Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®
Neurocomputing 56 (2004) 467–473

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Letters

A new fixed-point algorithm for independent component analysis

Zhenwei Shi[a], Huanwen Tang[a], Yiyuan Tang[b,c,d,*]

[a] *Institute of Computational Biology and Bioinformatics, Dalian University of Technology, Dalian 116023, PR China*

[b] *Institute of Neuroinformatics, Dalian University of Technology, Dalian 116023, PR China*

[c] *Laboratory of Visual Information Processing, The Chinese Academy of Sciences, Beijing 100101, PR China*

[d] *Key Lab for Mental Health, The Chinese Academy of Sciences, Beijing 100101, PR China*

Abstract

A new fixed-point algorithm for independent component analysis (ICA) is presented that is able blindly to separate mixed signals with sub- and super-Gaussian source distributions. The new fixed-point algorithm maximizes the likelihood of the ICA model under the constraint of decorrelation and uses the method of Lee et al. (Neural Comput. 11(2) (1999) 417) to switch between sub- and super-Gaussian regimes. The new fixed-point algorithm maximizes the likelihood very fast and reliably. The validity of this algorithm is confirmed by the simulations and experimental results.

Keywords: Independent component analysis; Blind source separation; Fixed-point algorithm

1. Introduction

Independent component analysis (ICA) [9] is a statistical technique whose main applications are blind source separation, blind deconvolution, and feature extraction. The following noise-free linear model of ICA is used in this paper:

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

* Corresponding author. Institute of Neuroinformatics, Dalian University of Technology, Dalian 116023, China.

E-mail addresses: szw1977@yahoo.com.cn (Z. Shi), imestang@dlut.edu.cn (H. Tang), yy2100@163.net (Y. Tang).

0925-2312/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2003.09.002

where $\mathbf{x} = (x_1, x_2, \ldots, x_N)^{\mathrm{T}}$ is an N-dimensional observation vector, $\mathbf{A}$ is a nonsingular $N \times N$ mixing matrix and $\mathbf{s} = (s_1, s_2, \ldots, s_N)^{\mathrm{T}}$ is an N-dimensional original source vector having independent components. The basic problem of ICA is then to estimate both the mixing matrix $\mathbf{A}$ and the realizations of the independent components s_i using only observations of the mixtures x_i $(i = 1, 2, \ldots, N)$, i.e., to find an $N \times N$ linear mapping $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N)^{\mathrm{T}}$ such that the unmixed signals $\mathbf{u} = (u_1, u_2, \ldots, u_N)^{\mathrm{T}}$,

$$\mathbf{u} = \mathbf{W}\mathbf{x} \tag{2}$$

are statistically independent. The sources are recovered up to scaling and permutation.

Several estimation methods for ICA have been proposed recently [1–8,10]. This paper presents a new fixed-point algorithm for ICA. The learning algorithm can be derived using the maximum likelihood estimation. The new fixed-point algorithm maximizes the likelihood under the constraint of decorrelation and uses the method of Lee et al. [10] to switch between sub- and super-Gaussian regimes.

2. The likelihood of the ICA model

From (2), the probability density function of the observations $\mathbf{x}$ can be expressed as [10]

$$p(\mathbf{x}) = |\det(\mathbf{W})|\, p(\mathbf{u}), \tag{3}$$

where $p(\mathbf{u}) = \prod_{i=1}^{N} p_i(u_i)$ is the hypothesized distribution of $p(\mathbf{s})$. Assume that we have T observations of $\mathbf{x}$, denoted by $\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(T)$. Then the likelihood can be obtained as the product of this density evaluated at the T points. This is denoted by L and considered as a function of $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N)^{\mathrm{T}}$. The log-likelihood is given by [7]

$$\log L(\mathbf{W}) = \sum_{t=1}^{T} \sum_{i=1}^{N} \log p_i(\mathbf{w}_i^{\mathrm{T}} \mathbf{x}(t)) + T \log |\det(\mathbf{W})|. \tag{4}$$

To simplify notation, we can denote the sum over the sample index t by an expectation operator, and divide the likelihood by T to obtain

$$\frac{1}{T} \log L(\mathrm{W}) = E\left\{ \sum_{i=1}^{N} \log p_i(\mathbf{w}_i^{\mathrm{T}} \mathbf{x}(t)) \right\} + \log |\det(\mathbf{W})|. \tag{5}$$

The expectation here is not the theoretical expectation, but an average computed from the observed samples. Of course, in the algorithms the expectations are eventually replaced by sample averages, so the distinction is purely theoretical. And maximization of likelihood (5) make the output components independent [7].

A useful preprocessing strategy in ICA is to first whiten the observed variables. This means that before the application of the ICA algorithm (and after centering, i.e. making $\mathbf{x}$ a zero mean variable), we transform the observed vector $\mathbf{x}$ linearly so that we obtain a new vector $\tilde{\mathbf{x}}$ which is white, i.e. its components are uncorrelated and their variances

equal unity. In other words, $\mathbf{x}$ is linearly transformed into a random vector

$$\tilde{\mathbf{x}} = \mathbf{V}\mathbf{x} = \mathbf{V}\mathbf{A}\mathbf{s}, \tag{6}$$

whose covariance matrix equals the identity matrix: $E\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathrm{T}}\} = \mathbf{I}$. Thus the ICA model still holds. Consider the new vector $\tilde{\mathbf{x}}$ and constrain the estimates of the independent components $\tilde{\mathbf{u}}_i = \mathbf{w}_i^{\mathrm{T}}\tilde{\mathbf{x}}$ $(i = 1, 2, \ldots, N)$ $(\tilde{\mathbf{u}} = \mathbf{W}\tilde{\mathbf{x}})$ to be uncorrelated and to have unit variance. Now, uncorrelatedness and unit variance of the $\tilde{\mathbf{u}}_i$ means

$$E\{\tilde{\mathbf{u}}\tilde{\mathbf{u}}^{\mathrm{T}}\} = \mathbf{W}E\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathrm{T}}\}\mathbf{W}^{\mathrm{T}} = \mathbf{W}\mathbf{W}^{\mathrm{T}} = \mathbf{I}. \tag{7}$$

This implies that $\mathbf{W}$ is orthogonal and $|\det \mathbf{W}|$ must be constant. On the other hand, $\tilde{\mathbf{u}}_i$ $(i = 1, 2, \ldots, N)$ are uncorrelated, if $\mathbf{W}$ is orthogonal. And we can restrict our search for the matrix $\mathbf{W}$ to the space of orthogonal matrices.

It is here assumed that the data is preprocessed by centering and whitening. Maximizing likelihood (5), and taking into account the constraint decorrelation, one obtains the following optimization problem (i.e. maximizing the likelihood in the space of orthogonal matrices):

$$\begin{aligned} &\max \quad \Psi(\mathbf{W}) = E\left\{\sum_{i=1}^{N} \log p_i(\mathbf{w}_i^{\mathrm{T}}\tilde{\mathbf{x}}(t))\right\} \\ &\text{s.t.} \quad \mathbf{W}\mathbf{W}^{\mathrm{T}} = \mathbf{I}. \end{aligned} \tag{8}$$

That is, we should find the vectors $\mathbf{w}_i (i = 1, 2, \ldots, N)$ which are orthogonal to each other, such that

$$\begin{aligned} &\max \quad \Psi(\mathbf{W}) = E\left\{\sum_{i=1}^{N} \log p_i(\mathbf{w}_i^{\mathrm{T}}\tilde{\mathbf{x}}(t))\right\} \\ &\text{s.t.} \quad \|\mathbf{w}_i\|^2 = 1, \quad i = 1, 2, \ldots, N. \end{aligned} \tag{9}$$

3. A fixed-point algorithm for ICA

In this section, we derive a new fixed-point algorithm for ICA. We note that at a stable point of the optimization problem (9), the partial derivative of $\Psi(\mathbf{W})$ at $\mathbf{w}_i$ (i.e. $\partial\Psi(\mathbf{W})/\partial\mathbf{w}_i$) must point in the direction of $\mathbf{w}_i$ $(i = 1, 2, \ldots, N)$, that is, the partial derivative must be equal to $\mathbf{w}_i$ multiplied by some scalar constant (according to the Kuhn–Tucker conditions [11]). Only in such a case, adding the partial derivative to $\mathbf{w}_i$ does not change its direction, and we can have convergence. This means that we should have

$$\mathbf{W} \leftarrow \frac{\partial\Psi(\mathbf{W})}{\partial\mathbf{W}} = E\{\mathbf{g}(\tilde{\mathbf{u}})\tilde{\mathbf{x}}^{\mathrm{T}}\} = E\{\mathbf{g}(\tilde{\mathbf{u}})\tilde{\mathbf{u}}^{\mathrm{T}}\}\mathbf{W}, \tag{10}$$

where $\tilde{\mathbf{u}} = \mathbf{W}\tilde{\mathbf{x}}(t)$, $\mathbf{W}\mathbf{W}^{\mathrm{T}} = \mathbf{I}$, $\mathbf{g}(\tilde{\mathbf{u}}) = (g_1(\tilde{\mathbf{u}}_1), g_2(\tilde{\mathbf{u}}_2), \ldots, g_N(\tilde{\mathbf{u}}_N))^{\mathrm{T}}$, $g_i(\tilde{\mathbf{u}}_i) = (\log p_i(\tilde{\mathbf{u}}_i))' = (p_i(\tilde{\mathbf{u}}_i))'/p_i(\tilde{\mathbf{u}}_i)$ $(i = 1, 2, \ldots, N)$. The parametric density estimate $p_i(\tilde{\mathbf{u}}_i)$ plays an essential role in the success of the learning rule in Eq. (10). As Lee et al. [10] introduced,

the switching between the sub- and super-Gaussian nonlinearities is

$$g_i(\tilde{\mathbf{u}}_i) = \begin{cases} -\tilde{\mathbf{u}}_i - \tanh(\tilde{\mathbf{u}}_i), & \text{super-Gaussian,} \\ -\tilde{\mathbf{u}}_i + \tanh(\tilde{\mathbf{u}}_i), & \text{sub-Gaussian.} \end{cases}$$

Thus, the switching between the sub- and super-Gaussian learning rule is

$$\mathbf{W} \leftarrow E\{-\mathbf{K}\tanh(\tilde{\mathbf{u}})\tilde{\mathbf{u}}^{\mathrm{T}} - \tilde{\mathbf{u}}\tilde{\mathbf{u}}^{\mathrm{T}}\}\mathbf{W} \begin{cases} k_i = 1, & \text{super-Gaussian,} \\ k_i = -1, & \text{sub-Gaussian,} \end{cases} \tag{11}$$

where k_i are elements of the N-dimensional diagonal matrix $\mathbf{K}$ (a switching matrix based on stability criteria [10]),

$$k_i = \operatorname{sign}(E\{\operatorname{sech}^2(\tilde{\mathbf{u}}_i)\}E\{\tilde{\mathbf{u}}_i^2\} - E\{[\tanh(\tilde{\mathbf{u}}_i)]\tilde{\mathbf{u}}_i\}). \tag{12}$$

After every fixed-point iteration, orthogonalization of $\mathbf{W}$ can be done by the symmetric orthogonalization methods [6]. In other words

$$\mathbf{W} \leftarrow (\mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1/2}\mathbf{W}. \tag{13}$$

Thus, we obtain the new fixed-point algorithm as follows:

(i) center the data to make its mean zero;
(ii) whiten the data to give $\tilde{\mathbf{x}}(t)$;
(iii) choose an initial orthogonal matrix for $\mathbf{W}$;
(iv) compute $\tilde{\mathbf{u}} = \mathbf{W}\tilde{\mathbf{x}}(t)$, the N-dimensional diagonal matrix $\mathbf{K} = \operatorname{diag}(k_i)$, $k_i = \operatorname{sign}(E\{\operatorname{sech}^2(\tilde{\mathbf{u}}_i)\}E\{\tilde{\mathbf{u}}_i^2\} - E\{[\tanh(\tilde{\mathbf{u}}_i)]\tilde{\mathbf{u}}_i\})$, for $i = 1, 2, \ldots, N$;
(v) update the separating matrix by $\mathbf{W} \leftarrow E\{-\mathbf{K}\tanh(\tilde{\mathbf{u}})\tilde{\mathbf{u}}^{\mathrm{T}} - \tilde{\mathbf{u}}\tilde{\mathbf{u}}^{\mathrm{T}}\}\mathbf{W}$;
(vi) do a symmetric orthogonalization of the matrix $\mathbf{W}$ by $\mathbf{W} \leftarrow (\mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1/2}\mathbf{W}$;
(vii) if not converged, go back to step (iv).

Note that convergence means that the old and new values of $\mathbf{w}_i$ point in the same direction, i.e. their dot-product is (almost) equal to 1 ($i = 1, 2, \ldots, N$).

4. Simulations and experimental results

Five random signals (5000 data points); one Gaussian, two sub-Gaussian and two super-Gaussian available with MATLAB were used for simulations (available at http://www.cis.hut.fi/projects/ica). The new fixed-point algorithm separated the random signals exactly as the originals. The waveforms (200 data points) of five original signals, their mixtures, and the separated output signals are shown in Fig. 1. Furthermore, the following algorithms were included in the comparison:

- the FastICA fixed-point algorithm [6], using the tanh nonlinearity with symmetric orthogonalization (FPsymth);
- natural gradient algorithm for maximum likelihood estimation [10], i.e. extended infomax algorithm (ExtICA);
- the new fixed-point algorithm (NewFP).

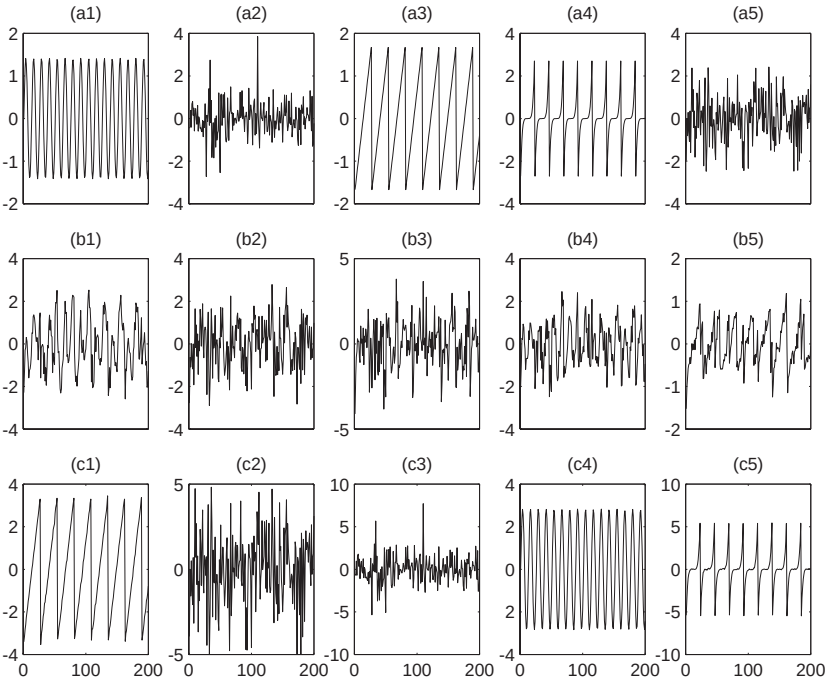

Fig. 1. Exact separation of five random signals by the new fixed-point algorithm. ((a1)–(a5)) The five original signals: one Gaussian source, two sub-Gaussian sources and two super-Gaussian sources. ((b1)–(b5)) The input mixtures. ((c1)–(c5)) The recovered signals.

The algorithms were compared along the two sets of criteria, statistical and computational. The computational load was measured as iteration needed for convergence. The statistical performance, or accuracy, was measured using a performance index, defined as [1]

$$E = \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \frac{|p_{ij}|}{\max_k |p_{ik}|} - 1 \right) + \sum_{j=1}^{N} \left(\sum_{i=1}^{N} \frac{|p_{ij}|}{\max_k |p_{kj}|} - 1 \right), \tag{14}$$

where p_{ij} is the ijth element of $N \times N$ matrix $\mathbf{P} = \mathbf{WA}$. The larger the value E is, the poorer the statistical performance of a separation algorithm. We performed experiments with the five random signals above. For achieving statistical reliability, the experiment was repeated over 100 different realizations of the input data. For each of the 100 realizations, the accuracy was measured using the error index E. The computational load was measured as iteration needed for convergence. Fig. 2 shows a diagram of the computational load vs. the statistical performance. One sees clearly that the FPsymth requires the smallest amount of computation, but its statistical performance is not as good as the other two. The ExtICA provides the better separated results, but

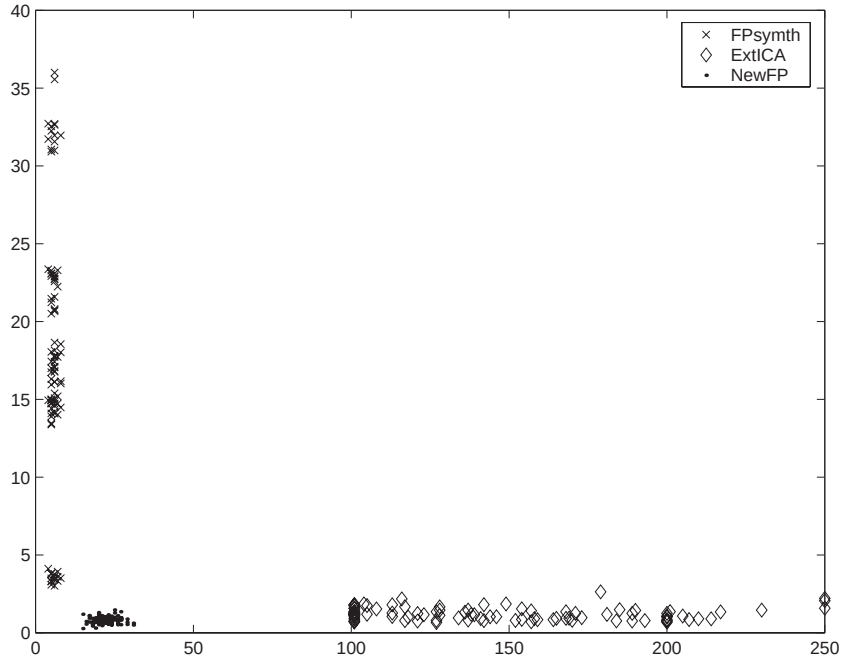

Fig. 2. The computational load vs. the performance index E.

its computational load is the largest. The new fixed-point algorithm is much faster than the ExtICA and obtains the best statistical performance. Averagely, the three algorithms took 0.69 s (FPsymth), 2.82 s (NewFP), 10.16 s (ExtICA) in the computations (2 GHz Pentium4). If the Gaussian signal is not included, the statistical performance of the FPsymth is as good as the other two.

5. Conclusions

A new fixed-point algorithm for ICA is presented that is able blindly to separate mixed signals with sub- and super-Gaussian source distributions. The independent component analysis model is estimated by maximum likelihood estimation. The new fixed-point algorithm maximizes the likelihood under the constraint of decorrelation, and uses the method of Lee et al. [10] to switch between sub- and super-Gaussian regimes. The new fixed-point algorithm is different from the fixed-point algorithm of Hyvärinen [6] which is in fact an approximative Newton method. The new fixed-point algorithm maximizes the likelihood very fast and reliably. The validity of this algorithm is confirmed by the simulations and experimental results.

Acknowledgements

The work was supported by NSFC(30170321, 90103033), MOE(KP0302) and MOST(2001CCA00700). The authors would like to thank the reviewers for their useful comments and suggestions.

References

[1] S.-I. Amari, A. Cichocki, H. Yang, A new learning algorithm for blind source separation, Advances in Neural Information Processing System, Vol. 8, MIT Press, Cambridge, MA, 1996, pp. 757–763.
[2] A. Bell, T. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Comput. 7 (6) (1995) 1129–1159.
[3] J.-F. Cardoso, B.H. Laheld, Equivariant adaptive source separation, IEEE Trans. Signal Process. 44 (12) (1996) 3017–3030.
[4] A. Cichocki, R. Unbehauen, Robust neural networks with on-line learning for blind identification and blind separation of sources, IEEE Trans. Circuits Systems 43 (11) (1996) 894–906.
[5] P. Comon, Independent component analysis—a new concept? Signal Process. 36 (1994) 287–314.
[6] A. Hyvärinen, Fast and robust fixed-point algorithm for independent component analysis, IEEE Trans. Neural Networks 10 (3) (1999) 626–634.
[7] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.
[8] A. Hyvärinen, E. Oja, A fast fixed-point algorithm for independent component analysis, Neural Comput. 9 (7) (1997) 1483–1492.
[9] C. Jutten, J. Herault, Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture, Signal Process. 24 (1991) 1–10.
[10] T.-W. Lee, M. Girolami, T. Sejnowski, Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources, Neural Comput. 11 (2) (1999) 417–441.
[11] D.G. Luenberger, Optimization by Vector Space Methods, Wiley, New York, 1969.