Talanta 179 (2018) 538–545

Contents lists available at ScienceDirect

Talanta

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

Determination of the optimal number of components in independent components analysis

Amine Kassouf[a], Delphine Jouan-Rimbaud Bouveresse[b], Douglas N. Rutledge[b,*]

[a] Department of Chemistry and Biochemistry, Faculty of Sciences Ii, Lebanese University, Jdeideth El Matn 90656, Fanar, Lebanon
[b] UMR Ingenierie Procedes Aliments, AgroParisTech, Inra, Universite Paris-saclay, F-91300 Massy, France

ARTICLE INFO

Keywords:
Independent components analysis (ICA)
Random_ICA
KMO_ICA_Residuals
ICA_corr_y
Optimal number
Validation

ABSTRACT

Independent components analysis (ICA) may be considered as one of the most established blind source separation techniques for the treatment of complex data sets in analytical chemistry. Like other similar methods, the determination of the optimal number of latent variables, in this case, independent components (ICs), is a crucial step before any modeling. Therefore, validation methods are required in order to decide about the optimal number of ICs to be used in the computation of the final model. In this paper, three new validation methods are formally presented. The first one, called Random_ICA, is a generalization of the ICA_by_blocks method. Its specificity resides in the random way of splitting the initial data matrix into two blocks, and then repeating this procedure several times, giving a broader perspective for the selection of the optimal number of ICs. The second method, called KMO_ICA_Residuals is based on the computation of the Kaiser-Meyer-Olkin (KMO) index of the transposed residual matrices obtained after progressive extraction of ICs. The third method, called ICA_corr_y, helps to select the optimal number of ICs by computing the correlations between calculated proportions and known physico-chemical information about samples, generally concentrations, or between a source signal known to be present in the mixture and the signals extracted by ICA. These three methods were tested using varied simulated and experimental data sets and compared, when necessary, to ICA_by_blocks. Results were relevant and in line with expected ones, proving the reliability of the three proposed methods.

1. Introduction

Nowadays, following recent progress in analytical chemistry and its instrumentation, ever more complex data is being produced, requiring advanced mathematical and statistical tools in order to extract hidden information. For the purpose of simplifying the interpretation of such complex data sets, blind source separation (BSS) techniques have gained considerable attention in the last years, in particular Independent Components Analysis (ICA), which may be considered as one of the most reliable techniques in this field. ICA relies on the assumption that most measured signals must be mixtures of independent signals or what may be called “source signals”. Given a set of such mixed signals, ICA works by finding a linear transformation of those mixtures, allowing to recover source signals or independent components (ICs), using a criterion that measures statistical independence between sources. Each of these calculated ICs will be associated to a different physical process, making it easier to give more meaningful interpretation to highlighted discriminations [1–3].

In fact, the applicability of ICA was proven to be successful in several analytical chemistry domains and in the processing of different types of data. Without being exhaustive, these include the use of ICA as a pretreatment method to eliminate artefacts from multiway data [4], for the resolution of overlapping GC-MS signals [5–8], for robust optimization in liquid chromatography [9] and for processing spectroscopic data such as: visible spectroscopy [10], NIR [11,12] and MIR [13,14], NMR [15], Raman images [16], 3D-front face fluorescence [17–19], laser-induced breakdown spectroscopy (LIBS) [20], etc. In addition to what has been cited as qualitative application of ICA for discrimination, identification and classification purposes, this chemometric tool has been applied as a multivariate regression method, providing useful information for quantitative analysis of components in mixtures [21–26].

One crucial fact about ICA is related to the determination of the optimal number of ICs which must be carefully selected to perform this analysis. In fact, extracting too few ICs may result in non-pure signals, still consisting of mixtures, whereas calculating too many may excessively decompose source signals and introduce noise. Moreover, there is no natural order for the extraction of ICs, meaning that a more informative IC could be extracted after less informative ones [3]. For

* Correspondence to: AgroParisTech, 16 Rue Claude Bernard, 75005 Paris, France.
E-mail address: douglas.rutledge@agroparistech.fr (D.N. Rutledge).

https://doi.org/10.1016/j.talanta.2017.11.051
Received 8 September 2017; Received in revised form 17 November 2017; Accepted 23 November 2017
Available online 26 November 2017
0039-9140/

this reason, methods commonly used to decide the number of significant components in Principal Components Analysis can rarely be applied in ICA; one exception being the use of permutation tests [27].

In order to overcome this obstacle, several methods have been proposed in the literature, but most of them rely on prior knowledge of the data [28,29]. Other methods were however developed, among which ICA_by_blocks [30] may be considered as the most commonly applied one [10,12,15,17,18,25,31,32]. Briefly, this method consists in splitting the data matrix into two (or more) blocks of approximately equal number of rows in a way that each block should be representative of the whole data matrix. For each of these predefined blocks, ICA models with an increasing number of ICs are computed. ICs corresponding to true source signals should be found in each of the blocks. Such true ICs are highly correlated while noisy ICs or non-characteristic signals will have low correlations. The model with the highest number of highly correlated ICs indicates the optimal number of ICs to be extracted [3,30].

In this context, the objective of this article is to present a formal description of three new methods for the determination of the optimal number of ICs and hence push for further development in this field. The first method, called Random_ICA, is a generalization of the ICA_by_blocks method. The second one, called KMO_ICA_Residuals, is based on the computation of the KMO index and the third, called ICA_corr_y, is proposed to be used specifically on data corresponding to quantitative information. Tests were performed on simulated and experimental data sets and compared, when necessary, to ICA_by_blocks.

2. Theoretical background

2.1. Independent components analysis (ICA)

Given that observed signals are organized into a data matrix **X** (n,p) with ***n*** measured signals, corresponding to the different analyzed samples, each containing ***p*** data points or variables for which the intensities have been measured, and assuming that these measured signals are linear mixtures of source signals, the general model of ICA can be described as:

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

where **A** is the matrix of coefficients (proportions), specifying the relative contributions of the source signals to each mixture, or the so-called “mixing matrix”; and **S** is the matrix of source signals (the independent components, ICs). In short, ICA aims to determine both **A** and **S**, knowing only **X**. It attempts to achieve this objective by estimating a demixing matrix $\mathbf{W} = \mathbf{A}^{-1}$, so that the source signals (ICs) may be recovered from **X** according to the following equation:

$$\mathbf{S} = \mathbf{WX} \tag{2}$$

The mixing matrix **A** can then be calculated as:

$$A = X\mathbf{S}^{\mathbf{T}}(\mathbf{SS}^{\mathbf{T}})^{-1} \tag{3}$$

In the present study, the Joint Approximate Diagonalization of Eigenmatrices (JADE) algorithm was used, which aims to extract independent sources from signal mixtures by maximizing their non-Gaussianity [3].

The measured signals are assumed to be combinations of several independent signals, and so, as a consequence of the Central Limit Theorem, they should present a "more Gaussian" distribution than the individual sources. The objective of JADE is to rotate the loadings vectors obtained by applying a Principal Components Analysis (PCA) to the complete set of observed signals, **X**, so as to maximise their non-Gaussianity. The PCA is applied to the row-centered X_c matrix to obtain a reduced number of orthogonal loadings vectors of equal variance, $\mathbf{P}_w$.

The fourth-order cumulants of these $\mathbf{P}_w$ with themselves (*auto*-cumulants), as well as the cumulants of all combinations of $\mathbf{P}_w$ (*cross*-cumulants), are calculated and placed in a fourth order tensor, of dimensions $k \times k \times k \times k$ (where k is the number of PCA loadings in $\mathbf{P}_w$, corresponding to the number of ICs to be calculated).

The fourth-order *auto*-cumulant of a vector **x** can be defined as its kurtosis, κ:

$$Cum_4\{\mathbf{x}, \mathbf{x}, \mathbf{x}, \mathbf{x}\} = E\{\mathbf{x}^4\} - 3.E^2\{\mathbf{x}^2\}; \tag{4}$$

The fourth-order *auto*- and *cross*-cumulants are given by:

$$\kappa_4\{\mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k, \mathbf{v}_l\} = E\{\mathbf{v}_i\mathbf{v}_j\mathbf{v}_k\mathbf{v}_l\} - E\{\mathbf{v}_i\mathbf{v}_j\}..E\{\mathbf{v}_k\mathbf{v}_l\} - E\{\mathbf{v}_i\mathbf{v}_k\}..E\{\mathbf{v}_j\mathbf{v}_l\} - E\{\mathbf{v}_i\mathbf{v}_l\}..E\{\mathbf{v}_j\mathbf{v}_k\} \tag{5}$$

where the 4 vectors $\mathbf{v}_i$, $\mathbf{v}_j$, $\mathbf{v}_k$, and $\mathbf{v}_l$ are different combinations of the vectors in the matrix $\mathbf{P}_w$.

If the vectors are independent, their fourth-order *cross*-cumulant will be zero and their *auto*-cumulants maximal. JADE therefore calculates a rotation matrix to diagonalize the initial cumulants tensor so that the vectors become statistically independent, to give the source signals, **S**. The proportions in the mixing matrix, **A**, are obtained by projecting **X** onto **S**:

$$\mathbf{A} = \mathbf{X}..\mathbf{S}^T(\mathbf{S}\,\mathbf{S}^T)^{-1} \tag{3'}$$

2.2. Random_ICA

Random_ICA can be considered as a generalization of the ICA_by_blocks method [30]. As previously mentioned, ICA_by_blocks starts by splitting the data matrix into a certain number of blocks, generally two. Moreover, it was clearly stated that care must be taken in the construction of these blocks so that each block is, as much as possible, representative of the complete data matrix. Four options were introduced with the ICA_by_blocks method in order to build these blocks, the choice of the option to apply being determined by the way the signals are ordered in the original matrix: Venetian blind (where regularly spaced samples are attributed to each block, e.g., if two blocks are to be defined, all samples with an even index are placed in the first block, and all samples with an odd index are placed in the second block), successive blocks (the first half of the samples constitutes the first block, and the second half constitutes the second block), random repartition of the samples into the two blocks, and predefined groups (the user chooses, according to his own criteria, which sample goes into which group). However, in the case of complex and unstructured data sets, a certain bias may be introduced during the distribution of rows into the different blocks. Therefore, the Random_ICA method introduced in this paper starts by randomly distributing the rows of the data matrix **X** into **2** blocks of approximately equal sizes. As in ICA_by_blocks, for each of these predefined blocks, F_{max} ICA models are computed, with from 1 to F_{max} ICs. To avoid the possibility of a bias being introduced by a particular distribution of the rows into the blocks, the whole procedure is repeated k times resulting in different sets of blocks, producing a broader perspective for the selection of the optimal number of ICs.

2.3. KMO_ICA_Residuals

The Kaiser-Meyer-Olkin (KMO) index [33,34] was developed to check whether the factorial analysis of a data set is pertinent. In fact, if the original variables are orthogonal, it is useless to perform a Principal Components Analysis (PCA) of the data.

The KMO index is calculated as follows:

$$\mathrm{KMO} = \frac{\sum_i \sum_{j \neq i} r_{ij}^2}{\sum_i \sum_{j \neq i} r_{ij}^2 + \sum_i \sum_{j \neq i} a_{ij}^2} \tag{6}$$

where r_{ij} is the correlation between variables i and j, and the partial correlation a_{ij} is defined as: