012.

[8] S. A. Vavasis, "On the Complexity of Nonnegative Matrix Factorization," SIAM Journal on Optimization, vol. 20, no. 3, pp. 1364–1377, 2009.

[9] A. Makarov, "Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis," Analytical Chemistry, vol. 72, no. 6, pp. 1156–62, 2000.

[10] A. Makarov, E. Denisov, A. Kholomeev, W. Balschun, O. Lange, K. Strupat, and S. Horning, "Performance Evaluation of a Hybrid Linear Ion Trap/Orbitrap Mass Spectrometer," Analytical Chemistry, vol. 78, no. 7, pp. 2113–2120, 2006.

[11] E. Werner, V. Croixmarie, T. Umbdenstock, E. Ezan, P. Chaminade, J.-C. Tabet, and C. Junot, "Mass Spectrometry-Based Metabolomics: Accelerating the Characterization of Discriminating Signals by Combining Statistical Correlations and Ultrahigh Resolution," Analytical Chemistry, vol. 80, no. 13, pp. 4918–4932, 2008.

[12] M. Anderle, S. Roy, H. Lin, C. Becker, and K. Joho, "Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum." Bioinformatics, vol. 20, no. 18, pp. 3575–3582, 2004.

[13] P. Du, G. Stolovitzky, P. Horvatovich, R. Bischoff, J. Lim, and F. Suits, "A noise model for mass spectrometry based proteomics," Bioinformatics, vol. 24, no. 8, pp. 1070–1077, 2008.

[14] H. P. Benton, D. M. Wong, S. A. Trauger, and G. Siuzdak, "XCMS2: processing tandem mass spectrometry data for metabolite identification and structural characterization," Analytical Chemistry, vol. 80, no. 16, pp. 6382–6389, Aug. 2008.

[15] T. Pluskal, S. Castillo, A. Villar-Briones, and M. Oresic, "MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data." BMC Bioinformatics, vol. 11:395, pp. 1–11, 2010.

[16] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," Advances in neural information processing systems, vol. 13, no. 1, pp. 556–562, 2001.

[17] E. F. Gonzalez and Y. Zhang, "Accelerating the Lee-Seung Algorithm for Nonnegative Matrix Factorization," Dept. Comput. Appl. Math., Rice University, Tech. Rep. TR05-02, 2005.

[18] J. Rapin, J. Bobin, A. Larue, and J.-L. Starck, "Sparse and Non-Negative BSS for Noisy Data," IEEE Transactions on Signal Processing, vol. 61, pp. 5620–5632, 2013.

[19] P. O. Hoyer, "Non-negative sparse coding," in Proceedings of Neural Networks for Signal Processing, 2002, pp. 557–565.

[20] J. Kim and H. Park, "Sparse Nonnegative Matrix Factorization for Clustering," Georgia Institute of Technology, Tech. Rep. GT-CSE-08-01, 2008.

[21] A. Cichocki, R. Zdunek, and S.-i. Amari, "Hierarchical ALS Algorithms for Nonnegative Matrix and 3D Tensor Factorization," in Proceedings of ICA. Springer, 2007, pp. 169–176.

[22] N. Gillis and F. Glineur, "Accelerated Multiplicative Updates and Hierarchical ALS Algorithms for Nonnegative Matrix Factorization," Neural Computation, vol. 24, no. 4, pp. 1085–1105, 2012.

[23] N. Gillis, "Sparse and Unique Nonnegative Matrix Factorization

Through Data Preprocessing," Journal of Machine Learning Research, vol. 13, pp. 3349–3386, 2012.

[24] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," IEEE Transactions on Audio, Speech & Language Processing, vol. 14, no. 4, pp. 1462–1469, 2006.

[25] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" Journal of the ACM, vol. 58, no. 3, p. 11, 2011.

[26] J. Rapin, J. Bobin, A. Larue, and J.-L. Starck, "NMF with Sparse Regularizations in Transformed Domains," SIAM Journal on Imaging Sciences, vol. 7, no. 4, pp. 2020–2047, 2014, sIAM J. Imaging Sci., 7(4), 20202047. (28 pages).

[27] Y. Moudden and J. Bobin, "Hyperspectral BSS Using GMCA With Spatio-Spectral Sparsity Constraints," IEEE Transactions on Image Processing, vol. 20, no. 3, pp. 872–879, 2011.