Equi-Convergence Algorithm for Blind Separation of Sources with Arbitrary Distributions

L.-Q. Zhang, S. Amari and A. Cichocki

Brain-style Information Systems Research Group
RIKEN Brain Science Institute
Wako shi, Saitama 351-0198, JAPAN
http://www.bsp.brain.riken.go.jp

Abstract. This paper presents practical implementation of the equi-convergent learning algorithm for blind source separation. The equi-convergent algorithm [4] has favorite properties such as isotropic convergence and universal convergence, but it requires to estimate unknown activation functions and certain unknown statistics of source signals. The estimation of such activation functions and statistics becomes critical in realizing the equi-convergent algorithm. It is the purpose of this paper to develop a new approach to estimate the activation functions adaptively for blind source separation. We propose the exponential type family as a model for probability density functions. A method of constructing an exponential family from the activation (score) functions is proposed. Then, a learning rule based on the maximum likelihood is derived to update the parameters in the exponential family. The learning rule is compatible with minimization of mutual information for training demixing models. Finally, computer simulations are given to demonstrate the effectiveness and validity of the proposed approach.

1 Introduction

Blind source separation or independent component analysis [14, 12] introduces a novel paradigm for signal processing and has attracted considerable attention in signal processing society. A number of neural networks and statistical methods [14, 12, 11, 5, 16, 9, 10, 17, 18] have been developed for blind signal separation. There are a number of factors which are likely to affect separation results in applications, such as the number of active sources, the distributions of source signals, time-variable mixtures and noise.

It is the main purpose of this paper to develop a learning algorithm with certain uniform convergence in the sense that all components converge to true solution at the same rate regardless of what types of the distributions the source signals have. Such an equi-convergent algorithm has been proposed in [4]. However, because the algorithm includes unknown activation functions and certain statistics of source signals, its practical implementation still remains open. The

problem attributes how to estimate the statistics of source signals. If the activation functions are not suitably chosen, the learning algorithm will not converge to the true solution. Thus the online estimators of the statistics of the source signals might not be accurate enough to realize the equi-convergence algorithm.

Stability of learning algorithms [4,10], is critical to successful separation of source signals from measurements. One way is to estimate those statistics adaptively [17,13]. Some other statistical models, such as the Gaussian mixture model [8,15] are also employed to estimate the distributions of source signals. Usually, such methods are computing demanding.

In contrast to the previous works on the estimation of the distributions for source signals, this paper attempts to avoid to estimate source distributions, but rather to online adapt activation functions for source signals. The adaptation of activation functions has two purposes: one is to modify the activation functions such that the true solution is the stable equilibrium of the learning system; the second is to estimate the sparseness of source signals. This simplification makes it easy to estimate the parameters in the generative models and reduce computing cost. Usually, the adaptation of activation functions needs only a very few parameters for each component. There are some advantages by using the exponential family to estimate activation functions. It is easy to reveal the relation between the distributions and activation functions (score functions). And also we can easily construct a linear connection of the score functions for the exponential family if we want to separate signals with specific distributions. Another advantage of the approach is its compatibility, i.e. both the updating rules for the demixing model and for the free parameters in the distribution models make the cost function decrease to its minimum, if the learning rate is sufficiently small.

Assume that source signals are stationary zero-mean processes and mutually statistically independent. Let $\mathbf{s}(k) = (s_1(k), \cdots, s_n(k))^T$ be the vector of unknown independent sources and $\mathbf{x}(k) = (x_1(k), \cdots, x_m(k))^T = \mathbf{A}\mathbf{s}(k) + \mathbf{v}(k)$ a sensor vector, which is a linear instantaneous mixture of the sources with additive noises $\mathbf{v}(k)$. The blind separation problem is to recover original signals $\mathbf{s}(k)$ from observations $\mathbf{x}(k)$ without prior knowledge on the source signals and the mixing matrix, but the assumption of mutual independence of source signals. The demixing model here is a linear transformation of the form

$$\mathbf{y}(k) = \mathbf{W}\mathbf{x}(k), \tag{1}$$

where $\mathbf{y}(k) = (y_1(k), \cdots, y_m(k))^T, \mathbf{W} \in \mathbf{R}^{m \times m}$ is the demixing matrix to be determined during training. Assume that $m \geq n$, i.e. the number of the sensor signals is larger than the number of the source signals. We intend to train the demixing model $\mathbf{W}$ such that n components are designed to recover the n source signals and the rests correspond to the zeros or noise.

2 Learning algorithms

The estimation of demixing model $\mathbf{W}$ can be formulated in the framework of a semiparametric statistical model [3]. The probability density function of $\mathbf{x}$ can

be expressed as

$$p_X[\mathbf{x}, \mathbf{W}, p(\mathbf{s})] = |\det(\mathbf{W})| p(\mathbf{W}\mathbf{x}), \tag{2}$$

which depends on two unknowns separating matrix $\mathbf{W} = \mathbf{A}^{-1}$ and pdf $p(\mathbf{s})$. Here, the statistical model (2) includes not only an unknown matrix $\mathbf{W}$ to be estimated but also an unknown pdf function $p(\mathbf{s})$ which we do not need to estimate. Such a model is said to be semi-parametric. The semi-parametric model theory suggests to use an estimating function to estimate demixing matrix $\mathbf{W}$. In blind source separation, the estimating function is a matrix function $\mathbf{F}(\mathbf{y}, \mathbf{W})$, which does not depend on $p(\mathbf{s})$, provided it satisfies certain regularity conditions which can be found in [6]. Amari and Kawanabe [6] proposed an information geometric theory of estimating functions by extending the differential geometry of statistics [1], and gave the set of all the estimating functions. It is also proved that the effective part of the off-diagonal elements of estimating functions F_{ij} for $(i \neq j)$ is spanned by the functions of the form $\varphi(y_i)y_j$ and $\varphi(y_j)y_i$ and the diagonal part by $f(y_i) = \psi(y_i)y_i - 1$, where φ and ψ are arbitrary functions. The optimal one for φ and ψ is to choose $\varphi_i(y) = -\frac{d}{dy_i}\log p_i(y_i) = -\frac{\dot{p}_i(y_i)}{p_i(y_i)}$, if we can estimate the true source probability distribution $p_i(y_i)$ adaptively. From the theory of estimating functions, we have the general form of estimating functions [3]

$$\mathbf{F}(\mathbf{y}, \mathbf{W}) = \boldsymbol{\mathcal{K}}(\mathbf{W}) \circ [\mathbf{I} - \boldsymbol{\varphi}(\mathbf{y})\mathbf{y}^T]\mathbf{W}, \tag{3}$$

where $\varphi(\mathbf{y}) = [\varphi_1(y_1), \ldots, \varphi_m(y_m)]^T$ is an arbitrary vector function and $\boldsymbol{\mathcal{K}}(\mathbf{W})$ is an arbitrary linear operator which maps matrices to matrices. An online learning algorithm based on the estimating function is described by

$$\mathbf{W}(k+1) = \mathbf{W}(k) - \eta(k)\mathbf{F}(\mathbf{y}, \mathbf{W}). \tag{4}$$

Different linear operator $\boldsymbol{\mathcal{K}}(\mathbf{W})$ leads to different existing algorithms such as the natural gradient algorithm [5]. It should be noted that different algorithms have different stability regions. Therefore, the choice of the nonlinear activation functions and the linear operator $\boldsymbol{\mathcal{K}}(\mathbf{W})$ is vital to successful separation of source signals.

2.1 Equi-convergence Learning Algorithm

Given an estimating function, learning algorithm (4) may have different convergence rate for different component $y_i(k)$ of $\mathbf{y}(k)$. It is the purpose of this paper to present an algorithm such that all the components of the output $\mathbf{y}(k)$ have the same convergence rate in the sense of

$$\frac{\partial E[\mathbf{F}(\mathbf{y}, \mathbf{W})]}{\partial \mathbf{W}} = \boldsymbol{\mathcal{I}}, \tag{5}$$

where $\boldsymbol{\mathcal{I}}$ is the $n^2 \times n^2$ identity matrix, that is, the Hessian of the learning algorithm is the identity. This implies that algorithm (4) converges equally well in

any directions. In paper [4], a universally convergent learning algorithm was developed with an estimating function $\mathbf{F}(\mathbf{y}, \mathbf{W}) = \mathbf{G}(\mathbf{y})\mathbf{W}$, where $\mathbf{G}(\mathbf{y}) = [g_{ij}(\mathbf{y})]$ is given by

$$g_{ij} = \frac{1}{\sigma_i^2 \sigma_j^2 \kappa_i \kappa_j - 1} \left\{ \varphi(y_j) y_i - \sigma_j^2 \kappa_i \varphi(y_i) y_j \right\}, \qquad i \neq j \tag{6}$$

$$g_{ii} = \frac{1}{m_i + 1} \left\{ \varphi(y_i) y_i - 1 \right\}, \tag{7}$$

where $\sigma_i^2 = E[y_i^2]$, $\kappa_i = E[\dot{\varphi}_i(y_i)]$, $m_i = E[y_i^2 \dot{\varphi}_i(y_i)]$, where $\dot{\varphi} = d\varphi/dy$. However, the estimating function includes the unknown activation functions and a number of unknown statistics σ_i, κ_i, m_i, which depend on the source signals. How to estimate these statistics becomes the key problem to realize the equi-convergence learning algorithm. In this paper, we suggest to estimate both the activation functions and the statistics σ_i, κ_i, m_i. The motivation is if the activation functions are not well chosen, the algorithm may not converge to the true solution and the online estimators for the statistics will be not accurate enough to realize the equi-convergence algorithm.

3 Exponential Family

In order to model the probability distributions of the source signals, we suggest to use an exponential family [7], which is expressed in term of certain functions $\{C(y), \psi_1(y), \cdots, \psi_N(y)\}$ as $p(y, \boldsymbol{\theta}) = \exp\left[-C(y) - \sum_{i=1}^N \theta_i \psi_i(y) + \mathcal{N}(\boldsymbol{\theta})\right]$, $\boldsymbol{\theta}$ is the vector of free parameters, and $\mathcal{N}(\boldsymbol{\theta})$ is a normalization term such that the integral of $p(y, \boldsymbol{\theta})$ over the whole interval $(-\infty, \infty)$ is equal to one. There are some good properties in the exponential family, such as *flatness*, as a statistical model. Refer to [7] for further properties of the exponential family.

We provide here a feasible way to construct the exponential family for blind source separation. First, we define a parametric family for activation functions $\varphi(y, \boldsymbol{\theta}) = \sum_{i=1}^N \theta_i \varphi_i(y)$, where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_N)^T$ are the parameters to be determined adaptively and φ_i are certain activation functions. From the definition of activation functions, $\varphi(y, \boldsymbol{\theta}) = -\frac{\partial \log p(y, \boldsymbol{\theta})}{\partial y}$, the probability density function is equivalently given by

$$p(y, \boldsymbol{\theta}) = \exp\left\{-\boldsymbol{\theta}^T \boldsymbol{\psi}(y) + \mathcal{N}(\boldsymbol{\theta})\right\}, \tag{8}$$

where $\mathcal{N}(\boldsymbol{\theta})$ is the normalization term and $\boldsymbol{\psi}(y) = (\int_0^y \varphi_1(\tau) d\tau, \cdots, \int_0^y \varphi_N(\tau) d\tau)$.

In blind separation, it is well known that the hyperbolic tangent function $\varphi(y) = tanh(y)$ is a good activation function for super-Gaussian sources and the cubic function $\varphi(y) = y^3$ is a good activation function for sub-Gaussian sources. Thus, we define the exponential family by linearly combining the three activation functions in the form of $\varphi(y, \boldsymbol{\theta}) = \theta_1 \alpha \ tanh(\alpha y) + \theta_2 4\beta \ y^3 + \theta_3 \ y$, where $\alpha = \frac{\pi}{2}$, $\beta = (\Gamma(0.75)/\Gamma(0.25))^2$. The purpose of such choice of α and β is to ensure

the distributions have the uni-variance. The corresponding exponential family is expressed as

$$p(y, \boldsymbol{\theta}) = \exp\left(\theta_1 \log sesh(\alpha y) - \theta_2 \beta y^4 - \theta_3 \frac{y^2}{2} + \mathcal{N}(\boldsymbol{\theta})\right). \tag{9}$$

In this exponential family, $\boldsymbol{\psi}(y) = (-\log sesh(\frac{\pi}{2}y), \beta y^4, \frac{y^2}{2})^T$ and $\mathcal{N}(\boldsymbol{\theta})$ is the normalization term. This exponential family includes three typical distributions: the super-Gaussian, sub-Gaussian and Gaussian distributions. In particular, the exponential family covers the homotopy family $p(y, \theta, 1-\theta, 0)$ which connects the following two density functions $\frac{sech(\alpha y)}{2}$ and $exp(-\beta y^4 + \mathcal{N}(0, 1, 0))$. Figure 1 shows the waveform of the homotopy family varying θ from -1 to 1. This family will be used to model the distributions of source signals in this paper.

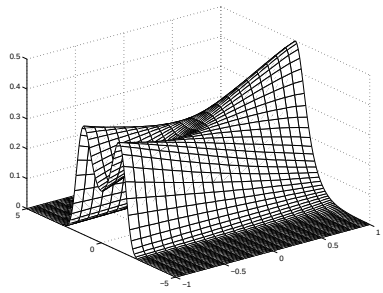

Fig. 1. The waveform of the homotopy family varying θ from -1 to 1

4 Adaptation of Activation Functions

Assume that $q_i(y_i, \boldsymbol{\theta}_i) = \exp\left\{-\boldsymbol{\theta}_i^T \boldsymbol{\psi}(y_i) + \mathcal{N}(\boldsymbol{\theta}_i)\right\}$ is a model for the marginal distribution of y_i, $(i = 1, \cdots, m)$. Various approaches such as entropy maximization and minimization of mutual information lead to the following cost function,

$$l(\mathbf{y}, \boldsymbol{\theta}, \mathbf{W}) = -\log(|det(\mathbf{W}|) - \sum_{i=1}^{m} \log q_i(y_i, \boldsymbol{\theta}_i), \tag{10}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \cdots, \boldsymbol{\theta}_m^T)^T$ is the parameters determined adaptively. The main purpose of the adaptation of activation functions is to modify the activation function such that the true solution is the stable equilibrium of learning dynamics.

4.1 Natural Gradient Learning

By minimizing the cost function (10) with respect to $\boldsymbol{\theta}$ by using the gradient descent approach, we derive learning algorithms for training parameters $\boldsymbol{\theta}$. When

the parameter space has a certain underlying structure, the ordinary gradient of a function does not represent its steepest direction. The steepest descent direction in a Riemannian space is given by the natural gradient [2], which takes the form of

$$\tilde{\nabla}_\theta l(\mathbf{y}, \boldsymbol{\theta}, \mathbf{W}) = \mathcal{G}^{-1} \nabla_\theta l(\mathbf{y}, \boldsymbol{\theta}, \mathbf{W}) \tag{11}$$

where $\mathcal{G}$ is the Riemannian metric of the parameterized space, $\nabla_\theta l(\mathbf{y}, \boldsymbol{\theta}, \mathbf{W}) = (\frac{\partial l}{\partial \boldsymbol{\theta}_1}^T, \cdots, \frac{\partial l}{\partial \boldsymbol{\theta}_m}^T)^T$ and $\frac{\partial l}{\partial \boldsymbol{\theta}_i} = -\boldsymbol{\psi}(y_i) + \mathcal{N}'(\boldsymbol{\theta}_i)$. The Riemannian structure of the parameter space of statistical model $\{q_i(y_i, \boldsymbol{\theta}_i)\}$ is defined by the Fisher information [1]

$$\mathcal{G}_i(\boldsymbol{\theta}_i) = E\left[\frac{\partial \log q_i(y_i, \boldsymbol{\theta}_i)}{\partial \boldsymbol{\theta}_i} \frac{\partial \log q_i(y_i, \boldsymbol{\theta}_i)^T}{\partial \boldsymbol{\theta}_i}\right] \tag{12}$$

in the component form. The learning algorithm based on the natural gradient descent approach is described as

$$\boldsymbol{\theta}_i(k+1) = \boldsymbol{\theta}_i(k) - \eta(k) \mathcal{G}_i^{-1}(\boldsymbol{\theta}_i(k)) \frac{\partial l(\mathbf{y}, \boldsymbol{\theta}, \mathbf{W})}{\partial \boldsymbol{\theta}_i}. \tag{13}$$

From the cost function (10), we see that the minimization of mutual information is equivalent to maximizing the likelihood for parameters $\boldsymbol{\theta}_i$ because the first term in (10) does not depend on $\boldsymbol{\theta}_i$. Thus it should be noted that the above learning rule is actually equivalent to the maximum log-likelihood rule for each component. Both learning rules for updating parameters $\boldsymbol{\theta}$ and the demixing model $\mathbf{W}$ make cost function $L(\boldsymbol{\theta}, \mathbf{W}) = E[l(\mathbf{y}, \boldsymbol{\theta}, \mathbf{W})]$ smaller and smaller, provided the learning rate is sufficiently small.

5 Implementation of Equi-Convergence Algorithm and Simulation

In this section, we present how to implement the equi-convergence algorithm and give computer simulations to demonstrate the effectiveness and the performance of the proposed approach.

In order to implement the equi-convergence learning algorithm, we need to estimate the statistics of the output signals, which estimate source signals. The statistics σ_i, κ_i, m_i are evaluated by the following iterations

$$\kappa_i(k+1) = (1-\mu)\kappa_i(k) + \mu \dot{\varphi}(y_i(k), \boldsymbol{\theta}_i(k)), \tag{14}$$

where μ is a learning rate. The other statistics are estimated in a similar way.

We choose $\mathbf{F}(\mathbf{y}, \mathbf{W}) = \mathbf{G}(\mathbf{y})\mathbf{W}$, defined by (6) and (7), as the estimating function. Therefore, the equi-convergent algorithm (4) is realized by the following two stages: the first stage is to train demixing matrix $\mathbf{W}$ using the natural gradient algorithm (4), to train $\boldsymbol{\theta}_i$ using algorithm (13) and simultaneously, to estimate statistics σ_i, κ_i, m_i adaptively. After certain times iteration, we shift the

learning algorithm to the equi-convergence algorithm (4). In order to remove the fluctuation caused by the non-stationarity of source signals, we use the averaged version of the algorithm (4) over certain time windows.

A large number of simulations have been done to show the validity and performance of the proposed algorithm. Here, we give one simulation example. The five source signals consist of two super-Gaussian, two sub-Gaussian and one Gaussian signals. We set the initial values $\theta = 0.5$ for all the five components. The mixing matrix is randomly generated by computer. For each batch iteration, we take 20 sample data as the output signals $\mathbf{y}(k)$. Figure 2 shows on-line estimation of $\mathbf{y}(k)$ and $\boldsymbol{\theta}$ by using the equi-convergent learning algorithm. It can be proved that the parameter θ converges to 1 for the super-Gaussian signal with distribution $sech(\alpha\ y)/2$ and to 0 for the sub-Gaussian signal with distribution $exp(-\beta|y|^4 + \mathcal{N}(0,1,0))$.

The computer simulations show that the proposed approach can separate both super-Gaussian and sub-Gaussian source signals simultaneously without presuming any knowledge on the source signals.

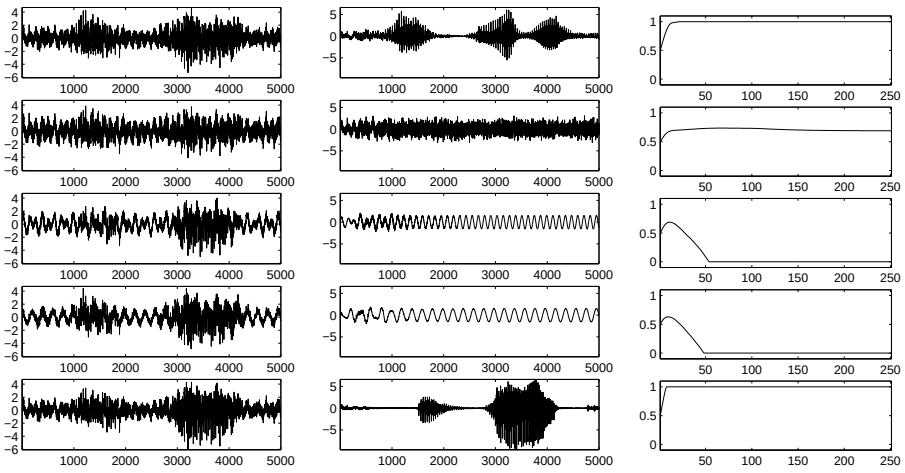

Fig. 2. a) The sensor signals; b) Online estimation of $\mathbf{y}(k)$; c) Online estimation of $\theta(k)$

6 Conclusion

In this paper, we present a new approach to realize the equi-convergence learning algorithm for blind source separation. An exponential family is employed as a model for the distributions of the source signals. A adaptation rule is developed for updating the parameters in the model of distributions of source signals. The

equi-convergence algorithm is implemented by updating the demixing matrix and the parameters in the distributions simultaneously. The equi-convergence algorithm has two favorite properties: the isotropic convergence and universal convergence. Computer simulations verify such properties.

References

1. S. Amari. *Differential–geometrical methods in statistics, Lecture Notes in Statistics*, volume 28. Springer, Berlin, 1985.
2. S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998.
3. S. Amari and J.-F. Cardoso. Blind source separation– semiparametric statistical approach. *IEEE Trans. Signal Processing*, 45:2692–2700, Nov. 1997.
4. S. Amari, T. Chen, and A. Cichocki. Stability analysis of adaptive blind source separation. *Neural Networks*, 10:1345–1351, 1997.
5. S. Amari, A. Cichocki, and H.H. Yang. A new learning algorithm for blind signal separation. In G. Tesauro, D.S. Touretzky, and T.K. Leen, editors, *Advances in Neural Information Processing Systems 8 (NIPS*95)*, pages 757–763, 1996.
6. S. Amari and M. Kawanabe. Information geometry of estimating functions in semiparametric statistical models. *Bernoulli*, 3(1):29–54, 1997.
7. S. Amari and H. Nagaoka. *Methods of Information Geometry*. AMS and Oxford University Press, 2000.
8. H. Attias. Indepedent factor analysis. *Neural Computatation*, 11(4):803–851, 1999.
9. A.J. Bell and T.J. Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.
10. J.-F. Cardoso and B. Laheld. Equivariant adaptive source separation. *IEEE Trans. Signal Processing*, SP-43:3017–3029, Dec 1996.
11. A. Cichocki and R. Unbehauen. Robust neural networks with on-line learning for blind identification and blind separation of sources. *IEEE Trans Circuits and Systems I : Fundamentals Theory and Applications*, 43(11):894–906, 1996.
12. P. Comon. Independent component analysis: a new concept? *Signal Processing*, 36:287–314, 1994.
13. S. Douglas, A. Cichocki, and S. Amari. Multichannel blind separation and deconvolution of sources with arbitrary distributions. In *Proc. of NNSP'97*, pages 436–445, Florida, US, September 1997.
14. C. Jutten and J. Herault. Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.
15. T. Lee and M. Lewicki. The generalized gaussian mixture model using ica. In P. Pajunen and J. Karhunen, editors, *Proc. ICA'2000*, pages 239–244, Helsinki, Finland, June 2000.
16. E. Oja and J. Karhunen. Signal separation by nonlinear hebbian learning. In M. Palaniswami, Y. Attikiouzel, R. Marks II, D. Fogel, and T. Fukuda, editors, *Computational Intelligence - A Dynamic System Perspective*, pages 83–97, New York, NY, 1995. IEEE Press.
17. W. Lee T, M. Girolami, and T. Sejnowski. Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources. *Neural Computation*, 11(2):606–633, 1999.
18. L. Zhang, A. Cichocki, and S. Amari. Natural gradient algorithm for blind separaiton of overdetermined mixture with additive noise. *IEEE Signal Processing Letters*, 6(11):293–295, 1999.