I&ANN'98

Proceedings of the

International **ICSC** Workshop on

Independence & Artificial Neural Networks
I&ANN'98

February 9-10, 1998
University of La Laguna, Tenerife, Spain

Editor: C. Fyfe

Publication by **ICSC** Academic Press
International Computer Science Conventions
Canada / Switzerland

ISBN 3-906454-13-4

Table of Contents

FURTHER RESULTS ON NONLINEARITY AND SEPARATION CAPABILITY OF A LINEAR MIXTURE ICA METHOD AND LEARNED PARAMETRIC MIXTURE ALGORITHM

Lei Xu [1], *Chi Chiu Cheung*[1], *and Shun-ichi Amari* [2]

1. Computer Science and Engineering Department
The Chinese University of Hong Kong, Shatin, NT, Hong Kong, P.R.China
2. Frontier Research Program, RIKEN,
Hirosawa, 2-1, Wako-shi, Saitama, 351-01, Japan.

ABSTRACT

Further results on the nonlinearity and separation capability of the classic maximum likelihood-information theoretic ICA method have been obtained. The idea of 'loose matching' can be further addressed into a specific conjecture assertion that sources can be separated from others as long as the kurtosis signs of the densities specified by the nonlinear transfer functions match the kurtosis signs of sources. Moreover, the previously proposed learned parametric mixture algorithm has been simplified with only two densities and their position parameters adjusted. Experiments are given to support this assertion and the success of the simplified learned parametric mixture algorithm.

1. INSTANTANEOUS LINEAR MIXTURE AND MAXIMUM LIKELIHOOD-INFORMATION THEORETIC ICA

We consider the most widely studied *Instantaneous invertible linear mixture* ICA problem. That is, we have x from k independent sources $s^{(1)}, \cdots, s^{(k)}$ via a so called *mixing matrix* $k \times k$ invertible matrix A with

$$x = As, \quad A = [a_{i,j}],\ i = 1, \cdots, d,\ j = 1, \cdots, n;$$
$$Es = 0, \quad s = [s^{(1)}, \cdots, s^{(k)}]^T, \tag{1}$$

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

$$y = Wx = WAs = Vs, \quad V = WA, \tag{2}$$

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

This work was supported by HK RGC Earmarked Grants CUHK250/94E, CUHK484/95E and CUHK 339/96E and by Ho Sin-Hang Education Endowment Fund for Project HSH 95/02.

We focus on a simply iterative model

$$W^{new} = W^{old} + \eta \Delta W,$$
$$\Delta W = \begin{cases} (a)\ gradient: \\ W^{-T} + \phi(y)x^T; \\ (b)\ natural\ gradient: \\ (I + \phi(y)y^T)W \end{cases},$$
$$\phi(y) = [\phi_1(y_1), \cdots, \phi_k(y_k)]^T. \tag{3}$$

The form with the choice (a) for ΔW can be equivalently obtained in the terms of *maximum likelihood learning (ML)* [7, 12], *information-maximization (INFORMAX)* [10, 2] and *minimum mutual information(MMI)* [1], as well as a special case of Bayesian-Kullback Ying-Yang learning [19]. For simplicity, we call the above eq.(3) *Maximum Likelihood-Information Theoretic ICA Method*.

The choice (b) for ΔW, called the natural gradient algorithm and proposed by [1], will produce an equivalent result but with an improved convergence property. Although the same iterative model is studied in these different references, their detail implementation algorithms are actually different in the use of the nonlinear function $\phi(y)$, with successes on different special types of source distributions. For examples, in [10, 2] a prefixed sigmoid function $s(r)$ is used, which is equivalent to specify a fixed $\phi_j(y^{(j)})$ via eq.(4). In this case, eq.(3) works for the cases that all the sources are *super-Gaussian* (e.g., for human speech signals with highly peaked density [2]) but fails at least for some *sub-Gaussian* sources of [18].

$$\phi_j(y^{(j)}) = \frac{\partial p(y^{(j)})/\partial y^{(j)}}{p(y^{(j)})},$$
$$p(y^{(j)}) = s'(y^{(j)}) = \frac{ds(y^{(j)})}{dy^{(j)}}. \tag{4}$$

In [7, 1], $\phi_j(y^{(j)})$ is given via approximating the marginal densities by fixed truncated Gram-Charlier series, it works for the cases that all the sources are sub-Gaussian, but

fails at least for some sup-Gaussian sources [5]. In [12], a fixed $\phi_j(r) = -r^3$ is used with success for an experiment on uniform sources, where are sub-Gaussian. In [4], experiments have also shown that the algorithm eq.(3) works for the cases that all the sources are uniform or gamma (both are *sub-Gaussian*), but fails for sources of human speech signals (it is sup-Gaussian). Moreover, for the special cases of two channels $k = 2$, it has been mathematically proved that eq.(3) with $\phi_j(r) = -r^3$ works for the cases that all the sources are sub-Gaussian, but may fail for sources of sup-Gaussian in [4] too. The need of optimizing the nonlinearity to the true source densities was first addressed in [12], where $\phi_j(y^{(j)}) = -\sum_{n=1}^{N} c_{j,n} h_n(y^{(j)})$ is used with $h_1(r) \cdots, h_N(r)$ being a set of bases, e.g. $h_n(r) = sign[r]|r|^{n-1}$ and $c_{j,n}$ is estimated to minimize the covariance matrix of the estimation on W^{-1}, with efficiency of the estimator discussed. These existing efforts share a common point of attempting directly or indirectly to approximate the unknown source densities or equivalently the marginal densities $p(y^{(j)})$.

In [19, 14, 15, 16, 17, 18], the effort on a quite different direction has been made with a belief that although a very good fit between the densities specified by $\phi_j(y^{(j)})$ via eq.(4) and the unknown source densities can produce a separation performance, it is difficult and also not absolutely necessary, and that we may use densities $g_j(y^{(j)})$ with

$$\phi_j(y^{(j)}) = \frac{\partial g_j(y^{(j)})/\partial y^{(j)}}{g_j(y^{(j)})}, \qquad (5)$$

to replace $p(y^{(j)})$ in eq.(4) for "loosely matching" the unknown marginal densities. This loose matching may mean that $g_j(y^{(j)})$ differs from the unknown marginal densities considerably and the choice of $\{g_j(y_j)\}$ is much wider than $\{p_{y_j}(y_j)\}$. In other words, we attempt to explore the relation between the nonlinearity and separation capacity. Some results are obtained in [14, 15, 16, 17, 18] and an implementation algorithm that uses a learned parametric mixture for $g_j(y^{(j)})$ is proposed and shown by experiments to work for sources of both sub-Gaussian and sup-Gaussian as well as their mixtures. However, the specific relation between the nonlinearity and separation capacity is still not totally clear yet and the number of densities needed in the learned parametric mixture is given heuristically.

In this paper, a further step forward is reported. The idea of 'loose matching' is further addressed into a specific conjecture assertion that a source can be separated from others as long as its kurtosis sign is the same as the kurtosis sign of one $g_j(y^{(j)})$ and that all the k sources can be separated when there is an one-to-one same-sign-correspondence between the kurtosis signs of sources and $g_j(y^{(j)}), j = 1, \cdots, k$. Moreover, from this assertion we simplify the previously proposed learned parametric mixture algorithm with only two densities and their position parameters adjusted. Experiments are provided to demonstrate the results.

(1) $g_j(y_j) = \frac{\exp(-\frac{3}{4}y_j^{4/3})}{2(3/4)^{1/4}\gamma(3/4)}$	Super-Gaussian
(2) $g_j(y_j) = \frac{\exp(-y_j)}{(1+\exp(-y_j))^2}$	Super-Gaussian
(3) $g_j(y_j) = \frac{1}{\sqrt{2\pi}}\exp(-y_j^2/2)$	Gaussian
(4) $g_j(y_j) = \frac{\sqrt{2}}{\gamma(1/4)}\exp(-y_j^4/4)$	Sub-Gaussian

$g_j(y_j)$	$\phi_j(y_j)$	kurtosis $\frac{\mu_4}{(\mu^2)^2} - 3$
Case (1)	$-\sqrt[3]{y_j}$	1.2216
Case (2)	$1 - 2logsig(y_j)$	1.2
Case (3)	$-y_j$	0
Case (4)	$-y_j^3$	-0.8118

Table 1: Properties and separation capabilities of several non-linearities.

Before closing this section, it deserves to mention that in the sense of asymptotic stability of an averaged equation the nonlinearity has been also analyzed by [3] and [11], respectively in the connection of a non-Gaussian criterion with $\phi_j(y^{(j)})$ used in their corresponding ICA algorithms. It is interesting to further explore the relation between their studies and the conjecture on $g_j(y^{(j)})$ suggested in this paper for the ICA algorithm eq.(3).

2. LOOSELY MATCHING BETWEEN NONLINEARITY AND SOURCE DENSITY

As shown in Tab.1, we consider the relation between the kurtosis of sources and $g_j(y^{(j)}), j = 1, \cdots, k$ in typical cases as follows:

(1) When the prefixed $g_j(y^{(j)})$ is super-gaussian, i.e., the standardized kurtosis is positive, the algorithm eq.(3) works for sources of *super-Gaussian*, but fails for sources of *sub-Gaussian*. For example, in [2] the fixed sigmoid

$$g_j(y^{(j)}) = \frac{ds(y^{(j)})}{dy^{(j)}}, \quad s(r) = \frac{1}{1+e^{-r}}, \qquad \phi(r) = 1 - 2logsig(r) \qquad (6)$$

corresponds to a positive kurtosis 1.2 as shown in the second column of Tab.1, and it reported in [2] that the algorithm eq.(3) works for human speech signals with highly peaked density (i.e., super-gaussian signals). However, the experiments given in [18] has shown that it fails at sub-gaussian sources (e.g., uniform density or *gamma* density). For another example, when we use the fixed nonlinearity as shown in the first column of Tab.1 with a positive standardized kurtosis 1.2216, experiment in [5]

has shown that the algorithm eq.(3) works for the sources of *super-Gaussian*, but fails for sources of *sub-Gaussian*.

(2) When the prefixed $g_j(y^{(j)})$ is sub-gaussian, i.e., the kurtosis is negative, the algorithm eq.(3) 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 Tab.1 with its kurtosis -0.8118, experiments have also shown that eq.(3) works for the cases that all the sources are uniform or gamma (both are *sub-Gaussian*), but fails for sources of speech signals (it is sup-Gaussian) [18]. Also, for the case of two channels $k = 2$, it has been mathematically proved that it works for the cases that all the sources are sub-Gaussian, but may fail for sources of sup-Gaussian in [4]. In [1], by truncated Gram-Charlier series the following fixed

$$\begin{aligned} g_j(y^{(j)}) &= \frac{1}{C'} exp[-\frac{1}{16}(y^{(j)})^{12} - \frac{5}{8}(y^{(j)})^{10} \\ &+ \frac{7}{12}(y^{(j)})^8 + \frac{47}{24}(y^{(j)})^6 - \frac{29}{16}(y^{(j)})^4], \\ \phi_j(r) &= -\frac{3}{4}(y^{(j)})^{11} - \frac{25}{4}(y^{(j)})^9 + \frac{14}{3}(y^{(j)})^7 \\ &+ \frac{47}{4}(y^{(j)})^5 - \frac{29}{4}(y^{(j)})^3, \end{aligned} \tag{7}$$

is used in the algorithm eq.(3). Although the function is quite complicated, as shown in [4], $g_j(y^{(j)})$ has a negative standardized kurtosis and thus is sub-gaussian. Therefore, experiments have shown that it indeed works for the cases that all the sources are sub-Gaussian, but fails at least for some sup-Gaussian sources.

In summary of the above experimental facts, we make *a conjecture assertion that a source can be separated from others as long as there is one j with the kurtosis sign of $g_j(y^{(j)})$ being the same as the the kurtosis signs of source, or in other word, the product of the kurtosis of source and the kurtosis of $g_j(y^{(j)})$ is positive. As a whole, all the sources can be separated as long as there is an one-to-one same-sign-correspondence between the kurtosis signs of k sources and the kurtosis signs of $g_j(y^{(j)})$, $j = 1, \cdots, k$. That is, there is at least one way to pair all the sources and $\{g_j(y_j)\}$ such that for each pair the source and $g_j(y^{(j)})$ have a positive product in their kurtosis.*

We are still seeking a mathematical proof for the above conjecture assertion. In the rest of this paper, we support it via another different angle by experiments again.

3. A SIMPLIFIED LEARNED PARAMETRIC MIXTURE ALGORITHM

In order to let each $g_j(y^{(j)})$ loosely matching a source density, we need to use a flexible density function form for $g_j(y^{(j)})$. For this purpose, a finite mixture of densities is used in [14, 15, 16, 17, 18]. The finite mixture model is a general tool for density estimation, which has been widely used in the literature, e.g., see [6] as well as those in the Ref. List of [21, 20]. The special use in [14, 15, 16, 17, 18] is for forming a flexible parametric learnable $g_j(y^{(j)})$ such that it can loosely match a source density.

It deserves to mention that the use of a mixture of logistic densities for modeling $g_j(y^{(j)})$ has been previously proposed in 1995 by the first author of the present paper and implemented in a joint paper with his colleague under the name of entropy maximization [9] as well as in 1996 by .Pearlmutter and Parra [13] under the name of maximum likelihood density estimation. However, it should be noted that both the previous uses are different from each other and also different from the one used in [14, 15, 16]. *First*, the one [1] given in [14, 15, 16] is suggested with a clear motivation obtained in April 1996: *we can use a flexibly adjusting density form to loosely match a source density*, instead of attempting to estimate as accurately as possible a density by a finite mixture, which is implicitly granted in [9] and [13][2]. The significance of this loosely matching motivation can be further understood in the rest of this paper. *Second*, there is some detailed differences in the algorithms given in [9], [13] and [14, 15, 16]. *Third*, the bias parameter of logistic function used in [13] is modeled by an auto-regression which is more general than simply using a constant in [9] and [14, 15, 16]. Moreover, the learned parametric mixtures used in [14, 15, 16] includes other densities such as gaussian and also the learning is implemented by the well known EM algorithm, instead of gradient technique used in [9] and [13].

For the convenience of comparing with the previous results, here we just introduce the one used in [17, 18]. That is,

$$g_j(y_j) = \sum_{i=1}^{p_j} \alpha_{ji} q(u_{ji}), \quad u_{ji} = b_{ji}(y_j - a_{ji}),$$
$$\sum_{i=1}^{p_j} \alpha_{ji} = 1, \ \ \alpha_{ji} = \frac{\exp(\gamma_{ji})}{\sum_{k=1}^{p_j} \exp(\gamma_{ik})}, \tag{8}$$

where $q(\cdot)$ is some density function, p_j is the number of components in the mixture, α_{ji} is the weight of the component, b_{ji} controls the variant of the j^{th} density and a_{ji} is the bias, or location of the center, of the j^{th} density.

The density function is given by

$$\begin{aligned} q(u_{ji}) &= b_{ji} h'(u_{ji}), \\ h(u_{ji}) &= logsig(u_{ji}) = \frac{1}{1 + \exp(-u_{ji})}, \\ h'(u_{ji}) &= \frac{\exp(-u_{ji})}{(1 + \exp(-u_{ji}))^2} \end{aligned} \tag{9}$$

[1] In fact, the basic idea of the methods given in [14, 15, 16] was first proposed in April 1996 during a two weeks visit of the first author to the third author's Lab and presented in a formal seminar in RIKEN, Japan in that period.

[2] Actually, in [13], the use of a mixture of logistic densities was not discussed in its regular text part, but adopted directly in its Appendix A for its simulation in Sec.4.

The gradient algorithm is used to adapt parameters as follows:

$$\Delta\gamma_{ji} \propto \frac{1}{g_j(y_j)} \sum_{k=1}^{p_j} b_{ik} h'(u_{ik}) \alpha_{ik} (\delta_{kj} - \alpha_{ji}),$$
$$\Delta b_{ji} \propto \frac{\alpha_{ji}}{g_j(y_j)} \{h'(u_{ji}) + h''(u_{ji}) u_{ji}\},$$
$$\Delta a_{ji} \propto -\frac{1}{g_j(y_j)} \alpha_{ji} b_{ji}^2 h''(u_{ji}),$$
$$for\ h(u_{ji}) = logsig(u_{ji}),$$
$$q'(u_{ji}) = \frac{(\exp(-u_{ji}) - 1)}{(1 + \exp(-u_{ji}))^3},$$
$$\phi_j(y_j) = \frac{1}{g_j(y_j)} \sum_{i=1}^{p_j} \alpha_{ji} b_{ji} q'(u_{ji}) \qquad (10)$$

where δ_{ji} is the Kroniker delta function. The above are updated together with the learning eq.(3).

In [17, 18], the experiments demonstrated that the algorithm with mixture of densities with $p_j = 7$ can approximate the marginal densities 'quite well' and perform separation in all experiments tried. In those experiments, samples were mixed from the bi-modal beta distribution $\beta(0.5, 0.5)$ in [-0.5,0.5], uniformly distribution in [-1,1] and a permuted speech signal, with the first two being sub-gaussian and the last one being super-gaussian. However, th algorithm 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 uniformly distribution in [-1,1]; while the algorithm in [1] works well for the bi-modal beta distribution $\beta(0.5, 0.5)$ in [-0.5,0.5] and the uniformly distribution in [-1,1], but fails for the permuted speech signal.

In [17, 18], the number of components $p_j = 7$ was arbitrarily chosen. From the conjecture statement given in Sec.2, we can easily see that it is able to change the kurtosis sign of $g_j(y_j) = 0.5 \sum_{i=1}^{2} q(u_{ji}), u_{ji} = y_j - a_{ji}$ from positive to negative by just adjusting the position parameter a_{ji}. Therefore, if the the conjecture statement given in Sec.2 is true, this simple mixture should be already enough for our purpose. That is, we can set $p_j = 2$, $\alpha = 0.5, b_{ji} = 1$. Thus, the algorithm eq.(10) is simplified into $\Delta a_{ji} \propto -\frac{1}{g_j(y_j)} 0.5 q''(u_{ji})$. We have tried it on all the experiments made in [17, 18] with successes.

In the following we introduce one example of three channel mixture. The first is an artificially generated bi-modal 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:

$$A = \begin{bmatrix} 1 & 0.6 & 0.3 \\ 0.8 & 1 & 0.3 \\ 0.4 & 0.9 & 1 \end{bmatrix} \qquad (11)$$

After the system has stabilized, a snapshot of V and a are:

$$V = \begin{bmatrix} 10.2583 & 0.0205 & -0.1169 \\ -0.0085 & 5.0408 & -0.0813 \\ -0.0132 & -0.0095 & 9.3977 \end{bmatrix},$$
$$a = \begin{bmatrix} -3.3378 & 3.3657 \\ -2.4460 & 2.4672 \\ 0.0241 & 0.0241 \end{bmatrix} \qquad (12)$$

From this V, we see that three channels have been successfully separated with signal/noise ratio being around 1000. Some detailed experiment results are given in Fig.1. The first row given are the densities (histograms) of the three sources. The second row given are the obtained $\{g_j(y_j)\}$, which are far from the densities (histograms) except that their kurtosis signs are the same as those given in the first row. This fact actually provides a further support of our conjecture assertion made in Sec.2. The third row given are the CDF functions $s(y_j) = \int_{-\infty}^{y_j} g_j(y_j) dy_j$ and the last row given are the histograms of the nonlinear transformation $z_j = s(y_j)$ as did in [2], which are again quite different from the histograms of the first row, although they are roughly similar in their configurations.

4. CONCLUSIONS

An interesting conjecture assertion has been proposed. According to this assertion, a source can be separated from others as long as there is one j such that the product of the kurtosis of source and the kurtosis of $g_j(y^{(j)})$ is positive. As a whole, all the sources can be separated as long as there at least one way to pair all the sources and $\{g_j(y_j)\}$ such that for each pair the corresponding source and $g_j(y^{(j)})$ have a positive product in their kurtosis. This assertion has not only been supported by several experiments on the typically used $g_j(y_j)$ but also applied to simplify the previous proposed learned parametric mixture algorithm significantly, which is again verified by experimental successes. This latter fact also provides a further support for the proposed conjecture assertion. Currently, further work is still undergoing on seeking a mathematical proof of this conjecture assertion.

5. REFERENCES

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

[2] A.J. Bell and T.J. Sejnowski, "An information-maximatization approach to blind separation and blind deconvolution", *Neural Computation* 7 (1995) 1129-1159.

[3] J.F.Cardoso and B.Laheld, "Equivalent Adaptive source separation", *IEEE Tans. Signal Processing*, Vol.44, pp3017-3030, 1996.

[4] C.C. Cheung and L. Xu, " Separation of Two Independent Sources by the Information-theoretic Approach with Cubic Nonlinearity", *Proc. of 1997 IEEE International Conference on Neural Networks (IEEE-INNS IJCNN97)*, June 9-12, 1997, Houston, TX, USA, Vol.4, pp2239-2244.

[5] C.C. Cheung, " Adaptive Blind Signal Processing", Master Thesis, Dept of Computer Science and Engineering, The Chinese University of Hong Kong, June, 1997.

[6] Dempster, A., Laird, N. M., & Rubin, D. B. "Maximum-likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society, B*, **39** (1), 1-38, (1977).

[7] M.Gaeta and J.-L Lacounme, " Source Separation without a priori knowledge: the maximum likelihood solution", in *Proc. European Signal Processing Conf EUSIPCO90*, pp621-624, 1990.

[8] C.Jutten, "From source separation to Independent component analysis: An introduction to special session", Invited special session on Blind Signal Separation, *Proc. of 1997 European Symp. on Artificial Neural Networks*, Bruges, April 16-18, pp243-248.

[9] I. King and L. Xu, "Adaptive Contrast Enhancement by Entropy Maximization with a 1-K-1 Constrained Network", *Proc. 1995 Intl Conf. on Neural Information Processing (ICONIP95)*, Oct 30 - Nov. 3, 1995, Beijing, Vol. II, pp703-706.

[10] J.-P. Nadal and N.Parga, "Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer", *Network 5*, 1994, 565-581.

[11] E.Oja, Karhunen, J. and Hyvärinen, A, "From neural principal components to neural independent components", *Proc. Int. Conf. on Artificial Neural Networks ICANN'97*, October 8 - 10, Lausanne, Switzerland, pp. 519 - 528 (1997).

[12] D.T.Pham, P. Garat a nd C. Jutten, "Separation of a mixture of independent sources through a maximum likelihood approach", in *Singal Processing VI: Theories and Applications*, J. Vandewalle et al (eds), Elsevier Science Publishing, 1992, pp771-774.

[13] B.A.Pearlmutter and L.C.Parra, "A Context-Sensitive generalization of ICA", in *Progress in Neural Information Processing: Proc. Intl. Conf. on Neural Information Processing (ICONIP 96)* (Hong Kong, Sept 24-27, 1996; Springer-Verlag: Singapore 1996) 1235-1239.

[14] L. Xu, "Bayesian Ying-Yang System and Theory as A Unified Statistical Learning Approach: (I) Unsupervised and Semi-Unsupervised Learning", Invited paper, S. Amari and N. Kassabov eds., *Brain-like Computing and Intelligent Information Systems*, 1997, Springer-Verlag, pp241-274.

[15] 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", Invited paper, *Lecture Notes in Computer Science, Proc. of Intl Workshop on Theoretical Aspects of Neural Computation,*, May 26-28, Hong Kong, Springer-Verlag, 1997, pp25-60.

[16] L. Xu, "Bayesian Ying-Yang Learning Based ICA Models" , *Proc. 1997 IEEE Workshop on Neural Networks and Signal Processing*, 24-26 Sept., Florida, (1997).

[17] L. Xu, C.C. Cheung, H.H. Yang and S.-I. Amari, " Independent component analysis by the information-theoretic approach with Mixture of Density ", *Proc. of 1997 IEEE Intl. Conf on Neural Networks (IEEE-INNS IJCNN97)*, June 9-12, Houston, TX, USA, Vol. III, pp1821-1826(1997).

[18] L. Xu, C.C. Cheung, J. Ruan, and 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. of 1997 European Symp. on Artificial Neural Networks*, Bruges, April 16-18, pp291-296(1997).

[19] L. Xu and S.-I. Amari, "A general independent component analysis framework based on Bayesian-Kullback Ying-Yang Learning", in *Progress in Neural Information Processing: Proc. Intl. Conf. on Neural Information Processing (ICONIP 96)* (Hong Kong, Sept 24-27, 1996; Springer-Verlag: Singapore 1996) 1235-1239.

[20] L. Xu and M.I. Jordan, " On convergence properties of the EM algorithm for Gaussian mixtures" , *Neural Computation 8*, (2), 129–151(1996).

[21] L. Xu and M.I. Jordan, " Unsupervised learning by EM algorithm based on finite mixture of Gaussians", *Proc. of WCNN'93*, Portland, OR, Vol. II, 431-434.

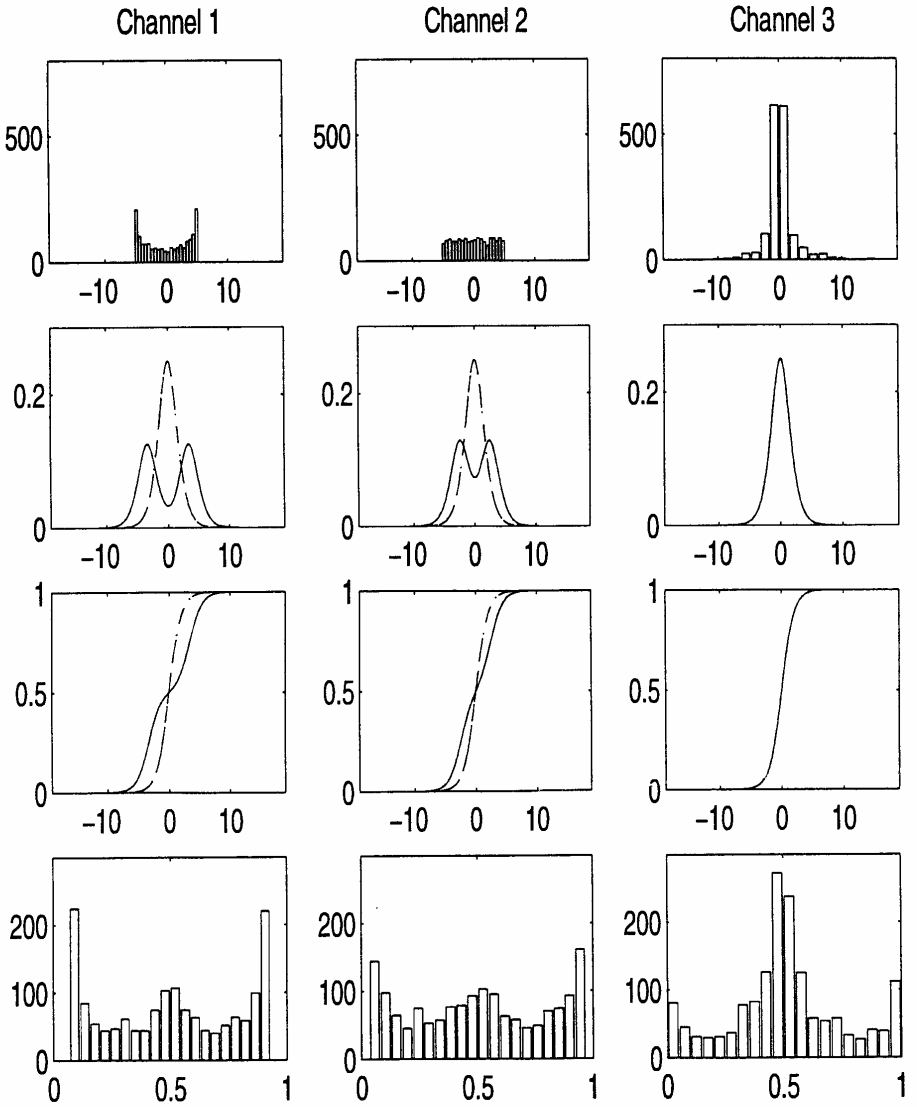

Figure 1: The experiment result of the proposed algorithm with mixture of two densities.