ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 291-296

# Nonlinearity and Separation Capability:
## Further Justification for the ICA Algorithm with A Learned Mixture of Parametric Densities*

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

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

**Abstract.** We discuss the relation between nonlinearity and separation capability in the information-theoretic ICA scheme. We propose with justification that a 'loose matching' between the nonlinearity and source distribution is needed. These results give further support to the implementation technique by a learned mixture of parametric densities.

## 1 Introduction

Nonlinearity is an essential element in adaptive ICA algorithms since it picks up and controls some high order statistics. This issue was previously discused in the maximum likelihood approach preoposed in [5]. In the information-theoretic ICA approaches (e.g., MMI, INFORMAX)[4, 1, 2, 9, 11], the choice of nonlinearity is also a critical issue. Actually, it determines on which class of source distributions the ICA algorithm can work. In contrary to 'strict matching' proposed in previous works [1, 2], we propose that only a 'loose matching' is needed between the nonlinearity and source distribution, justified by the theoretical and experimental analysis on several cases. Also, these results support the use of technique of learning a flexible mixture of parametric densities in implementation† [10, 11].

## 2 Problem and the information-theoretic ICA scheme

Suppose there are $n$ unknown independent *sources* $\mathbf{s} = [s_1, \dots, s_n]^T$ with $E\mathbf{s} = \mathbf{0}$. The sources are mixed by an unknown static nonsingular *mixing matrix* $\mathbf{A}$ as $\mathbf{x} = \mathbf{A}\mathbf{s}$. Given only the *observed signals* $\mathbf{x}$, the ICA problem is to determine the *de-mixing matrix* $\mathbf{W}$ which gives the *recovered signals* $\mathbf{y} = \mathbf{W}\mathbf{x}$, such that $\mathbf{y}$ resembles $\mathbf{s}$ as far as possible. Theoretically $\mathbf{s}$ can only be

*This project was supported by the HK RGC Earmarked Grants CUHK250/94E and CUHK484/95E and by Ho Sin-Hang Education Endowment Fund for Project HSH 95/02.

†On one reviewing feedback of the present paper, it is mentioned that a paper in French on SRETSI95 by Pham used mixture of densities via Parzen estimation for a block MMI ICA algorithm. We are sorry to be unable to make clear comments here since we are not clear the source SRETSI95 and also unfortunely can not read French, and thus are not sure what kind of that algorithm exactly is. From that piece of message, seemly the densities in that mixture are nonparametric estimations based on the observations and can not be changed together with the change of the de-mixing matrix to optimize the MMI criterion. Differently, the key point of our approach[11] is the used of a flexible mixture of parametric densities with their parameters learned together with the learning of the de-mixing matrix to optimize the MMI criterion.

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 291-296

determined up to an arbitrary permutation and scaling. That is, if we obtain $\mathbf{V} = \mathbf{WA} = \mathbf{PD}$, where $\mathbf{D}$ is a diagonal matrix and $\mathbf{P}$ is a permutation matrix, separation is said to be achieved.

Recently, a general information-theoretic ICA scheme has been suggested [9, 11] from the YING-YANG Learning Scheme [7, 8]. With $\{g_i(r)\}$ used to model the scale families of pdf's of the sources $\{p_{s_j}(s_j)\}$, the following cost function is formulated:

$$\begin{aligned} J(\mathbf{W}) &= \int_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{|\det[\mathbf{W}]| \prod_{i=1}^{n} g_i(\mathbf{w}_i^T \mathbf{x})} d\mathbf{x} \\ &= \int_{\mathbf{s}} p(\mathbf{s}) \log \frac{p(\mathbf{s})}{|\det[\mathbf{V}]| \prod_{i=1}^{n} g_i(\mathbf{v}_i^T \mathbf{s})} d\mathbf{s} = J(\mathbf{V}) \end{aligned} \tag{1}$$

The natural gradient decent algorithm [1] is used to perform $\min_{\mathbf{W}} J(\mathbf{W})$:

$$\Delta \mathbf{W} \propto [\mathbf{I} + \mathbf{h}(\mathbf{y})\mathbf{y}^T]\mathbf{W} \tag{2}$$

where $\mathbf{h}(\mathbf{y}) = [h_1(y_1), \ldots, h_n(y_n)]^T$, $h_i(y_i) = g_i'(y_i)/g_i(y_i)$ and $g_i(y_i) = f_i'(y_i)$.

## 3 Nonlinearity and Separation Capability

The separation capability of the algorithm is determined by $\{h_i(y_i)\}$, which follows from the choice of $\{g_i(y_i)\}$. If $\{g_i(y_i)\}$ models the scale families of $\{p_{s_j}(s_j)\}$ appropriately, the system can perform separation. If $g_i(y_i)$ is designated to be equal to $p_{y_i}(y_i)$, $J(\mathbf{W})$ will reduce to the mutual information[4, 1, 6]

$$J(\mathbf{W}) = \int_{\mathbf{y}} p_y(\mathbf{y}) \log \frac{p_y(\mathbf{y})}{\prod_{i=1}^{n} p_{y_i}(y_i)} d\mathbf{y} \tag{3}$$

The minimization of this $J(\mathbf{W})$ can always yield a correct solution $\mathbf{W}$ because $J = 0$ when $p_y(\mathbf{y}) = \prod_{i=1}^{n} p_{y_i}(y_i)$. Hence, theoretically $g_i(y_i) = p_{y_i}(y_i)$ can work on any source distribution but this choice bears some implementation difficulty as $p_{y_i}(y_i)$ is not known in advance.

On the other hand, it has been proposed recently that the use of a set of *pre-fixed* $g_i(y_i)$ may also separate sources with *a particular class* of distribution [3, 11]. We consider the following cases:

(i) In [2], $f_i(y_i)$ are chosen to be $\text{logsig}(y_i) = 1/[1 + exp(-y_i)]$, etc, and are shown to be able to separate sources with sharply peaked super-gaussian pdf. In experiments it works on human speech signals [2] but fails on uniformly or beta(0.5,0.5) distributed signals, which are sub-gaussian [11].

(ii) A more general choice for $f_i(y_i)$ is $\tilde{f}_i(\tilde{y}_i) = \text{logsig}(b\tilde{y}_i) = 1/[1+exp(-b\tilde{y}_i)]$ where the steepness $b$ is a positive real number. However, we can easily prove:

**Lemma** Consider an information-theoretic ICA system A with $\tilde{f}_i(\tilde{y}_i) = \text{logsig}(b\tilde{y}_i)$ and a system B with $f_i(y_i) = \text{logsig}(y_i)$. $\tilde{\mathbf{V}} = \mathbf{V}/b$ is a solution

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 291-296

of the equilibrium equation $\nabla_{\widetilde{W}} J(\widetilde{\mathbf{W}}) = \mathbf{0}$ for system A if and only if $\mathbf{V}$ is a solution of this same equilibrium equation for system B.

Which says that $b$ is just an arbitrary scaling factor for the measuring unit of $\mathbf{y}$ and cannot affect the properties of the nonlinearity.

(iii) In [3], $h_i(y_i)$ is directly chosen as $h_i(y_i) = c_i y_i^3$ with $c_i < 0$. It has been theoretically proved the system can separate two sub-gaussian sources but cannot separate two super-gaussian sources.

(iv) **THEOREM 1** *Consider the case $h_1(y_1) = c_{11} y_1$ and $h_2(y_2) = c_{23} y_2^3$ with $c_{11} < 0$ and $c_{23} < 0$ acting on two channels of signals. If:*

*(a) One source is sub-gaussian and one source is super-gaussian, or*

*(b) One source is gaussian and one source is non-gaussian,*

*for any initial value, $\mathbf{V}$ will converge to and stay stably at one of the following eight correct solutions for signal separation:*

$$\textit{Solution } \mathbf{A}_{\mathrm{I}}\textit{:}\quad \mathbf{V} = \begin{bmatrix} \pm(-c_{11} E[s_1^2])^{-\frac{1}{2}} & 0 \\ 0 & \pm(-c_{23} E[s_2^4])^{-\frac{1}{4}} \end{bmatrix} \tag{4}$$

$$\textit{Solution } \mathbf{A}_{\mathrm{II}}\textit{:}\quad \mathbf{V} = \begin{bmatrix} 0 & \pm(-c_{11} E[s_2^2])^{-\frac{1}{2}} \\ \pm(-c_{23} E[s_1^4])^{-\frac{1}{4}} & 0 \end{bmatrix} \tag{5}$$

*such that the resulting $y_2$ recovers the channel of $s$ that has a flatter pdf.*

**Proof** The equilibrium equation for the algorithm is $\nabla_W J(\mathbf{W}) = [\nabla_V J(\mathbf{V})]\mathbf{A}^{-1} = 0$, which implies (provided that $\det \mathbf{V} \neq 0$):

$$E[\mathbf{I} + \mathrm{h}(\mathbf{Vs})(\mathbf{Vs})^T] = \mathbf{0} \tag{6}$$

The equations for the non-diagonal elements can be written as:

$$\begin{bmatrix} \mu_1^2 & \mu_2^2 \\ v_{21}^2 \mu_1^4 + 3 v_{22}^2 \mu_1^2 \mu_2^2 & v_{22}^2 \mu_2^4 + 3 v_{21}^2 \mu_1^2 \mu_2^2 \end{bmatrix} \begin{bmatrix} v_{11} v_{21} \\ v_{12} v_{22} \end{bmatrix} = \mathbf{0} \tag{7}$$

where $\mu_i^p = E[s_i^p]$. Denote the left matrix in eq. (7) as M, then $\det \mathbf{M} = v_{22}^2 \mu_1^2 (\mu_2^4 - 3[\mu_2^2]^2) - v_{21}^2 \mu_2^2 (\mu_1^4 - 3[\mu_1^2]^2)$. Under the stated condition, we have $\det \mathbf{M} \neq 0$, and hence $[v_{11}v_{21}\ v_{12}v_{22}]^T = \mathbf{0}$. Coping it with the equations for the diagonal elements of eq. (6), we get solution groups $\mathrm{A_I}$ and $\mathrm{A_{II}}$ exhaustively.

For Solution group $\mathrm{A_I}$, the Hessian matrix $\nabla_V^2 J(\mathbf{V})$ is negative definite (stable) if $s_2$ is sub-gaussian, negative semi-definite (stability not determined) if $s_2$ is gaussian and neither negative/positive definite/semi-definite (saddle point) if $s_2$ is super-gaussian. Similarly, for Solution group $\mathrm{A_{II}}$, $\nabla_V^2 J(\mathbf{V})$ is negative definite if $s_1$ is sub-gaussian, negative semi-definite if $s_1$ is gaussian and neither negative/positive definite/semi-definite if $s_1$ is super-gaussian. It can be shown that there is no local maxima in $J(\mathbf{V})$ and that on singular subspace $\det \mathbf{V} = 0$, $J(\mathbf{V}) \to +\infty$ as there is deterministic linear dependancy between channels.

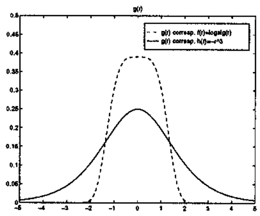

Figure 1: Solid: g(r)=exp(-r)/(1+exp(-r))$^2$ , corresp. to $f(r) = logsig(r)$ in case (i). Dash: g(r)=($\pi/\gamma(3/4)$) exp(-r$^4$/4), corresp. to $h(r) = -r^3$ in case (iii).

Thus, $J(\mathbf{V})$ is monotonic increasing around the local minima, as $v_{ij} \to \pm\infty$, $J(\mathbf{V}) \to +\infty$.

Hence, we have *global convergence* to the stable solutions as follows:

| $s_1$ | $s_2$ | Stable Solution | $y_1$ | $y_2$ |
|---|---|---|---|---|
| super-gaussian | sub-gaussian | $A_I$ | $s_1$ | $s_2$ |
| sub-gaussian | super-gaussian | $A_{II}$ | $s_2$ | $s_1$ |
| gaussian | sub-gaussian | $A_I$ | $s_1$ | $s_2$ |
| sub-gaussian | gaussian | $A_{II}$ | $s_2$ | $s_1$ |
| super-gaussian | gaussian | $A_I$ | $s_1$ | $s_2$ |
| gaussian | super-gaussian | $A_{II}$ | $s_2$ | $s_1$ |

In all cases, the pdf of $y_2$ is flatter than that of $y_1$. □

In figure 1, the $g_i(y_i)$ in case (i) is more sharply peaked (have greater kurtosis) and the $g_i(y_i)$ in case (iii) is flatter. The fact that the $g_i(y_i)$ in case (i) cannot separate signals with flat pdf and the $g_i(y_i)$ in case (iii) cannot separate super-gaussian signals suggests that some matching of $\{g_i(y_i)\}$ to the scale families of $\{p_{s_j}(s_j)\}$ is needed. However, the fact that one fixed $g_i(y_i)$ can work on a broad class of source distribution suggests that the matching needed is not so strict. Hence, these results suggest that only a 'loose matching' between $\{g_i(y_i)\}$ and the scale families of $\{p_{s_j}(s_j)\}$ is needed. In case (iv), the cubic nonlinearity in channel 2 selects the $s_i$ with flatter pdf to recover. This fact further supports the suggestion of 'loose matching'.

## 4 Implementation with mixture of densities

A flexible mixture of parametric densities is suggested to achieve the loose matching [10, 11]:

$$g_i(y_i) = \sum_{j=1}^{p_i} \alpha_{ij}\psi(u_{ij}), \quad u_{ij} = b_{ij}(y_i - a_{ij}) \quad \alpha_{ij} = \frac{\exp(\gamma_{ij})}{\sum_{k=1}^{p_i} \exp(\gamma_{ik})} \tag{8}$$

with $\sum_{j=1}^{p_i} \alpha_{ij} = 1$ and $\psi(\cdot)$ being some density function in the form of $\psi(u_{ij}) = b_{ij}\phi'(u_{ij})$ and $\phi(u_{ij}) = \text{logsig}(u_{ij})$. Thus, we have:

$$h_i(y_i) = \frac{1}{g_i(y_i)} \sum_{j=1}^{p_i} \alpha_{ij} b_{ij} \psi'(u_{ij}) \tag{9}$$

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 291-296

which is substituted into eq. (2) as the algorithm for $\mathbf{W}$. Together with eq. (2), the parameters $\{\gamma, \mathbf{a}, \mathbf{b}\}$ of $g_i(y_i)$ are also learned to minimize the $J(\mathbf{W})$ given by eq.(3) via the following descending algorithm :

$$\Delta\gamma_{ij} \propto \frac{1}{g_i(y_i)} \sum_{k=1}^{p_i} b_{ik}\phi'(u_{ik})\alpha_{ik}(\delta_{kj} - \alpha_{ij}), \tag{10}$$

$$\Delta b_{ij} \propto \frac{\alpha_{ij}}{g_i(y_i)}\{\phi'(u_{ij}) + \phi''(u_{ij})u_{ij}\}, \ \Delta a_{ij} \propto -\frac{1}{g_i(y_i)}\alpha_{ij}b_{ij}^2\phi''(u_{ij}) \tag{11}$$

## 5 Experiment

As shown in Figure 2, three channels of signals are used: samples from bimodal beta distribution beta(0.5, 0.5) in [-0.5,0.5], uniformly distribution in [-1,1] and a permuted speech signal. They are mixed with the mixing matrix:

$$\mathbf{A} = \begin{bmatrix} 1 & 0.6 & 0.2 \\ 0.8 & 1 & 0.3 \\ 0.4 & 0.9 & 1 \end{bmatrix} \tag{12}$$

In the simulation with the learned mixture of parametric densities with $p_i =$ 5, all sources are successfully separated, where all $\gamma_{ij}$ and $a_{ij}$ are initialized as 1/5 and 0 respectively. $b_{i1}, \cdots, b_{i5}$ are initialized in the interval $[10^{-0.3}, 10^{1.2}]$. The histograms of $y_i$ and $z_i$, and the shape of $g_i(y_i)$ and $f_i(y_i)$ are plotted in Figure 2. The simulation with $f_i(y_i) = \text{logsig}(y_i)$ can only separate the speech signal but failed on the other two sub-gaussian signals as did in [11].

## 6 Conclusion

The relation between the nonlinearity and separation capability is discussed and a 'loose matching' requirement is proposed. Cases on different situation have been presented to support this proposal. This justification can support the the technique of learning a flexible mixture of parametric densities for implementation.

# References

[1] S.-I. Amari, A. Cichocki, H. Yang, A new learning algorithm for blind separation of sources, in *Advances in Neural Information Processing 8* 757-763, (1996).

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

[3] C.C. Cheung and L. Xu, Independent component analysis on two channels by the information-theoretic approach with cubic nonlinearity, submitted to *Neurocomputing*.

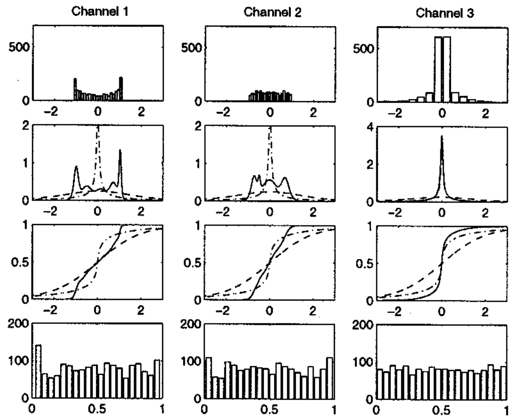


Figure 2: Result of the experiment. Row 1: histograms of $y_i$. Row 2 & 3: $g_i(y_i)$ and $f_i(y_i)$ respectively. (— adapted mixture of densities, –·– initial, – – $f_i(\cdot) = \text{logsig}(\cdot)$ for comparison.) Row 4: histograms of $z_i$.

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

[5] D.T.Pham, P. Garat and 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.

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

[7] L. Xu, A unified learning scheme: Bayesian-Kullback YING-YANG Machine, in *Advances in Neural Information Processing 8* 444-450, (1996).

[8] L. Xu, Bayesian-Kullback YING-YANG Machine: reviews and new results, in *Progress in Neural Information Processing: Proc. Intl. Conf. on Neural Information Processing (ICONIP 96)* 59-67 (1996).

[9] L. Xu and S.-I. Amari, A general independent component analysis framework based on Bayesian-Kullback Ying-Yang Learning, in the above same source, pp1235-1239 (1996).

[10] L. Xu, C.C. Cheung, H.H. Yang and S.-I. Amari, "Maximum Equalization by Entropy Maximization and Mixture of Cumulative Distribution Functions", to appear on *Proc. of 1997 IEEE International Conference on Neural Networks/*.

[11] L. Xu, C.C. Cheung, J. Ruan and S.-I. Amari, A general information-theoretic ICA scheme with an implementation technique by a learned mixture of parametric densities, submitted to a Journal.