Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Blind decomposition of infrared spectra using flexible component analysis

Ivica Kopriva [a,*], Ivanka Jerić [b], Andrzej Cichocki [c,d]

[a] Division of Laser and Atomic Research and Development, Croatia
[b] Division of Organic Chemistry and Biochemistry Ruđer Bošković Institute, Bijenička cesta 54, HR-10000, Zagreb, Croatia
[c] Laboratory for Advanced Brain Signal Processing Brain Science Institute, RIKEN 2-1, Hirosawa, Wako-shi, Saitama, 351-0198, Japan
[d] Warsaw University of Technology and Systems Research Institute, PAN, Poland

ARTICLE INFO

Article history:
Received 14 January 2009
Received in revised form 20 February 2009
Accepted 3 April 2009
Available online 8 April 2009

Keywords:
Chemometrics
Blind source separation
Flexible component analysis
Sparse component analysis
Wavelet translation
FT-IR spectroscopy

ABSTRACT

The paper presents flexible component analysis-based blind decomposition of the mixtures of Fourier transform of infrared spectral (FT-IR) data into pure components, wherein the number of mixtures is less than number of pure components. The novelty of the proposed approach to blind FT-IR spectra decomposition is in use of hierarchical or local alternating least square nonnegative matrix factorization (HALS NMF) method with smoothness and sparseness constraints simultaneously imposed on the pure components. In contrast to many existing blind decomposition methods no *a priori* information about the number of pure components is required. It is estimated from the mixtures using robust data clustering algorithm in the wavelet domain. The HALS NMF method is compared favorably against three sparse component analysis algorithms on experimental data with the known pure component spectra. Proposed methodology can be implemented as a part of software packages used for the analysis of FT-IR spectra and identification of chemical compounds.

1. Introduction

Extraction of the pure component spectra from the mixtures of their linear combinations is of great interest in many applications. Classical approach to extraction of the spectra of pure components is to match the mixture's spectra with a library of reference compounds. This approach is ineffective with the accuracy strongly dependent on the library's content of the pure component spectra and cannot reflect the variation of the spectral profile due to environmental changes. Alternatives to library matching approach are blind decomposition methods, wherein pure components' spectra are extracted using mixtures spectra only. Blind approaches to pure components spectra extraction have been reported in NMR spectroscopy [1], infrared (IR) [2–4] and near infrared (NIR) spectroscopy [4–6], EPR spectroscopy [7,8], mass spectrometry [1,4,9,10] Raman spectroscopy [11,12] etc. In a majority of blind decomposition schemes independent component analysis (ICA) [13–15] is employed to solve related blind source separation (BSS) problem. ICA assumes that: (*i*) pure components are statistically independent, (*ii*) at most one is normally distributed and (*iii*) number of mixtures is greater than or equal to the unknown number of pure components. The two requirements: to have more linearly independent mixtures than pure components and to have statistically independent pure components seem to be most critical for the success of the BSS approach to blind decomposition of the mixtures spectra into pure components spectra [4,5,8,10]. Statistical independence assumption is certainly not fulfilled in the case of IR spectra [2–6] because they are highly correlated i.e. overlapped. Raw data preprocessing technique by first or second order derivative has been used in FT-IR spectra analysis to reduce level of statistical dependence among pure components, [2–6]. This technique actually belongs to the generalization of the ICA known as dependent component analysis (DCA), [14,16–18]. An algorithm for blind decomposition of EPR spectra has been derived in [8] minimizing contrast function that exploits sparseness rather than statistical independence among the pure components. Unfortunately, sparseness criterion cannot be used in the case of FT-IR spectra due to high degree of overlap between them, especially in wavelength or wavenumber domain. All discussed blind spectra decomposition methods require the number of mixtures spectra to be equal to or greater than the unknown number of pure components spectra. In a number of real world situations it is however not easy to acquire mixtures spectra with different concentrations of the pure components spectra. In this regard it is desirable property of blind decomposition methods to solve related BSS problem with as few mixtures as possible. Here, we demonstrate flexible component analysis (FCA) approach to blind decomposition of more than two pure components FT-IR spectra from two mixtures only. To solve related underdetermined BSS (uBSS) problem we use recently developed nonnegative matrix factorization (NMF) algorithm that is known as local or hierarchical alternating least squares (HALS) NMF algorithm [19,20,40]. Its unique property is to estimate concentration or mixing matrix globally and pure components spectra locally, wherein smoothness and sparseness constraints are simultaneously imposed on the pure components spectra. Unlike

[*] Corresponding author. Tel.: +385 1 4571 286; fax: +385 1 4680 104.
E-mail address: ikopriva@irb.hr (I. Kopriva).

0169-7439/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2009.04.002

majority of the BSS algorithms that assume the number of pure components to be known, proposed approach estimates it from the mixtures spectra in the wavelet domain by means of data clustering algorithm, [21]. Transformation of the mixtures spectra in wavelet domain yields representation that is significantly sparser than in original wavenumber domain. This enables more accurate estimation of the number of pure components spectra, especially due to the fact that used data clustering algorithm requires that pure components spectra are in average sparse in the chosen basis. Comparison of the HALS NMF approach against sparse component analysis (SCA) based approach [22–25] on experimental uBSS problem, which is presented in Section 3, yields favorable results. Therefore, it is believed that proposed FCA-based approach to blind extraction of the FT-IR pure components spectra is practically important. The rest of the paper is organized as follows. We introduce data clustering algorithm, SCA and FCA concepts in Section 2. Results and discussion of the experimental comparative performance analysis of the FCA and SCA approaches on two mixtures of IR spectra containing three pure components are given in Section 3. Conclusions are presented in Section 4.

2. Computational methods

Like many decomposition methods proposed approach is based on static linear mixture model

$$\mathbf{X} = \mathbf{AS} \tag{1}$$

where $\mathbf{X} \in \mathrm{R}_{0+}^{N\times T}$ represents matrix of N measured mixtures spectra across T wavenumbers, $\mathbf{A} \in \mathrm{R}_{0+}^{N\times M}$ represents the matrix of concentration profiles also called the mixing matrix and matrix $\mathbf{S} \in \mathrm{R}_{0+}^{M\times T}$ contains M pure components spectra across T wavenumbers. Due to the nature of the problem all quantities in Eq. (1) are nonnegative. As already pointed out, the number of pure components M is in principle unknown although many BSS/ICA algorithms assume that it is either known in advance or can be easily estimated. This does not seem to be true in practice, especially when the BSS problem is underdetermined. Here, we shall treat M as unknown parameter that will be estimated by the clustering algorithm to be described in Section 2.1. In addition to estimate the number of pure components used data clustering algorithm also estimates the concentration matrix. This is necessary for the SCA approach described in Section 2.2, but is not necessary for HALS NMF approach described in Section 2.3. In overall, the BSS problem related to blind FT-IR spectra decomposition consists of: (*i*) estimating the number of pure components spectra; (*ii*) estimating the matrix of the pure components spectra S; (*iii*) estimating the concentration matrix A. All three tasks are executed using matrix of mixtures spectra X only. In addition to that, we allow the number of pure components spectra M to be greater than the number of mixtures spectra N. Hence, blind FT-IR spectra decomposition problem becomes uBSS problem.

2.1. Data clustering

In FT-IR spectra decomposition problem considered in this paper we shall assume that pure components spectra are in average $k = M$-1 sparse in wavelet domain. This implies that at each coordinate in wavelet domain in average only one pure component is active i.e. nonzero. This assumption allows to reduce number of mixtures to $N = 2$, hence reducing the computational complexity of to be used data clustering algorithm [21] by reducing dimension of the concentration subspaces, that equals average number of active components, to 1. However, we are aware that it is not realistic to demand that pure components FT-IR spectra do not overlap in any representation domain including wavelet domain used here. That is why we expect that pure components spectra are only in average $k = M - 1$ sparse in wavelet domain. Under such assumption the appropriately chosen function, see Eq. (3), will effectively cluster data, wherein the number of clusters corresponds with the estimate of the number of pure components M. If the number of coordinates that violates $k = M - 1$ sparseness assumption in wavelet domain is relatively large this will influence accuracy of the estimation of the concentration matrix due to the repositioning of the cluster centers. It will not however influence in the same amount the accuracy of the estimation of the number of clusters. Thus, performance of the SCA algorithms that require the estimate of the concentration matrix in order to proceed to the next phase and solve underdetermined system of linear equations will be affected significantly if FT-IR spectra are not sparse enough in the chosen basis. On the other hand proposed FCA approach will be significantly less sensitive to the level of sparseness of the FT-IR spectra because it only requires from the clustering algorithm the estimate of the number of pure components spectra.

Because solution of the BSS problem is generally characterized by scale indeterminacy we shall assume the unit norm constraint (in the sense of ℓ_2 norm) on the columns of the concentration matrix A, i.e., $\{\|\mathbf{a}_m\|_2 = 1\}_{m=1}^{M}$. As already pointed out, in this paper we do assume the number of mixtures to be $N = 2$. Thus, the normalized mixing vectors $\{\mathbf{a}_\mathrm{m}\}_{\mathrm{m}=1}^{\mathrm{M}}$lie in the first quadrant on the unit circle, i.e., they are parameterized as:

$$\mathbf{a}_m = [\cos(\varphi_m) \quad \sin(\varphi_m)]^{\mathrm{T}} \qquad m = 1, \ldots, M \tag{2}$$

where φ_m represents mixing angle that is confined in the interval $[0, \pi/2]$. We do assume that mixtures are transformed into wavelet domain through wavelet transform

$$X_n(a,b) = \tfrac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x_n(t)\psi\left(\frac{t-b}{a}\right)dt \quad n = 1, \ldots, N$$

where the a and b represent respectively scale (resolution level) and time shift and $\Psi(t)$ represents wavelet function. After extensive experiments we have found out that symmlets with two to eight vanishing moments yield best results in terms of sparseness of X(a,b). Thus, the results reported in Section 3 were obtained with the symmlets with the four vanishing moments. The fact that symmlets performed best is just experimental finding. We have also tried Daubechie's wavelet of different order, Haar wavelet, Morlet wavelet, Mexican hat wavelet, Coiflets and some biorthogonal wavelets. From the sparse representation point of view the key property of the wavelet is to match well the waveform of the particular signal of interest (in this case the FT-IR spectra). It is however very hard to find such a wavelet in case of FT-IR signals. Perhaps, the optimal solution would be to design new wavelet that will reflect better morphological properties of FT-IR data than standard wavelets do. Wavelet transform above can be used either as continuous or as discrete. In the results presented in Section 3 we have used discrete shift invariant wavelet transform with the resolution levels corresponding to $a = 2^1$ or $a = 2^2$. By assuming *1*-dimensional concentration subspaces the clustering algorithm [21] is outlined by the following steps:

1) We remove all data points close to the origin for which applies: $\{\|\mathbf{x}(a, b_t)\|_2 \leq \varepsilon\}_{t=1}^{T}$, where ε represents some predefined threshold. This corresponds with the case when pure components spectra are close to zero.
2) Normalize to unit $\ell 2$ norm remaining data points $\mathbf{x}(a,b_t)$, i.e., $\{\mathbf{x}_,(a, b_t) \leftarrow \mathbf{x}(a,b_t)/\|\mathbf{x}\,(a,b_t)\|_2\}_{t=1}^{\overline{T}}$, where $\overline{T} \leq T$ denotes number of data points that remained after the elimination process in step 1.
3) Calculate function f(a), where a is defined with Eq. (2):

$$f(\mathbf{a}) = \sum_{t=1}^{\overline{T}} exp\left(-\frac{d^2(\mathbf{x}(a,b_t),\mathbf{a})}{2\sigma^2}\right) \tag{3}$$

where $d(\mathbf{x}(a,b_t),\mathbf{a}) = \sqrt{1-(\mathbf{x}(a,b_t)\cdot\mathbf{a})^2}$ and $(\mathbf{x}(a,b_t)\cdot\mathbf{a})$ denotes inner product. Parameter σ in Eq. (3) is called dispersion. If set to sufficiently small value, in our experiments this turned out to be