Signal Processing 123 (2016) 75–83

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Application of non-negative matrix factorization to LC/MS data

Jérémy Rapin [a,b], Antoine Souloumiac [a,*], Jérôme Bobin [b], Anthony Larue [a], Chistophe Junot [c], Minale Ouethrani [c], Jean-Luc Starck [b]

[a] CEA, LIST, 91191 Gif-sur-Yvette Cedex, France
[b] CEA, IRFU, Service d'Astrophysique, 91191 Gif-sur-Yvette, France
[c] CEA, DSV/iBiTec-S, Service de Pharmacologie et d'Immunoanalyse, Laboratoire d'Étude du Métabolisme des Médicaments, 91191 Gif-sur-Yvette Cedex, France

ARTICLE INFO

Article history:
Received 18 March 2015
Received in revised form
9 December 2015
Accepted 14 December 2015
Available online 7 January 2016

Keywords:
BSS
NMF
Sparsity
Multiplicative noise
LC/MS

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 their processing. These data are however 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 paper, we propose an adaptation of a state-of-the-art NMF algorithms so as to specifically be able to deal with LC/MS data, by using a non-stationary noise model and a stochastic term. We finally perform experiments and compare standard NMF algorithms on both simulated data and an annotated LC/MS dataset. The results of these experiments highlight the significant improvement obtained by our adaptation over other NMF algorithms.

1. Introduction

1.1. 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 solvent. 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 paper, and each of the m lines of this matrix is a n-sample long spectrum at a given acquisition time.

1.2. 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 Tauler et al. [1], Paatero and Tapper [2], and Lee and Seung [3]. The non-negative assumption arises naturally in many applications such as hyperspectral imaging [4], nuclear magnetic resonance [5,6] or LC/MS [7–9]. Indeed, in LC/MS, the mass spectra are non-negative, and the mixtures are related to

* Corresponding author.
E-mail address: antoine.souloumiac@cea.fr (A. Souloumiac).

http://dx.doi.org/10.1016/j.sigpro.2015.12.014
0165-1684/

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 $\mathbf{S}_{j,\cdot} \in \mathbb{R}^{1\times n}$:

$$\mathbf{Y}_{i,\cdot} = \sum_{j=1}^{r} \mathbf{A}_{i,j}\mathbf{S}_{j,\cdot} + \mathbf{Z}_{i,\cdot}, \ \forall i \in \{1,\dots,m\}, \tag{1}$$

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:

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

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 [10] and finding an optimal solution is therefore very difficult. Different NMF algorithms or even different initializations of a same algorithm therefore yield different factorizations.

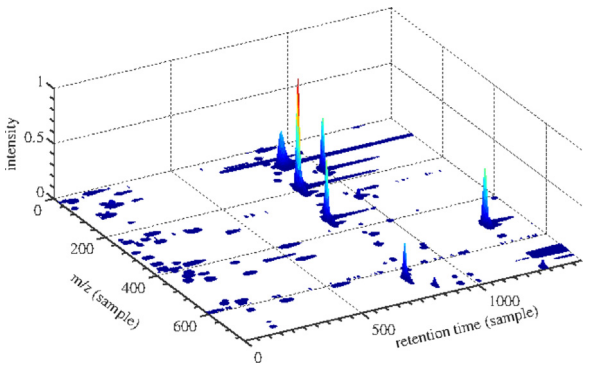

Fig. 1. LC/MS chromatogram of a small sample (low intensity peaks have been removed to ease the visualization).

1.3. Contribution

Although NMF seems particularly well suited for processing LC/MS data, it is seldom used on this type of data. In this paper, we examine LC/MS data and test reference and state-of-the-art NMF algorithms on an annotated dataset so as to understand the difficulties of these algorithms in dealing with LC/MS data. We then propose an adaptation of a state-of-the-art NMF algorithm, the non-negative generalized morphological component analysis (nGMCA), aiming at overcoming these difficulties. 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.

2. LC/MS dataset

The data considered in this paper were acquired from a mixture of eleven commercial chemical compounds for which the mass spectrum is known [11] and retention times are annotated (see Fig. 5 for the list of compounds). This sample was analyzed in an LC/MS pipeline using an orbitrap mass analyzer [12,13]. We focus on a time range going from 2 to 18 min 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.

2.1. 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. 2(a). 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, the ions are produced

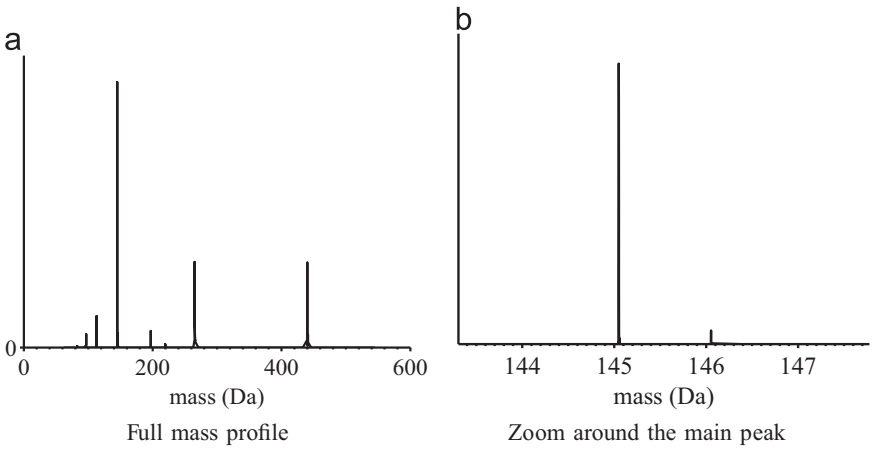

Fig. 2. Examples of mass profile at a time t.