INDEPENDENT COMPONENTS ANALYSIS :

THEORY, APPLICATIONS AND DIFFICULTIES

Douglas N. Rutledge

UMR Genial, AgroParisTech, INRA, Université Paris-Saclay
91300 Massy, France
douglas.rutledge@agroparistech.fr

Abstract

Independent Components Analysis (ICA) is a blind source separation method that has been developed to extract the underlying source signals from a set of observed signals where they are mixed in unknown proportions. This is made possible by making the assumption that the "pure" sources are by definition totally unrelated ("independent") with non-gaussian intensity distributions, whereas their mixtures have more gaussian distributions due to the Central Limit Theorem. ICA can thus be used not just to analyse 2D datasets (e.g. IR spectra) but also multiway datasets (e.g. 3D Excitation-Emission fluorescence spectra) after unfolding. Like other methods based on latent variables, a careful investigation has to be carried out to determine which components are significant and which are not. Therefore, it is important to dispose of valid procedures to decide on the optimal number of independent components (ICs) to extract in the final ICA model.

The objective of this workshop is to introduce ICA through the study of several real cases, and to show how it performs compared to other more classical multivariate methods such as Principal Components Analysis (PCA). In this way, the relative advantages and disadvantages of ICA will be pointed highlighted.

Keywords - Independent Components Analysis, Durbin-Watson, Multi-way data

INTRODUCTION

Since its development in the 1980s in the domain of signal processing, the use of Independent Components Analysis (ICA) has increased, and has spread to several different scientific domains, including analytical chemistry. ICA is applied to a set of observed signals, composing the rows of a data matrix X, and aims at finding the source signals in the mixtures, as well as their proportions in each signal, with no knowledge other than X. The assumption underlying ICA is the statistical independence of the underlying "source" signals.

Mathematically speaking, the matrix **X** ($n \times p$) can be expressed as the product of two matrices **A** ($n \times k$) and **S** ($k \times p$), such that:

$$\mathbf{X} = \mathbf{A} \times \mathbf{S}$$

where **S** is the matrix of source signals (on the rows), and **A** is the mixing matrix, describing the proportions of the pure signals in each observed, mixed signal of **X**. n is the number of signals on the rows of the X matrix, described over p variables, while k is the number of pure sources. To make a parallel with other multivariate methods, the columns of **A** can be assimilated to score vectors, while the rows of **S**, which are the Independent Components (ICs), can be assimilated to loadings vectors.

The objective of ICA is therefore to find **A** and **S** from **X**. Several algorithms exist to perform ICA, depending on the approach employed to assess the statistical independence. Among the set of existing algorithms, the Joint Approximate Diagonalization of Eigenmatrices (JADE) algorithm has our preference, as it is fast and gives stable results. It is based on the construction of a fourth-order cumulant array from the data, which is then diagonalised.

In most real cases, the number of sources, *k*, is not known a priori, and has to be determined before, during or after the ICA procedure. Several methods developed to enable the determination of the number of ICs (*k*) will be presented briefly.

PCA

In Principal Components Analysis (PCA), the data matrix **X** is seen as a collection of objects (the rows of **X**) in a multidimensional space defined by the original variables. Objects 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 the 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. 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 source 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 dangerous.

ICA

The assumption underlying ICA is that each row of the data matrix is a weighted sum of source signals, the weights being proportional to the contribution of the corresponding source signals to that particular mixture. Neither the original source signals, nor their proportions in the analyzed mixtures, are known. In ICA, **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.

Several hypotheses underlie the use of ICA :

1) Variations in source signals are not in any way related. Therefore, by calculating independent latent variables, it should be possible to recover the pure source signals;

2) The intensity distributions of source signals are not random, and therefore do not present a Gaussian histogram;

3) Thanks to the Central Limit Theorem, the measured signals, being combinations of several independent sources, are "more Gaussian" than the sources.

The objective of ICA is therefore to search for the least Gaussian possible sources.

EXAMPLES OF APPLICATIONS

The analysis of several very different types of data using ICA will be presented.