Expert Systems With Applications 75 (2017) 56–62

Contents lists available at ScienceDirect

Expert Systems With Applications

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

An unsupervised method to determine the optimal number of independent components

Angel Mur[*], Raquel Dormido, Natividad Duro, Daniel Mercader

Department of Computer Sciences and Automatic Control, UNED, Juan del Rosal 16 - 28040 Madrid, Spain

ARTICLE INFO

Article history:
Received 20 October 2016
Revised 17 January 2017
Accepted 18 January 2017
Available online 20 January 2017

Keywords:
Distance correlation
ICA
JADE algorithm
Linear correlation
Optimal number of independent components

ABSTRACT

In this paper, we present a new method to determine the optimal number of independent components after applying an Independent Component Analysis (*ICA*) to a set of mixed signals. The proposed method, called Linear Correlations between Components (*LCC*), uses the *JADE* algorithm to calculate the independent components. The *LCC* method allows to automatically select the optimal number of independent components in an unsupervised way without any previous knowledge. It has been tested using synthetic mixed signals where the number of pure (or independent) signals is known. This method is very simple, fast and easy to implement.

1. Introduction

Given a set of mixed signals that have been created by combining a set of pure signals in unknown proportions, the Independent Components Analysis (*ICA*) (Bouveresse & Rutledge, 2016; Hyvärinen & Oja, 2000) is a blind-source separation method that enables the extraction of the pure signals, as well as their proportions, from the set of mixed signals. *ICA* has been applied in many domains in which signals have to be analysed (Hao et al., 2009; He, Clifford, & Tarassenko, 2006; Krishnaveni, Jayaraman, Manoj Kumar, Shivakumar, & Ramadoss, 2005; Wang, Ding, & Hou, 2008). In particular, it is used to separate significant components from signals and remove artefacts.

ICA is based on the construction of the Independent Components (*IC*). The procedure for determining the optimal number of *IC*'s, *k*, is an important issue when developing an *ICA* model. The value of *k* corresponds to the number of pure signals where each of them explains an independent phenomenon. In general, when the number of desired *IC*'s, *NIC*, is smaller than the optimal one, some significant components are mixed together in the small number of extracted *IC*'s. On the other hand, if a *NIC* bigger than *k* is required, some of the significant components are decomposed into subcomponents (it means that not all the components are independent). In both cases, the components obtained neither represent nor explain correctly the independent phenomena. Therefore, it is necessary to use a validation method to determine *k*. This is the main goal of this paper.

Some methods make use of some theoretical knowledge of the pure signals to determine *k*, such as the Amari index (Rutledge & Jouan-Rimbaud Bouveresse, 2013). However, in practice this information is not available and consequently any method to determine *k* should work without a priori knowledge. This fact involves the need to find an unsupervised method to determine the optimal number of independent components when analyzing a set of mixed signals.

Bouveresse and Rutledge (2016) show a review of the most interesting algorithms developed to find *k*: the Durbin–Watson criterion (Rutledge & Barros, 2002), the *ICA*_by_Blocks (Jouan-Rimbaud Bouveresse, Moya-González, Ammari, & Rutledge, 2012), the Random_*ICA*_by_Blocks, the *RV*_*ICA*_by_Blocks and the *ICA_corr_Y*. Unlike *ICA_corr_Y*, the first four methods do not require any specific prior knowledge. The Durbin–Watson criterion can only be used in structured signals although the other methods can be applied to any type of data. The methods that use blocks of data need to pay particular attention in selecting representative and comparable data blocks. Furthermore, these "blocks methods" cannot find *k* when the number of mixed signals is less than 2 × *k*.

In addition to these methods, there is another approach based on the Principal Components Analysis (*PCA*) (Semmlow, 2004; Jolliffe, 2002) applied to the mixed signals. This algorithm considers the optimal number of *IC*'s equal to the number of significant Principal Components. This method is simple but it is based on hypotheses that are not always valid. Moreover, the optimal number of *IC*'s is chosen from a scree plot. This plot is a descending curve

[*] Corresponding author.
E-mail addresses: a.r.m.g@outlook.fr (A. Mur), raquel@dia.uned.es (R. Dormido), nduro@dia.uned.es (N. Duro), mercaderd@yahoo.es (D. Mercader).

http://dx.doi.org/10.1016/j.eswa.2017.01.015
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

representing the eigenvalues vs. the Principal Component index. The optimal number of *IC*'s is the value at which the eigenvalues start to level off.

In the present work, *JADE* is the selected *ICA* algorithm (Cardoso & Souloumiac, 1993). *JADE* consists of an eigenmatrix decomposition of a higher-order cumulant tensor. The cumulants give a measure of the non-Gaussianity of the components. For each *NIC*, *JADE* provides a stable result. This algorithm starts by restricting the operation of *JADE* to the *NIC* first principal components obtained from a *PCA* (Cardoso Resources, 2016).

In this paper, a simple and effective method to calculate k named Linear Correlations between Components (*LCC*) is shown. This algorithm uses *JADE* and it takes advantages of both the concept of independence and the fact that a decomposed *IC* appears when *NIC* is superior to k. The *LCC* method allows to automatically select k in an unsupervised way.

Unlike *LCC*, every method described above needs at least a specific condition to find k that reduces its capacity of being generic. In general, these methods require a graphical representation to select k.

On the other hand, the main characteristics of *LCC* are the following: it is an unsupervised method, it determines k automatically, it does not need the support of a graphical representation, there is no restriction with respect to the type of pure signals, the mixed signals do not need to be ordered in blocks and it can calculate k when the number of mixed signals is inferior to $2 \times k$. However, as any method, it needs a number of mixed signals superior to k.

The only limitation (similar to others methods) comes from the computational resources that *JADE* needs. Depending on the available memory and *CPU* time, *JADE* cannot create a very large number of components (Cardoso Resources, 2016).

In practice, the k found cannot be validated by using external information. One way to evaluate the result is to compare the results of different algorithms. The *LCC* is a different method and can contribute to the evaluation process. However, *LCC* is robust enough and its results can be trusted.

In Section 2, *ICA* and some basic concepts used in *LCC* are reviewed. In Section 3, the proposed method *LCC* is described. In Section 4, the *LCC* and other methods are tested using synthetic data. Finally, in Sections 5 and 6, the discussion and conclusions of the paper are respectively presented.

2. Background

This section explains the *ICA* algorithm and reviews the concepts of independence and correlation between two random variables.

2.1. ICA

Given a set of n mixed signals formed by combining k independent signals with p samples each, its *ICA* model is defined as $M = A \times S$, where M is a $n \times p$ data matrix, A is a $n \times k$ mixing matrix and S is a $k \times p$ matrix of independent signals.

The objective of *ICA* is to calculate A and S knowing only M. *ICA* does not need any knowledge concerning the nature of the source signals or their proportions. To estimate *A*, *ICA* requires the pure signals in S to be truly independent and non-Gaussian. Both conditions are usually met when the sources are real signals.

The independence in *ICA* can be reached by maximizing the non-Gaussianity of the components or by minimizing the mutual information (Wang et al., 2008). Around this concept, different *ICA* algorithms have been developed: FastICA (Hyvärinen & Oja, 1997), *JADE* (Cardoso & Souloumiac, 1993; Cardoso Resources, 2016), InfoMax (Bell & Sejnowski, 1995), Mutual Information Least Dependent Component Analysis (Stögbauer, Kraskov, Astakhov, & Grassberger, 2004), Stochastic Non-Negative Independent Components Analysis (Astakhov, Stögbauer, Kraskov, & Grassberger, 2006), *RADICAL* (Learned-Miller & Fisher III, 2003), etc.

2.2. Independence and linear correlation

Suppose a random variable X that can take v different values, with the probability that $X{=}x_{\mathrm{i}}$ defined to be $P(X{=}x_{\mathrm{i}}){=}p_i$. Then the expectation ($E$) of X is defined as

$$E(X) = x_1 p_1 + x_2 p_2 + \cdots + x_v x_v \tag{1}$$

and its variance (*Var*)

$$Var(X) = E(X^2) - (E(X))^2 \tag{2}$$

Suppose that X and Y are two random variables with expected values $E(X)$, $E(Y)$ and variances $Var(X)$, $Var(Y)$, respectively. The covariance (*Cov*) of (X,Y) is defined by

$$Cov(X, Y) = E(XY) - E(X)E(Y) \tag{3}$$

and its correlation (*Corr*)

$$Corr(X, Y) = \frac{Cov(X, Y)}{(Var(X)Var(Y))^{\frac{1}{2}}} \tag{4}$$

The $Corr(X,Y)$ measures the linear dependence between two variables X and Y, giving a value between -1 and $+1$ both inclusive.

Two random variables X and Y are uncorrelated when their correlation coefficient is zero: $Corr(X,Y){=}0$. Being uncorrelated is the same as having zero covariance and therefore, from (3):

$$E(XY) = E(X)E(Y) \tag{5}$$

If X and Y are independent, then they are uncorrelated and consequently $Corr(X,Y) = 0$. If $Corr(X,Y) \neq 0$, then X and Y present some grade of dependence. If $Corr(X,Y) = 0$, then X and Y can be either dependent or independent.

Suppose that (x_m, y_m) with $m = 1,2, \ldots.\ v$ form a sample from a pair of random variables X and Y. The distance covariance (*dCov*) (Székely, Rizzo, & Bakirov, 2007) is defined as:

$$dCov^2(X, Y) = \frac{1}{v^2} \sum_{j,h=1}^{v} A_{j,h} B_{j,h} \tag{6}$$

where

$$A_{j,h} := a_{j,h} - \bar{a}_{j.} - \bar{a}_{.h} + \bar{a}_{..}, \quad a_{j,h} = \|x_j - x_h\| \quad j, h = 1, 2, \ldots..v$$

$$B_{j,h} := b_{j,h} - \bar{b}_{j.} - \bar{b}_{.h} + \bar{b}_{..}, \quad b_{j,h} = \|y_j - y_h\| \quad j, h = 1, 2, \ldots..v$$

$$\bar{a}_{j.} = \frac{1}{v} \sum_{h=1}^{v} a_{jh}, \quad \bar{a}_{.h} = \frac{1}{v} \sum_{j=1}^{v} a_{jh}, \quad \bar{a}_{..} = \frac{1}{v^2} \sum_{j,h=1}^{v} a_{jh}$$

The notation is similar for $\bar{b}_{j.}, \bar{b}_{.h}$ and $\bar{b}_{..}$.

The distance variance (*dVar*) of X is

$$dVar^2(X) = dCov^2(X, X) = \frac{1}{v^2} \sum_{j,h=1}^{v} A_{j,h}^2 \tag{7}$$

The distance correlation (*dCorr*) (Székely, Rizzo, & Bakirov, 2007) of two random variables X and Y is obtained by dividing their distance covariance by the product of their distance standard deviations. The distance correlation is

$$dCorr^2(X, Y) = \frac{dCov^2(X, Y)}{(dVar^2(X) dVar^2(Y))^{\frac{1}{2}}} \tag{8}$$

The distance correlation is a measure of the statistical dependence between two random variables X and Y. This measure of dependence is zero if and only if X and Y are independent. $Corr(X,Y) = 0$ does not imply independence while $dCorr(X,Y) = 0$ implies independence.