ELSEVIER

Neurocomputing 30 (2000) 79–102

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Some global and local convergence analysis on the information-theoretic independent component analysis approach☆

Chi Chiu Cheung*, Lei Xu

Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, NT., Hong Kong, P.R. China

Received 23 November 1996; accepted 22 March 1999

Abstract

In this paper, we present a detailed theoretical analysis on the information-theoretic Independent Component Analysis (IT-ICA) approach. We first provide a number of lemmas and theorems on properties of the corresponding cost function in the general n-channel case with differentiable, odd, monotonic decreasing nonlinearity. A theorem on behaviour of the cost function along a radially outward line is given for characterizing the global configuration of the cost function in the parameter space. Furthermore, on the 2-channel IT-ICA system with cubic nonlinearity, we not only exhaustively solve out all equilibrium points and the condition for stability, but also give a global convergence theorem.

Keywords: Independent component analysis; Nonlinearity; Information theoretic; Global convergence; High-order statistics

1. Introduction

The Blind Source Separation (BSS) problem has been a popular problem in this decade because it not only has many applications but also involves interesting problems in high-order statistics and nonlinear systems. A number of different

☆This project was supported by the HK RGC Earmarked Grant CUHK 339/96E.

* Corresponding author.

E-mail address: cheungcc@cse.cuhk.edu.hk (C.C. Cheung)

0925-2312/00/$ - see front matter
PII: S0925-2312(99)00118-6

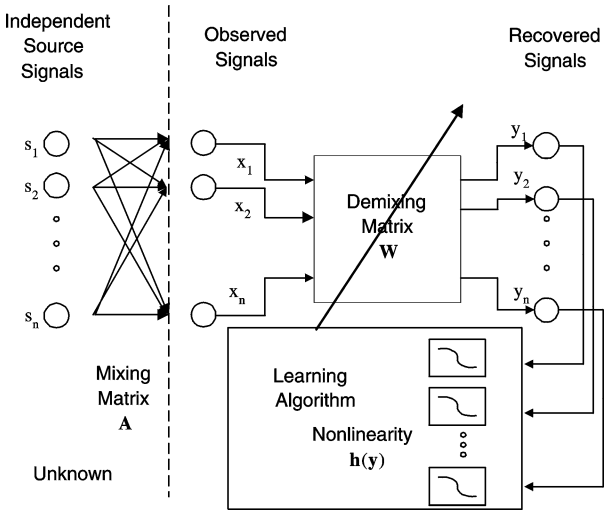


Fig. 1. The problem definition and network used.

approaches have been proposed to solve the problem. The literature is so vast that only a few are cited [12,7,9,1–3,13,23,24,11,21]. The noiseless instantaneous linear source separation problem (also known as the ICA problem) with equal number of sources and mixed signals is considered in this paper and defined as follows. Suppose there are n channels of statistically independent *source signals* $\boldsymbol{s} = [s_1, \ldots, s_n]^{\mathrm{T}}$ with zero mean. They are instantaneously and linearly mixed by a static non-singular $n \times n$ mixing matrix $\boldsymbol{A}$ to give the *observed signals* $\boldsymbol{x} = [x_1, \ldots, x_n]^{\mathrm{T}} = \boldsymbol{As}$. The ICA problem is to tune the $n \times n$ *de-mixing matrix* $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n]^{\mathrm{T}}$ so as to give the *recovered signals* $\boldsymbol{y} = [y_1, \ldots, y_n]^{\mathrm{T}} = \boldsymbol{Wx} = \boldsymbol{Vs}$ which are desired to be as similar as possible as $\boldsymbol{s}$ (Fig. 1). Theoretically, the source signals can only be determined up to arbitrary permutation of channels and scaling factors. Therefore, if we obtain $\boldsymbol{V} = \boldsymbol{WA} = \boldsymbol{PD}$, where $\boldsymbol{P}$ is a permutation matrix and $\boldsymbol{D}$ is a diagonal matrix, then source separation is said to be successful.

Among different approaches, the maximum likelihood (ML) approach [17] and information-theoretic approaches like the information maximization (INFORMAX/ME) [2], and the minimum mutual information (MMI) [1], all essentially involve the minimization of the following cost function:

$$
\begin{aligned}
J &= \int_{\boldsymbol{y}} p_{\boldsymbol{y}}(\boldsymbol{y}) \ln \frac{p_{\boldsymbol{y}}(\boldsymbol{y})}{\prod_{i=1}^{n} g_i(y_i)} \, \mathrm{d}\boldsymbol{y} \\
&= \int_{\boldsymbol{x}} p_{\boldsymbol{x}}(\boldsymbol{x}) \ln \frac{p_{\boldsymbol{x}}(\boldsymbol{x})}{|\det \boldsymbol{W}| \prod_{i=1}^{n} g_i(\boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{x})} \, \mathrm{d}\boldsymbol{x} \\
&= \int_{\boldsymbol{s}} p_{\boldsymbol{s}}(\boldsymbol{s}) \ln \frac{p_{\boldsymbol{s}}(\boldsymbol{s})}{|\det \boldsymbol{V}| \prod_{i=1}^{n} g_i(\boldsymbol{v}_i^{\mathrm{T}} \boldsymbol{s})} \, \mathrm{d}\boldsymbol{s}
\end{aligned} \tag{1}
$$

where the functions $\{g_i(y_i)\}$ are a set of 'model pdf's' satisfying

$$g_i(r) > 0, \quad r \in R, \quad \text{and} \quad \int_{-\infty}^{+\infty} g_i(r)\,\mathrm{d}r = 1 \tag{2}$$

[20][1] from which the nonlinearity $\{h_i(y_i)\}$ of the algorithm is derived and discussed. In [20], the same cost function is rederived from the Bayesian–Kullback Ying–Yang learning scheme. A gradient descent algorithm is used in [2] to minimize J but later the stochastic *natural gradient descent* algorithm [1,24] is found to be much faster and bearing good convergence properties:

$$\begin{aligned}\Delta \boldsymbol{W} &= -\varepsilon(t)[\nabla_{\boldsymbol{W}} J(\boldsymbol{W})]\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} \\ &= \varepsilon(t)[\boldsymbol{I} + \boldsymbol{h}(\boldsymbol{y})\boldsymbol{y}^{\mathrm{T}}]\boldsymbol{W},\end{aligned} \tag{3}$$

where $\boldsymbol{h}(\boldsymbol{y}) = [h_1(y_1), \ldots, h_n(y_n)]^{\mathrm{T}}$ and $h_i(y_i) = g_i'(y_i)/g_i(y_i)$, $\varepsilon(t)$ is the learning rate commonly set as a sufficiently small constant. The choice of nonlinearity $h_i(y_i)$ is crucial to the separation capability of the algorithm. In [2], a 'reversed sigmoid' function like

$$h_i(y_i) = 1 - 2\,\mathrm{logsig}(y_i) = \frac{\exp(-y_i) - 1}{\exp(-y_i) + 1} \tag{4}$$

or

$$h_i(y_i) = -2\tanh(y_i) \tag{5}$$

is used and it is experimentally found that they can separate super-Gaussian signals. In [1], the following polynomial is used:

$$h_i(y_i) = -\frac{3}{4}y^{11} - \frac{25}{4}y^9 + \frac{14}{3}y^7 + \frac{47}{4}y^5 - \frac{29}{4}y^3. \tag{6}$$

The relation between the nonlinearity used and the separation capability is further discussed in [5,21–23]. Flexible nonlinearity that is adaptable on-line is proposed [23]. As the cost function J is an information-theoretic quantity, for convenience, we regard all the algorithms featured by Eq. (3) as the 'information-theoretic ICA (IT-ICA) approach'.

In this paper, we first investigate the properties of the cost function J in the parameter space, with $h_i(y_i)$ being any differentiable, odd, monotonic decreasing function. We prove several theorems and lemmas on its singularity, continuity, absence of local maximum, behaviour of the cost function along a radially outward line, scale of the equilibrium points, etc. Then we specialize to the 2-channel sources with a cubic $h_i(y_i)$ as follows:

$$h_i(y_i) = -c_i y_i^3, \quad c_i > 0. \tag{7}$$

[1] In [20], $g_i(.)$ is only restricted to be positive and integrable. However, the restriction of integral being one adds convenience and more intuitive understanding of $g_i(.)$ being 'model pdf's' for the sources.

We exhaustively solve all separating and non-separating equilibrium points of the cost function, find out the condition for stability, and finally, with the help of results proved above, give a global convergence theorem.

The cubic nonlinearity or control of fourth-order cross-moments $E[y_i^3 y_j]$ appeared in various literature in the past [4,10,8,12,14,16,18], in either feedback HJ network or direct (feedforward) networks. The condition of stability of the separating states in our work (Theorem 23 in Section 3.3) turns out to be the same as those obtained in the previous works [4,14,16,18]. However, our works differ from these previous works at least in two aspects. First, in [4,14,16,18], the diagonal elements of $\boldsymbol{W}$ were set as 1, and thus, there were only two tunable variables in the 2-channel case. In our work, the whole de-mixing matrix is to be tuned and thus the parameter space is four-dimensional in the 2-channel case. The learning equation (3) of the off-diagonal elements of $\boldsymbol{W}$ is coupled with the diagonal elements, and the algorithm cannot be trivially reduced to a form equivalent to those in the mentioned literature. Hence, the analysis in this work and the mentioned literature are on different network structures and different algorithms. Second, and most important of all, a *global convergence theorem* is proved in our work while only local convergence is proved in the mentioned literature. In particular, Theorem 5 in this paper, which is on the behaviour of the cost function along a radially outward line in the parameter space, clearly depicts the global configuration of the cost function in the parameter space and is of great importance. It provides a direct tool for global convergence analysis, while investigation on stability of each equilibrium point can only prove local convergence within its basin of attraction.

This paper is organized as follows. Section 2 presents theoretical results on the properties of the cost function and algorithm. Section 3 investigates the 2-channel cubic nonlinearity case and arrives at a global convergence theorem. Section 4 presents experimental verification on the 2-channel cubic nonlinearity case. Section 5 gives the conclusions.

2. Properties of the cost function and algorithms

In this section, we shall present some theoretical results on the properties of the cost function $J(\boldsymbol{V})$ in the n^2-dimensional $\boldsymbol{V}$-parameter space. The analysis of the information-theoretic ICA approach follows an idea proposed by Xu and Amari in [20] that investigates the cost function J in the $\boldsymbol{V}$-parameter space $(v_{11}, \ldots, v_{1n}, v_{21}, \ldots, v_{nn})$ rather than in the $\boldsymbol{W}$-parameter space $(w_{11}, \ldots, w_{1n}, w_{21}, \ldots, w_{nn})$. This is because $\boldsymbol{V} = \boldsymbol{WA}$ bears a one-to-one mapping to $\boldsymbol{W}$ and is a set of parameters that completely characterizes the system. The mathematical analysis using $\boldsymbol{V}$ is simpler since it does not explicitly involve $\boldsymbol{A}$ and $\boldsymbol{x}$. Any result obtained in terms of $\boldsymbol{V}$ can by transformed to that in terms of $\boldsymbol{W}$ by $\boldsymbol{W} = \boldsymbol{VA}^{-1}$.

In particular, the analysis often involves determination of equilibrium points of the cost function J. A standard method to do so is to solve the equilibrium equation for $\boldsymbol{W}$:

$$\nabla_{\boldsymbol{W}} J(\boldsymbol{W}) = E_{\boldsymbol{x}}[I + \boldsymbol{h}(\boldsymbol{Wx})(\boldsymbol{Wx})^{\mathrm{T}}][\boldsymbol{W}^{\mathrm{T}}]^{-1} = \boldsymbol{0}. \tag{8}$$

However, this equation is difficult to be solved directly due to the involvement of mixture $\boldsymbol{x}$. On the other hand, it is equivalent to solving the equilibrium equation for $\boldsymbol{V}$:

$$\nabla_{\boldsymbol{V}} J(\boldsymbol{V}) = E_s[I + \boldsymbol{h}(\boldsymbol{V}\boldsymbol{s})(\boldsymbol{V}\boldsymbol{s}^{\mathrm{T}})][\boldsymbol{V}^{\mathrm{T}}]^{-1} = \boldsymbol{0} \tag{9}$$

since $\nabla_{\boldsymbol{W}} J(\boldsymbol{W}) = \nabla_{\boldsymbol{V}} J(\boldsymbol{V})\boldsymbol{A}^{\mathrm{T}}$ and $\boldsymbol{A}$ is non-singular [20]. This equation is easier to solve since sources $\boldsymbol{s}$ are independent. Hence, it is more appropriate to investigate the J scalar field in the $\boldsymbol{V}$-parameter space rather than in the $\boldsymbol{W}$-parameter space.

In [20], it is pointed out that after associating the CDF-like transformation function

$$f_i(y_i) = \int_{-\infty}^{y_i} g_i(r)\,\mathrm{d}r \tag{10}$$

to obtain the transformed vector $\boldsymbol{z} = [f_1(y_1), \ldots, f_n(y_n)]^{\mathrm{T}}$ with Jacobian of transformation

$$\det\left[\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}^{\mathrm{T}}}\right] = \prod_{i=1}^{n} g_i(y_i), \tag{11}$$

the negative (differential) entropy of $\boldsymbol{z}$, $-H(\boldsymbol{z})$ in the INFORMAX approach [2], is equivalent to the cost function J:

$$J = \int_{\boldsymbol{y}} p_{\boldsymbol{y}}(\boldsymbol{y}) \ln \frac{p_{\boldsymbol{y}}(\boldsymbol{y})}{\prod_i^n g_i(y_i)}\,\mathrm{d}\boldsymbol{y} = \int_{\boldsymbol{z}} p_{\boldsymbol{z}}(\boldsymbol{z}) \ln p_{\boldsymbol{z}}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} = -H(\boldsymbol{z}). \tag{12}$$

This equivalence provides good intuitive understanding of some properties of the cost function to be introduced.

The following lemmas and theorem investigate the information-theoretic ICA approach with any differentiable, odd, monotonic decreasing nonlinearity $h_i(y_i)$ (some lemmas even have less restrictions on $h_i(y_i)$). This broad class includes, but is not restricted to:

- the cubic nonlinearity equation (7), and other $h_i(y_i) = -c_i y_i^p$, $c_i > 0$, p being a positive odd integer,
- $h_i(y_i)$ being reversed sigmoids like Eqs. (4) and (5)
- $h_i(y_i) = -c_i y_i^{1/p}$, $c_i > 0$, p being a positive odd integer,

2.1. Singularity and continuity of $J(\boldsymbol{V})$

Lemma 1 (Singularity). *For the information-theoretic ICA approach with any* $g_i(.)$ *satisfying Eq.* (2) *on any number of channels*, $J(\boldsymbol{V}) \to +\infty$ *as* $\det \boldsymbol{V} \to 0$.

Proof. If $\det \boldsymbol{V} = \det \boldsymbol{W} = 0$, there is a deterministic linear dependence on the recovered signals, that is, any y_i can be written as $y_i = L_i(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$ where $L_i(.)$ is a linear function. Hence, $z_i = f_i(y_i)$ also bears a deterministic relationship with

$\{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}$ and the differential conditional entropy $H(z_i|z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n) \to -\infty$. As the joint entropy $H(\boldsymbol{z}) = H(z_i) + H(z_i|z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$, we have $H(\boldsymbol{z}) \to -\infty$, and $J(\boldsymbol{V}) \to +\infty$. □

The (n^2-1)-dimensional subspace defined by $\det \boldsymbol{V} = 0$ ($\det \boldsymbol{W} = 0$) in the n^2-dimensional $\boldsymbol{V}$-parameter space ($\boldsymbol{W}$-parameter space) is called the 'singular subspace'.

Remark 2. For the natural gradient algorithm, Eq. (3), if $\boldsymbol{W}$ is initialized as a singular matrix, it will subsequently be trapped in the singular subspace $\det \boldsymbol{W} = 0$ because

$$\det \boldsymbol{W}_{t+1} = (\det[\boldsymbol{I} + \varepsilon(t)[\boldsymbol{I} + \boldsymbol{h}(\boldsymbol{y})\boldsymbol{y}^{\mathrm{T}}]])(\det \boldsymbol{W}_t) = 0. \tag{13}$$

Surely, it cannot perform source separation.

Lemma 3 (Continuity). *For the information-theoretic ICA approach with odd, monotonic decreasing $h_i(y_i)$ nonlinearity on any number of channels of signals, $J(\boldsymbol{V})$ is continuous at any non-singular $\boldsymbol{V}$.*

Proof. $J(\boldsymbol{V})$ is continuous at some finite point $\boldsymbol{V}^*$ if and only if the gradient $\nabla_V J(\boldsymbol{V})$ exists and is finite at $\boldsymbol{V}^*$. Consider

$$\nabla_V J(\boldsymbol{V}) = -\left\{\frac{(\mathrm{adj}\, \boldsymbol{V})^{\mathrm{T}}}{\det \boldsymbol{V}} + E_s[\boldsymbol{h}(\boldsymbol{V}\boldsymbol{s})(\boldsymbol{s})^{\mathrm{T}}]\right\}, \tag{14}$$

where adj $\boldsymbol{V}$ denotes the adjoint of $\boldsymbol{V}$. It is obvious that $E_s[h_i(\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{s})s_j]$ at a finite point $\boldsymbol{V}^*$ is finite for odd, monotonic decreasing $h_i(y_i)$. The magnitudes of the elements in the first term are infinitely large only when $\det \boldsymbol{V} = 0$, hence $\nabla_V J(\boldsymbol{V})$ exists and is finite for any non-singular $\boldsymbol{V}$. Therefore, $J(\boldsymbol{V})$ is continuous anywhere except on the singular subspace. □

2.2. Absence of local maximum of $J(\boldsymbol{V})$

Lemma 4. *For the information-theoretic ICA approach with differentiable, monotonic decreasing $h_i(y_i)$ nonlinearity, on any number of channels of signals, there is no local maximum of $J(\boldsymbol{V})$ in the whole $\boldsymbol{V}$-parameter space.*

Proof. The stability of equilibrium points are determined by checking the Hessian matrix:

$$\nabla_V^2 J(\boldsymbol{V}) = \boldsymbol{Q} = \begin{bmatrix} \frac{\partial^2 J}{\partial v_{11}\partial v_{11}} & \cdots & \frac{\partial^2 J}{\partial v_{11}\partial v_{1n}} & \frac{\partial^2 J}{\partial v_{11}\partial v_{21}} & \cdots & \frac{\partial^2 J}{\partial v_{11}\partial v_{nn}} \\ \vdots & \ddots & \vdots & & & \vdots \\ \frac{\partial^2 J}{\partial v_{1n}\partial v_{11}} & \cdots & \frac{\partial^2 J}{\partial v_{1n}\partial v_{1n}} & & & \vdots \\ \frac{\partial^2 J}{\partial v_{21}\partial v_{11}} & & & \ddots & & \vdots \\ \vdots & & & & \ddots & \vdots \\ \frac{\partial^2 J}{\partial v_{nn}\partial v_{11}} & \cdots & \cdots & \cdots & \cdots & \frac{\partial^2 J}{\partial v_{nn}\partial v_{nn}} \end{bmatrix}. \tag{15}$$

If the Hessian is positive definite, the equilibrium point is a local minimum; if the Hessian is negative definite, the equilibrium point is a local maximum; if the Hessian is neither positive definite nor negative definite, i.e., some of the eigenvalues are/is positive and some are/is negative, the equilibrium point is a saddle point.

The diagonal elements of the Hessian matrix of $J(\boldsymbol{V})$ are

$$\frac{\partial^2 J}{\partial v_{ij}\partial v_{ij}} = \frac{(\operatorname{cof} v_{ij})^2}{(\det \boldsymbol{V})^2} + E_s[-h'_i(\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{s})s_j^2] \quad i,j = 1,2,\ldots,n. \tag{16}$$

For any monotonic decreasing $h_i(y_i)$, $h'_i(y_i) < 0$, $E_s[-h'_i(\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{s})s_j^2] > 0$, and hence $\partial^2 J/\partial v_{ij}\partial v_{ij} > 0$. The first leading principal minor of the Hessian matrix is a diagonal element, and hence is positive. Therefore, the Hessian matrix cannot be negative definite at any $\boldsymbol{V}$ and there is no local maximum in the whole $\boldsymbol{V}$-parameter space. □

2.3. Behaviour of $J(\boldsymbol{V})$ along a radially outward line

Theorem 5. *For the information-theoretic ICA approach with differential, odd, monotonic decreasing $h_i(y_i)$ nonlinearity on any number of channels of signals, consider a radially outward line $\boldsymbol{V} = N\hat{\boldsymbol{V}}$ passing through a non-singular $\hat{\boldsymbol{V}} = [\hat{\boldsymbol{v}}_1,\ldots,\hat{\boldsymbol{v}}_n]^{\mathrm{T}}$, where $N = \|\boldsymbol{V}\| = [(\mathrm{Vec}(\boldsymbol{V}^{\mathrm{T}}))^{\mathrm{T}}\cdot\mathrm{Vec}(\boldsymbol{V}^{\mathrm{T}})]^{1/2} = \sqrt{\sum_{i=1}^n\sum_{j=1}^n v_{ij}^2}$ is the Euclidean norm of $\boldsymbol{V} = N\hat{\boldsymbol{V}}$, and $\hat{\boldsymbol{V}} = \boldsymbol{V}/N$ is a point on the sphere of unit norm. For any non-singular $\hat{\boldsymbol{V}}$, the cost function along the line, written as $J_{\hat{\boldsymbol{V}}}(N) = J(N\hat{\boldsymbol{V}})$, $N \in [0, +\infty)$ is convex and has a unique finite local minimum $N_{\hat{\boldsymbol{V}}}^{\min}$.*

Proof. The directional derivative of $J(\boldsymbol{V})$ along the radially outward direction $\mathrm{Vec}(\hat{\boldsymbol{V}}^{\mathrm{T}})$ is

$$\begin{aligned}
J'_{\hat{\boldsymbol{V}}}(N) &= \frac{\mathrm{d}J(N\hat{\boldsymbol{V}})}{\mathrm{d}N} \\
&= \frac{\partial J}{\partial \mathrm{Vec}(\boldsymbol{V}^{\mathrm{T}})}\cdot\mathrm{Vec}(\hat{\boldsymbol{V}}^{\mathrm{T}}) \\
&= -\frac{1}{N}\left\{\mathrm{Vec}\left[\frac{(\operatorname{adj}\boldsymbol{V})^{\mathrm{T}}}{\det\boldsymbol{V}} + E_s[\boldsymbol{h}(\boldsymbol{V}\boldsymbol{s})(\boldsymbol{s})^{\mathrm{T}}]\right]^{\mathrm{T}}\right\}^{\mathrm{T}}\cdot\mathrm{Vec}(\boldsymbol{V}^{\mathrm{T}}) \\
&= -\frac{1}{N}\left\{\frac{1}{\det\boldsymbol{V}}\sum_{i=1}^n\sum_{j=1}^n v_{ij}\operatorname{cof} v_{ij} + \sum_{i=1}^n E_s[h_i(\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{s})(\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{s})]\right\} \\
&= -\frac{1}{N}\left\{\frac{1}{\det\boldsymbol{V}}(n\det\boldsymbol{V}) + \sum_{i=1}^n E_s[h_i(\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{s})(\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{s})]\right\} \\
&= -\frac{1}{N}\sum_{i=1}^n\{1 + NE_s[h_i(N\hat{\boldsymbol{v}}_i^{\mathrm{T}}\boldsymbol{s})(\hat{\boldsymbol{v}}_i^{\mathrm{T}}\boldsymbol{s})]\}
\end{aligned} \tag{17}$$

where the factor

$$F(N) = -\sum_{i=1}^{n} \{1 + NE_s[h_i(N\hat{\boldsymbol{v}}_i^{\mathrm{T}}\boldsymbol{s})(\hat{\boldsymbol{v}}_i^{\mathrm{T}}\boldsymbol{s})]\} \tag{18}$$

is a monotonic increasing function of N since $\{h_i(y_i)\}$ are odd and monotonic decreasing. Noting that $F(0) = -n$ and $F(N) \to +\infty$ as $N \to +\infty$, we deduce that $F(N)$ must change sign at some unique, finite $N_{\hat{\boldsymbol{V}}}^{\min}$. Hence $J'_{\hat{\boldsymbol{V}}}(N_{\hat{\boldsymbol{V}}}^{\min}) = 0$.

Now consider

$$J''_{\hat{\boldsymbol{V}}}(N) = \frac{\mathrm{d}J'_{\hat{\boldsymbol{V}}}(N)}{\mathrm{d}N} = \frac{n}{N^2} + \sum_{i=1}^{n} E_s[-h'_i(N\hat{\boldsymbol{v}}_i^{\mathrm{T}}\boldsymbol{s})(\hat{\boldsymbol{v}}_i^{\mathrm{T}}\boldsymbol{s})^2]. \tag{19}$$

As $h_i(.)$ are monotonic decreasing, $h'_i(y_i) < 0$, and $J''_{\hat{\boldsymbol{V}}}(N) > 0 \ \forall N \in [0, +\infty)$. Hence, $J_{\hat{\boldsymbol{V}}}(N)$ is convex and $N_{\hat{\boldsymbol{V}}}^{\min}$ is a local minimum. □

Corollary 6 (Asymptotic behaviour). *For the information-theoretic ICA approach with differentiable, odd, monotonic decreasing $h_i(y_i)$ on any number of channels of signals, $J(\boldsymbol{V}) \to +\infty$ as the norm $\|\boldsymbol{V}\| \to +\infty$.*

Noting also that $J(N\hat{\boldsymbol{V}}) \to +\infty$ as $N \to 0$, since $\boldsymbol{V} = \boldsymbol{0}$ is a singularity, We can illustrate the behaviour of $J(\boldsymbol{V})$ along a radially outward line in Fig. 2.

As the information-theoretic ICA algorithm performs descent (stochastically) in $J(\boldsymbol{V})$ while $J(\boldsymbol{V}) \to +\infty$ as $\|\boldsymbol{V}\| \to +\infty$, the probability for $\boldsymbol{V}$ to continuously move in any outward direction tends to zero. Hence we have the following corollary:

Corollary 7 (Impossibility of divergence). *For the information-theoretic ICA algorithm with differentiable, odd, monotonic decreasing $h_i(y_i)$ nonlinearity acting on any number of channels of signals, if $\boldsymbol{W}$ ($\boldsymbol{V}$) is initialized at some non-singular point, then any v_{ij}, $i,j = 1, \ldots, n$ of $\boldsymbol{V}$ will not diverge to $\pm\infty$.*

Remark 8. The increase of $J(\boldsymbol{V})$ in the outward direction of $\boldsymbol{V}$ can be understood intuitively by considering entropy $H(\boldsymbol{z})$. The CDF-like transformation function $f_i(y_i)$ is limiting to the lower bound as $y_i \to -\infty$ and limiting to the upper bound as

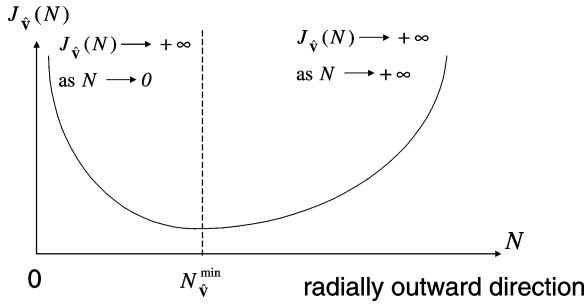


Fig. 2. An illustration of the behavior of $J(\boldsymbol{V})$ along a radially outward line.

$y_i \to +\infty$. Both ends of $f_i(y_i)$ tend to be more flat as $y_i \to \pm\infty$, that is, large ranges of y_i with large magnitudes are mapped into small ranges of $z_i = f_i(y_i)$. As $\|\boldsymbol{V}\|$ tends to be large, the random variable y_i generally becomes large and $z_i = f_i(y_i)$ is squeezed to be more concentrated at the regions near the upper and lower bounds. Hence, there is less randomness in $\boldsymbol{z}$, the entropy $H(\boldsymbol{z})$ decreases or equivalently, $J(\boldsymbol{V})$ increases. In the limit $\|\boldsymbol{V}\| \to +\infty$, some z_i is squeezed to the discrete upper and lower bounds of $f_i(.)$ and the (differential) entropy $H(\boldsymbol{z}) \to -\infty$ or equivalently $J(\boldsymbol{V}) \to +\infty$.

2.4. Effect of scale parameter of the nonlinearity

If a family of nonlinear functions $g_i(y_i)$ can be written as $(1/\theta_i)\check{g}_i(y_i/\theta_i)$, θ_i is called the scale parameter of $g_i(y_i)$. The corresponding $h_i(y_i) = (1/\theta_i)\check{h}_i(y_i/\theta_i)$. We shall state that this scale parameter only has the effect of controlling the scale of the equilibrium points and does not affect the separation capability of the nonlinearity. As the proofs are straightforward but tedious (refer to [5]), and they are not contained in this paper.

Lemma 9. *Consider an information-theoretic ICA system A using $\tilde{g}_i(y_i)$ and another information-theoretic ICA system B using $g_i(y_i) = (1/\theta_i)\tilde{g}_i(y_i/\theta_i)$, $i = 1, \ldots, n$. For any $\boldsymbol{V}^A* = [\boldsymbol{v}_1^A*, \ldots, \boldsymbol{v}_n^A*]^T$, let $\boldsymbol{V}^B* = [\boldsymbol{v}_1^B*, \ldots, \boldsymbol{v}_n^B*]^T$ such that $\boldsymbol{v}_i^B* = \theta_i \boldsymbol{v}_i^A*$, $i = 1, \ldots, n$. Then, $\boldsymbol{V}^B*$ is an equilibrium point of system B if and only if $\boldsymbol{V}^A*$ is an equilibrium point of system A.*

Corollary 10. *For an information-theoretic ICA system using $g_i(y_i)$ from a scale family $\{(1/\theta_i)\tilde{g}_i(y_i/\theta_i)\}$, $i = 1, \ldots, n$, the magnitude of $\boldsymbol{v}_i^T$ of the equilibrium points is controlled by the scale parameter θ_i as $v_{ij} \propto \theta_i$, $j = 1, \ldots, n$, and after the system has converged to some equilibrium point, the magnitude of recovered signal $E[|y_i|]$ is proportional to θ_i.*

Corollary 11. *For an information-theoretic ICA system using $g_i(y_i)$ from a scale family $\{(1/\theta_i)\tilde{g}_i(y_i/\theta_i)\}$, $i = 1, \ldots, n$, the values of scale parameters $\{\theta_i\}$ do not affect the number and forms of the solutions of the equilibrium equation.*

Lemma 12. *Consider systems A and B in Lemma 9. The stability of $\boldsymbol{V}^B*$ in system B is the same as the stability of $\boldsymbol{V}^A*$ in system A.*

Corollary 13. *The scale parameter cannot affect the stability of the equilibrium points.*

A nonlinearity is said to be capable of separating some sources if $\boldsymbol{V}$ converge to a correct solution $\boldsymbol{PD}$. The separation capability hence depends on the number of equilibrium points, forms (whether they are equal to some $\boldsymbol{PD}$) of the equilibrium points and the stability of the equilibrium points. Since the scale parameter can affect none of the factors, we have the following corollary:

Corollary 14. *The scale parameter cannot affect the separation capability and a family of information-theoretic ICA systems using members of a scale family $\{(1/\theta_i)\tilde{g}_i(y_i/\theta_i)\}$ with different θ_i has the same separation capability.*

2.5. Comparison of cost function values of equilibrium points

Lemma 15. *For the information-theoretic ICA algorithms with* $h_i(y_i) = -c_i y_i^p$, *where* p *is a positive integer greater than* 1 *or a fraction* $1/q$ *with positive integer* q *greater than* 1 ($p = 3, 5, \ldots$ *or* $p = 1/3, 1/5, \ldots$), *on any number of channels, if* $\boldsymbol{V}_A$ *and* $\boldsymbol{V}_B$ *are two equilibrium points, the equation*

$$J(\boldsymbol{V}_A) - J(\boldsymbol{V}_B) = \ln \frac{|\det \boldsymbol{V}_B|}{|\det \boldsymbol{V}_A|} \tag{20}$$

always holds.

Proof. From Eq. (1),

$$\begin{aligned} J(\boldsymbol{V}_A) - J(\boldsymbol{V}_B) &= E_s\left[\ln \frac{p_s(\boldsymbol{s})}{|\det \boldsymbol{V}_A| \prod_{i=1}^n g_i(\boldsymbol{v}_{Ai}^{\mathrm{T}} \boldsymbol{s})}\right] - E_s\left[\ln \frac{p_s(\boldsymbol{s})}{|\det \boldsymbol{V}_B| \prod_{j=1}^n g_j(\boldsymbol{v}_{Bj}^{\mathrm{T}} \boldsymbol{s})}\right] \\ &= \ln \frac{|\det \boldsymbol{V}_B|}{|\det \boldsymbol{V}_A|} - \sum_{i=1}^n E_s[\ln g_i(\boldsymbol{v}_{Ai}^{\mathrm{T}} \boldsymbol{s})] + \sum_{j=1}^n E_s[\ln g_j(\boldsymbol{v}_{Bj}^{\mathrm{T}} \boldsymbol{s})]. \end{aligned} \tag{21}$$

For the mentioned nonlinearity, $g_i(y_i) = C_i \exp(-c_i y_i^{p+1}/(p+1))$ where C_i are normalizing constants. Then,

$$J(\boldsymbol{V}_A) - J(\boldsymbol{V}_B) = \ln \frac{|\det \boldsymbol{V}_B|}{|\det \boldsymbol{V}_A|} + \sum_{i=1}^n E_s[c_i |\boldsymbol{v}_{Ai}^{\mathrm{T}} \boldsymbol{s}|^{p+1}] - \sum_{j=1}^n E_s[c_j |\boldsymbol{v}_{Bj}^{\mathrm{T}} \boldsymbol{s}|^{p+1}]. \tag{22}$$

However, we have $E_s[c_i(|\boldsymbol{v}_i^{\mathrm{T}} \boldsymbol{s}|)^{p+1}] = 1$ for any equilibrium point by the self-coupling equilibrium equations (equations for the diagonal elements of matrix equation (9), see Eq. (29)). Hence, the second term in the above equation equals n and the third term equals $-n$. They cancel each other and thus Eq. (20) holds. □

2.6. Number and stability of correct solutions in the 2-channel case

Lemma 16. *For the information-theoretic ICA approach with odd, monotonic decreasing* $h_i(y_i)$ *on* 2 *channels of signals, the correct solutions have the form*

$$\textit{Solutions A1} - \text{A4}, \quad \boldsymbol{V} = \begin{bmatrix} \pm |v_{11}^*| & 0 \\ 0 & \pm |v_{22}^*| \end{bmatrix} \tag{23}$$

or

$$\textit{Solutions A5} - \text{A8}, \quad \boldsymbol{V} = \begin{bmatrix} 0 & \pm |v_{12}^*| \\ \pm |v_{21}^*| & 0 \end{bmatrix}, \tag{24}$$

where, for Solutions A1–A4, $|v_{11}^*|$ *and* $|v_{22}^*|$ *are respectively the unique magnitudes of the solutions for the self-coupling equilibrium equations*

$$1 + E_{s_1}[h_1(v_{11} s_1) s_1 v_{11}] = 0, \tag{25}$$

$$1 + E_{s_2}[h_2(v_{22} s_2) s_2 v_{22}] = 0 \tag{26}$$

and for Solutions A5–A8, $|v_{12}^*|$ and $|v_{21}^*|$ are respectively the unique magnitude of the solutions for the self-coupling equilibrium equations:

$$1 + E_{s_2}[h_1(v_{12}s_2)s_2v_{12}] = 0, \tag{27}$$

$$1 + E_{s_1}[h_2(v_{21}s_1)s_1v_{21}] = 0. \tag{28}$$

There are totally 8 correct solutions in the 2-channel case.

Proof. The matrix equilibrium equation (9) can be explicitly written as

Self-coupling equilibrium equations:

$$E_s[-h_i(\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{s})\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{s}] = 1, \quad i = 1, \ldots, n. \tag{29}$$

Cross-coupling equilibrium equations:

$$E_s[-h_i(\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{s})\boldsymbol{v}_j^{\mathrm{T}}\boldsymbol{s}] = 0, \quad i, j = 1, \ldots, n, \ i \neq j. \tag{30}$$

Substituting $v_{12} = v_{21} = 0$ for solutions A1–A4, it can be easily seen that the cross-coupling equilibrium equations are automatically satisfied. As $h_i(y_i)$ are monotonic decreasing and odd, the magnitudes $|v_{11}^*|$ and $|v_{22}^*|$ have unique solutions for the self-coupling equations (25) and (26), respectively. The case for solutions A5–A8 is similar. Counting the combination of signs of the elements, we conclude that there are exactly eight correct solutions. □

This lemma can be easily extended to the n-channel case to give the number of correct solutions that separate all the n sources as being equal to $2^n \times n!$.

Remark 17. We can have some intuitive understanding on the functionality of the self-coupling and cross-coupling equilibrium equations. The cross-coupling equilibrium equation restricts high-order cross-moments $E[-h_i(y_i)y_j]$, $i \neq j$, to be zero and determines possible forms of equilibrium points. The self-coupling equations control the magnitudes of the recovered signals and hence the magnitudes of the equilibrium points. Note that there can also be non-separating $\boldsymbol{V}$ which satisfies the cross-coupling equilibrium equations and hence a spurious solution can exist.

Lemma 18. *For the information-theoretic ICA approach with differentiable, odd, monotonic decreasing $h_i(y_i)$ on 2 channels of signals, the sufficient and necessary condition for Solutions A1 − A4 in Eq.* (23) *to be stable is*

$$E[s_1^2]E[s_2^2]E_{s_1}[-h'_1(v_{11}^*s_1)]E_{s_2}[-h'_2(v_{22}^*s_2)] - \frac{1}{v_{11}^{*2}v_{22}^{*2}} > 0. \tag{31}$$

The sufficient and necessary condition for Solutions A5–A8 in Eq. (24) *to be stable is*

$$E[s_1^2]E[s_2^2]E_{s_2}[-h'_1(v_{12}^*s_2)]E_{s_1}[-h'_2(v_{21}^*s_1)] - \frac{1}{v_{12}^{*2}v_{21}^{*2}} > 0. \tag{32}$$

Proof. From Eq. (15), the Hessian matrix for Solutions A1–A4 is

$$\nabla_V^2 J(V) = \boldsymbol{Q} =$$

$$\begin{bmatrix} \frac{1}{v_{11}^{*2}} + E_{s_1}[-h'_1(v_{11}^* s_1)s_1^2] & 0 & 0 & 0 \\ 0 & E_s[-h'_1(v_{11}^* s_1)s_2^2] & \frac{1}{v_{11}^* v_{12}^*} & 0 \\ 0 & \frac{1}{v_{11}^* v_{12}^*} & E_s[-h'_2(v_{22}^* s_2)s_1^2] & 0 \\ 0 & 0 & 0 & \frac{1}{v_{22}^{*2}} + E_{s_2}[-h'_2(v_{22}^* s_2)s_2^2] \end{bmatrix}. \tag{33}$$

The sufficient and necessary condition for the Hessian to be positive definite is that all leading principal minors must be positive. The elements q_{11}, q_{44} are always positive. Hence, the condition for all leading principal minors to be positive is

$$\det\begin{bmatrix} q_{22} & q_{23} \\ q_{32} & q_{33} \end{bmatrix} = E_s[-h'_1(v_{11}^* s_1)s_2^2]E_s[-h'_2(v_{22}^* s_2)s_1^2] - \frac{1}{v_{11}^{*2} v_{22}^{*2}} > 0. \tag{34}$$

By the independence assumption of source signals, condition (31) is reached. Similarly, the condition for stability of Solutions A5–A8 can be found. □

Lemmas 16 and 18 are actually partial results. To apply them to a particular nonlinearity, we have to solve Eqs. (25)–(28), (31) and (32), and get the condition of stability in terms of the statistics or distribution of the sources only. However, one difficulty is to pick the elements $\{v_{ij}^*\}$ out of the expectation operation. Only in simple cases like the cubic nonlinearity case can we pick $\{v_{ij}^*\}$ out. For the reversed sigmoid, as the expectation of terms involving $h_i(\boldsymbol{v}_i\boldsymbol{s})$ and $h'_i(\boldsymbol{v}_i\boldsymbol{s})$ are difficult to be broken down to separate $\boldsymbol{v}_i$ from the expectation (at least, the Taylor expansion of $h_i(y_i)$ involves infinite number of terms), the equilibrium equations and condition for stability are difficult to solve and study. Moreover, investigation on other non-separating equilibrium points and stability of them are needed for global convergence analysis.

3. Investigation in the 2-channel cubic nonlinearity case and global convergence analysis

The global convergence behaviour of the information-theoretic ICA algorithms with the cubic nonlinearity equation (7) is investigated through the following three steps: (1) Explicitly and exhaustively determine all equilibrium points of the cost function. (2) Determine, for each equilibrium point, whether or under what condition it is a local minimum or saddle point. (3) Incorporate with the proved global properties on the cost function scalar field to give the global convergence theorem.

3.1. Exhaustive determination of equilibrium points

Writing the set of equilibrium equations (29) and (30) out, we have

Self-coupling equilibrium equations:

$$E[y_1^4] = v_{11}^4\mu_1^4 + 2v_{11}^2v_{12}^2m + v_{12}^4\mu_2^4 = \frac{1}{c_1}, \tag{35}$$

$$E[y_2^4] = v_{21}^4\mu_1^4 + 2v_{21}^2v_{22}^2m + v_{22}^4\mu_2^4 = \frac{1}{c_2}. \tag{36}$$

Cross-coupling equilibrium equations:

$$E[y_1^3y_2] = v_{11}v_{21}(v_{11}^2\mu_1^4 + v_{12}^2m) + v_{12}v_{22}(v_{12}^2\mu_2^4 + v_{11}^2m) = 0, \tag{37}$$

$$E[y_2^3y_1] = v_{11}v_{21}(v_{21}^2\mu_1^4 + v_{22}^2m) + v_{12}v_{22}(v_{22}^2\mu_2^4 + v_{21}^2m) = 0, \tag{38}$$

where $\mu_i^p = E[s_i^p]$ and $m = 3\mu_1^2\mu_2^2$.

They are a system of four equations with four unknowns $\{v_{11}, v_{12}, v_{21}, v_{22}\}$. A key step to solve them is to write the cross-coupling equations (37) and(38) in the following form (suggested by Jiong Ruan):

$$\begin{bmatrix} v_{11}^2 & v_{12}^2 \\ v_{21}^2 & v_{22}^2 \end{bmatrix}\begin{bmatrix} \mu_1^4 & m \\ m & \mu_2^4 \end{bmatrix}\begin{bmatrix} v_{11}v_{21} \\ v_{12}v_{22} \end{bmatrix} = \mathbf{0}. \tag{39}$$

Denote

$$\boldsymbol{M} = \begin{bmatrix} v_{11}^2 & v_{12}^2 \\ v_{21}^2 & v_{22}^2 \end{bmatrix}\begin{bmatrix} \mu_1^4 & m \\ m & \mu_2^4 \end{bmatrix}. \tag{40}$$

Eq. (39) implies

$$\begin{bmatrix} v_{11}v_{21} \\ v_{12}v_{22} \end{bmatrix} = \mathbf{0} \quad \text{or} \quad \det \boldsymbol{M} = 0. \tag{41}$$

We treat these two exhaustive possibilities in case A and case B, respectively.

Case A:

$$\begin{bmatrix} v_{11}v_{21} \\ v_{12}v_{22} \end{bmatrix} = \mathbf{0}. \tag{42}$$

By self-coupling equilibrium equations (35) and(36), we find that elements in any row of $\boldsymbol{V}$ cannot be all zeros simultaneously. Hence, putting $v_{12} = 0$ and $v_{21} = 0$ into the self-coupling equations (35) and(36), we get:

Solutions A1–A4:

$$\boldsymbol{V} = \begin{bmatrix} \pm (c_1\mu_1^4)^{-1/4} & 0 \\ 0 & \pm (c_2\mu_2^4)^{-1/4} \end{bmatrix}. \tag{43}$$

Putting $v_{11}=0$ and $v_{22}=0$ into Eqs. (35) and (36), we get:

Solutions A5 – A8:

$$\boldsymbol{V}=\begin{bmatrix} 0 & \pm(c_1\mu_2^4)^{-1/4} \\ \pm(c_2\mu_1^4)^{-1/4} & 0 \end{bmatrix}. \tag{44}$$

Solutions A1–A8 are the eight and only eight solutions in case A and they are the correct solutions that can perform source separation mentioned in Lemma 16.

Case B: Now we consider the case

$$\det \boldsymbol{M}=(v_{11}^2v_{22}^2-v_{12}^2v_{21}^2)(\mu_1^4\mu_2^4-m^2)=0. \tag{45}$$

Assume that

$$\mu_1^4\mu_2^4-m^2=\mu_1^4\mu_2^4-[3(\mu_1^2)^2][3(\mu_2^2)^2]\neq 0, \tag{46}$$

i.e., the two sources are not 'globally Gaussian', Eq. (45) becomes

$$v_{11}^2v_{22}^2-v_{12}^2v_{21}^2=(v_{11}v_{22}+v_{12}v_{21})(v_{11}v_{22}-v_{12}v_{21})=0. \tag{47}$$

Since $\det \boldsymbol{V}=v_{11}v_{22}-v_{12}v_{21}\neq 0$ for equilibrium points ($|\nabla_{\boldsymbol{v}}J(\boldsymbol{V})|\to+\infty$ if $\det \boldsymbol{V}\to 0$), we have

$$v_{11}v_{22}+v_{12}v_{21}=0. \tag{48}$$

Notice the possible combinations of the signs $s_{ij}=v_{ij}/|v_{ij}|$ of v_{ij}. From Eq. (48), we find that three of the four v_{ij} must be of the same sign, and the remaining one the opposite sign. Hence we have the constraint

$$s_{11}s_{12}s_{21}s_{22}=-1. \tag{49}$$

putting Eq. (48) back into Eqs. (35)–(38), we get:

Solutions B1–B8:

$$\boldsymbol{V}=\begin{bmatrix} s_{11}(2c_1\eta_1)^{-1/4} & s_{12}(2c_1\eta_2)^{-1/4} \\ s_{21}(2c_2\eta_1)^{-1/4} & s_{22}(2c_2\eta_2)^{-1/4} \end{bmatrix} \tag{50}$$

where

$$s_{ij}=1 \text{ or } -1 \quad \text{satisfying } s_{11}s_{12}s_{21}s_{22}=-1, \tag{51}$$

with totally eight combinations,

$$\eta_1=\mu_1^4+3\sqrt{\mu_1^4/\mu_2^4}\mu_1^2\mu_2^2, \tag{52}$$

$$\eta_2=\mu_2^4+3\sqrt{\mu_2^4/\mu_1^4}\mu_1^2\mu_2^2. \tag{53}$$

Solutions B1–B8 are the eight and only eight solutions in case B. However, for these solutions $\boldsymbol{V}\neq\boldsymbol{PD}$. They are spurious solutions that cannot perform source separation but still satisfy the four equilibrium equations.

The exhaustive determination of the equilibrium points can be summarized in the following lemma:

Lemma 19. *The cost function of information-theoretic ICA approach $J(\boldsymbol{V})$ with cubic nonlinearity Eq.* (7), *has exactly sixteen equilibrium points*, *Solutions* A1–A8 (*Eqs.* (43) *and* (44)) *and Solutions* B1–B8 (*Eq.* (50)). *Solutions* A1–A8 (*Group A*) *are correct solutions that can perform source separation*, *while Solutions* B1–B8 (*Group B*) *are spurious solutions that cannot perform source separation.*

Remark 20. Note the scale of the equilibrium points are controlled by $c_i^{-1/4}$. It can be easily seen that c_i is related to the scale parameter by $c_i^{-1/4} \propto \theta_i$. Hence the lemmas and corollary in Section 2.4 are verified.

3.2. Stability of the equilibrium points

The condition for stability of correct Solution Group A is easily found by substituting Eq. (7) into Lemma 18. The condition for stability of spurious Solution Group B is found in the Appendix. They are summarized in the following lemma:

Lemma 21. *For source signals satisfying the following condition:*

$$E[s_1^4]E[s_2^4] - [3(E[s_1^2])^2][3(E[s_2^2])^2] < 0. \tag{54}$$

Solutions A1–A8 are minima and Solutions B1–B8 are saddle points of $J(\boldsymbol{V})$. For source signals satisfying

$$E[s_1^4]E[s_2^4] - [3(E[s_1^2])^2][3(E[s_2^2])^2] > 0. \tag{55}$$

Solutions B1–B8 are minima and Solutions A1–A8 are saddle points of $J(\boldsymbol{V})$.

Remark 22. A signal s is called sub-Gaussian if the kurtosis $E[s^4] - 3(E[s^2])^2$ is negative, called super-Gaussian if the kurtosis is positive and called Gaussian if the kurtosis is zero. A super-Gaussian signal has sharply peaked pdf with long tail and a sub-Gaussian signal has a flat pdf with a short tail. The term $E[s_1^4]E[s_2^4] - [3(E[s_1^2])^2][3(E[s_2^2])^2]$ in Eqs. (54) and (55) is defined as the 'joint kurtosis' for the two signals. Two signals that satisfy Eq. (54), i.e., with negative joint kurtosis, are called 'globally sub-Gaussian' [10]. Similarly, two signals that satisfy Eq. (55), i.e. with positive joint kurtosis, are called 'globally super-Gaussian', and two signals with zero joint kurtosis are called 'globally Gaussian'.

3.3. Global convergence analysis

By Corollary 7, $\boldsymbol{V}$ will not diverge to infinity. By Lemmas 1, 3 and Corollary 6, the singular subspace divides the whole parameter space into several semi-infinite continuous regions. Hence, when the algorithm is running, $\boldsymbol{V}$ must converge to one of the local minima in the region it is initialized in. Therefore, we have the following theorem on the global convergence behaviour of the algorithm.

Theorem 23. *For the information-theoretic ICA algorithm, Eq.* (3), *with the cubic nonlinearity equation* (7) *acting on two channels of signals,* $\boldsymbol{W}$ *being initialized at some non-singular point,*

- *If the two source signals satisfy Eq.* (54), *then* $\boldsymbol{V}$ *will converge to one of the Solutions* $A1$–$A8$, *Eqs.* (43) *and* (44).
- *If the two source signals satisfy Eq.* (55), *then* $\boldsymbol{V}$ *will converge to one of the Solutions* $B1$–$B8$, *Eq.* (50).

In a word, Theorem 23 means that the information-theoretic ICA algorithms with cubic nonlinearity can separate two globally sub-Gaussian sources but cannot separate two globally super-Gaussian sources.

Remark 24. The cross-coupling equilibrium condition, Eqs. (37) and (38), are equivalent to controlling the fourth-order cross-moments $E[y_i^3 y_j]$, $i \neq j$ to be zero. Due to their simplicity, cubic nonlinearity and cancellation of these fourth-order cross-moments are popular in ICA approaches with different network structures [8,12,18,15,19,10,3,14,4]. From the theoretical analysis in this section, it can be seen that for algorithms with cubic nonlinearity or fourth-order cross-moments $E[y_i^3 y_j]$, $i \neq j$, the conditions on successful separation naturally involve fourth order statistics of the sources (e.g. the joint kurtosis). It is not surprising that the condition for successful separation for these approaches have similar forms [19,3] or even exactly the same form [18,4,10,14]. (However, this work is not equivalent to those previous work as mentioned in Section 1.)

4. Experimental demonstration

The experiments are aimed at demonstrating the theoretical results. The natural gradient descent algorithm, Eq. (3), with cubic nonlinearity, Eq. (7), is used. It is chosen that $c_1 = c_2 = 1$. For all experiments, the learning rate is kept at 0.0001. The following mixing matrix is used:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0.6 \\ 0.7 & 1 \end{bmatrix}. \tag{56}$$

The experiments are run for a number of scans through the data set long enough that $\boldsymbol{W}$ seems to have converged to a stable point.

The performance of the separation is determined by how close to $\boldsymbol{PD}$ the matrix $\boldsymbol{V} = \boldsymbol{WA}$ is. The element v_{ij} determines the amplitude of source signal s_j goes into recovered signal y_j and v_{ij}^2 determines the power. The greatest v_{ij}^2 in a row in $\boldsymbol{V}$ is regarded as the power of the 'signal' and the sum of other v_{ij}^2 of the row is regarded as the power of the 'interference'. Hence, we define the interference-to-signal power ratio

of channel i in decibel (dB) units as

$$I/S_i = 10 \times \log_{10}\left(\frac{\sum_{j \neq k} v_{ij}^2}{v_{ik}^2}\right), \qquad k = \arg\max_l v_{il}^2. \tag{57}$$

We use the mean of the interference-to-signal ratio over the channels as the performance index of source separation.

It should be noted that for convergence to the correct solution, the performance index depends on the mean magnitude of fluctuation around the solution, as a fixed learning rate is used. Since the magnitude of fluctuation decreases as the magnitude of the learning rate decreases, better performance (more negative performance index in dB) can be obtained by using a smaller learning rate (with slower convergence).

4.1. Experiments on two sub-Gaussian sources

In this experiment, two channels of artificially generated independently and identically distributed (iid) source signals with uniform distribution in $[-1, 1]$ are used. Each channel consists of 100,000 data points. Statistics of the data set are

$$\mu_1^2 = 0.3326, \qquad \mu_1^4 = 0.1991, \qquad \mu_2^2 = 0.3337, \qquad \mu_2^4 = 0.2007. \tag{58}$$

Both channels are sub-Gaussian, the standardized joint kurtosis $\mu_1^4\mu_2^4/[(\mu_1^2)^2(\mu_2^2)^2] - 9$ is -5.756 and they are obviously globally sub-Gaussian. We tried two initializations of $\boldsymbol{W}$. The first one is the identity matrix, which is a natural choice when no supplementary information is provided, and means $\boldsymbol{V}$ starts from the original mixture $\boldsymbol{A}$. The second initialization is at one of the spurious solutions B, $\boldsymbol{W}_{\text{init}} = \boldsymbol{V}_B\boldsymbol{A}^{-1}$, where

$$\boldsymbol{V}_B = \begin{bmatrix} -0.9851 & 0.9832 \\ 0.9851 & 0.9832 \end{bmatrix} \tag{59}$$

to test the stability of Solution Group B.

For the case $\boldsymbol{W}$ is initialized as an identity matrix, the system converges in about 30,000 data points. The performance graph, interference-to-signal ratio versus number of data points scanned, is plotted in Fig. 3. For the case initialization is at the Solution B, the system converges in about 200,000 data points. The four-dimensional trajectories of the convergence are plotted in two two-dimensional graphs, Figs. 4(a) and (b), each of which being the projection to two coordinates.

$\boldsymbol{V}$ in the two cases converges to two of the correct Solution A's:

$$\boldsymbol{V}_A = \begin{bmatrix} \pm 1.4970 & 0 \\ 0 & 1.4941 \end{bmatrix}. \tag{60}$$

The interference-to-signal ratio reaches -40 dB in both case.

Hence it is experimentally verified that Solution Group A is stable and Solution Group B is not stable in this case.

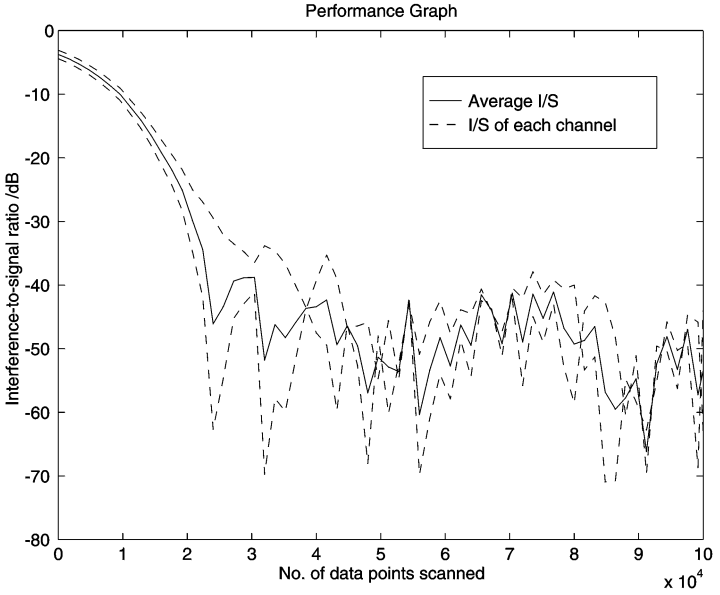


Fig. 3. The performance graph of the algorithm with cubic nonlinearity acting on uniformly distributed sources.

4.2. Experiments on two super-Gaussian sources

In this experiment, two channels of human speech signals are used. The first channel is recorded from a man telling a story and the second channel is recorded from a woman reading news. Both signals are recorded at 8 kHz and consist of 100,000 data points (12.5 s). The signals are randomly permuted to get rid of non-stationarity. Statistics of the data set are

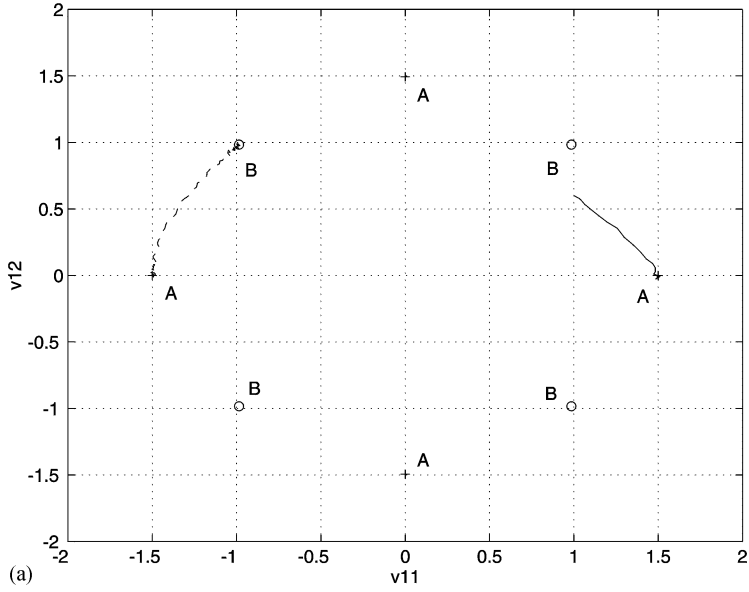
$$\mu_1^2 = 0.0625, \qquad \mu_1^4 = 0.0435, \qquad \mu_2^2 = 0.2500, \qquad \mu_2^4 = 0.3342. \tag{61}$$

Both signals are super-Gaussian. The standardized joint kurtosis $\mu_1^4\mu_2^4/[(\mu_1^2)^2(\mu_2^2)^2] - 9$ is 50.55 and obviously the source is globally super-Gaussian. We have also tried two initializations of $\boldsymbol{W}$. The first one is the identity matrix. The second initialization is at one of the correct Solution A's, $\boldsymbol{W}_{\text{init}} = \boldsymbol{V}_A\boldsymbol{A}^{-1}$, where

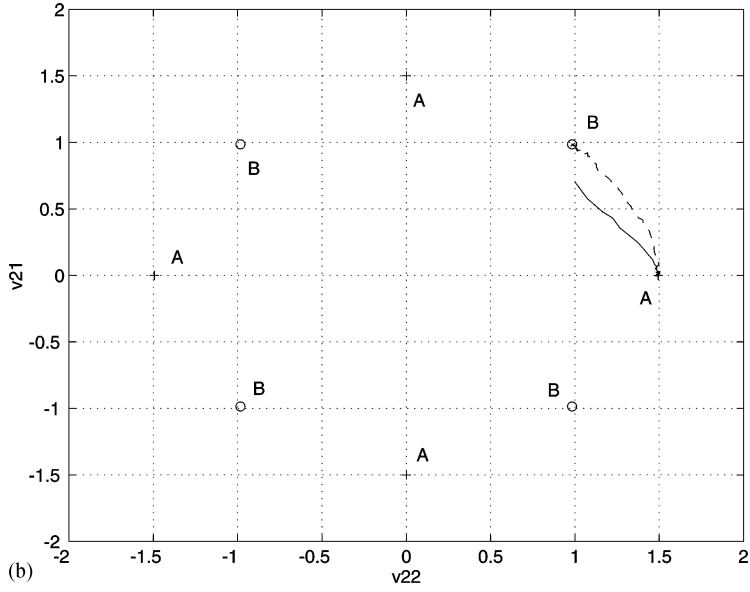
$$\boldsymbol{V}_A = \begin{bmatrix} 2.1897 & 0 \\ 0 & 1.3152 \end{bmatrix} \tag{62}$$

to test the stability of Solution Group A.

For the case $\boldsymbol{W}$ initialized as an identity matrix, the system converges in about 80,000 data points. For the case initialization is at the solution A, the system converges in about 160,000 data points. The four-dimensional trajectories of the convergence are plotted in two two-dimensional graphs, Figs. 5(a) and (b), each of which being the projection to two coordinates.



(a)



(b)

Fig. 4. The trajectories of convergence of $\boldsymbol{V}$ of the information-theoretic ICA algorithm with cubic nonlinearity on uniformly distributed sources. Solid: $\boldsymbol{W}_{\text{init}} = \boldsymbol{I}$. Dashed: $\boldsymbol{W}_{\text{init}} = \boldsymbol{V}_B \boldsymbol{A}^{-1}$. Solution A's and B's are marked by 'A' and 'B', respectively. The convergence points are solution A's.

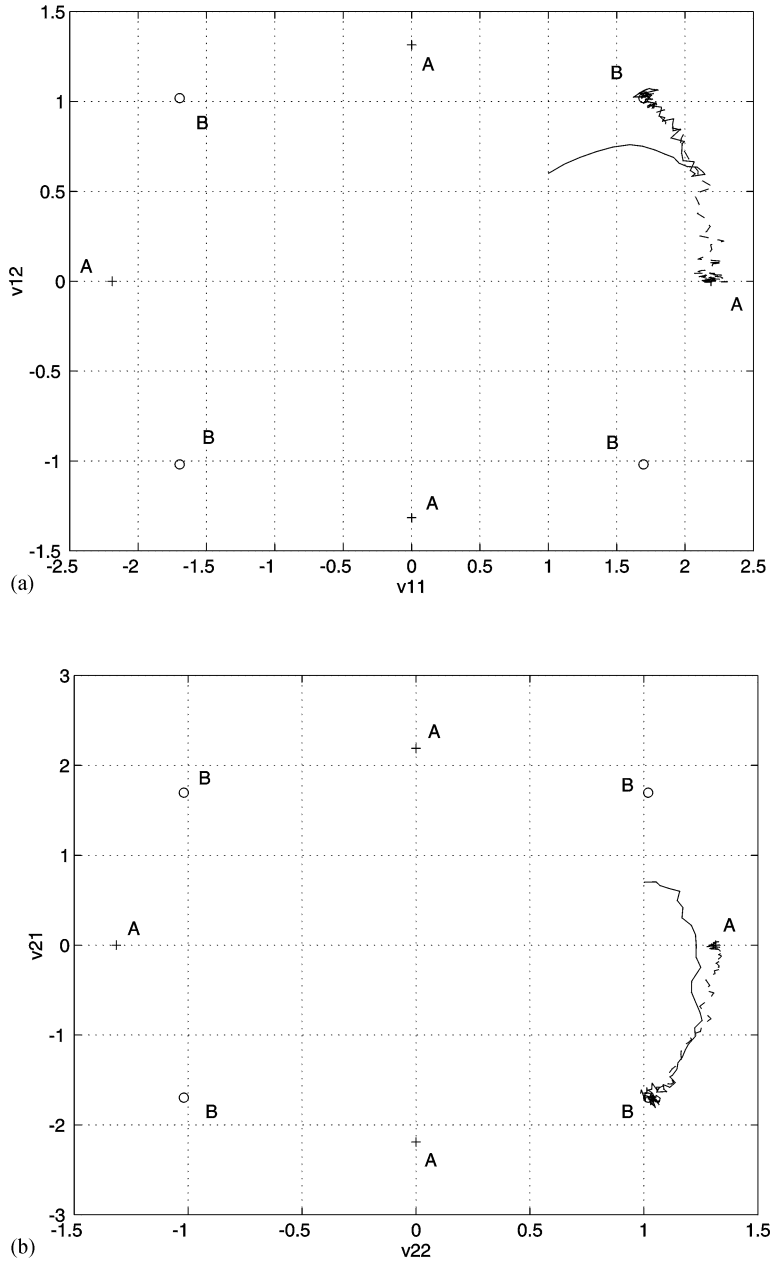


Fig. 5. The trajectories of convergence of $\boldsymbol{V}$ of the information-theoretic ICA algorithm with cubic nonlinearity on permuted speech signals. Solid: $\boldsymbol{W}_{\text{init}} = \boldsymbol{I}$. Dashed: $\boldsymbol{W}_{\text{init}} = \boldsymbol{V}_A \boldsymbol{A}^{-1}$. The convergence points are solution B.

V in both cases happens to converge to one of the spurious Solution B's:

$$V_B = \begin{bmatrix} 1.6964 & 1.0187 \\ -1.6964 & 1.0187 \end{bmatrix}. \tag{63}$$

Hence it is experimentally verified that Solution Group B is stable and Solution Group A is not stable in this case.

5. Conclusions

In addition to using standard techniques for determining equilibrium points and their stability for proving local convergence, we put a great deal of effort in investigating the global behaviour of the cost function in the whole parameter space. We successfully obtained theoretical results on global properties of the cost function of the information-theoretic ICA approach with any differentiable, odd, monotonic decreasing nonlinearity in the general n-channel case. In the simple 2-channel cubic nonlinearity case, we solve all equilibrium points with the condition on their stability found, and give a global convergence theorem. The global convergence theorem is verified by computer simulation. This theoretical work provides a solid foundation to investigate how nonlinearity affects separation capability in ICA algorithms [21–23].

Acknowledgements

We would like to thank Prof. Ruan Jiong for the initial step in solving the equilibrium equations in the 2-channel cubic nonlinearity case, and Philip Fu for suggesting the mathematical technique for the refinement of the result in asymptotic behaviour of the cost function from its original version [5]. We also thank the anonymous reviewers for their valuable comments.

Appendix. Stability of solution group B

This part is to find out the condition for Solutions B1–B8 to be local minima of $J(V)$ in the 2-channel cubic nonlinearity case.

For *Solution B1–B8*,

$$(\det V)^2 = D^2 = \frac{2}{c\eta}, \tag{A.1}$$

where

$$\begin{aligned} \eta &= \sqrt{\eta_1\eta_2} = \sqrt{\mu_1^4\mu_2^4} + m > 0, \\ c &= \sqrt{c_1c_2} > 0, \end{aligned} \tag{A.2}$$

and η_1, η_2 and m are defined in Section 3.1.

From Eq. (15), the Hessian matrix is

$$\nabla_v^2 J(V) = Q = \frac{1}{2\sqrt{2}} \begin{bmatrix} \sqrt{c_1}p_1 & -s_{21}s_{22}\sqrt{c_1}(\sqrt{n}+\frac{4m}{\sqrt{n}}) & -s_{12}s_{22}\sqrt{c}\sqrt{\eta_1} & s_{12}s_{21}\sqrt{c}\sqrt{n} \\ -s_{21}s_{22}\sqrt{c_1}(\sqrt{n}+\frac{4m}{\sqrt{n}}) & \sqrt{c_1}p_2 & s_{11}s_{22}\sqrt{c}\sqrt{\eta} & -s_{11}s_{21}\sqrt{c}\sqrt{\eta_2} \\ -s_{12}s_{22}\sqrt{c}\sqrt{\eta_1} & s_{11}s_{22}\sqrt{c}\sqrt{\eta} & \sqrt{c_2}p_1 & -s_{11}s_{12}\sqrt{c_2}(\sqrt{\eta}+\frac{4m}{\sqrt{\eta}}) \\ s_{12}s_{21}\sqrt{c}\sqrt{\eta} & -s_{11}s_{21}\sqrt{c}\sqrt{\eta_2} & -s_{11}s_{12}\sqrt{c_2}(\sqrt{\eta}+\frac{4m}{\sqrt{\eta}}) & \sqrt{c_2}p_2 \end{bmatrix}, \tag{A.3}$$

where

$$p_1 = \sqrt{\eta_1} + \frac{6\mu_1^4}{\sqrt{\eta_1}} + \frac{2m}{\sqrt{\eta_2}},$$
$$p_2 = \sqrt{\eta_2} + \frac{6\mu_2^4}{\sqrt{\eta_2}} + \frac{2m}{\sqrt{\eta_1}}. \tag{A.4}$$

After a great deal of tedious symbolic manipulation [5], the characteristic polynomial is found to be

$$\begin{aligned} \det(\boldsymbol{Q} - \lambda \boldsymbol{I}) &= \lambda^4 + A_3\lambda^3 + A_2\lambda^2 + A_1\lambda + A_0 \\ &= \lambda^4 \\ &\quad + \lambda^3\left\{ -\frac{1}{2\sqrt{2}}(\sqrt{c_1}+\sqrt{c_2})\frac{7\sqrt{\mu_1^4\mu_2^4}+3m}{\sqrt{\eta(\eta-m)}}(\sqrt{\mu_1^4}+\sqrt{\mu_2^4})\right\} \\ &\quad + \lambda^2\left\{ c\frac{(2\sqrt{\mu_1^4\mu_2^4}+m)(3\sqrt{\mu_1^4\mu_2^4}+m)}{\eta(\eta-m)}(\sqrt{\mu_1^4}+\sqrt{\mu_2^4})^2 \right. \\ &\quad \left. + 2(c_1+c_2)(3\sqrt{\mu_1^4\mu_2^4}-m)\right\} \\ &\quad + \lambda\left\{ -2\sqrt{2}c(\sqrt{c_1}+\sqrt{c_2})\frac{5\mu_1^4\mu_2^4-m^2}{\sqrt{\eta(\eta-m)}}(\sqrt{\mu_1^4}+\sqrt{\mu_2^4})\right\} \\ &\quad + 32c^2\sqrt{\mu_1^4\mu_1^4}(\sqrt{\mu_1^4\mu_1^4}-m). \end{aligned} \tag{A.5}$$

For globally super-Gaussian signals, $\mu_1^4\mu_2^4 - m^2 > 0$, we have $A_3 < 0$, $A_2 > 0$, $A_1 < 0$ and $A_0 > 0$. Hence, every eigenvalue, being the root of the characteristic polynomial, must be positive. The Hessian matrix will be positive definite and therefore Solutions B1–B8 are local minima.

For globally sub-Gaussian signals, $\mu_1^4\mu_2^4 - m^2 < 0$, we have $A_0 < 0$. Noting that $A_0 = \lambda_1\lambda_2\lambda_3\lambda_4$, where λ_1, λ_2, λ_3, λ_4 are the eigenvalues, we must have three eigenvalues of the same sign and the remaining eigenvalues must be of the opposite sign. Hence Solutions B1–B8 are saddle points. □

References

[1] S.-I. Amari, A. Cichocki, H. Yang, A new learning algorithm for blind separation of sources, in: D.S. Touretzky, M.C. Mozer, d Michael E. Hasselmo (Eds.,) Advances in Neural Information Processing 8, MIT Press, Cambridge, MA, 1996, pp. 757–763.

[2] A.J. Bell, T.J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Comput. 7 (1995) 1129–1159.

[3] J.F. Cardoso, B. Laheld, Equivariant adaptive source separation, IEEE Trans. Signal Process. 44 (12) (1996) 3017–3030.

[4] T. Chen, R. Chen, A neural network approach to blind identification of stochastic and deterministic signals, in: Proceedings of the 28th Asilomar Conference on Signals, Systems & Computers, 1994, pp. 892–896.

[5] C.C. Cheung, Adaptive blind signal separation, Master's Thesis, Department of Computer Science & Engineering, The Chinese University of Hong Kong, Hong Kong, People's Republic of China, 1997.

[7] P. Comon, Independent component analysis – a new concept?, Signal Processing 36 (1994) 287–314.

[8] P. Comon, C. Jutten, J. Hérault, Blind separation of sources, Part II: problems statement, Signal Processing 24 (1) (1991) 11–20.

[9] N. Delfosse, P. Loubaton, Adaptive blind separation of independent sources: a deflation approach, Signal Processing 45 (1995) 59–83.

[10] Y. Deville, L. Andry, Application of blind source separation techniques to multi-tag contactless identification systems, in: Proceedings of the International Symposium on Nonlinear Theory and its Applications (NOLTA 95), Las Vegas, USA, December 10–14, 1995, pp. 73–78.

[11] A. Hyvärinen, E. Oja, A fast fixed-point algorithm for independent component analysis, Neural Comput. 9 (1997) 1483–1492.

[12] C. Jutten, J. Hérault, Blind separation of sources, Part I: an adaptive algorithm based on neuromimetic architecture, Signal Processing 24 (1) (1991) 1–10.

[13] J. Karhunen, Neural approaches to independent component analysis and source separation, invited paper, Proceedings of the 4th European Symposium on Artificial Neural Networks (ESANN'96), Bruges, Belgium, April 24–26, 1996.

[14] O. Macchi, E. Moreau, Self-adaptive source separation, Part I: convergence analysis of a direct linear network controlled by the Herault–Jutten algorithm, IEEE Trans. Signal Process. 45 (1997) 918–926.

[15] E. Moreau, O. Macchi, New self-adaptive algorithms for source separation based on contrast functions, in: Proceedings of the IEEE Signal Processing Workshop on Higher Order Statistics, 1993, pp. 215–219.

[16] E. Moreau, O. Macchi, Self-adaptive source separation – Part II: comparison of the direct, feedback, and mixed linear network, IEEE Trans. Signal Process. 46 (1998) 39–50.

[17] D.T. Pham, P. Garat, C. Jutten, Separation of a mixture of independent sources through a maximum likelihood approach, in: J. Vandewalle, R. Boite, M. Moonen, A. Oosterlinck (Eds.), Signal Processing VI: Theories and Applications, Elsevier, Amsterdam, 1992, pp. 771–774.

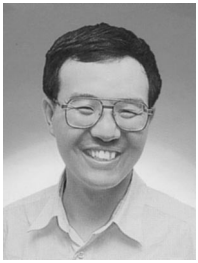
[18] E. Sorouchyari, Blind separation of sources, Part III: stability analysis, Signal Processing 24 (1) (1991) 21–29.

[19] L. Wang, J. Karhunen, E. Oja, A bigradient optimization approach for robust PCA, MCA, and source separation, in: Proceedings of the IEEE International Conference on Neural Networks, Perth, Australia, November 1995, pp. 1684–1689.

[20] L. Xu, S.-I. Amari, A general independent component analysis framework based on Bayesian-Kullback Ying-Yang Learning, in: Progress in Neural Information Processing: Proceedings of the International Conference on Neural Information Processing (ICONIP 96), Hong Kong, September 24–27, 1996; Springer-Verlag: Singapore, 1996, pp. 1235–1239.

[21] L. Xu, C.C. Cheung, S.-I. Amari, Learned parametric mixture based ICA algorithm, Neurocomputing 22 (1998) 69–80.

[22] L. Xu, C.C. Cheung, J. Ruan, S.-I. Amari, Nonlinearity and separation capability: Further justification for the ICA algorithm with mixture of densities, invited special session on Blind Signal Separation, in Proceedings of the 5th European Symposium on Artificial Neural Networks (ESANN 97), Bruges, Belgium, April 16–18, 1997, pp. 291–296.
[23] L. Xu, C.C. Cheung, H.H. Yang, S.-I. Amari, Independent component analysis by the information-theoretic approach with mixture of densities, in: Proceedings of the 1997 International Conference on Neural Networks (ICNN 97), Houston, USA, June 9–12, 1997, pp. 1821–1826.
[24] H.H. Yang, S.-I. Amari, Adaptive online learning algorithms for blind separation: Maximum entropy and minimum mutual information, Neural Comput. 9 (1997) 1457–1482.



Lei Xu (Ph.D., IEEE Senior member) is currently a professor in the Department of Computer Science and Engineering at the Chinese University of Hong Kong where he joined in 1993 as a senior lecturer first and then took the current position in 1996. He is also a professor at Peking University since 1992, where he started as a postdoc in the Department of Maths in 1987 and then became one of the ten exceptionally promoted young associate professors of Peking Univ in 1988. During 1989–1993, he worked as a postdoc or senior research associate at several universities in Finland, Canada and USA, including Harvard and MIT. He is a past president of Asian-Pacific Neural Networks Assembly, an associate editor for six renowned international academic journals on neurocomputing, including Neural Networks, IEEE Trans. on Neural Networks. He has published over 180 academic papers; given over ten keynote/invited/tutorial talks as well as served as their program committee member and session cochairs in international major Neural Networks conferences in recent years, including WCNN, IEEE-ICNN, ENNS-ICANN, ICONIP, IJCNN, NNCM. Also, he was a program committee chair of ICONIP'96 and a general chair of IDEAL'98. He has received several prestigious Chinese national academic awards (including National Nature Science Award and State Education Council FOK YING TUNG Award) and also some international awards, and is listed in several major Who's Who and the First Five Hundreds publications by CIBC, ABI and Marquis Who's Who.

Chi Chiu Cheung received his B.Sc. (Hons.) in Physics and received his M.Phil. in Computer Science and Engineering, both from the Chinese University of Hong Kong in 1995 and 1997, respectively. He received the 1997 Outstanding M.Phil. Award of Faculty of Engineering, the Chinese University of Hong Kong. He is currently a Ph.D. student of the Department of Computer Science and Engineering at the Chinese University of Hong Kong.