*4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), April 2003, Nara, Japan*

# INDEPENDENT COMPONENT ANALYSIS WITH JOINT SPEEDUP AND SUPERVISORY CONCEPT INJECTION: APPLICATIONS TO BRAIN fMRI MAP DISTILLATION

*Yasuo Matsuyama, Ryo Kawamura, and Naoto Katsumata*

Department of Computer Science, School of Science and Engineering,
Waseda University, Tokyo 169-8555 Japan
yasuo2@waseda.jp, ryo@wizard.elec.waseda.ac.jp, katsu@ruri.waseda.jp

## ABSTRACT

Methods to combine speedup terms and supervisory concept injection are presented. The speedup is based upon iterative optimization of the convex divergence. The injection of supervisory information is realized by adding a term which reduces an additional cost for a specified concept. Since the convex divergence includes usual logarithmic information measures, its direct application gives faster algorithms than existing logarithmic methods. This paper first shows a list of newly obtained general properties of the convex divergence. Then, these properties are used to derive faster algorithms for the independent component analysis. Then, an additional term for incorporating supervisory information is introduced. The efficiency of the total algorithm is tested using a set of real data - - brain fMRI time series. Successful results in view of convergence speed, software complexity, and extracted brain maps are reported. Finally, another class of the convex divergence optimization, the $\alpha$-EM algorithm, is commented upon.

## 1. INTRODUCTION

Computing and optimizing information measures comprise many important problems both in theory and in applications. Independent component analysis (ICA) [1] is has dual aspects: It is theoretically interesting due to its semiparametric nature, and it is rich in applications due to its independence of physical entity. This paper covers both of such aspects.

Usually, the target information measure for optimization is based upon logarithm [2] and [3]. But, the information measure to be optimized in this paper is the convex divergence [4]. Since the convex divergence includes usual logarithmic information measures as special cases, we can expect better performance than the logarithmic ones. In the problem of the ICA, the merit appears in convergence speed without losing the algorithm's flexibility. This is a featuring aspect of this paper in theory.

The other side of a coin, the application aspect, is related to the brain functional MRI analysis (fMRI) [5]. Since the derived algorithm using the convex divergence maintains flexibility to create variants, an injection of supervisory information [6] is possible. Therefore, the organization of this paper becomes as follows. Section II gives basic properties of the convex divergence and their relationships to the extended class of logarithm. Section III gives a formulation of the independent component analysis as a minimization of the convex divergence. Then, concrete algorithms are derived. On the convergence speed, the proposed method is faster than traditional or logarithmic methods. This is examined in Section IV through brain fMRI map distillation. The separation of brain map's active areas is quite successful using the supervisory information. Section V gives general remarks on the use of the convex divergence. Other problems coined into the convex divergence minimization, e.g., expectation-maximization are commented on.

## 2. PROPERTIES OF THE CONVEX DIVERGENCE

### 2.1. Definition and Differential Properties

The convex divergence, or f-divergence [4] (as its forerunner, Eq. (4.20) of [7]), is defined as follows. Let $\psi$ and $\varphi$ be generic parameters for probability density functions. The convex divergence between two probability densities $p_\psi$ and $p_\varphi$ is defined by the following equation.

$$\begin{aligned} D_f(\psi\|\varphi) &= \int_{\mathcal{Y}} p_\varphi(y) f(p_\psi(y)/p_\varphi(y))dy \\ &= \int_{\mathcal{Y}} p_\psi(y) g(p_\varphi(y)/p_\psi(y))dy \\ &= D_g(p_\varphi\|p_\psi) \geq g(1) = f(1). \end{aligned} \quad (1)$$

Here, $\mathcal{Y}$ is chosen to be a $N$-dimensional Euclidian space. The function $f(r)$ is convex for $r \in (0, \infty)$. Its dual func-

This work was supported by the Productive ICT Academia Program in the 21st Century COE Programs, and by the Grant-in-Aid for Scientific Research.

tion $g(r)$ is defined by

$$g(r) = rf(1/r), \tag{2}$$

which is also convex for $r \in (0, \infty)$. We normalize the constant $f(1) = 0$. Then, the convex divergence is zero if and only if $p_\psi(y) = p_\varphi(y)$, $y$-a.e.

We consider the case that $f(r)$ is twice continuously differentiable. Let $\partial^{ij}$ mean that $i$-times partial differentiation with respect to $\psi$ and $j$-times partial differentiation with respect to $\varphi$. Then, we have the following relationships.

$$D_f(\varphi\|\varphi) = 0, \tag{3}$$
$$\partial^{10} D_f(\varphi\|\varphi) = 0, \tag{4}$$
$$\partial^{20} D_f(\varphi\|\varphi) = f''(1) F_Y(\varphi). \tag{5}$$

Here, $F_Y(\varphi)$ is the Fisher information matrix. Because of the relationship (5), the convex divergence can be regarded as a fundamental amount of information. Then, we pay attention to the following ratio.

$$c \stackrel{\text{def}}{=} \frac{f''(1)}{f'(1)} = -\frac{g''(1)}{g'(1)} \in \boldsymbol{R} \tag{6}$$

By using this constant, the following expansions can be obtained:

$$\frac{f''(r)}{f'(1)} = \frac{1}{c(1-c)}(r - r^c) + o(1), \tag{7}$$
$$\frac{g''(r)}{g'(1)} = \frac{-1}{c(1-c)}(r^{1-c} - 1) + o(1). \tag{8}$$

From equations (7) and (8), we find that

$$L^{(c)}(r) = \frac{1}{1-c}(r^{1-c} - 1) \tag{9}$$

can be regarded as an extended class of the logarithm. We call this the $c$-logarithm. In fact, we have

$$L^{(1)}(r) = \log r. \tag{10}$$

If we add a set of assumptions that

$$f(xy) = kf(x)f(y) \tag{11}$$

and

$$f''(1) = g''(1) = 1, \tag{12}$$

then the $\alpha$-divergence [8], [9] is obtained. In this case, the constants $c$ and $\alpha$ have the following relationships:

$$c = \frac{1-\alpha}{2} \tag{13}$$

and

$$1 - c = \frac{1+\alpha}{2}. \tag{14}$$

### 2.2. Information Matrix and Cramér-Rao Bound

By using the $c$-logarithm, we have the following equality on the information matrices.

$$M^{(c)}(\varphi) \stackrel{\text{def}}{=} E_p\left[c\, p^{-2(1-c)} \left(\frac{\partial L_c}{\partial \varphi}\right)\left(\frac{\partial L_c}{\partial \varphi^T}\right)\right] \tag{15}$$
$$= -E_p\left[p^{-(1-c)} \left(\frac{\partial^2 L_c}{\partial \varphi \partial \varphi^T}\right)\right] = cF_Y(\varphi), \tag{16}$$

whose early versions are found in [10], [11], [12]. We consider the case that the information matrices are positive definite, i.e., $F_Y(\varphi) > 0$, $c > 0$, and hence $M_Y^{(c)}(\varphi) > 0$. Because of Equations (9), (15) and (16), we have that the Cramér-Rao bound is independent of $c$. This means that the general convex divergence can be used in estimation problems without sacrificing the performance in comparison with the logarithmic methods. Guaranteed by this fact, we discuss iterative minimizations of the convex divergence for the independent component analysis.

## 3. INDEPENDENT COMPONENT ANALYSIS USING CONVEX DIVERGENCE

### 3.1. Problem Formulation

In the problem of the convex divergence, a set of vector random data is given.

$$x(n) = [x_1(n), \cdots, x_K(n)]^T = As(n), \quad (n = 1, \cdots, N). \tag{17}$$

Here, the $K$ by $K$ matrix $A$ and the source vector

$$s(n) = [s_1(n), \cdots, s_K(n)]^T \tag{18}$$

are unknown. Additional assumptions are the following.

1. The components $s_i(n)$ and $s_j(n)$ are independent each other for $i \neq j$.

2. The unknown components $s_i(n)$, $(i = 1, \cdots, K)$, are non-Gaussian except for at most one specific $i$.

Therefore, we want to find a demixing matrix

$$W = \Lambda \Pi A^{-1} \tag{19}$$

so that the components of

$$Wx(n) \stackrel{\text{def}}{=} y(n) = [y_1(n), \cdots, y_K(n)]^T \tag{20}$$

are independent each other for every $n$. Here, $\Lambda$ is a nonsingular diagonal matrix and $\Pi$ is a permutation matrix. These two matrices are also unknown.

Using the convex divergence $D_f$, this ICA problem is formulated as a minimization of the following cost function.

$$\begin{aligned} I_f(\bigwedge_{i=1}^K Y_i) &\stackrel{\text{def}}{=} D_f\left(p(y_1,\cdots,y_K)\|\prod_{i=1}^K q_i(y_i)\right) \\ &\stackrel{\text{def}}{=} D_f(p(y)\|q(y)) \\ &= D_g(q(y)\|p(y)) \\ &= I_g(\bigwedge_{i=1}^K Y_i) \\ &= \int_{\mathcal{X}} p(x) g\left(\frac{|W|q(y)}{p(x)}\right) dx. \end{aligned} \tag{21}$$

### 3.2. Update Equations

The generalized gradient [13], relative gradient [14], or natural gradient [15] denoted by $\tilde{\nabla}$ is obtained by multiplying $cW^TW$ after partially differentiating (21) with respect to $W$. Then, we have the following equality.

$$\begin{aligned} &-\tilde{\nabla} I_g(\bigwedge_{i=1}^K Y_i) \stackrel{\text{def}}{=} -\frac{\partial I_g(\bigwedge_{i=1}^K Y_i)}{\partial W}(cW^TW) \\ &= -c\int_{\mathcal{X}} q(y) g'\left(\frac{|W|q(y)}{p(x)}\right)\left\{I - \vartheta(y)x^TW^T\right\}|W| dx\, W \\ &= -c\int_{\mathcal{Y}} q(y) g'\left(\frac{q(y)}{p(y)}\right)\left\{I - \vartheta(y)y^T\right\} dy\, W. \end{aligned} \tag{22}$$

Here, $\vartheta(y)$ is a vector

$$-\vartheta(y) = \mathrm{col}\left[\{\frac{q_i'(y_i)}{q_i(y_i)}\}_{i=1}^K\right]. \tag{23}$$

We assume that $\vartheta(y_i)$ be an odd function such as $y_i^3$ and $tanh(y_i)$. Note that $\tilde{\nabla} I_g = \tilde{\nabla} I_f$. Equation (22) is not yet in a realizable form as a concrete algorithm since it contains an unknown probability density $q(y)$. Therefore, the next step is to find a realizable approximation to (22). Since

$$qg'(q/p) = -g''(1)p + \{g'(1) + g''(1)\}q + o(1) \tag{24}$$

holds around $p \approx q$, we have the following update value for an iterative minimization.

$$\begin{aligned} &-\frac{\partial I_f}{\partial W}\left(cW^TW\right) \\ &= -\frac{\partial I_g}{\partial W}\left(cW^TW\right) \\ &= f''(1)\Big[c\left\{I - E_{p(y)}[\vartheta(y)y^T]\right\}W \\ &\quad +(1-c)\left\{I - E_{q(y)}[\vartheta(y)y^T]\right\}W\Big] + o(1), \end{aligned} \tag{25}$$

and

$$\tilde{\Delta}_f W = -\rho_t \frac{\partial I_f}{\partial W} W^TW \tag{26}$$

Here, $\rho_t$ is a small positive number called the learning rate. Thus, $0 < c \leq 1$ is a region for faster convergence with the ratio of

$$r = 1 + (\frac{1-c}{c})\frac{q}{p}. \tag{27}$$

Note that $c = 1$ is the case of the minimum mutual information ICA because of (10).

### 3.3. Utilization of Past and Future Information

Equation (25) still requires the unknown probability density function $q(y)$. Therefore, we need to give an interpretation of $q(y)$ in iterative updates. Since $p(y)$ is expected to converge to $q(y)$ as the matrix $W$ is updated, we interpret $p(y)$ and $q(y)$ as follows.

1. [Use of the past information]

   For the current iteration index $t$, the interpretation is $p^{(t-\tau)}(y) := p(y)$ and $p^{(t)}(y) := q(y)$.

2. [Use of future estimation]

   In this case, the interpretation is $p^{(t)}(y) := p(y)$ and $p^{(t+\tau)}(y) := q(y)$.

Here, $\tau$ is a natural number.

### 3.4. Algorithms

The first version utilizes a set of past update information.

**[Momentum f-ICA]**

If we use $p(y)$ as $p^{(t-\tau)}(y)$ and $q(y)$ as $p^{(t)}(y)$ at the $t$-th iteration, then the sample-based learning is as follows.

$$\begin{aligned} &\tilde{\Delta}_f W(t) = \tilde{\Delta} W(t) + \mu_f \tilde{\Delta} W(t-\tau) \\ &= \rho_t\Big[\{I - \varphi(y(t))y(t)^T\}W(t) \\ &\quad +\mu_f\{I - \varphi(y(t-\tau))y(t-\tau)^T\}W(t-\tau)\Big] \end{aligned} \tag{28}$$

Here, $\mu_f = \frac{c}{1-c}$. Thus, we added a momentum term $\tilde{\Delta} W(t-\tau)$ weighted by $\mu_f$. Note that the case of $\mu_f = \frac{1-\alpha}{1+\alpha}$ corresponds to the $\alpha$-ICA [16]. Further special case of $\alpha = 1$, i.e., $\mu_f = 0$ is the plain minimum mutual information method of [2], [3].

The second version utilizes estimation of a future value.

**[Turbo (Look-ahead) f-ICA]**

$$\begin{aligned} &\tilde{\Delta}_f W(t) = \tilde{\Delta} W(t) + \nu_f \tilde{\Delta} W(t+\tau) \\ &= \rho_t\Big[\{I - \varphi(y(t))y(t)^T\}W(t) \\ &\quad +\nu_f(t)\{I - \varphi(\hat{y}(t+\tau))\hat{y}(t+\tau)^T\}\hat{W}(t+\tau)\Big] \end{aligned} \tag{29}$$

Here, $\nu_f = \frac{1}{\mu_f} = \frac{1-c}{c}$.

The look-ahead terms $\hat{W}(t+\tau)$ and $\hat{y}(t+\tau)$ are estimations of $W(t+\tau)$ and $y(t+\tau)$ using the usual log-version. Thus, we added a predicted term $\tilde{\Delta}\hat{W}(t+\tau)$ weighted by $\nu_f$.

We give the following comments on the above two update methods.

1. Equation (28) is the result of a weighted superposition of convex functions: Positively weighted superposition of convex functions gives another convex function.

2. There is a duality between Equations (28) and (29).

3. $\tau = 1$ works effectively enough for both anticipatory and non-anticipatory methods.

4. On the use of the look-ahead method, a semi-batch mode is recommended to show the merit of speedup.

### 3.5. Partial Supervision

Because of the unknown permutation matrix $\Pi$, the resulting matrix $W$ forces users to identify which source is which. This enhances undesirable off-line nature of the algorithm. Therefore, we consider injection of partially supervisory information so that the target information is recovered as the top source.

From Equation (20), the observed signal $x(n)$ is expressed by a mixture of $y(n)$ by

$$x(n) = W^{-1}y(n) \stackrel{\text{def}}{=} Uy(n). \tag{30}$$

Let

$$U \stackrel{\text{def}}{=} [u_1, \cdots, u_K]. \tag{31}$$

Here,

$$u_j = [u_{1j}, \cdots, u_{Kj}]^T. \tag{32}$$

Then,

$$x(n) = u_1 y_1(n) + \cdots + u_K y_K(n). \tag{33}$$

Thus, the vector $u_k$ possesses the information on the mixture. Therefore, we consider to control the ordering of $\{u_k\}_{k=1}^K$ and each vector's components. Suppose we have a set of teacher signals or a target pattern, say $\hat{R}$. Then, this teacher information can be incorporated into the iterative minimization by adding a descent cost term obtained from

$$F(U, \hat{R}) = tr\{(\hat{R} - U)^T(\hat{R} - U)\}. \tag{34}$$

For this cost function, the gradient descent term is

$$\Delta U = \lambda(\hat{R} - U), \tag{35}$$

where $\lambda$ is a small positive constant. If $\hat{R}$ is nonsingular, the following approximation can be used

$$\Delta U = \lambda\hat{R}\{I - (W\hat{R}^{-1})\} \approx \lambda\hat{R}(W\hat{R} - I). \tag{36}$$

Since we have to use the effect of $\Delta U$ with the increment $\Delta W$, the following transformed version is used.

$$\Delta V = -W\{\Delta U\}W. \tag{37}$$

This equation comes from an expansion of an the update matrix $U^{-1}$ [17], [18].

## 4. APPLICATIONS TO BRAIN MAP ESTIMATION

### 4.1. Assigned Task and Teacher Pattern

Experiments in this section is used to evaluate convergence speed and accuracy of extracted brain maps. An important feature here is that the test data is a real world one - - not a simulation.

As was explained in the previous section, the supervisory information is injected to the matrix $U$ by specifying the task pattern $\hat{R}$. This supervision is column-wise. Let a column vector

$$\hat{a}_1 = \text{col}[0, 0, 0, 1, 1, 1, 0, 0, 0, \cdots, 1, 1, 1, 0, 0, 0], \tag{38}$$

be an on-off pattern of the assigned task to the subject. Then, we compute its power-matched version $\hat{r}_1$ where the column sum is zero and the variance is the same as $u_1$. Then, $\Delta u_1$ was computed by using Equation (33). Since the rest vectors $\{\hat{r}_j\}$, $j > 1$, are arbitrary, i.e., unsupervised, this freedom was interpreted as $\Delta u_j = 0$ for $j > 1$. Note that

(i) Selecting $k = 1$ is a process of finding an appropriate permutation.

(ii) The power matching reduces the amplitude's uncertainty in $\Lambda$.

### 4.2. Experiments

The presented algorithm was tested for visual area separation experiments on human brain fMRI data. Time and spatial axes are transposed so that independent areas are obtained [5]. Figure 1 illustrates the comparison of the convergence speed. This figure shows that

[Presented method with a constant learning rate]
$\vee$
[Presented method with
Hestenes-Stiefel type learning rate adjustment]
$\vee$
[Minimum mutual information method].

Figure 2 illustrates an obtained activation pattern. Because of the partially supervised learning, this pattern is obtained as the first column of the matrix $U$. Figure 3 is the extracted brain map which gives separation of V1 and V2 areas. This result is compatible with the one obtained from the t-test.

## 5. CONCLUDING REMARKS

In this paper, we discussed the utilization of the convex divergence for iterative optimization. Besides the theoretical interest as a generalization, there is a concrete merit of speedup of convergence in comparison with usual optimization of logarithmic information measures. In the problem

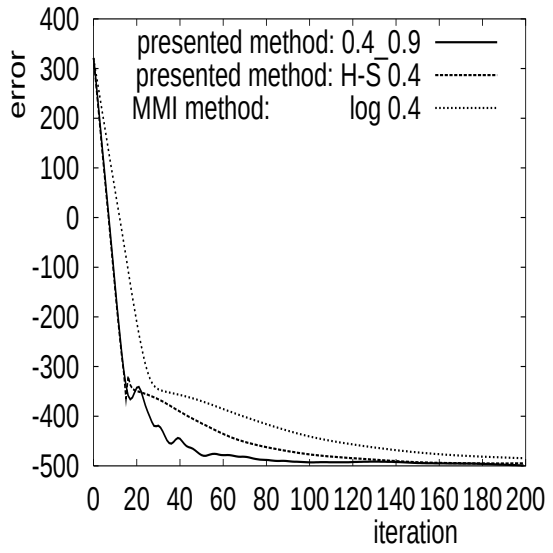


Figure 1: Speed comparison of the presented methods and the traditional method.

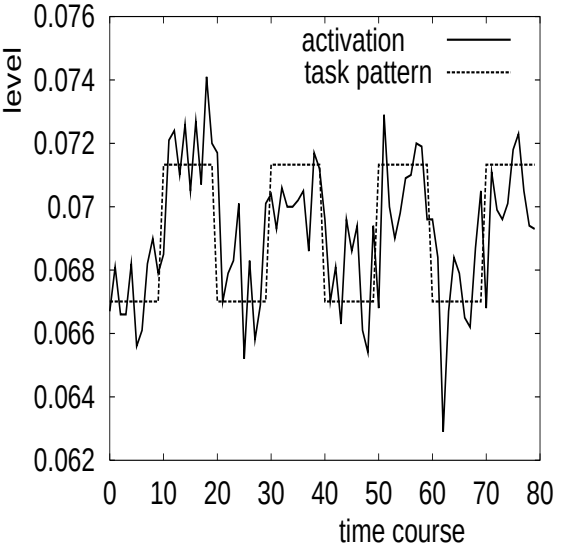


Figure 2: Estimated activation pattern $u_1$.

of ICA, the convex function $f(r)$ was directly used for the derivation of concrete algorithms. This is because, in the problem of ICA, the effect of the convex divergence is concentrated on the number $c$ of Eq. (6). In a different class of convex divergence minimizations, however, a restriction on the function $f$ is necessary so that a closed form update can be obtained. The $\alpha$-EM algorithm is such a case, where the $\alpha$-divergence is used. Interested readers are requested to refer to [11], [12].


## ACKNOWLEDGMENT

The authors are grateful to Dr. K. Tanaka and Dr. R.A. Waggoner of RIKEN Brain science Institute for permitting them to try out the test data set. The authors are also grateful to Mr. Shuichiro Imahara (now, with Toshiba), for his discussions and experiments.


## 6. REFERENCES


[1] C. Jutten and J. Herault, Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture, Signal Processing, vol. 24, pp. 1-20, 1991.

[2] 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.

[3] H.H. Yang and S. Amari, Adaptive online learning algorithm for blind separation: Maximum entropy and minimum mutual information, Neural Computation, vol. 9, pp. 1457-1482, 1997.

[4] I. Csiszár, Information-type measures of difference of probability distributions and indirect observations, Studia Sci. Math. Hungarica, vol. 2, pp. 299-318, 1967.

[5] M.J. McKeown, T-P. Jung, S. Makeig, G. Brown, S.S. Kindermann, T-W. Lee and T.J. Sejnowski, Spatially independent activity patterns in functional MRI data during the stroop color-naming task, Proc. National Academy of Sci. USA, vol. 95, pp. 803-810, 1998.

[6] Y. Matsuyama and S. Imahara, The $\alpha$-ICA algorithm and brain map distillation from fMRI images, Proc. Int. Conf. on Neural Information Processing, vol. 2, pp. 708-713, 2000.

[7] A. Rényi, On measures of entropy and information, Proc. 4th Berkeley Symp. Math. Stat. and Pr., vol. 1, pp. 547-561, 1960.

[8] S. Amari, Differential geometry of statistics, Institute of Mathematical Statistics Lecture Notes, vol. 10, pp. 21-94, 1985.

[9] S. Amari and H. Nagaoka, *Methods of Information Geometry*, Iwanami, 1993 (Translation by D. Harada, AMS, 2000).

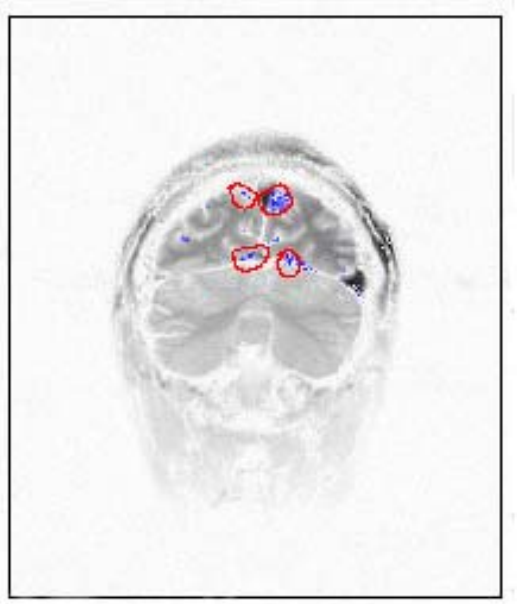

Figure 3: Obtained brain map.


[10] Y. Matsuyama, The $\alpha$-EM algorithm: A block connectable generalized learning tool for neural networks, Lecture Notes in Computer Science, No. 1240, pp. 483-492, Berlin, Germany: Springer-Verlag, June, 1997.

[11] Y. Matsuyama, The $\alpha$-EM algorithm and its basic properties, Transactions of Institute of Electronics, Information and Communications Engineers, vol. J82-D-I, pp. 1347-1358, 1999.

[12] Y. Matsuyama, The $\alpha$-EM algorithm: Surrogate likelihood optimization using $\alpha$-logarithmic information measures, IEEE Trans. on Information Theory, vol. 49, pp. x-y, 2003.

[13] M. Jamshidian and R.I. Jennrich, Conjugate gradient acceleration of the EM algorithm, J. ASA, vol. 88, pp. 221-228, 1993.

[14] J.-F. Cardoso and B.H. Laheld, Equivariant adaptive source separation, IEEE Trans. on SP, vol. 44, pp. 3017-3030, 1996.

[15] S. Amari, Natural gradient works efficiently in learning, Neural Computation, vol. 10, pp. 252-276, 1998.

[16] Y. Matsuyama, N. Katsumata, Y. Suzuki and S. Imahara, The $\alpha$-ICA algorithm, Proc. Int. Workshop on Independent Component Analysis, pp. 297-302, 2000.

[17] Y. Matsuyama, S. Imahara and N. Katsumata, Optimization transfer for computational learning, Proc. Int. Joint Conf. on Neural Networks, vol. 3, pp. 1883-1888, 2002.

[18] Y. Matsuyama and R. Kawamura, Supervised map ICA: Applications to brain functional MRI, Proc. Int. Conf. on Neural Information Processing, vol. 5, pp. 2259-2263, 2002.

[19] T. Kailath, The divergence and Bhattacharyya distance measures in signal selection, IEEE Tranc. Comm., vol. COM-15, pp. 52-60, 1967.

[20] R. Beran, Minimum Hellinger distance estimates for parametric models, Annnals of Statistice, vol. 5, pp. 445-463, 1977.

[21] S. Amari, A. Cichocki and H.H. Yang, A new learning algorithm for blind signal separation, In: Advences in Neural Infromation Processing Systems, MIT Press, pp. 757-763, 1996.


## ADDITIONAL REMARKS

Reviewers gave comments on this paper. The authors are quite thankful to them. Since the comments were diverse in their contents, the authors decided to summarize their replies here so that the page space can be effectively used.

1. "Speedup" in this paper is used to indicate the comparison of this paper's method and its subclass, the minimum mutual information.

2. The method of ICA is different from EM. The ICA is semi-parametric.

3. Experiments using simulated data are givn in [16]. It is observed that the speedup is the effect beyond the increase of the learning rate.

4. For the momentum ICA, $c = 0.7$ is a recommended rule-of-thumb. Note that $c = 0.5$, or $\alpha = 0$, is the case of the Bhattacharyya distance [19], and equivalently the Hellinger distance [20]. Thus, properties obtaind therein will be beneficial to readers.

5. The vertical axis of Figure 1 shows

$$D(W) + H(X) - \frac{n}{2}\log(2\pi e)$$

[21]. Therefore, the value can be a negative number.

6. Additive regularization term to the main function can be used in a wide variety of gradient-style ICA algorithms. It is necassary to decrease this effect as the iteration proceeds so that the independence of estimated components is the main target. The term used in this paper worked effectively beuause of its simplicity.