ELSEVIER

Neurocomputing 22 (1998) 69–80

NEUROCOMPUTING

Learned parametric mixture based ICA algorithm[1]

Lei Xu[a,*], Chi Chiu Cheung[a], Shun-ichi Amari[b]

[a] *Computer Science and Engineering Department, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, People's Republic of China*

[b] *Frontier Research Program, RIKEN, Japan*

Accepted 3 July 1998

Abstract

The learned parametric mixture method is presented for a canonical cost function based ICA model on linear mixture, with several new findings. First, its adaptive algorithm is further refined into a simple concise form. Second, the separation ability of this method is shown to be qualitatively superior to its original model with prefixed nonlinearity. Third, a heuristic way is suggested for selecting the number of densities in a learned parametric mixture. Finally, experiments have been conducted to show the success of this method on the sources that can either be sub-Gaussian or super-Gaussian, as well as a combination of both the types. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Independent component analysis; Parametric density mixture; Learning; Information theoretic; Maximum likelihood; Blind separation; Nonlinearity

1. Introduction

We consider the classic instantaneous linear mixture ICA problem. That is, we have x from k independent sources $s = [s^{(1)}, \ldots, s^{(k)}]^{\mathrm{T}}$ via a linear mixing $n \times k$ matrix A with

$$x = As, \quad A = [a_{i,j}], \quad i = 1, \ldots, n, j = 1, \ldots, k; \quad n \geq k, \quad Es = 0. \tag{1}$$

The objective is to find a so-called de-mixing matrix W to get

$$y = Wx = WAs = Vs, \qquad V = WA, \qquad y = [y^{(1)}, \ldots, y^{(k)}]^{\mathrm{T}}, \tag{2}$$

[*] Corresponding author. E-mail: lxu@cse.cuhk.edu.hk

[1] This work was supported by HK RGC Direct Grant CUHK 2050162.

PII S0925-2312(98)00050-2

such that either $y = s$ or y recovers s only up to constant unknown scales and any permutation of indices.

This problem has been studied in the literature by many researchers for many years with a number of results [8,4,12,13]. Here, we only concentrate on one stream that relates to the minimization of the following cost:

$$J(W) = -\ln|W| - \sum_{j=1}^{k} \int p(x) \ln p_j(w_j^{\mathrm{T}} x|\xi_j)\, \mathrm{d}x, \quad W = [w_1, \ldots, w_k]^{\mathrm{T}}, \tag{3}$$

for solving the problem when A and W are $k \times k$ invertible matrices, where $p(y^{(j)}|\xi_j)$ is a parametric estimation of the marginal density of $y^{(j)}$. This minimization can be simply implemented by the gradient approach, but the following natural gradient algorithm proposed by Arrari et al. [1] has a better convergence property:

$$W^{\mathrm{new}} = W^{\mathrm{old}} + \eta(I + \phi(y)y^{\mathrm{T}})W, \quad \phi(y) = [\phi_1(y^{(1)}), \ldots, \phi_k(y^{(k)})]^{\mathrm{T}},$$

$$\phi_j(r) = \frac{\mathrm{d} \ln p_j(r|\xi_j)}{\mathrm{d}r}. \tag{4}$$

This cost Eq. (3) was previously obtained using the name of maximum likelihood (ML) approach [9,15] and later repeatedly revised by the approaches of *information-maximization* (*INFORMAX*) [11,2], minimum mutual information (MMI) [1], and Bayesian Kullback Ying–Yang learning [20]. For convenience, in this paper we use the name of the $J(W)$ cost based ICA to denote all these studies, because we concentrate only on the issue of how its performance is affected by the form of $p_j(r|\xi_j)$.

The current paper was initiated at [23] during a visit of the first author to the RIKEN Lab of the third author, for a short period in April 1996, during which all the authors of [1,4] happened to be there. There were some discussions between the third author of [1] and the first author of the present paper, as well as the authors of [3,4]. The discussions clarified the relationship between INFORMAX and MMI.[2] Also, the present first author was attracted to the cost Eq. (3) and came to the proposal [23], with two basic ideas which were regarded as new at that time.

First, in the studies of both [1,2] as well as in those previous efforts related to the cost Eq. (3), the function form of $p(y^{(j)} \mid \xi_j)$ or, equivalently, the nonlinearity of $\phi_j(r)$, is prefixed during the learning on W. However, with a fixed $p(y^{(j)} \mid \xi_j)$, Eq. (4) works only when all the sources are either sub-Gaussian or super-Gaussian. Motivated by this fact, a basic idea in [23] suggests that $p(y^{(j)} \mid \xi_j)$ should be learned together with the learning on W, and the learning on $p(y^{(j)} \mid \xi_j)$ can be done by learning the parameters in a finite mixture of logistic distributions or Gaussian densities.[3] Second, the success

[2] Which have been further well explored from different perspectives and also in a more broad sense by [24,3,20]. An even earlier result was made in [11].

[3] In fact, this idea was also motivated by a previous study. In 1995, the present first author had also suggested an idea of using a mixture of logistic distributions as a flexibly adjustable density for image histogram equalization, which was implemented with success in a joint conference paper with his colleague using the name of entropy maximization [10]. Moreover, the success of using a mixture of logistic distributions for modeling the marginal density had also been reported in 1996 by [14] using the name of maximum likelihood density estimation.

of [2] on super-Gaussian sources by simply using the conventional sigmoid function indicates that an accurate learning of $p(y^{(j)} \mid \xi_j)$ seems unnecessary. Motivated by this fact, it is suggested in [23] to learn $p(y^{(j)} \mid \xi_j)$ such that they loosely match the source densities. That is, instead of attempting to estimate marginal densities of y or source densities of s as accurately as possible, we can get $p_j(\,.\mid \xi_j)$ among a general family of $g_j(r)$ with $0 < g_j(r) < \infty$, $\int_{-\infty}^{\infty} g_j(r)\mathrm{d}r < \infty$ plus some mild constraints [20]. To know how loosely matching this $g_j(r)$ should be, we need to study how the separation ability of Eq. (4) is affected by using different types of $g_j(r)$ in place of a marginal density estimate on $p(y^{(j)} \mid \xi_j)$.

The first idea was later implemented by mixtures of logistic distributions, with success on the sources that could either be sub-Gaussian or super-Gaussian as well as a combination of both the types [22]. Moreover, this work has been further extended to the cases of the full row rank non-invertible W in Section 5 of [18] with an EM-like new adaptive algorithm proposed that was based on using Gaussian mixtures for $p(y^{(j)} \mid \xi_j)$. Some results that relate to the second ideas have also been obtained with the help of theoretical and experimental analyses [21,6]. Since these previous results are distributed in various conference papers, in this paper we elaborate on this method in a more systematic way, to include the latest findings.

Though there are several good algorithms in literature [4,24,13] for the problem Eqs. (1) and (2), we believe that this paper is still of some value to the literature on the $J(W)$ cost based ICA. First, it records some results developed from a different but ignored perspective. Second, the adaptive algorithm obtained from this perspective is actually very simple and easy to implement. Third, it applies to cases where the kurtosis of the original sources does not exist or is zero as it is not based on kurtosis estimation.

2. Learned parametric mixture based ICA method

2.1. The basic idea and implementing algorithm

The basic idea first given in [23], suggests that W can be obtained by

$$\min_{\{W,p(y)\in\mathscr{P}\}} J(W,p(\,y)),\quad J(W,p(y)) = -\ln|W| - \int p(x)\ln p(y)\,\mathrm{d}x,$$
$$p(y) = \prod_{j=1}^{k} p(w_j^{\mathrm{T}}x|\xi_j), \tag{5}$$

where $\mathscr{P}$ is a parametric family, parameterized by $\{\xi_j\}$ of independent densities $p(y) = \prod_{j=1}^{k} p(y^{(j)}|\xi_j)$. $\min_W J(W)$ with $J(W)$ given by Eq. (3) is a constrained special case where $\min_{W,\mathrm{s.t.}p(y)=\hat{p}(y)} J(W,\ p(y))$ with $\hat{p}(y)$ prefixed through $p(y^{(j)} \mid \xi_j)$ or $\phi_j(r)$.

A simple and powerful representation for $\mathscr{P}$ is to let $p(y^{(j)} \mid \xi_j)$ to be modeled by a parametric mixture [23,22]:

$$p(y^{(j)}|\xi_j) = \sum_{i=1}^{n_j} \alpha_i^{(j)} q(y^{(j)}|\xi_{i\cdot j}),$$
$$1 \geq \alpha_i^{(j)} > 0, \quad \sum_{i=1}^{n_j} \alpha_i^{(j)} = 1, \xi_j = \{\alpha_i^{(j)}, \xi_{ij}\}_{i=1}^{n_j}, \tag{6}$$

where n_j is a given number of densities in the mixture, and $q(y^{(j)} \mid \xi_{ij})$ is a parametric density with its function form prespecified, e.g., it can be a Gaussian density or logistic distribution by

$$q(y^{(j)}|\xi_{ij}) = G(y^{(j)}, a_i^{(j)}, b_i^{(j)}), \quad \xi_{ij} = \{b_i^{(j)}, a_i^{(j)}\},$$
$$q(y^{(j)}|\xi_{ij}) = \frac{\partial s(y^{(j)}, \xi_{ij})}{\partial y^{(j)}}, \quad s(y^{(j)}, \xi_i j) = \frac{1}{1 + \exp(-b_i^{(j)}(y^{(j)} - a_i^{(j)}))}, \tag{7}$$

where $G(r, m, \sigma^2)$ denotes a Gaussian with mean m and variance σ^2.

By putting Eq. (6) into Eq. (5), we have [23,22]:

$$\min_{\{W,\xi\}} J(W,\xi), \quad J(W,\xi) = -\ln|W| - \sum_{j=1}^{k} \int p(x) \ln\left[\sum_{i=1}^{n_j} \alpha_i^{(j)} q(w_j^{\mathrm{T}} x|\xi_{ij})\right] \mathrm{d}x,$$
$$\xi = \{\xi_j\}_{j=1}^{k}. \tag{8}$$

Interestingly, we can also directly get Eq. (8) as a special case of the so-called Bayesian Kullback Ying–Yang dependence reduction through its forward architecture [19].

In implementation, $\min_{\{W,\xi\}} J(W, \xi)$ can be made by gradient approach through alternatively fixing one of W, ξ and updating the other to reduce $J(W, \xi)$. This alternative minimization procedure will guarantee convergence to a local minimum of $J(W, \xi)$. Moreover, it can be implemented in an on-line way via stochastic gradient approach, as follows:

Step 1: For each x, fix ξ and update W to reduce $-\ln|W| - \sum_{j=1}^{k} \ln\sum_{i=1}^{nj} \alpha_i^{(j)} q(w_j^{\mathrm{T}} x \mid \xi_{ij})$ by Eq. (4).

Step 2: Then fix W and update $\xi = \xi^{\mathrm{old}} + \eta\Delta\xi$ with $\Delta\xi$ being the gradient descent direction of $-\sum_{j=1}^{k} \ln\sum_{i=1}^{nj} \alpha_i^{(j)} q(w_j^{\mathrm{T}} x|\xi_{ij})$. For example, the detailed form of $\Delta\xi$ in the case of logistic distribution is given below:

$$\alpha_i^{(j)} = \frac{\exp(\gamma_i^{(j)})}{\sum_{k=1}^{n_j} \exp(\gamma_k^{(i)})}, \quad y^{(j)} = w_j^{\mathrm{T}} x, \quad h_i^{(j)} = \frac{\alpha_i^{(j)} q(y^{(j)}|\xi_{i,j})}{\sum_{r=1}^{n_j} \alpha_i^{(r)} q(y^{(r)}|\xi_{ir})},$$
$$u_i^{(j)} = b_i^{(j)}(y^{(j)} - a_i^{(j)}), \quad t_i^{(j)} = \frac{1 - \exp(-u_i^{(j)})}{1 + \exp(-u_i^{(j)})},$$
$$\Delta\gamma_i^{(j)} = \sum_{k=1}^{n_j} h_k^{(j)}(\delta_{kj} - \alpha_i^{(j)}), \quad \Delta b_i^{(j)} = h_i^{(j)}\left(\frac{1}{b_i^{(j)}} - y^{(j)} t_i^{(j)}\right), \quad \Delta a_i^{(j)} = h_i^{(j)} b_i^{(j)} t_i^{(j)}, \tag{9}$$

where δ_{ij} is the Kronecker delta function.

In cases where the Gussian mixtures are used, the detailed form of $\Delta\xi$ can be obtained in a similar way. Moreover, we can also get an EM-like adaptive algorithm, as given in [18].

In the following subsections, we discuss some major issues in relation to the performance of this learned parametric mixture based ICA method.

2.2. Directly workable for non-invertible linear mixture

As shown first in [18], all the discussions and the algorithm in Section 2.1 are applicable to a full row rank non-invertible W for the case Eq. (1) with A being $n \times k$, $n > k$. In [18], this problem is considered as a degenerated case of $\sigma^2 \to 0$ in a noisy mixture $x = As + e_x$, with e_x obtained from a Gaussian $G(e_x, 0, \sigma^2 I_n)$. The so-called Bayesian Kullback Ying–Yang (BKYY) learning [16,17] is used for ICA with its Ying space being $\prod_{j=1}^{k} p(y^{(j)}|\xi_j)$, its Ying passage being $G(x, As, \sigma^2 I_n)$ (i.e., $x = As + e_x$) and Yang passage being $G(s, Wx, \Sigma)$. Also, it is assumed that $WA = I$ or $A = W^- = W^{\mathrm{T}}(WW^{\mathrm{T}})^{-1}$, from which we have $\Sigma = WW^{\mathrm{T}}\sigma^2 I_k$.

When $\sigma_x^2 \to 0$, as shown in [18], the BKYY learning in this special case is equivalent to the minimization of the following cost function:

$$J(W,\xi) = -0.5 \ln|WW^{\mathrm{T}}| - \sum_{j=1}^{k} p(x) \ln p(w_j^{\mathrm{T}} x|\xi_j)\,\mathrm{d}x, \tag{10}$$

which is different from $J(W, \xi)$ of Eq. (8) only in the sense that $\ln|W|$ is replaced by 0.5 $\ln|WW^{\mathrm{T}}|$. When W is invertible, the two become the same. In other words, Eq. (10) is an extension of Eq. (8), that is applicable to the full row rank non-invertible W.

The minimization of $J(W, \xi)$ in Eq. (10) can be accomplished by either the gradient descent or the natural gradient descent updating:

$$W^{\mathrm{new}} = W^{\mathrm{old}} + \eta\,\Delta W, \quad \Delta W = \begin{cases} (WW^{\mathrm{T}})^{-1}W + \phi(y)x^{\mathrm{T}} & \text{gradient,} \\ (I + \phi(y)y^{\mathrm{T}})W & \text{natural gradient,} \end{cases} \tag{11}$$

where $\phi(y)$ is the same as in Eq. (4). Interestingly, the equation formed by the natural gradient actually remains unchanged as that in Eq. (4). Therefore, all the discussions and the algorithm in Section 2.1 apply to the cases of full row rank non-invertible W in exactly the same way as invertible W.

The detailed derivation of Eq. (10) from BKYY learning was first given in Section 5 of [18], where its Eqs. (19) and (20) were exactly the same as the above Eq. (10) and the gradient equation in Eq. (11), respectively. Subsequently, Eq. (10) was also given in [7] though neither the derivation nor the reference to [18].

2.3. Improved separation performance

The problem of whether the algorithm in Section 2.1 can perform separation is equivalent to asking whether all the local minima of $\min_{\{W,\,\xi\}} J(W, \xi)$ are separation solutions. If yes, the problem can be solved by using any converged solution of the algorithm. Moreover, if only a part of the local minimums of $\min_{\{W,\,\xi\}} J(W, \xi)$ are separation solutions, the algorithm works if it is initialized at a nearby region of one such local minimum. However, if it is initialized randomly, the algorithm still works only with a probability that depends on the area of attraction of each local minimum and the ratio of the local minima which are separation solutions to non-separation solutions. In addition, if we also have additional information to check whether a converged solution by the algorithm is a separation solution or not, we can continuously run the algorithm until a separation solution is finally obtained, i.e., the source separation can be achieved in probability one at least theoretically. In practice, we encounter computational complexity, which depends on the area of attraction of each local minimum and the ratio of the local minimums which are separation solutions to non-separation solutions. It can be impractical if the computational complexity is too high.

As long as $\mathscr{P}$ includes the original source density $p_o(y)$ (e.g., for Eq. (6) we can let n_j to be large enough), $\min_{\{W,\ p(y)\in\mathscr{P}\}} J(W,\ p(y))$ will have at least one separation solution with $p_o(y)$ and its corresponding separation W_o. However, for its constrained special case $\min_{\{W,\ \text{s.t.}\ p(y)=\hat{p}(y)\}} J(W, p(y))$, the local minimums may not include this W_o and the number of its local minimums may increase because of the imposed constraint $p(y) = \hat{p}(y)$. On the other hand, by relaxing such a constraint, $\min_{\{W,\ p(y)\in\mathscr{P}\}} J(W, p(y))$ not only includes one separation solution but may also reduce the number of local minimums of $\min_{\{W,\ \text{s.t.}\ p(y)=\hat{p}(y)\}} J(W,\ p(y))$, which qualitatively explains why the learned parametric mixture based ICA method can improve the success in separation.

2.4. Selecting the number of densities in a mixture

The larger the number n_j is, the larger the family $\mathscr{P}$ is, and the better the chance of achieving success in separation. This may not be necessary because this will also increase the computing cost. Moreover, it is still possible for a fixed $\hat{p}(y) \neq p_o(y)$, $\min_{\{W\}} J(W, \hat{p}(y))$ to result in a separation solution W. The success of [2] on the super-Gaussian sources by simply using the conventional sigmoid function is a typical example. Thus, to find out the necessary n_{j} in a finite mixture, the second idea in [23] on "loosely matching" needs to be considered.

To do so, we need to analyze the condition on a prefixed $\hat{p}(y)$ such that all or at least a part of local minimums of $\min_{\{W\}} J(W, \hat{p}(y))$ are separation solutions. This task is not easy. In the following, we briefly summarize some related progresses:

(1) In [15], a fixed $\phi_j(y^{(j)}) = -(y^{(j)})^3$ is used with success for experiments on uniform sources, which are sub-Gaussian. In [6], experiments have also shown Eq. (4) to work for cases where all the sources are uniform or gamma (both are sub-Gaussian), but to fail for sources of human speech signals (it is super-Gaussian). Moreover, for the special cases of two channels (i.e., $k = 2$), a systematic study has been made with $\phi_j(y^{(j)}) = c_j(y^{(j)})^3$, $c_i < 0$ in Eq. (4). It is proved that sub-Gaussian sources can be separated because all the stable converging points are shown to be separation solutions. However, super-Gaussian source cannot be separated. In [21], for the two source cases with $\phi_1(y^{(1)}) = c_{11}y^{(1)}$ and $\phi_2(y^{(2)}) = c_{23}\ (y^{(2)})^3$ with $c_{11} < 0$ and $c_{23} < 0$ in Eq. (4), it has been proved that Eq. (4) works sucessfully when one

source is sub-Gaussian and the other is super-Gaussian, or when one source is Gaussian and the other is non-Gaussian. Moreover, it has also been shown experimentally that Eq. (4) works for super-Gaussians but fails to separate sub-Gaussian sources when $s_j(r)$ is the conventional sigmoid as used in [11,2].

(2) When the prefixed $p_j(y^{(j)})$ is super-Gaussian, i.e., its standardized kurtosis is positive, Eq. (4) is experimentally shown to work well for the sources of super-Gaussian, but fails for sources of sub-Gaussian. For example, in [2] the fixed sigmoid $p_j(y^{(j)}) = \mathrm{d}s(y^{(j)})/\mathrm{d}y^{(j)}$, $s(r) = 1/(1 + \mathrm{e}^{-r})$ is used, which corresponds to a positive kurtosis 1.2 as shown in the second column of Table 1. It is reported in [2] that Eq. (4) works for human speech signals with highly peaked density (i.e., super-Gaussian signals). However, the experiments given in [21] have shown that it fails at sub-Gaussian sources (e.g., uniform density or *gamma* density). Stating another example, when we use the fixed nonlinearity as shown in the first column of Table 1 with a positive standardized kurtosis 1.2216, the experiment in [5] has shown that Eq. (4) works for the sources of super-Gaussian, but fails for sources of sub-Gaussian.

(3) When the prefixed $p_j(y^{(j)})$ is sub-Gaussian, i.e., if its kurtosis is negative, Eq. (4) works for sources of sub-Gaussian, but fails for sources of super-Gaussian. When we use the fixed nonlinearity as shown in the third column of Table 1 with its kurtosis -0.8118, experiments have shown that Eq. (4) works for the cases where all the sources are uniform or gamma (both are sub-Gaussian), but fails for sources of speech signals (which is super-Gaussian) [21]. Particularly, for the case $k = 2$, it has been mathematically proven that it works for cases where all the sources are sub-Gaussian, but may fail for sources of super-Gaussian in [6]. For another example, as shown in

Table 1
Properties and separation capabilities of several non-linearities

(1)	$g_i(y^{(j)}) = \dfrac{\exp(-\frac{3}{4}(y^{(j)})^{4/3})}{2(3/4)^{1/4}\gamma^{(3/4)}}$	Super-Gaussian
(2)	$g_i(y^{(j)}) = \dfrac{\exp(-y^{(j)})}{(1+\exp(-y^{(j)}))^2}$	Super-Gaussian
(3)	$g_i(y^{(j)}) = \dfrac{1}{\sqrt{2\pi}}\exp(-(y^{(j)^2}/2))$	Gaussian
(4)	$g_i(y^{(j)}) = \dfrac{\sqrt{2}}{\gamma(1/4)}\exp(-(y^{(j)^4}/4))$	Super-Gaussian

$g_i(y^{(j)})$	$\phi_j(y^{(j)})$	Kurtosis $\dfrac{\mu^4}{(\mu^2)^2} - 3$
Case (1)	$-\sqrt[3]{y^{(j)}}$	1.2216
Case (2)	$1 - 2\mathrm{logsig}_{\mathrm{j}}(y^{(j)})$	1.2
Case (3)	$-y^{(j)}$	0
Case (4)	$-(y^{(j)})^3$	-0.8118

[6], the prefixed $p_j(y^{(j)})$ obtained by truncated Gram-Charlier series in [1] also has a negative standardized kurtosis, and experiments have shown that it indeed works for cases where all the sources are sub-Gaussian, but fails at least for some super-Gaussian sources.

The above results hint that whether the sources can be separated relates to whether there is a loose matching between the kurtosis of sources and the used $p_j(y^{(j)})$, which is consistent to the findings in the well-studied stream of those kurtosis estimation based ICA.

Therefore, the heuristic for selecting n_j is to consider such a loose match. Roughly, we can expect a simple mixture with $n_j = 2$ to be generally workable since by changing its parameters we can change the kurtosis of $p_j(y^{(j)})$ from positive to negative to match the values of the kurtosis of a source, by quite a large range. However, we remark that even in this simple case, the learned parametric mixture method is different from those kurtosis estimation based ICA, since it considers not only kurtosis but also the configuration, and thus it may still work even in a case where the kurtosis of the original sources does not exist or is zero.

3. Experimental examples

We have three sources. The first is an artificially generated bimodal symmetric $\beta(0.5,0.5)$ distributed i.i.d. source, the second is an artificially generated uniform $(-0.5, 0.5)$ distributed i.i.d. source, the third one is a permuted speech signal. The mixing matrix used is:

$$A = \begin{bmatrix} 1 & 0.6 & 0.3 \\ 0.8 & 1 & 0.3 \\ 0.4 & 0.9 & 1 \end{bmatrix}. \tag{12}$$

For this example, the experiments in [22,21] demonstrated that the use of $p_j(y^{(j)})$ as a fixed sigmoid as in [2] works well for the permuted speech signal, but fails for the bi-modal beta distribution $\beta(0.5, 0.5)$ in $[-0.5, 0.5]$ and the uniform distribution in $[-1, 1]$; while the use of $p_j(y^{(j)})$ as a fixed sigmoid as in [1] works well for the bi-modal beta distribution $\beta(0.5, 0.5)$ in $[-0.5, 0.5]$ and the uniform distribution in [-1, 1], but fails for the permuted speech signal.

Shown in Fig. 1 are the results of the algorithm given in Section 2.1. The settings are such that $n_j = 5$, all $\gamma_i^{(j)}$ and $a_i^{(j)}$ are initialized as $\frac{1}{5}$ and 0, respectively, as well as $b_i^{(1)}, \ldots, b_i^{(5)}$ are initialized in the interval $[10^{-0.3}, 10^{1.2}]$. The results obtained demonstrate that the algorithm can approximate sources 'quite well' and perform separation successfully. By comparing rows 1 and 2, we can observe that the basic configurations of $p_j(y^{(j)})$ and the sources are basically well matched. Row 3 gives a comparison between the corresponding $s_j(y^{(j)}) = \int_{-\infty}^{y^{(j)}} p_j(r)\,\mathrm{d}r$ and the fixed $s_j(r) = 1/(1 + \mathrm{e}^{-r})$ in order to observe the difference between the learned mixture approach and the fixed sigmoid approach used in [2]. Row 4 gives the histograms of $z^{(j)} = s_j(y^{(j)})$, which

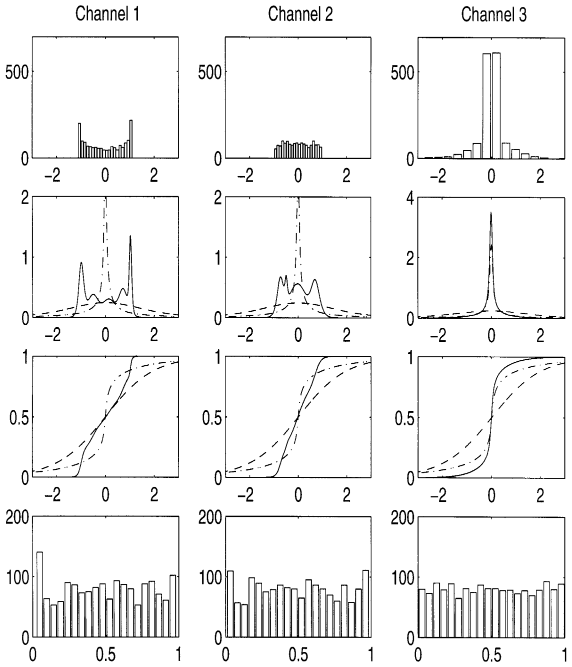


Fig. 1. Results of the experiment by the learned parametric mixture algorithm with a mixture of five densities. Row 1: histograms of $s^{(j)}$. Row 2 and 3: $g_j(y^{(j)})$ and $s_j(y^{(j)}) = \int_{-\infty}^{y^{(j)}} g_j(r)\,\mathrm{d}r$, respectively. (— learned mixture of densities, – · – initial, - - $s_j(r) = \mathrm{logsig}(r)$ for comparison.) Row 4: histograms of $z^{(j)} = s_j(y^{(j)})$.

shows how far $z^{(j)}$ is from the uniform density, from which we can again see a good match between $p_j(y^{(j)})$ and source densities.

The results in Fig. 2 are obtained in a simplified case with $n_j = 2$, $\alpha = 0.5$, $b_i^{(j)} = 1$. After the learning has stabilized, a brief display of V and a is

$$V = \begin{bmatrix} 10.2583 & 0.0205 & -0.1169 \\ -0.0085 & 5.0408 & -0.0813 \\ -0.0132 & -0.0095 & 9.3977 \end{bmatrix}, \quad a = \begin{bmatrix} -3.3378 & 3.3657 \\ -2.4460 & 2.4672 \\ 0.0241 & 0.0241 \end{bmatrix}. \tag{13}$$

From this V, we see that three channels have again been successfully separated with signal/noise ratio being around 1000. In Fig. 2, the histograms of the sources in row 1 are quite different from the resultant $p_j(y^{(j)})$ in row 2. This point can also be observed from row 4, where the histograms are obviously different from the uniform density. However, we can observe that the configurations (and thus kurtosis also) have a loose match between the first row and the second row. Moreover, we can observe from the second row and the third row how the configurations are different from the one specified by the fixed $s_j(r) = 1/(1 + \mathrm{e}^{-r})$.

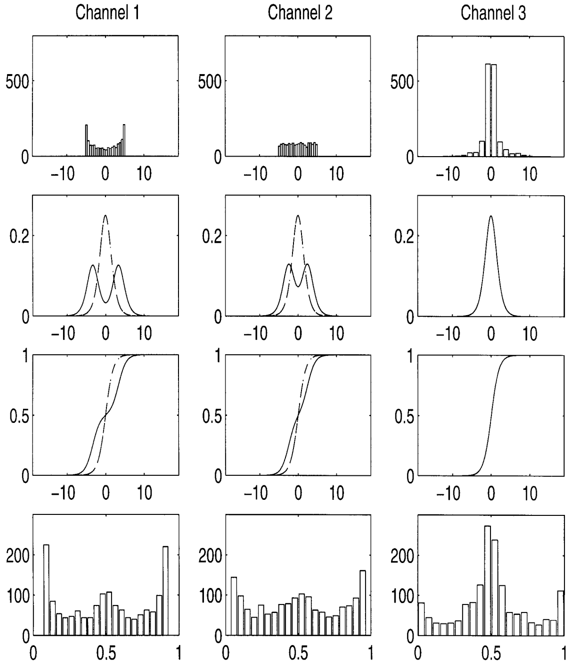


Fig. 2. Results of the experiment by the learned parametric mixture algorithm with a mixture of two densities. Row 1: histograms of $s^{(j)}$. Row 2 and 3: $g_j(y^{(j)})$ and $s_j(y^{(j)}) = \int_{-\infty}^{y^{(j)}} g_j(r)\,\mathrm{d}r$, respectively. (— learned mixture of densities, – · –, initial, - - $s_j(r) = \mathrm{logsig}(r)$ for comparison.) Row 4: histograms of $z^{(j)} = s_j(y^{(j)})$.

4. Conclusion

The learned parametric mixture method for ICA on linear mixture has been further elaborated, with a much simplified form in its adaptive algorithm along with new insights. It is shown to be qualitatively superior to its original model with prefixed nonlinearity. Also, a heuristic way is suggested for selecting the number of densities. Experiments have demonstrated the success of this method on sources of either sub-Gaussian or super-Gaussian as well as their combinations.

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, M.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-maximatization approach to blind separation and blind deconvolution, Neural Comput. 7 (1995) 1129–1159.

[3] J.F. Cardoso, Informax and maximum likelihood for blind source separation, IEEE Signal Process. Lett. 4, 109–111.

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

[5] C.C. Cheung, Adaptive blind signal separation, Master Thesis, Department of Computer Science and Engineering, The Chinese University of Hong Kong, June 1997.

[6] C.C. Cheung, L. Xu, Separation of two independent sources by the information-theoretic approach with cubic nonlinearity, Proc. IEEE Int. Conf. on Neural Networks (IEEE-INNS IJCNN97), 9–12 June 1997, Houston, TX, USA, vol. 4, pp. 2239–2244.

[7] Cichocki et al., Independent component analysis for noisy data in: C. Fyfe (Ed.), Proc. Int. ICSC Workshop on Independence and Artificial neural networks (I&ANN'98), 9–10 February, Tenerife, Spain, ICSC Academic Press, 1998, pp. 52–58.

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

[9] M.Gaeta, J.-L Lacounme, Source separation without a priori knowledge: the maximum likelihood solution, Proc. European Signal Process. Conf. EUSIPCO90 (1990) 621–624.

[10] I. King, L. Xu, Adaptive contrast enhancement by entropy maximization with a 1-K-1 constrained network, Proc. 1995 Int. Conf. on Neural Information Process. (ICONIP95), 30 October–3 November 1995, Beijing, vol. II, pp. 703–706.

[11] J.-P. Nadal, N. Parga, Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer, Network 5 (1994) 565–581.

[12] E. Oja, ICA Learning rules: stationarity, stability, and sigmoids, in: C. Fyfe (Ed.), Proc. of Int. ICSC Workshop on Independence and Artificial Neural Networks (I&ANN'98), 9–10 Febraury, Tenerife, Spain, ICSC Academic Press, New York, 1998, pp. 97–103.

[13] E. Oja, J. Karhunen, A. Hyvärinen, From neural principal components to neural independent components, Proc. Int. Conf. on Artificial Neural Networks ICANN'97, 8–10 October, Lausanne, Switzerland, 1997, pp. 519–528.

[14] B.A. Pearlmutter, L.C. Parra, A context-sensitive generalization of ICA, in: Progress in Neural Information Processing: Proc. Int. Conf. on Neural Information Processing (ICONIP 96), Hong Kong, 24–27 September 1996, Springer, Singapore, 1996, pp. 1235–1239.

[15] D.T. Pham, P. Garat, C. Jutten, Separation of a mixture of independent sources through a maximum likelihood approach, in: J. Vandewalle et al. (Eds.), Singal Processing VI: Theories and Applications, Elsevier, Amsterdam, 1992, pp. 771–774.

[16] L. Xu, Bayesian Ying–Yang System and Theory as A Unified Statistical Learning Approach: (I) Unsupervised and Semi-Unsupervised Learning, Invited paper, in: S. Amari, N. Kassabov (Eds.), Brain-like Computing and Intelligent Information Systems, Springer, Berlin, 1997, pp. 241–274.

[17] L. Xu, Bayesian Ying–Yang system and theory as a unified statistical learning approach (II): from unsupervised learning to supervised learning and temporal modeling and (III): models and algorithms for dependence reduction, data dimension reduction, ICA and supervised learning, in: K.W. Wong, I. King, D.Y. Yeung (Eds.), Theoretical Aspects of Neural Computation: A Multidisciplinary Perspective (TANC97), Springer, Berlin, 1997, pp. 25–60.

[18] L. Xu, Bayesian Ying–Yang learning based ICA models, Neural Networks for Signal Processing VII: Proc. IEEE Signal Processing Society Workshop, 24–26 September 1997, Florida, pp. 476–485.

[19] L. Xu, BYY dependence reduction theory and blind source separation, Proc. Intentional Joint Conf. on Neural Networks, 5–9 May 1998, Anchorage, Alaska, Vol. II, pp. 2495–2500.

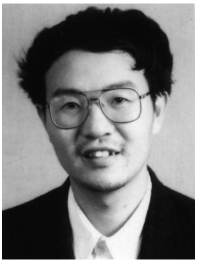
[20] L. Xu, S.-I. Amari, A general independent component analysis framework based on Bayesian-Kullback Ying–Yang Learning, in: Progress in Neural Information Processing: Proc. Int. Conf. on Neural Information Processing (ICONIP 96), Hong Kong, 24–27 September 1996, Springer, Singapore, 1996, pp. 1235-1239.

[21] L. Xu, C.C. Cheung, J. Ruan, S.-I. Amari, Nonlinearity and separation capability: further justification for the ICA algorithm with a learned mixture of parametric densities, Invited special session on Blind Signal Separation, Proc. European Symp. on Artificial Neural Networks, Bruges, 16–18 April 1997, pp. 291-296.

[22] L. Xu, C.C. Cheung, H.H. Yang, S.-I. Amari, Independent component analysis by the information-theoretic approach with mixture of density, Proc. IEEE Int. Conf. on Neural Networks (IEEE-INNS IJCNN97), 9–12 June, Houston, TX, USA, 1997, pp. 1821-1826.

[23] L.Xu, H.H. Yang, S.-I. Amari, Signal source separation by mixtures accumulative distribution functions or mixture of bell-shape density distribution functions, Research Proposal, Presented at FRONTIER FORUM (speakers: D. Sherrington, S. Tanaka, L. Xu, J. F. Cardoso), organized by S. Amari, S. Tanaka, A. Cichocki, The Institute of Physical and Chemical Research (RIKEN), Japan, 10 April 1996.

[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 of the Deptartment 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 since 1996. He is also a professor of Peking University since 1992, where he started as a postdoc of Deptartment of Math 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 in 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'sWho 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.

Shun-ichi Amari was graduated from the graduate School of the University of Tokyo in 1963 majoring in Mathematical Engineering. He then became associate professor at Kyushu University, the University of Tokyo and full professor at the University of Tokyo. He is now Professor Emeritus at the University of Tokyo. He is Director of Brain Style Information Systems Group, RIKEN Brain Science Institute. He received the IEEE Neural Networks Pioneer Award, IEEE Emanuel Piore Award, Japan Academy Award and so on. He served as President of the International Neural Network Society, Vice President of Institute of Electronics, Information and Communications Engineers and in many other positions.