Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Neurocomputing 69 (2006) 2416–2419
NEUROCOMPUTING
www.elsevier.com/locate/neucom

Letters

Non-combinatorial estimation of independent autoregressive sources

Barnabás Póczos, András Lőrincz*

Eötvös Loránd University, Pázmány Péter sétány 1/C, H-1117 Budapest, Hungary

Received 6 December 2005; received in revised form 7 February 2006; accepted 7 February 2006
Communicated by R.W. Newcomb
Available online 18 July 2006

Abstract

Identification of mixed independent subspaces is thought to suffer from combinatorial explosion of two kinds: the minimization of mutual information between the estimated subspaces and the search for the optimal number and dimensions of the subspaces. Here we show that independent autoregressive process analysis, under certain conditions, can avoid this problem using a two-phase estimation process. We illustrate the solution by computer demonstration.

Keywords: Independent component analysis; Combinatorial explosion; Autoregressive process

1. Introduction

Identification of linear dynamical systems (LDS) driven by non-Gaussian noise is important, it is considered hard and it has not been solved yet in general [3]. Special solutions are known, for example, one can solve the problem if the observation process of the LDS is noiseless and if the hidden processes are driven by independent non-Gaussian noises. In this case the independent subspaces can be identified by independent subspace analysis (ISA) on innovations (ISAI) [5] and then the AR processes can be identified within the subspaces. This method, however, is slow, because (i) the minimization of mutual information between the estimated subspaces and (ii) the search for the optimal number and dimensions of the subspaces are both subject to combinatorial explosions. Here, we introduce a two-phase procedure, which avoids the combinatorial explosion under certain conditions.

2. The IPA model

Assume that we have d pieces of m_i-dimensional first order AR processes

$$\mathbf{s}^i(t+1) = \mathbf{F}^i\mathbf{s}^i(t) + \mathbf{v}^i(t), \tag{1}$$

where $\mathbf{F}^i \in \mathbb{R}^{m_i \times m_i}$, $\mathbf{s}^i \in \mathbb{R}^{m_i}$, $i = 1, \ldots, d$, and $\mathbf{v}^i(t) \in \mathbb{R}^{m_i}$ are non-Gaussian, temporally independent and identically distributed (i.i.d.) noises. For the sake of simplicity, we shall assume that all m_i's are equal, $m_i = m\ \forall i$, but *all of the results* concern the general case. Sources $\mathbf{s}^i$ are the *hidden processes* of the external world. We cannot observe them directly, only their mixture is available for observation. Let us use the following notation $\mathbf{s}(t) = [(\mathbf{s}^1(t))^{\mathrm{T}}, \ldots, (\mathbf{s}^d(t))^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{dm}$, where superscript T denotes transposition. Our *observation model* is

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{md \times md}$ is the mixing matrix. Eqs. (1) and (2), together, form an LDS. Estimations for LDS parameters exist for Gaussian noise $\mathbf{v}(t)\,(= [(v^1(t))^{\mathrm{T}}, \ldots, (v^d(t))^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{dm})$ [3]. We, however, assume that noise $\mathbf{v}$ is non-zero, non-Gaussian, and i.i.d. We also assume that the matrix $\mathbf{A}$ is invertible. Then, as a result, we have an ISA problem on multi-dimensional AR processes [5]. Independent component analysis (ICA) is recovered if $\mathbf{F}^i = 0$ for all i and if $m = 1$ [2].

Let $\mathbf{F} \in \mathbb{R}^{md \times md}$ denote the block-diagonal matrix constructed from matrices $\mathbf{F}^1, \ldots, \mathbf{F}^d$, i.e., $\mathbf{F} =$ *blockdiag* $(\mathbf{F}^1, \ldots, \mathbf{F}^d)$. Then parameters of the model

$$\mathbf{s}(t+1) = \mathbf{F}\mathbf{s}(t) + \mathbf{v}(t), \tag{3}$$

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{4}$$

*Corresponding author. Tel.: +36 209 0555; +36 209 8473.
E-mail address: andras.lorincz@elte.hu (A. Lőrincz).

0925-2312/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2006.02.008

are to be approximated from the observations, e.g., by entropy estimation [5].

3. ISA model

Assume that we have d pieces of m-dimensional independent, non-Gaussian, i.i.d. sources denoted by $\mathbf{v}^i(t)$ $(i = 1, \ldots, d)$. Further, assume that only their mixture $\mathbf{e}(t) = \mathbf{A}\mathbf{v}(t)$ can be observed. The task is the estimation of matrix $\mathbf{A}$ and sources $\mathbf{v}^i(t)$ given the $\mathbf{e}(t)$ signals, where $\mathbf{e}(t)^{\mathrm{T}} = [(\mathbf{e}^1(t))^{\mathrm{T}}, \ldots, (\mathbf{e}^d(t))^{\mathrm{T}}]$, $\mathbf{e}^i(t) \in \mathbb{R}^m$. Let H denote Shannon's entropy. Then, one can solve this problem by minimizing the cost function $J = \sum_{i=1}^{d} H(\mathbf{e}^i)$. This method requires estimations of Shannon's entropy for *groups* of the coordinates [4].

3.1. Reduction of ISA to ICA + search for permutations

It has been observed experimentally that for certain ISA tasks ICA can estimate components of the subspaces [1,4,5]. Sufficient conditions that make the estimation exact are provided below:

Theorem 1 (*Reduction of ISA to ICA + permutation search*). *Assume that for the ISA task and for sources* $\mathbf{v}^i = [v_1^i, \ldots, v_m^i]^{\mathrm{T}}$ $(i = 1, \ldots, d)$ *we have that* $H\left(\sum_{j=1}^{m} w_j v_j^i\right) \geqslant \sum_{j=1}^{m} w_j^2 H(v_j^i)$, where $\sum_{j=1}^{m} w_j^2 = 1$ *holds for all* $\mathbf{w} = (w_1, \ldots, w_m)^{\mathrm{T}}$. *Then, if we execute an ICA algorithm that minimizes the sum of the individual entropies* (*i.e.*, $\sum_{i=1}^{d}\sum_{j=1}^{m} H(e_j^i)$) *on observed data* $\mathbf{e}(t) = \mathbf{A}\mathbf{v}(t)$, *and if the* $\mathbf{W}_{\mathrm{ICA}}$ *solution of the ICA algorithm is unique* (*up to permutation and the sign of the components*), *then the same matrix solves the ISA task* (*up to permutation and the sign of the components*). *In other words, it is sufficient to search for the* $\mathbf{W}_{\mathrm{ISA}}$ *matrix of the ISA task in the form of* $\mathbf{W}_{\mathrm{ISA}} = \mathbf{P}\mathbf{W}_{\mathrm{ICA}}$, *where* $\mathbf{P} \in \mathbb{R}^{md \times md}$ *is a permutation matrix.*

The proof and sufficient conditions of this theorem can be found in [7].

4. Parameter estimation

In this section we introduce a method, which is able to estimate the subspaces of our model (Eqs. (3) and (4)) without any combinatorial algorithm. Note that Eqs. (3) and (4) involve that the stochastic process $\{\mathbf{x}(t)\}$ is also an AR process:

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{s}(t+1) = \mathbf{A}\mathbf{F}\mathbf{s}(t) + \mathbf{A}\mathbf{v}(t) = \mathbf{A}\mathbf{F}\mathbf{A}^{-1}\mathbf{x}(t) + \mathbf{A}\mathbf{v}(t). \tag{5}$$

Let E denote the expectation operator. The innovation $\mathbf{e}(t)$ of the $\{\mathbf{x}(t+1)\}$ process is defined as $\mathbf{e}(t) = \mathbf{x}(t+1) - E(\mathbf{x}(t+1)|\mathbf{x}(t), \mathbf{x}(t-1), \ldots,)$. $\mathbf{A}\mathbf{v}(t)$ is independent of $\mathbf{x}(t)$, thus $E(\mathbf{x}(t+1)|\mathbf{x}(t), \mathbf{x}(t-1), \ldots,) = \mathbf{M}\mathbf{x}(t)$, where $\mathbf{M} = \mathbf{A}\mathbf{F}\mathbf{A}^{-1}\mathbf{x}(t)$, and the innovation of process $\{\mathbf{x}(t+1)\}$ is equal to $\mathbf{e}(t) = \mathbf{A}\mathbf{v}(t)$, which—according to our assumptions—is an i.i.d series. Therefore, any ISA algorithm can be applied to uncover the hidden noises of process $\mathbf{e}(t)$.

Under the condition that the components of the ISA task can be estimated by an ICA algorithm, we can take advantage of the hidden AR processes to uncover the unknown permutations of the coordinates of source $\mathbf{s}$. Also, the dimensions of the subspaces can be revealed by the estimation of matrix $\mathbf{F}$. Thus, the proposed algorithm has two phases:

- `Phase (1)`
  - (a) Estimate innovation $\mathbf{e}(t)$ from series $\{\mathbf{x}(t)\}$. Let $\hat{\mathbf{M}}$ denote an estimation of matrix $\mathbf{A}\mathbf{F}\mathbf{A}^{-1}$. For example, let $\hat{\mathbf{M}} := \arg\min_{\mathbf{M}} \sum_{t=1}^{T} \|\mathbf{x}(t+1) - \mathbf{M}\mathbf{x}(t)\|^2$ and $\hat{\mathbf{e}}(t) := \mathbf{x}(t+1) - \hat{\mathbf{M}}\mathbf{x}(t)$, where T stands for the number of observations and $\|\cdot\|$ denotes the Euclidean norm.
  - (b) Apply a traditional ICA on the estimated $\hat{\mathbf{e}}(t)$ innovations. Then, using the fact that $\mathbf{e}(t) = \mathbf{A}\mathbf{v}(t)$, we have estimations for matrix $\hat{\mathbf{A}}^{-1}$ and for vector $\hat{\mathbf{v}}(t)$. Namely, $\hat{\mathbf{A}}^{-1} := \mathbf{W}_{\mathrm{ICA}}$, $\hat{\mathbf{v}}(t) := \hat{\mathbf{A}}^{-1}\hat{\mathbf{e}}(t)$.
- `Phase (2)`
  - (a) $\hat{\mathbf{s}}(t) := \hat{\mathbf{A}}^{-1}\mathbf{x}(t)$.
  - (b) $\hat{\mathbf{F}} := \arg\min_{\mathbf{F}} \sum_{t=1}^{T} \|\mathbf{F}\hat{\mathbf{s}}(t) + \hat{\mathbf{v}}(t) - \hat{\mathbf{s}}(t+1)\|^2$.

The estimation of the optimal matrices $\mathbf{M}$ and $\mathbf{F}$ in (1a) and (2b) can be accomplished with standard mathematical tools [6], or via neural networks. For non-neural solutions, matrix $\hat{\mathbf{F}}$ can be computed directly, because $\hat{\mathbf{s}}(t+1) = \hat{\mathbf{A}}^{-1}\mathbf{x}(t+1) = \hat{\mathbf{A}}^{-1}\mathbf{A}\mathbf{F}\mathbf{A}^{-1}\hat{\mathbf{A}}\hat{\mathbf{s}}(t) + \hat{\mathbf{A}}^{-1}\mathbf{A}\mathbf{v}(t)$, thus

$$\hat{\mathbf{F}} \approx \hat{\mathbf{A}}^{-1}\hat{\mathbf{M}}\hat{\mathbf{A}}. \tag{6}$$

For a Hebbian estimation of the prediction matrix $\mathbf{F}$, note that the negative gradient of the objective $J = \frac{1}{2}\|\mathbf{s}(t+1) - \mathbf{F}\mathbf{s}(t)\|^2$ is proportional to $(\mathbf{s}(t+1) - \mathbf{F}\mathbf{s}(t))\mathbf{s}(t)^{\mathrm{T}}$, and thus the update rule for the estimation of $\mathbf{F}$ is

$$\Delta\hat{\mathbf{F}} = \mu_t(\mathbf{s}(t+1) - \mathbf{F}\mathbf{s}(t))\mathbf{s}(t)^{\mathrm{T}}, \tag{7}$$

where μ_t is the learning rate that may depend on time. This is the well-known Widrow–Hoff Delta-rule, also known as Adaline rule, which has neural network implementations.

Matrix $\hat{\mathbf{F}}$—apart from the permutation of components—has a block-diagonal structure. The corresponding components can be found without combinatorial efforts by grouping the hidden components that matrix $\hat{\mathbf{F}}$ connects to each other (see Fig. 1(e)). We say that two coordinates i and j are $\hat{\mathbf{F}}$-'connected' if $\max(|\hat{F}_{ij}|, |\hat{F}_{ji}|) > \varepsilon$). (In the ideal case $\varepsilon = 0$.) Then we can group the $\hat{\mathbf{F}}$-'connected' coordinates into separate subspaces using the following algorithm: (1) Choose an arbitrary coordinate i and group all $j \neq i$ coordinates to it which are $\hat{\mathbf{F}}$-'connected' with it. (2) Choose an arbitrary and not yet grouped coordinate. Find its connected coordinates. Group them together. (3)