Independent dynamics subspace analysis

Alexander Ilin

Helsinki University of Technology - Lab. of Computer and Information Science
P.O. Box 5400, FI-02015 TKK, Espoo - Finland

Abstract. The paper presents an algorithm for identifying the independent subspace analysis model based on source dynamics. We propose to separate subspaces by decoupling their dynamic models. Each subspace is extracted by minimizing the prediction error given by a first-order nonlinear autoregressive model. The learning rules are derived from a cost function and implemented in the framework of denoising source separation.

1 Introduction

Blind source separation (BSS) is a research problem which has recently received a lot of attention. The basic modeling assumption of linear BSS methods is that the observed signals form a linear combination of some underlying sources (also called components):

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)\,, \tag{1}$$

where $\mathbf{x}(t)$ denotes the vector of observations $x_k(t)$, $\mathbf{s}(t)$ is the vector of source signals $s_j(t)$ and $\mathbf{A}$ is called a mixing matrix. Both the sources $\mathbf{s}(t)$ and the mixing matrix $\mathbf{A}$ are unknown and have to be estimated from the measurements.

Independent component analysis (ICA) is a popular method for solving the BSS problem. ICA algorithms identify the model in (1) using the only assumption of the statistical independence of the sources: Each $s_j(t)$ is regarded as a sample from a random variable s_j and these variables are assumed mutually independent. Independence of random variables is a stronger assumption than uncorrelatedness and therefore ICA methods are typically based on higher-order statistics (see, e.g., [1] for introduction). Multidimensional ICA [2] or independent subspace analysis [3] is a natural extension of ICA. There, the source vector $\mathbf{s}$ is decomposed into several groups (linear subspaces):

$$\mathbf{s} = \begin{bmatrix}\mathbf{s}_1^T & \dots & \mathbf{s}_n^T\end{bmatrix}^T\,, \tag{2}$$

and the sources within one group $\mathbf{s}_i$ are generally assumed dependent while components from different groups are mutually independent.

The BSS problem can also be solved by analyzing the source temporal structure. Several researchers have shown that the sources can be separated by joint diagonalization of the autocorrelation matrices calculated for several time lags [4, 5] or by using linear predictors [6, 7]. The advantage of these methods is that they are typically based on second-order statistics and they can separate sources with Gaussian distributions provided that their time structure is different.

In this paper, we propose an algorithm which uses the source temporal structure to estimate independent subspaces. Each group of sources $\mathbf{s}_i$ in (2) is assumed to have independent dynamics which is modeled by a first-order nonlinear autoregressive process. The dimensionality of each group is assumed to be known. Similarly to [6], the subspaces are estimated so as to minimize the prediction error given by the subspace dynamic model. We call the resulting algorithm *independent dynamics subspace analysis* (IDSA).

The idea and motivation for the proposed method comes from the analysis of the complex climate data [8]. Climate phenomena constantly interact with each other and cannot be independent. They can usually be described by multi-dimensional dynamic processes and a meaningful separation criterion would be making the dynamics of different groups of sources decoupled as much as possible. A similar separation criterion was expressed using the variational Bayesian principles in [9].

2 Independent dynamics subspace analysis

2.1 The model

We consider the linear mixing model (1) in which the source vector is decomposed into groups as in (2). Each group $\mathbf{s}_i$ is assumed to be of a known dimensionality and to follow an independent first-order nonlinear dynamic model:

$$\mathbf{s}_i(t) = \mathbf{f}_i(\mathbf{s}_i(t-1)) + \mathbf{m}_i(t)\,, \quad i = 1, \dots, n\,, \tag{3}$$

where $\mathbf{f}_i$ is an unknown nonlinear function and $\mathbf{m}_i(t)$ is a noise vector. Thus, the IDSA model resembles linear dynamic factor analysis [10]. Assuming separate $\mathbf{f}_i$ in (3) means that the subspaces have decoupled dynamics, that is sources from one subspace do not affect the development of sources from other subspaces. In the special case of linear dynamics $\mathbf{s}(t) = \mathbf{B}\mathbf{s}(t-1) + \mathbf{m}(t)$, this is equivalent to having a block-diagonal matrix $\mathbf{B}$.

Without loss of generality, we can make the additional assumption that all the sources are mutually uncorrelated and of unit variance. The sources from different subspaces are uncorrelated due to independence and the correlations within the subspaces can always be removed by a linear transformation. Note that IDSA identifies the sources only up to linear rotations within the subspaces, which is a known indeterminacy of multidimensional ICA [2].

2.2 Preprocessing and separating structure

The first step of the IDSA algorithm is called whitening. It is a linear transformation of data (usually implemented by principal component analysis) which makes the data covariance matrix an identity matrix. After whitening, any orthogonal basis in the data space defines uncorrelated sources of unit variance. Therefore, whitening is a way to assure that the extracted components are mutually uncorrelated (see more details in, e.g., [1]).

The sources are extracted from the whitened data using a demixing matrix $\mathbf{W}$ as in denoising source separation (DSS) [11]:

$$\mathbf{S} = \mathbf{W}\mathbf{X}\,. \tag{4}$$

Here, we use the notation in which the matrices $\mathbf{X}$ and $\mathbf{S}$ contain respectively the whitened observations $\mathbf{x}(t)$ and the source vectors $\mathbf{s}(t)$ in their columns. Similarly to (2), the rows of matrices $\mathbf{W}$ and $\mathbf{S}$ are decomposed into groups and each group $\mathbf{W}_i$ estimates one independent subspace $\mathbf{S}_i = \mathbf{W}_i\mathbf{X}$. In the whitened space, $\mathbf{W}$ equals the inverse of the mixing matrix and it can therefore be restricted to be orthogonal. Note that this separating structure differs from [10] where the generative model is learned.

As in many ICA algorithms, the subspaces can be estimated either simultaneously (symmetric approach) or one after another (deflation approach). In the following, we assume that the deflation approach is used and we estimate one of the subspaces $\mathbf{S}_i = \mathbf{W}_i\mathbf{X}$. In order to simplify the notation, we drop the subscript i from the formulas and therefore we assume that $\mathbf{W} = \mathbf{W}_i$.

2.3 Derivation of the algorithm

2.3.1 Cost function

The demixing matrix $\mathbf{W}$ is estimated such that the prediction error of the corresponding subspace dynamic model (3) is minimized. Hence, the cost function

$$\mathcal{C} = \frac{1}{2}\sum_t \|\mathbf{s}(t) - \mathbf{f}(\mathbf{s}(t-1))\|^2 \,, \tag{5}$$

where $\mathbf{s}(t) = \mathbf{W}\mathbf{x}(t)$, is minimized w.r.t. $\mathbf{W}$ and $\mathbf{f}$.

2.3.2 Minimization w.r.t. $\mathbf{W}$

It follows from (4) that for whitened data $\mathbf{X}$ with $t = 1, \ldots, N$ it holds that

$$\mathbf{W} = \frac{1}{N}\mathbf{S}\mathbf{X}^T \,. \tag{6}$$

Using (6) and (4), the gradient descent step for $\mathbf{W}$ can be written as

$$\mathbf{W}_{\text{new}} = \mathbf{W} - \mu\frac{\partial\mathcal{C}}{\partial\mathbf{W}} = \frac{1}{N}\left(\mathbf{S} - N\mu\frac{\partial\mathcal{C}}{\partial\mathbf{S}}\right)\mathbf{X}^T \,, \tag{7}$$

where $\partial\mathcal{C}/\partial\mathbf{W}$, $\partial\mathcal{C}/\partial\mathbf{S}$ denote the matrices of the gradients of $\mathcal{C}$ w.r.t. the elements of $\mathbf{W}$ and $\mathbf{S}$ respectively, and μ is the step size. Hence, one step for optimizing $\mathbf{W}$ can be performed by first updating the sources with the step size μ_s:

$$\mathbf{S}_{\text{new}} = \mathbf{S} - \mu_s\frac{\partial\mathcal{C}}{\partial\mathbf{S}} \,, \tag{8}$$

then calculating the new value for $\mathbf{W}$ using (6) and finally projecting $\mathbf{W}$ onto the set of orthogonal matrices. This is the optimization strategy used in DSS.

Taking the derivative of the cost function (5) w.r.t. the sources is straightforward. The t-th column of the matrix gradient $\partial\mathcal{C}/\partial\mathbf{S}$ is equal to

$$\frac{\partial\mathcal{C}}{\partial\mathbf{s}(t)} = \mathbf{s}(t) - \mathbf{f}(\mathbf{s}(t-1)) - \left(\frac{\partial\mathbf{f}(\mathbf{s}(t))}{\partial\mathbf{s}}\right)^T(\mathbf{s}(t+1) - \mathbf{f}(\mathbf{s}(t))) \tag{9}$$

with the following exceptions: when $t = 1$, the term $\mathbf{s}(t) - \mathbf{f}(\mathbf{s}(t-1))$ is omitted; and when $t = N$, the term $(\partial \mathbf{f}(\mathbf{s}(t))/\partial \mathbf{s})^T (\ldots)$ is omitted. Here $\partial \mathbf{f}(\mathbf{s}(t))/\partial \mathbf{s}$ denotes the Jacobian matrix of $\mathbf{f}(\mathbf{s})$ calculated at $\mathbf{s}(t)$.

In the case of linear dynamics, (9) is simplified as $\mathbf{f}(\mathbf{s}(t)) = \mathbf{B}\mathbf{s}(t)$ and $\partial \mathbf{f}(\mathbf{s}(t))/\partial \mathbf{s} = \mathbf{B}$.

2.3.3 Minimization w.r.t. f

It is possible to use any model for $\mathbf{f}$, for example, radial-basis function or multilayer perceptron (MLP) networks [12]. In the experiments, we use a MLP

$$\mathbf{f}(\mathbf{s}) = \mathbf{D} \tanh(\mathbf{C}\mathbf{s} + \mathbf{c}) + \mathbf{d} \, . \tag{10}$$

The Jacobian matrix required in (9) is given by

$$\partial \mathbf{f}(\mathbf{s})/\partial \mathbf{s} = \mathbf{D} \, \mathrm{diag}(\tanh'(\mathbf{C}\mathbf{s} + \mathbf{c}))\mathbf{C} \, , \tag{11}$$

where $\mathrm{diag}(\mathbf{v})$ denotes a diagonal matrix with the elements of $\mathbf{v}$ on its main diagonal.

The parameters of the MLP (10) can be updated by minimizing the cost function (5) using any standard algorithm for training MLPs (see, e.g., [12]). It should be noted that regularization is crucial when updating a nonlinear model for $\mathbf{f}$. Too flexible a model would easily overfit to current source estimates and the algorithm would stop in a degenerate solution.

In the case of linear dynamics, minimizing the cost (5) w.r.t. $\mathbf{B}$ is straightforward.

2.3.4 Outline of the algorithm

All the steps of the IDSA algorithm are outlined in the following. Note that the presented implementation follows the framework of denoising source separation:

1: Whiten the data, initialize $\mathbf{W}$.

2: Update the sources $\mathbf{S}$ using (4).

3: Update the source dynamics $\mathbf{f}$ (see Section 2.3.3), calculate new values for $\mathbf{S}$ using (8)-(9). This is the denoising procedure in terms of DSS.

4: Reestimate $\mathbf{W}$ using (6) and project $\mathbf{W}$ onto the set of orthogonal matrices using symmetric or deflation orthogonalization (see, e.g., [1]).

5: Go to step 2 until convergence.

3 Experiments

We test the proposed algorithm on artificially generated data. The data is a linear mixture of two independent Lorenz processes with parameters [3 26.5 1] and [4 30 1], a harmonic oscillator and two white Gaussian noise signals. The original sources and their mixtures are shown in Fig. 1.

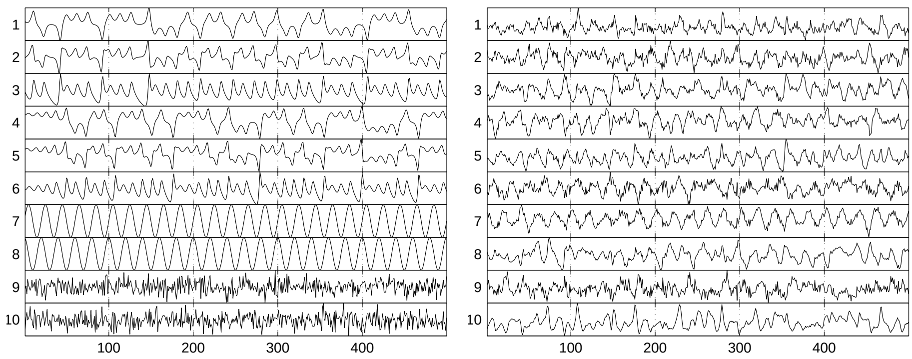

Fig. 1: Left: Original sources: two independent Lorenz processes (sources 1–3 and 4–6), a harmonic oscillator (7–8) and two Gaussian noise signals (9–10). Right: Linear mixture of the sources.

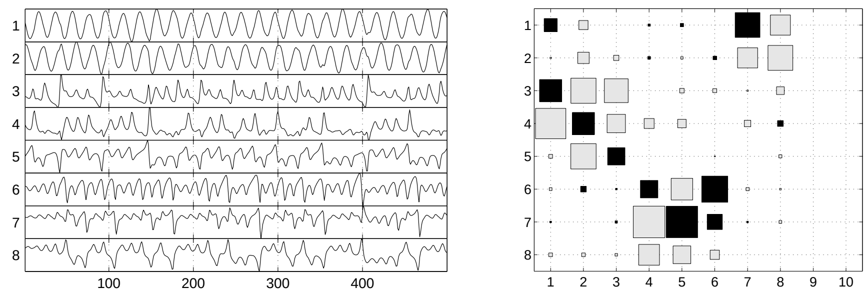

Fig. 2: Left: Sources extracted by IDSA using the symmetric approach. Right: Hinton diagram of the matrix reconstructing the found sources from the original ones. The three original subspaces (columns 1–3, 4–6, 7–8) have been recovered by the estimated components (rows 3–5, 6–8, 1–2 respectively).

The algorithm is set to extract three independent subspaces: a two-dimensional subspace with linear dynamics and two three-dimensional subspaces with non-linear dynamics. The recovered sources are shown in Fig. 2. The least-square reconstruction of the found sources using the original signals (see Fig. 2) indicates that the three original subspaces are correctly identified.

4 Discussion

We presented an algorithm called independent dynamics subspace analysis which identifies independent groups of sources by decoupling their dynamics. The current implementation of IDSA is based on the first-order autoregressive model for subspace dynamics. Including more time delays in the dynamic model can be useful when some of the subspace dimensions are not present in the data.

The independent subspaces can be estimated either symmetrically or by using deflation. The possibility to extract subspaces one by one provides a useful tool for finding components with the most predictable time course in multivariate time series such as climate data [8]. This is an important advantage compared to, for example, [10] where the model is learned for all data, which can be very difficult for highly multidimensional and noisy measurements.

The proposed algorithm is computationally very efficient compared to generative models estimated using the variational Bayesian principles, such as [10] or [9]. In practice, a frequency-based representation of data [8] might be useful before performing IDSA. Slower components are generally easier to predict and the algorithm can favor the slower solutions. Therefore, it is preferable that all subspaces in the data would have the same timescale of oscillations.

Acknowledgements

The author would like to thank Dr. Harri Valpola for his help in developing the method and the preparation of the manuscript.

References

[1] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. J. Wiley, 2001.

[2] J. F. Cardoso. Multidimensional independent component analysis. In *Proc. ICASSP'98*, Seattle, WA, 1998.

[3] A. Hyvärinen and P. O. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feafture subspaces. *Neural Computation*, 12(7):1705 – 1720, 2000.

[4] A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. on Signal Processing*, 45(2):434–44, 1997.

[5] A. Ziehe and K.-R. Müller. TDSEP — an effective algorithm for blind separation using time structure. In *Proc. 8th Int. Conf. on Artificial Neural Networks (ICANN'98)*, pages 675–680, Skövde, Sweden, 1998.

[6] A. Cichocki and R. Thawonmas. On-line algorithm for blind signal extraction of arbitrarily distributed, but temporally correlated sources using second order statistics. *Neural Processing Letters*, 12:91–98, 2000.

[7] A. Cichocki and S.-I. Amari. *Adaptive Blind Signal and Image Processing*. John Wiley & sons, 2002.

[8] A. Ilin, H. Valpola, and E. Oja. Exploratory analysis of climate data using source separation methods. *Neural Networks*, 2005. Accepted.

[9] H. Valpola and J. Karhunen. An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14(11):2647–2692, 2002.

[10] J. Särelä, H. Valpola, R. Vigário, and E. Oja. Dynamical factor analysis of rhythmic magnetoencephalographic activity. In *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA'2001)*, pages 451–456, San Diego, USA, 2001.

[11] J. Särelä and H. Valpola. Denoising source separation. *Journal of Machine Learning Research*, 6:233–272, 2005.

[12] S. Haykin. *Neural Networks – A Comprehensive Foundation, 2nd ed.* Prentice-Hall, 1999.