Trends in Analytical chemistry 50 (2013) 22–32

Contents lists available at SciVerse ScienceDirect

Trends in Analytical chemistry

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

Review

Independent Components Analysis with the JADE algorithm

D.N. Rutledge *, D. Jouan-Rimbaud Bouveresse

INRA, UMR 1145 Ingénierie Procédés Aliments, F-75005 Paris, France
AgroParisTech, UMR 1145 Ingénierie Procédés Aliments, F-75005 Paris, France

ARTICLE INFO

Keywords:
Chemometrics
Complex data set
Independence
Independent Components Analysis (ICA)
Interpretable signal
Joint Approximate Diagonalization of Eigenmatrices (JADE)
Multiway data array
Parallel Factor Analysis (PARAFAC)
Principal Components Analysis (PCA)
Three-way data

ABSTRACT

Independent Components Analysis (ICA) is a relatively recent method, with an increasing number of applications in chemometrics. Of the many algorithms available to compute ICA parameters, the Joint Approximate Diagonalization of Eigenmatrices (JADE) algorithm is presented here in detail. Three examples are used to illustrate its performance, and highlight the differences between ICA results and those of other methods, such as Principal Components Analysis. A comparison with Parallel Factor Analysis (PARAFAC) is also presented in the case of a three-way data set to show that ICA applied on an unfolded high-order array can give results comparable with those of PARAFAC.

Contents

* Corresponding author at: AgroParisTech, UMR 1145 Ingénierie Procédés Aliments, F-75005 Paris, France. Tel.: +33 (0) 1 44 08 16 48; Fax: +33 (0) 1 44 08 16 53.
E-mail address: douglas.rutledge@agroparistech.fr (D.N. Rutledge).

0165-9936/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trac.2013.03.013

1. Introduction

Independent Components Analysis (ICA) is becoming a method of choice in different scientific domains [1], including chemometrics. This method was first developed in the 1990s [2,3] in the field of signal processing in telecommunications [4,5], and its use has extended to all domains where the notion of "signal" is present; for example in the medical field, the analysis of electro-encephalograms [6], statistical process control [7], analytical chemistry [8], and metabolomics [9]. Several developments of ICA were recently reported [10]. In chemistry, it is frequent to analyze samples with spectroscopic techniques, such as infrared spectroscopy, fluorescence spectroscopy, and nuclear magnetic resonance. The analysis of such multi-dimensional and (highly-)correlated data sets relies on chemometric techniques [11].

So far, Principal Components Analysis (PCA) has been widely used in this respect. However, the interpretation of PCA loadings vectors is often not so straightforward, as these vectors may represent combinations of different phenomena described by the data. On the other hand, each ICA "loadings vector" (source signal) describes one independent phenomenon. This difference is due to the intrinsic properties of the two methods: while PCA is based on determining the orthogonal directions of maximum dispersion of the samples in the multidimensional space defined by the variables, the aim of ICA is to recover the pure source signals mixed together in the observed signals.

Several different algorithms to calculate Independent Components (ICs) are available, among which one can cite FastICA [12], InfoMax [13], and the Joint Approximate Diagonalization of Eigenmatrices (JADE) [14]. We have applied JADE successfully to several types of data (3D-fluorescence data, mid-infrared spectra and mass spectra). The major advantage of JADE over other algorithms is that it is based on matrix computation, involving matrix diagonalization, as is done in other "standard" chemometric methods, such as PCA or Factorial Discriminant Analysis (FDA). Other algorithms (e.g., FastICA) rely on an optimization procedure, and hence may yield variable results depending on the starting point and on the optimization path followed by the search algorithm. Comparisons of ICA algorithms have been published [15–18], but the conclusions differ, depending on the data analyzed and on the criteria used to evaluate the results.

The goal of this article is to present the concepts of ICA, and its advantages and disadvantages in chemometrics. For the reasons above, a thorough comparison of the results obtained from different ICA algorithms is not given in this article, which focuses on the JADE algorithm.

2. Theory

2.1. PCA and ICA: two different ways of approaching multivariate data

The mathematical methods used in multivariate data analysis are based on matrix algebra. The analyzed data are organized into a data matrix, $\mathbf{X}$ $(r \times c)$, the r samples corresponding to the rows of $\mathbf{X}$, while the c variables are the columns of $\mathbf{X}$. When signals are analyzed, it is usual to represent them as the rows of the data matrix, while the columns represent the variables for which the intensities have been measured.

In PCA, the data matrix $\mathbf{X}$ is seen as a collection of *objects* (the samples, in the rows of $\mathbf{X}$) in a multidimensional space defined by the original variables. Samples with similar values for the variables will be located close together in that space, whereas samples with very different values will be far apart. If the data matrix contains only Gaussian noise, the objects will be distributed spherically in the space of the variables. If, on the other hand, a non-spherical distribution is observed, it may be assumed there is information in the data.

The basic assumption of PCA is that the directions in which the samples are most dispersed are the most interesting and therefore the corresponding vectors are the most informative combinations of the original variables. Here, variability is assumed to be directly related to information.

PCA calculates new latent variables, called *Principal Components* (PCs), to describe these directions of maximum dispersion of the objects. The first PC is the vector describing the direction of maximum sample dispersion. Each following PC describes the maximal remaining variability, with the additional constraint that it must be orthogonal to all the earlier PCs to avoid it containing any of the information already extracted from the data matrix. In other words, each PC extracts as much remaining variance from the data as possible. The calculated PCs are weighted sums of the original variables, the weights being elements of a so-called *loadings vector*. Inspection of these loadings vectors may help determine which original variables contribute most to this PC direction. However, PCs being mathematical constructs describing the directions of greatest dispersion of the samples, there is no reason for the loadings vectors to correspond to underlying signals in the data set. Most of the time, PCs are combinations of pure source signals, and do not describe physical reality. For this reason their interpretation can be fraught with danger.

2.2. Independent Components Analysis

ICA is a method of Blind Source Separation (BSS) [1]. The assumption underlying ICA is that each row of the data matrix is a weighted sum of pure source signals, the weights being proportional to the contribution of the corresponding pure signals to that particular mixture. The original source signals and their proportions in the analyzed mixtures, are unknown. In ICA, $\mathbf{X}$ is not seen as a collection of points in a multidimensional space, but rather as a collection of signals (in the rows) with a certain number of common sources. ICA aims to extract these pure sources, underlying the observed signals, as well as their concentration in each mixture. For example, in chemistry, a signal can correspond to the spectrum of a mixture of several pure compounds: ICA may be used to find the pure spectra of the compounds and the concentration of each compound in each mixture.

Let us suppose that three "mixture-signals", $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$, are linear combinations of two pure signals, $\mathbf{s}_1$ and $\mathbf{s}_2$. These linear mixtures can be written as:

$$\mathbf{x}_1 = a_{11}\,\mathbf{s}_1 + a_{12}\,\mathbf{s}_2 \tag{1a}$$

$$\mathbf{x}_2 = a_{21}\,\mathbf{s}_1 + a_{22}\,\mathbf{s}_2 \tag{1b}$$

$$\mathbf{x}_3 = a_{31}\,\mathbf{s}_1 + a_{32}\,\mathbf{s}_2 \tag{1c}$$

In matrix notation:

$$\mathbf{X} = \mathbf{AS} \tag{2a}$$

where

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \tag{2b}$$

and

$$\mathbf{S} = \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \end{bmatrix} \tag{2c}$$