A NEW INFORMATION-THEORETIC BASED ICA ALGORITHM FOR BLIND SIGNAL SEPARATION

Y.-M. Cheung* and L. Xu**

Abstract

The typical information-theoretic approaches such as INFOMAX and MMI perform independent component analysis (ICA) by using a fixed nonlinearity function. Consequently, they can only separate either sub-Gaussian or super-Gaussian source signals, but not both. This article considers a flexible nonlinearity function that is a single polynomial term with the exponent learnable. The separation ability of this function is analysed, and a new ICA algorithm is proposed. The experiments have shown that this algorithm can successfully separate the mixture of sub-Gaussian and super-Gaussian sources.

Key Words

Independent component analysis, blind signal separation, flexible nonlinearity function, sub-Gaussian and super-Gaussian sources

1. Introduction

Due to attractive application on wireless communication, speech recognition, time series analysis, and the like, blind signal separation formulated as the independent component analysis (ICA) problem has recently received much attention. Here, we consider the instantaneous invertible linear mixture ICA problem. That is, each observed signal $\mathbf{x}$ with $\mathbf{x} = [x_1, \ldots, x_d]^T$ is a mixture of k hidden independently and identically distributed (i.i.d.) source signals (simply called *sources*) $s_1, s_2, \ldots, s_k$ through:

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

where $\mathbf{A}$ is an unknown full-column $d \times k$ mixing matrix and $\mathbf{s} = [s_1, s_2, \ldots, s_k]^T$. Hereafter, we assume that each source mean is zero without loss of generality. Otherwise, we can always make a transformation from (1) such that the assumption is held.

The objective of an ICA approach is to recover $\mathbf{s}$ from $\mathbf{x}$ up to an unknown constant scale and any permutation of indices by finding out an appropriate de-mixing matrix $\mathbf{W}$ such that:

$$\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{s} = \mathbf{P}\Lambda\mathbf{s} \tag{2}$$

where $\mathbf{P}$ is a permutation matrix and Λ is a diagonal matrix. It should be noted that the statistical independence among several Gaussian sources is invariant to the rotation transformation of them. Hence, ICA generally can separate non-Gaussian sources only with at most one Gaussian source, due to the fact that it exclusively uses the information of independence among the sources.

In the past, many information-theoretic-based ICA approaches have been proposed, such as INFOMAX [1, 3] and MMI [2]. In these algorithms, the nonlinearity functions are pre-assigned without any adjustability to match a variety of source signals in implementation, which may violate the ICA separating conditions as shown in [4]. A lot of experiments have shown that these algorithms with fixed nonlinearities can separate either sub-Gaussian or super-Gaussian signals, but not both. To circumvent this drawback, some improvements have been made. For example, the use of a mixture of logistic densities for modelling the marginal density has been made by Pearlmutter and Parra [5] under the name of maximum likelihood density estimation. Also, an approach called *learned parametric mixture-based ICA algorithm* (LPM) has been proposed in [6, 7], where the nonlinearity function is modelled by a parametric mixture of densities, and this mixture is learned together with the learning of $\mathbf{W}$. As the density parameters are adaptively learned, the nonlinearity function can be well chosen automatically such that the LPM successfully separates a combination of sub-Gaussian and super-Gaussian signals. In general, these mixture-of-density-based methods often generate a set of new extra parameters that need to be learned as well as $\mathbf{W}$. As a result, the increased algorithm complexity not only needs considerable extra efforts in implementation, but also makes the analysis of their separation abilities more difficult.

This article considers an alternative flexible nonlinearity function that is a single polynomial term with the exponent learned together with the de-mixing matrix $\mathbf{W}$.

* Department of Computer Science, Hong Kong Baptist University, Rm. 709, 7/F, Sir Run Run Shaw Building, Kowloon Tong, KLN, Hong Kong, PRC; e-mail: ymc@comp.hkbu.edu.hk

** Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong kong, PRC; e-mail: lxu@cse.cuhk.edu.hk

(paper no. 202-1198)

Not only is the separation ability of this function therefore analysed, but a new ICA algorithm is also proposed. Compared to INFOMAX and MMI approaches, this new algorithm has only one extra parameter to be learned. Hence, little extra effort is needed in implementation. The experiments have shown that this algorithm can successfully separate any combination of sub-Gaussian and super-Gaussian sources we have tried so far.

2. Terminology

This section outlines the terminology used in this work. Here, we consider the variables to be real-valued only. The definitions are similar in discrete cases so long as we replace the integral $\int$ by the summation $\sum$, and use probability function to instead of probability density function (pdf).

Statistical independence. Given a set of random variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, they are statistically independent if and only if:

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = \prod_{t=1}^{N} p(\mathbf{x}_i) \quad . \tag{3}$$

Note that $E(\mathbf{xy}) = E(\mathbf{x})E(\mathbf{y})$ if $\mathbf{x}$ and $\mathbf{y}$ are statistically independent. The reverse, however, does not generally hold.

Super-Gaussian and sub-Gaussian signals. A pdf $p(x)$ is said to be super-Gaussian if $\exists x_0 \in \Re^+$ such that:

$$\tilde{p}(x) < g(x) \tag{4}$$

for $\forall x \geq x_0$, where $\tilde{p}(x)$ is the normalized pdf of $p(x)$, that is, the pdf with zero mean and variance 1, and $g(x)$ is the standard Gaussian pdf. Similarly, a pdf $p(x)$ is said to be sub-Gaussian if $\exists x_0 \in \Re^+$ such that:

$$\tilde{p}(x) > g(x) \tag{5}$$

for $\forall x \geq x_0$. The signal with the super-Gaussian or sub-Gaussian distribution is called super-Gaussian signal or sub-Gaussian signal, respectively. Compared to the Gaussian signal, the shape of unimodal super-Gaussian signal is sharper near the mean point, whereas a unimodal sub-Gaussian signal is more even.

Kullback-divergence. The Kullback-divergence, also called *relative entropy* or *cross-entropy*, is a measure of the "distance" between two distributions $p(\mathbf{x})$ and $q(\mathbf{x})$. It is defined as:

$$KL(p(\mathbf{x}), q(\mathbf{x})) = \int p(\mathbf{x}) \ln \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} \tag{6}$$

The Kullback-divergence has three major properties:

1. $KL(p(\mathbf{x}), q(\mathbf{x})) \geq 0$ for $\forall p(\mathbf{x}), q(\mathbf{x})$. The "=" holds if and only if $p(\mathbf{x}) = q(\mathbf{x})$.
2. In general, $KL(p(\mathbf{x}), q(\mathbf{x})) \neq KL(q(\mathbf{x}), p(\mathbf{x}))$.
3. The Kullback-divergence is invariant to a linear transformation. That is, let $\mathbf{y} = \mathbf{Ax} + \mathbf{b}$; we have:

$$KL(p(\mathbf{y}), q(\mathbf{y})) = KL(p(\mathbf{x}), q(\mathbf{x})) \tag{7}$$

Rotation transformation. Suppose $\mathbf{R}$ is a $k \times k$ orthogonal matrix. Given a vector $\mathbf{x} \in \Re^k$, the rotation transformation of $\mathbf{x}$ is $\mathbf{R} \times \mathbf{x}$.

3. Review of Information-Theoretic ICA Approaches

The information-theoretic-based approaches obtain the de-mixing matrix $\mathbf{W}$ in (2) by minimizing the following Kullback-divergence function:

$$J(\mathbf{W}) = \int p(\mathbf{y}) \ln \frac{p(\mathbf{y})}{\prod_{j=1}^{k} g_j(y_j)} d\mathbf{y} \tag{8}$$

where $\mathbf{y} = \mathbf{Wx}$, and $g_j(y_j)$, $j = 1, \ldots, k$ is a marginal pdf of y_j. With some mathematical computation, the minimization of $J(\mathbf{W})$ is equivalent to maximize the function:

$$L(\mathbf{W}) = \ln |det(\mathbf{W})| + \sum_{j=1}^{k} \ln g_j(y_j) \tag{9}$$

where $det(\mathbf{W})$ denotes the determinant of $\mathbf{W}$.

It has been shown in [1] that $\mathbf{W}$ can be adaptively learned by the gradient descent method:

$$\mathbf{W}^{new} = \mathbf{W}^{old} + \eta[\mathbf{W}^{-T} + \boldsymbol{\phi}(\mathbf{y})\mathbf{x}^T] \tag{10}$$

or by the improved natural gradient descent [2]:

$$\mathbf{W}^{new} = \mathbf{W}^{old} + \eta[\mathbf{I} + \boldsymbol{\phi}(\mathbf{y})\mathbf{y}^T]\mathbf{W} \tag{11}$$

where η is a small positive learning rate, $\mathbf{I}$ is the $k \times k$ identity matrix, and $\boldsymbol{\phi}(\mathbf{y}) = [\phi_1(y_1), \phi_2(y_2), \ldots, \phi_k(y_k)]^T$ with $\phi_j(y_j) = (\partial \ln g_j(y_j)/\partial y_i)$ is the nonlinearity function.

Most of the learning rules in different information-theoretic-based algorithms can be unified as (11). The major distinction of these algorithms is the selection of different nonlinearity functions $\boldsymbol{\phi}(\mathbf{y})$. For example, INFOMAX [1] assigns $g_j(y_j)$ to be a sigmoid function, MMI [2] approximates $g_j(y_j)$ by Gram-Charlier expansion, and LPM [6, 7] uses an adaptively learned mixture of densities to estimate $g_j(y_j)$.

4. Polynomial Nonlinearity and Separation Ability

For a general form of nonlinearity functions, [4] has given the following results:

Theorem 1. The separating solution is a stable equilibrium of learning equation (11) if and only

$$m_i + 1 > 0 \tag{12}$$

$$k_i > 0 \tag{13}$$

$$\sigma_i^2 \sigma_j^2 k_i k_j > 1 \tag{14}$$

for all i, j $(i \neq j)$ under the normalization condition that $\forall r, E[\phi(y_r)y_r] = 1$, where $\sigma_i^2 = E[y_i^2]$, $k_i = E[\dot{\phi}_i(y_i)]$, $m_i = E[y_i^2 \dot{\phi}_i(y_i)]$, and $\dot{\phi}_i(y_i) = d\phi_i(y_i)/dy_i$.

In particular, when the nonlinearity function in (11) is the form:

$$\phi_j(y_j) = -sign(y_j)|y_j|^p \quad 1 \le j \le k \tag{15}$$

with $p > 0$, the conditions in Theorem 1 then become:

$$\left(\frac{E[|y_i|^{p+1}]}{E[|y_i|^2]E[|y_i|^{p-1}]}\right)\left(\frac{E[|y_j|^{p+1}]}{E[|y_j|^2]E[|y_j|^{p-1}]}\right) < p^2 \tag{16}$$

Here, we further generalize (15) to be:

$$\phi_j(y_j) = -sign(y_j)|y_j|^{p_j} \quad 1 \le j \le k \tag{17}$$

with $p_j > 0$; we then obtain the following result.

Corollary 1. Given a nonlinearity function $\phi(\mathbf{y})$ as shown in (17), if and only if:

$$\left(\frac{E[|y_i|^{p_i+1}]}{E[|y_i|^2]E[|y_i|^{p_i-1}]}\right)\left(\frac{E[|y_j|^{p_j+1}]}{E[|y_j|^2]E[|y_j|^{p_j-1}]}\right) < p_i p_j \tag{18}$$

for all i, j $(i \neq j)$, there must exist at least one stable separating solution that the converged $\mathbf{W}$ in (11) will successfully separate k sources $s_1, s_2, \ldots, s_k$.

The mathematical proof is given in Appendix A. Hence, for each source s_j, if there exists a positive exponent p_j such that:

$$\frac{E[|s_j|^{p_j+1}]}{E[|s_j|^2]E[|s_j|^{p_j-1}]} < p_j \tag{19}$$

there is at least one stable separating solution when $\mathbf{W}$ is learned by (11). In the following, we assume that the exponent set $\mathbf{P} = \{p_1, p_2, \ldots, p_k\}$ with p_j satisfying (19) always exists, which is also actually true for most sub-Gaussian and super-Gaussian sources. For example, when s_j is a unimodal sub-Gaussian source, we have $p_j = 3$, which satisfies (19).

5. New Approach: Adaptive Polynomial Power Learning Estimation-Based ICA Algorithm (APPLE-ICA)

As $\phi_j(y_j) = -sign(y_j)|y_j|^{p_j}$, the cost function in (9) becomes:

$$\begin{aligned} L(\mathbf{W}) &= Q(\mathbf{W},\ \mathbf{P}) \\ &= \ln|det(\mathbf{W})| + \sum_{j=1}^{k} \ln g_j(y_j) \\ &= \ln|det(\mathbf{W})| - \sum_{j=1}^{k} \frac{1}{p_j+1}|y_j|^{p_j+1} + C \end{aligned} \tag{20}$$

where C is a constant term. Hence, we can adaptively learn the parameter $\mathbf{P}$ as well as $\mathbf{W}$ by maximizing the cost function $Q(\mathbf{W}, \mathbf{P})$. To ensure each $p_j > 0$, we further let $p_j = \lambda e^{u_j}$, $1 \le j \le k$, where λ is a positive constant. Subsequently, we learn u_j's instead. The detailed APPLE-ICA algorithm is as follows:

Step 1. Initialize $\mathbf{W}$ and a parameter $\mathbf{U} = [u_1, u_2, \ldots, uk]^T$ with $1 \le j \le k$.

Step 2. Given an observed signal $\mathbf{x}$, let:

$$\begin{aligned} \mathbf{y} &= \mathbf{W}\mathbf{x} \\ p_j &= \lambda e^{u_j} \quad 1 \le j \le k \\ \phi_j(y_j) &= -sign(y_j)|y_j|^{p_j}. \end{aligned} \tag{21}$$

Step 3. Update $\mathbf{W}$ and $\mathbf{U}$ by:

$$\begin{aligned} \mathbf{W}^{new} &= \mathbf{W}^{old} + \eta[\mathbf{I} + \phi(\mathbf{y})\mathbf{y}^T]\mathbf{W} \\ u_j^{new} &= u_j^{old} + \eta\frac{\partial Q(\mathbf{W},\mathbf{P})}{\partial u_j} \end{aligned} \tag{22}$$

with:

$$\frac{\partial Q(\mathbf{W},\mathbf{P})}{\partial u_j} = \frac{p_j|y_j|^{p_j+1}}{p_j+1}\left(\frac{1}{p_j+1} - \ln|y_j|\right). \tag{23}$$

Steps 2 and 3 are repeated until both $\mathbf{W}$ and $\mathbf{P}$ converge.

6. Simulation Results

To investigate the performance of APPLE-ICA algorithm on all combinations of sub-Gaussian and super-Gaussian source signals, we consider $k = 3$, which results in four possible source combinations:

Combination 1. Three sub-Gaussian sources.

Combination 2. Two sub-Gaussian and one super-Gaussian sources.

Combination 3. One sub-Gaussian and two super-Gaussian sources.

Combination 4. Three super-Gaussian sources.

Therefore, we perform four experiments hereafter. In Experiment i with $1 \le i \le 4$, we use three source signals whose types are specified in Combination i.

In all experiments, we let sub-Gaussian sources be uniformly distributed and super-Gaussian ones be human speech signals. Furthermore, we set the true mixing matrix:

$$\mathbf{A} = \begin{pmatrix} 1.0 & 0.6 & 0.8 \\ 0.7 & 1.0 & 0.4 \\ 0.3 & 0.7 & 1.0 \end{pmatrix} \tag{24}$$

and fix the learning rate $\eta = 0.0001$, and $\lambda = 1.5$. In addition, the de-mixing matrix $\mathbf{W}$ was randomly initialized in each simulation run.

We measure the performance of APPLE-ICA algorithm by signal-to-noise ratio (SNR) defined by:

$$SNR(s_j, y_j) = 10\log_{10}\frac{\sigma_{s_j}^2}{MSE(s_j, y_j)} \quad 1 \le j \le k \tag{25}$$

where $\sigma_{s_j}^2$ is the variance of source signal s_j, and $MSE(s_j, y_j)$ is the mean square error between source signal s_j and its recovered signal y_j. As $\sigma_{s_j}^2$ is irrelevant to the algorithm performance, we can further ignore it and use a simplified SNR with:

$$SNR(s_j, y_j) = -10\log_{10} MSE(s_j, y_j) \quad 1 \le j \le k \tag{26}$$

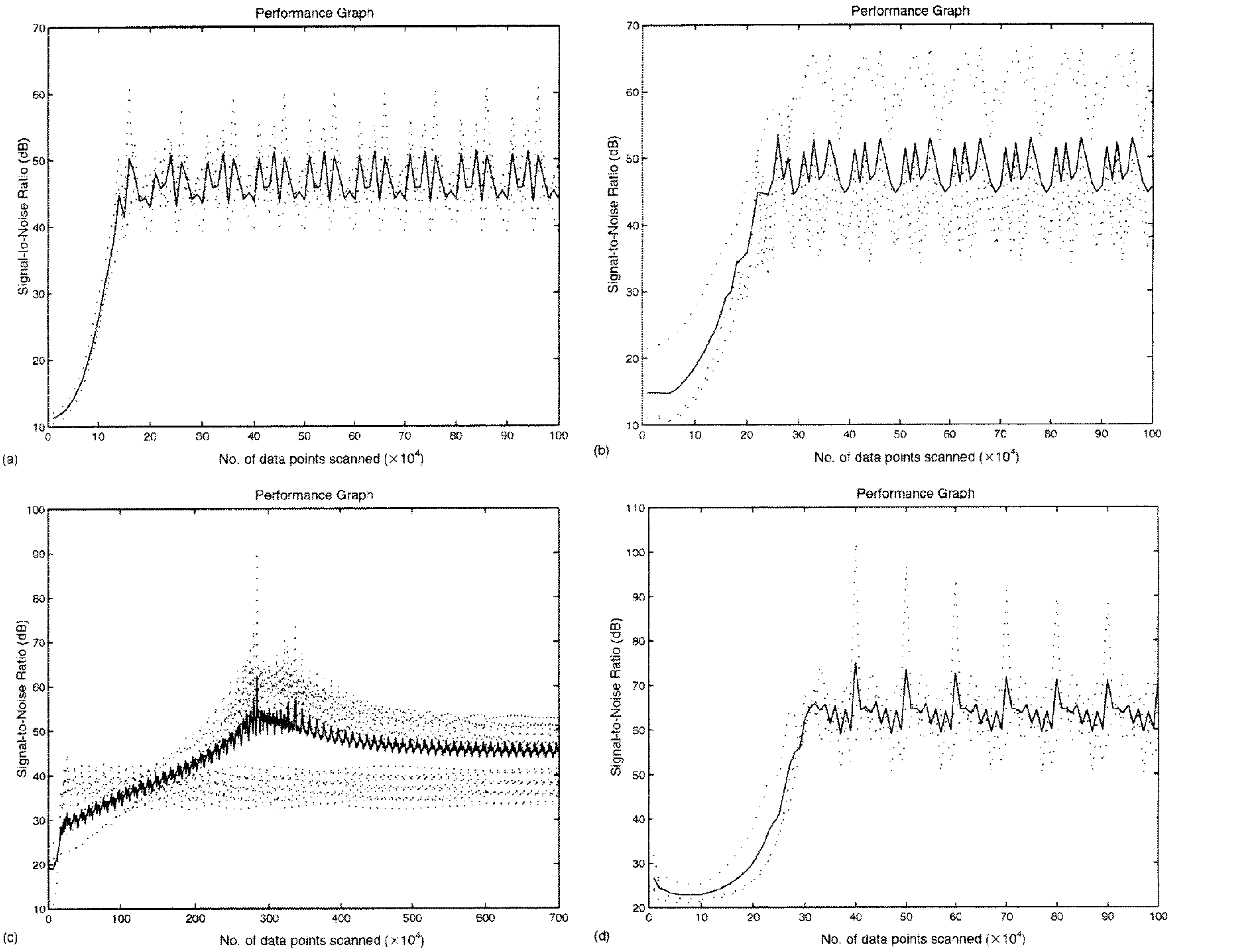


Figure 1. The SNR performance graphs of APPLE-ICA algorithm in: (a) Experiment 1; (b) Experiment 2; (c) Experiment 3 and (d) Experiment 4, where the dotted line is the SNR of an individual signal and the real line is the averaged SNR that is calculated by $(1/k)\sum_{j=1}^{k} SNR(s_j,\, y_j)$.

Table 1
Snapshot Values of $\mathbf{W}$ and $\mathbf{W} \times \mathbf{A}$ after APPLE-ICA Algorithm Converges

Experiment 1	$\mathbf{W} = \begin{pmatrix} 0.3919 & -0.0273 & -0.2999 \\ -0.3167 & 0.4129 & 0.0881 \\ 0.1021 & -0.2784 & 0.3131 \end{pmatrix}$ $\mathbf{W} \times \mathbf{A} = \begin{pmatrix} \mathbf{0.2828} & -0.0021 & 0.0027 \\ -0.0013 & \mathbf{0.2845} & -0.0001 \\ 0.0012 & 0.0020 & \mathbf{0.2834} \end{pmatrix}$	Experiment 2	$\mathbf{W} = \begin{pmatrix} 0.3877 & -0.0254 & -0.2965 \\ -0.3136 & 0.4132 & 0.0816 \\ 0.1947 & -0.5337 & 0.5915 \end{pmatrix}$ $\mathbf{W} \times \mathbf{A} = \begin{pmatrix} \mathbf{0.2810} & -0.0003 & 0.0036 \\ 0.0001 & \mathbf{0.2821} & -0.0041 \\ -0.0014 & -0.0028 & \mathbf{0.5338} \end{pmatrix}$
Experiment 3	$\mathbf{W} = \begin{pmatrix} 0.3850 & -0.0235 & -0.2933 \\ -0.5646 & 0.7217 & 0.2019 \\ 0.4090 & -1.0798 & 1.1741 \end{pmatrix}$ $\mathbf{W} \times \mathbf{A} = \begin{pmatrix} \mathbf{0.2805} & 0.0021 & 0.0052 \\ 0.0011 & \mathbf{0.5242} & 0.0389 \\ 0.0054 & -0.0125 & \mathbf{1.0694} \end{pmatrix}$	Experiment 4	$\mathbf{W} = \begin{pmatrix} 8.9010 & -0.4935 & -6.9232 \\ -2.6076 & 3.4556 & 0.7083 \\ 0.3776 & -1.0075 & 1.1299 \end{pmatrix}$ $\mathbf{W} \times \mathbf{A} = \begin{pmatrix} \mathbf{6.4786} & 0.0009 & 0.0002 \\ 0.0238 & \mathbf{2.3869} & 0.0045 \\ 0.0113 & 0.0100 & \mathbf{1.0290} \end{pmatrix}$

To make the SNR computation invariant to scaling, we normalize both the sources and its recovered counterpart into the range $[-1, 1]$.

The performance of APPLE-ICA algorithm in different cases is shown in Fig. 1, where we obtained the averaged SNR about 44.0 dB, 45.7 dB, 46.1 dB, and 70.9 dB in four experiments, respectively. Table 1 lists a snapshot of converged $\mathbf{W}$ as well as $\mathbf{W} \times \mathbf{A}$ in all experiments, where we found that the significant elements in $\mathbf{W} \times \mathbf{A}$ are just the diagonal ones in all cases. That is, the converged $\mathbf{W}$ has made $\mathbf{W} \times \mathbf{A}$ be a diagonal matrix. Hence, from ICA model in (2), we know that the wave forms of unknown sources $\mathbf{s}$ have been successfully recovered in all cases we have tried so far.

7. Conclusion

From the information-theoretic framework, we have presented an alternative ICA algorithm, which uses a flexible polynomial nonlinearity function with its exponent adaptively learned, as well as the de-mixing matrix $\mathbf{W}$. Consequently, an appropriate nonlinearity function for separating a variety of source signals can be automatically selected. As shown in the accompanied experiments, our proposed algorithm can successfully separate a mixture of sub-Gaussian and super-Gaussian source signals.

Acknowledgement

The work described in this paper was supported by a grant from the Research Grant Council of the Hong Kong SAR (Project No. CUHK4297/98E), and by a Faculty Research Grant of Hong Kong Baptist University (Project Code: FRG/01-02/II-24).

References

[1] A.J. Bell & T.J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, *Neural Computation*, *7*, 1995, 1129–1159.

[2] S.I. Amari, A. Cichocki, & H. Yang, A new learning algorithm for blind separation of sources, in D.S. Touretzky *et al.* (Eds.), *Advances in neural information processing*, *8* (Cambridge, MA: MIT Press, 1996).

[3] J.P. Nadal & N. Parga, Nonlinear neurons in the low-noise limit: A factorial code maximizes information transfer, *Network*, *5*, 1994, 565–581.

[4] S.I. Amari, T.P. Chan, & A. Cichocki, Stability analysis of adaptive blind source separation, *Neural Networks*, *10*(8), 1997, 1345–1351.

[5] B.A. Pearlmutter & L.C. Parra, A context-sensitive generalization of ICA, *Proc. of the International Conf. on Neural Information Processing (ICONIP'96)*, Hong Kong, 1996, 1235–1239.

[6] L. Xu, C.C. Cheung, & S.I. Amari, Learned parametric mixture based ICA algorithm, *Neurocomputing*, *22*, 1998, 69–80.

[7] L. Xu, C.C. Cheung, H.H. Yang, & S.I. Amari, Independent component analysis by the information-theoretic approach with mixture of density, *Proc. of the 1997 IEEE International Conf. on Neural Networks (IEEE-INNS IJCNN'97)*, Houston, TX, *3*, 1997, 1821–1826.

Appendix A

Proof of Corollary 1

Given (17), conditions (12) and (13) in Theorem 1 are always satisfied, as:

$$m_i = E[y_i^2 \dot{\phi}_i(y_i)] = p_i E[|y_i|^{p_i+1}] > 0 \tag{27}$$

$$k_i = E[\dot{\phi}_i(y_i)] = p_i E[|y_i|^{p_i-1}] > 0. \tag{28}$$

Furthermore, condition (14) becomes:

$$\sigma_i^2 \sigma_j^2 \kappa_i \kappa_j = p_i p_j E[y_i^2] E[|y_i|^{p_i-1}] E[y_j^2] E[|y_j|^{p_j-1}] > 1. \tag{29}$$

Because $\forall r$, $E[\phi(y_r) y_r] = 1$, we therefore have:

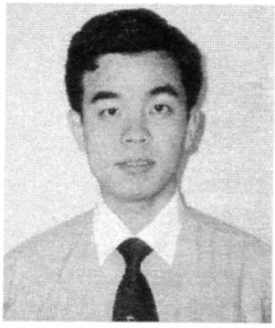
$$E[|y_r|^{p_r+1}] = 1 \quad r = 1, 2, \ldots, k. \tag{30}$$

Putting (30) into (29), we then have:

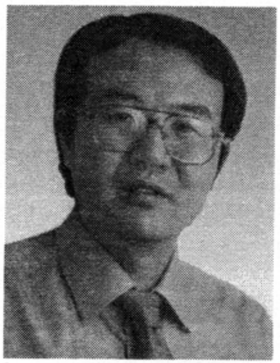
$$\left(\frac{E[|y_i|^{p_i+1}]}{E[|y_i|^2] E[|y_i|^{p_i-1}]}\right) \left(\frac{E[|y_j|^{p_j+1}]}{E[|y_j|^2] E[|y_j|^{p_j-1}]}\right) < p_i p_j. \tag{31}$$

□

Biographies



Yiu-ming Cheung received Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong in 2000. Currently, he is Assistant Professor in the Department of Computer Science, Hong Kong Baptist University. His research interests include independent component analysis, multivariate data clustering analysis, radial basis function networks, time series analysis, portfolio management, and automated trading system.



Lei Xu (IEEE Fellow) is a Professor of Department of Computer Science and Engineering at Chinese University of Hong Kong (CUHK). He is also a full Professor at Peking University, and an adjunct Professor at three other universities in China and UK. After receiving his Ph.D. from Tsinghua University in early 1987, he joined Peking University, where he became one of ten university-level exceptionally promoted young associate professors in 1988 and further been exceptionally promoted to a full Professor in 1992. During 1989–1993, he worked at several universities in Finland, Canada and USA, including Harvard and MIT. He joined CUHK in 1993 as a Senior Lecturer and then took the current professor position since 1996.