AN ADAPTIVE ENTROPY OPTIMIZATION ALGORITHM FOR BLIND SOURCE SEPARATION

Yi-Xiang Wang

Communication Research Laboratory
McMaster University, Hamilton, Ontario Canada L8S 4K1

ABSTRACT

Independent Component Analysis (ICA) or Blind Signal Separation (BSS) has become an increasing important research field due to its rapidly growing applications in various areas, such as telecommunication systems, sonar and radar systems, audio and acoustics, image enhancement and biomedical signal processing. In this paper, first a novel adaptive ICA (AICA) entropy optimization algorithm for finding pairs of simplified activation functions (SAF) will be introduced. Then the theoretical explanation is described. Finally we discuss the algorithm with a few of existing representative methods. It is worthy noting that the experimental simulation results prove the effective performance on separating signals for the algorithm.

1. INTRODUCTION

There have been a lot of papers discussing activation functions (AF) or contrast functions because they greatly influence the performance of ICA algorithms. Compared with those correlation-based algorithms such as Principal Component Analysis (PCA), ICA not only separates the second-order statistical signals but also reduces high-order statistical dependencies, attempting to make the separated signals as independent as possible.

Unsupervised learning rules were proposed to maximize the mutual information (MMI) between the inputs and the outputs of a neural network [1]. It was showed that in the low-noise case, MMI implied that the output probability density function (p.d.f.) can be factorized as a product of marginal p.d.f.s [2]. The stochastic gradient learning algorithms for MMI were derived [3] [4]. The BSS problem was put into an information-theoretic framework and demonstrated the separation and deconvolution of mixed sources [4] [5].

It showed that the infomax and the maximum likelihood estimation approaches are equivalent [6] [7]. The original infomax learning rule for blind separation [4] was suitable for super-Gaussian sources. [8] derived, by choosing negentropy as a projection pursuit index, a learning rule that is able to blindly separate not only mixed sub-Gaussian (sub-G) but also super-Gaussian (super-G) source distributions. This learning algorithm was showed to be an extension of the infomax principle satisfying a general stability criterion and preserving the simple architecture [9]. Natural gradient and relative gradient simplified the learning rules by eliminating the complex matrix inversion [10] [5]. Simulations and results on real-world physiological data showed the power of the proposed methods [11].

It is noted that [12] has developed an on-line learning algorithm which minimizes statistical dependence among outputs for blind separation of mixed signals. The dependence here is measured by the average mutual information (MI) of the outputs. The improved ICA (IICA) algorithm [13] extended the work by adopting a more precise AF.

2. PROBLEM STATEMENT

Assume that source signals $\mathbf{s}(t) = [s_1(t), ..., s_n(t)]^T$ are unknown and their components are, at each time instant, mutually statistically independent:

$$p(\mathbf{s}(t)) = \prod_{i=1}^{n} p(s_i(t)) \tag{1}$$

These source signals are also assumed to be stationary processes and each source has moments of any order with zero mean. Let $\mathbf{x}(t) = [x_1(t), ..., x_n(t)]^T$ be sensor signals, which is a linear instantaneous mixture of the source signals:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{2}$$

where $\mathbf{A} \in \mathbf{R}^{n \times n}$ is an unknown mixing matrix of full rank. The demixing model here is a linear transformation:

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t) \tag{3}$$

where $\mathbf{y}(t) = [y_1(t), ..., y_n(t)]^T$ is an output vector, $\mathbf{W} \in \mathbf{R}^{n \times n}$ is a demixing matrix to be identified. The system model is assumed to be noise-free. The number of source signals is the same as the number of sensors.

Note that our assumptions in the paper are the same as

[14] [5]. At most one source in the source signals is permitted to be Gaussian distribution. Sensor noise is considered to be an independent source itself before mixing.

Although there exist many ICA algorithms, their only objective is to find a precise way of weight update to recover the original sources. Let us view an effective weight update formula derived by [12] using Kullback-Leibler divergence (KLD) and entropy principles.

$$\frac{d\mathbf{W}}{dt} = \eta(t)\{\mathbf{W}^{-T} - f(\mathbf{y})\mathbf{x}^T\} \tag{4}$$

Substituting $\mathbf{y} = \mathbf{W}\mathbf{x}$ and $\mathbf{x}^T\mathbf{W}^T = \mathbf{y}^T$, we get

$$\frac{d\mathbf{W}}{dt} = \eta(t)\{\mathbf{I} - f(\mathbf{y})\mathbf{y}^T\}\mathbf{W}^{-T} \tag{5}$$

and using natural gradient (NG), we get the on-line algorithm

$$\frac{d\mathbf{W}}{dt} = \eta(t)(\mathbf{I} - f(\mathbf{y})\mathbf{y}^T)\mathbf{W} \tag{6}$$

where activation function (AF) $f(\mathbf{y}) = H(\mathbf{y}; k_i)$ is a vector which includes $f(y_i)$ and depends on the cumulants k_i and is necessary to compute the corresponding entropy.

3. SYMMETRY OF NON-GAUSSIAN

From our simulations we can prove that IICA algorithm [13] has very good performance in separating source signals blindly. However, we note that the activation function of IICA is too complex and we are interested in developing a simpler and more practical algorithm. Also the IICA algorithm works best for sub-G signals, and may fail for super-G signals. At the same time, we hope the algorithm design for AF can be canonical and solve both the super-G and sub-G problems. In this section we present a new approach on how to design a pair of ICA non-linear AFs which can work effectively not only for sub-G, but also for super-G signals. First let us introduce the following theorem.

Theorem 1: Consider a stochastic gradient on-line learning algorithm. If y is a random variable, $f_i(y)(i = 1, 2)$ are AFs, and $\phi(y)$ is a smooth non-linear interval monotonic odd function of y, then $f_i(y) = y - (-1)^i\phi(y)(i = 1, 2)$ is a non-linear super-G (sub-G) AF.

Proof. Let $p_i(y)(i = 1, 2)$ be two probability distribution functions of y, respectively, as shown by

$$-\frac{\partial \log p_i(y)}{\partial y} = y - (-1)^i\phi(y), (i = 1, 2) \tag{7}$$

Integrating the above equations, we get

$$p_i(y) = e^{-\frac{1}{2}y^2}e^{(-1)^i\int \phi(y)dy}, (i = 1, 2) \tag{8}$$

Note that $\int \phi(y)dy$ is an even function since $\phi(y)$ is defined to be odd. In addition, consider $e^{\int \phi(y)dy} > 1$, $\forall \int \phi(y)dy > 0$, it should have

$$p_1(y) < e^{-\frac{1}{2}y^2} < p_2(y) \tag{9}$$

The middle term of the above equation is the well known Gaussian p.d.f. with zero-mean and unit-variance. Therefore, $f_1(y)$ corresponding to $p_1(y)$ is a super-G AF, and $f_2(y)$ corresponding to $p_2(y)$ is a sub-G AF. If $\int \phi(y)dy < 0$, we have the reverse situation. The symmetry of super-G and sub-G, however, is always unchanged. ♢

Consider more general situations, we have this lemma:

Lemma 1: Consider a stochastic gradient on-line learning algorithm. If y is a random variable, $f_i(y)(i = 1, 2)$ are AFs, constant $a \in (0, 1]$ and $\phi(y)$ is a smooth non-linear interval monotonic odd function of y, then $f_i(y) = (-1)^{i+1}[(a + (-1)^{i+1} - 1)y + \phi(y)], (i = 1, 2)$ is a non-linear super-G (or sub-G) AF.

Proof. Using the same integration method in Theorem 1, we get:

$$p_i(y) = e^{-\frac{1}{2}y^2}e^{(-1)^{i+1}(\frac{1-a}{2}y^2 + \int \phi(y)dy)}, i = 1, 2. \tag{10}$$

Note that only the difference exists in the second term of the above equation; thus we get the following conclusions

$$p_1(y) < p(y) = e^{-\frac{1}{2}y^2} < p_2(y) \tag{11}$$

or

$$p_1(y) > p(y) = e^{-\frac{1}{2}y^2} > p_2(y) \tag{12}$$

The first p.d.f. relation equation tells us that $f_1(y)$ is a super-G function and $f_2(y)$ is a sub-G function, and vice versa for the second p.d.f. relation equation. Here the symmetrical feature of the two non-Gaussian p.d.f. is again proved. ♢

We also note that if $a = 1$, $p_1(y)$ and $p_2(y)$ in Lemma 1 are the same as that in Theorem 1. The above theorem and lemma can be easily expanded in vector form.

4. SIMPLIFIED ACTIVATION FUNCTIONS

The basic idea for simplifying AFs is that a non-Gaussian AF can be realized through combining a linear function and a non-linear function. First, if we choose y as the linear function, it is the easiest and natural choice. However, this choice reflects that the corresponding p.d.f. in probability space is Gaussian with zero-mean ($\mu = 0$) and one-variance ($\sigma^2 = 1$). Second, we have some choices for non-linear functions such as $\tanh(y)$, $\frac{1}{1+\exp(-ay)}$, and so on. Here we choose $\tanh(y)$ because of its good sigmoid and odd symmetrical properties. Then we use a polynomial $\hat{f(y)} = \sum_{i=1}^{n} a_i y^i$ to fit the data sets $\{y_k, \hat{f}_k(y)\}_{k=1}^{m}$ of the function $f(y) = y - \tanh(y)$(or $f(y) = y + \tanh(y)$).

The Taylor expansion is a useful approximation tool.

It can be represented approximately by an finite series of terms centered about a given point y_0 as follows: where $f^{(n)}(y_0)$ is the nth derivate of y at y_0.

$$\begin{aligned} f(y) &= \sum_{n=0}^{N} \frac{f^{(n)}(y_0)}{n!}(y-y_0)^n + R_N \\ &\approx \sum_{n=0}^{N} \frac{f^{(n)}(y_0)}{n!}(y-y_0)^n \end{aligned} \quad (13)$$

where $f^{(n)}(y_0)$ is the nth derivate of y at y_0, and R_N is a remainder assumed to be small.

The difference between a Taylor series representation and a general polynomial is that the Taylor series utilizes derivative information that is localized [15]. We can use Least Mean Square (LMS) method to best fit the data sets. Let error $\epsilon = f(y) - \sum_i^N a_i y^i$. We then look for minimizing

$$E[\epsilon\epsilon^T] = E[(f(y) - \sum_i^N a_i y^i)^2] \quad (14)$$

The minimum can be found by setting the partial derivatives with respect to $a_0, a_1, \cdots, a_N$ equal to zero. We thus get the normal equations $\mathbf{YA} = \mathbf{f}$ [15], where

$$\mathbf{Y} = \begin{pmatrix} N & \sum y_i & \cdots & \sum y_i^{N-1} \\ \sum y_i & \sum y_i^2 & \cdots & \sum y_i^N \\ \vdots & \vdots & \ddots & \vdots \\ \sum y_i^{N-1} & \sum y_i^N & \cdots & \sum y_i^{2(N-1)} \end{pmatrix} \quad (15)$$

$\mathbf{A} = (a_0, \cdots, a_N)^T$, $\mathbf{f} = (\sum f_i, \sum y_i f_i, \cdots, \sum y_i^{N-1} f_i)^T$. let $\mathbf{Y}$ be full rank. Then it is not difficult to solve the equation to get $a_0, a_1, \cdots, a_N$ through $\mathbf{A} = \mathbf{Y}^{-1}\mathbf{f}$. Using Theorem 1 and Lemma 1, we may find pairs of non-Gaussian AFs and simplify them. Let us view the following example.

A non-Gaussian AF of the fifth-order can be easily gotten through LMS polynomial fitting with actual real time signal.

$$f_1(y) = 0.0023y + 0.3107y^3 - 0.0755y^5 \quad (16)$$

Then using the above theorem and lemma, the other symmetrical non-Gaussian AF can be gotten immediately as

$$f_2(y) = 1.9977y - 0.3107y^3 + 0.0755y^5 \quad (17)$$

SAF is tested as sub-G AF and it works surprisingly well for separating source signals blindly through the on-line learning simulations. Table 1 lists its performance index (PI) value err_3 compared with that of err_1 [16] and the extended infomax ICA (EICA) err_2 [9]. The separation performance is evaluated through error measure err(·) [12] that is equal to:

$$\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\frac{|p_{ij}|}{max_k|p_{ik}|} - 1\right) + \sum_{j=1}^{N}\left(\sum_{i=1}^{N}\frac{|p_{ij}|}{max_k|p_{kj}|} - 1\right) \quad (18)$$

Table 1. PI values of activation functions

TEST	1	2	3	4	5	6	7
$\mathbf{err_1}$	1.37	1.97	1.77	2.26	1.37	2.26	1.77
$\mathbf{err_2}$	1.79	1.80	2.34	2.16	1.79	2.16	2.34
$\mathbf{f_{13}(y)}$	2.13	1.11	1.78	2.04	2.13	2.04	1.78
$\mathbf{f_{15}(y)}$	2.15	1.06	1.63	2.33	2.15	2.33	1.63

$$= \sum_{i=1}^{N}\left(\frac{\sum_{j=1,j\neq k}^{N}|p_{ij}|}{max_k|p_{ik}|}\right) + \sum_{j=1}^{N}\left(\frac{\sum_{i=1,i\neq k}^{N}|p_{ij}|}{max_k|p_{kj}|}\right) \quad (19)$$

where performance matrix $\mathbf{P} = (p_{ij})_{i,j} = \mathbf{WA}$. Note that the inputs are three sources and $\mathbf{A}$ is a random mixing matrix that we do not need to know.

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} 0.0680 & -0.0703 & 0.0726 \\ -0.0520 & -0.0049 & 0.0992 \\ 0.0955 & 0.0756 & 0.0515 \end{pmatrix} \\ s_1(t) &= 0.1\sin(400t)\cos(30t) \\ s_2(t) &= 0.01\text{sign}\{\sin(500t + 9\cos(40t))\} \\ s_3(t) &= 2\text{rand}(1, \text{length}(t)) - 1.0 \end{aligned} \quad (20)$$

and f_s = 10K Hz, N=200, η=0.1. We get $\text{err}_1 = 1.7730$, $\text{err}_2 = 2.3364, \text{err}_3 = 1.6344$.

In Table 1, $\mathbf{f_{13}(y)}$ denotes a third-order sub-G AF, and $\mathbf{f_{15}(y)}$ denotes a fifth-order sub-G AF. We have done more than fifteen simulations; however, we have found that the group PI results are almost the same for each four simulations if we ignore very small computational errors. The repeated PI values of two groups are shown in the table. The repeating order is arbitrary. We believe that when the algorithm converges, the local optima of these functions may have optimized the network parameter $\mathbf{WA}$ and then trapped PI into limited fixed values.

SAF and corresponding p.d.f. are compared with sub-G and super-G [11].The simulation results are excellent, especially for values of $y \in [-1, 1]$. The fitting errors are almost equal to zero. The small values of the first and the second moments indicate the fitting efficiency [17]. In fact, in our simulations and calculations y values are always small enough since the sensor signals (including uniform noise) are weak and the mixing weights are set to be low.

In more compact notation, the AFs can be rewritten as one formula, with a sign function of kurtosis switching between sub-G and super-G, as

$$f(y) = y + K(y - (ay + \phi(y))) \quad (21)$$

When $K = 1$, the formula switches to super-G, and when $K = -1$, it switches to sub-G. If there exist $K = -1$ and $\phi(y) = 0$ at the same time, then we have $f(y) = ay$, the SAF is used for Gaussian. Thus the fifth-order SAF formula of the example can be expressed as

$$f(y) = y - K(-0.9977y + 0.3107y^3 - 0.0755y^5) \quad (22)$$

where $K = \pm 1$ switches between super-G and sub-G, depending on the kurtosis measure.

What on-line learning is looking for includes a nonlinear function, from which the actual nonlinearity can be well described from a package of real time signal. The varying ratio of Gaussian in the differential sub-space is a line with the direction determined by constant a. Those absolute values of high varying ratio called super-G are always greater than ay since it must respond to some sharp and powerful information such as voice and music signals with noise environment. This is why its curve of p.d.f. always lies below the Gaussian because if it shares the same peak point with sub-G, it can get bigger varying ratio only by hiding under the Gaussian, vice versa.

From the above equations it is easy to get $y = (f_1(y) + f_2(y))/2$, it means that y is equal to the arithmetical average of super-G and sub-G. Therefore, it is sufficient for us to say that super-G and sub-G are symmetrical with respect to each other with the linear axis $f(\text{Gaussian}) = ay$.

If the precision can be accepted, the simplified polynomial approximation for the AFs may give us a new exploratory direction in analytical BSS because it has many advantages such as the ease of use and simple to differentiate and integrate. It also has the advantages of odd and even transformation with each differentiation (or integration) and excellent one-to-one correspondence through differentiation and integration.

5. CONCLUSIONS

In this paper a new simplified adaptive ICA algorithm has been proposed for the solution of BSS problem. The method is proved to be effective through our theoretical analysis, calculation results and experimental simulations.

6. REFERENCES

[1] R. Linsker, "Deriving receptive fields using an optimal encoding criterion," *Advances in Neural Information Processing Systems*, vol. 5, pp. 953-960, 1993.

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

[3] Z. Roth and Y. Baram, "Multi-dimensional density shaping by sigmoids," *IEEE Trans. on Neural Networks*, vol. 7, pp. 1291-1298, 1996.

[4] A.J. Bell and T.J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129-1159, 1995.

[5] J.-F. Cardoso and B.H. Laheld, "Equivariant adaptive source separation," *IEEE Trans. on Signal Processing*, vol. 44, no. 12, pp. 3017-3030, 1996.

[6] B.A. Pearlmutter and L.C. Parra, "Maximum likelihood blind source separation: a context-sensitive generalization of ica," *Advances in Neural Information Processing Systems*, vol. 9, pp. 613-619, 1997.

[7] J.-F. Cardoso, "Informax and maximum likelihood for source separation," *IEEE Letters on Signal Processing*, vol. 4, pp. 112-114, 1997.

[8] M. Girolami and C. Fyfe, "An extended exploratory projection pursuit network with linear and non-linear anti-Hebbian connections applied to the cocktail party problem", *Neural Networks*, vol. 10, pp. 1607-1618, 1998.

[9] T.-W. Lee, B.U. Koehler, and R. Orglmeister, "Blind source separation of nonlinear mixing models," *Neural Networks for Signal Processing VII*, pp. 406-415, 1997.

[10] S.-I. Amari, "Neural learning in structured parameter spaces - natural Riemannian gradient," *Advances in Neural Information Processing Systems*, vol. 9, pp. 127-133, 1997.

[11] T.-W. Lee, M. Girolami, and T.J. Sejnowski, "Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources," *Neural Computation*, vol. 11, pp. 417-441, 1999.

[12] S.-I. Amari, A. Cichocki, and H.H. Yang, "A new learning algorithm for blind separation," *Advances in Information Processing Systems*, vol. 8, pp. 757-763, 1996.

[13] S. Haykin and P. Gupta, "On the implementation of an idealized activation function for blind source separation," *Internal Report, Communication Research Laboratory, McMaster University, Canada*, 1999.

[14] P. Comon, "Independent Component Analysis - a new concept?," *Signal Processing*, vol. 36, pp. 287-314, 1994.

[15] P.Y. Papalambros, *Principles of Optimal Design*, Cambridge University Press, 1988.

[16] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *Proc. IEE Proc.-F*, vol. 140, no. 6, pp. 362-370, 1993.

[17] M.R. Spiegel and L.J. Stephens, *Statistics*, McGraw-Hill, 3rd, ed., 1999.