CONVEX DIVERGENCE AS A SURROGATE FUNCTION FOR INDEPENDENCE: THE f - DIVERGENCE ICA

Yasuo Matsuyama, Naoto Katsumata and Shuichiro Imahara

Department of Electrical, Electronics and Computer Engineering,
Waseda University, Tokyo 169-8555, Japan
{yasuo, katsu, shoe16}@wizard.elec.waseda.ac.jp

ABSTRACT

The convex divergence is used as a surrogate function for obtaining independence of random variables described by the joint probability density. If the kernel convex function is twice continuously differentiable, this case reveals a class of generalized logarithm. This class of logarithms gives generalizations of the score function and the Fisher information matrix which are related to the Cramér-Rao bound. Guided by these properties, independent component analysis (ICA) using the convex divergence is presented. Obtained algorithms use the past and/or future data. Software implementation is easy and beats the minimum mutual information ICA in the speed. Real world experiments on brain fMRI are also performed.

1. INTRODUCTION

The convex divergence [7] measures difference of two probabilities by using a class of convex functions. By choosing the convex function appropriately, this measure is non-negative, and is zero if and only if two probability measures are equal almost everywhere.

Consider the case that one probability measure is a joint probability density and the other is a product probability density. Then, the convex divergence reflects a degree of independence inherent in the random variables described by the joint probability density. Thus, an iterated minimization of the convex divergence on the structural parameters of the joint probability density can be interpreted as a learning process towards source signal separation into independent components. Such an issue using the mutual information was addressed in [3], [5], [8], [15] and many others. In [9] and [12], a super class of the mutual information, i.e., the α-divergence, was used from an interest as a generalization. In this paper, we start with the aforementioned convex divergence from theoretical interests. It is found that the derived algorithms have a merit of speedup. The algorithms are simple and easily applicable to real world data. Early versions can be found in [10], [11].

Organization of this paper is as follows. Section 2 reviews fundamental properties of the convex divergence. It is newly found that the case of twice continuously differentiable convex functions brings about a class of generalized logarithms. Discussions therein also give generalizations of the score function and the Fisher information matrix. A relationship to Cramér-Rao inequality is also revealed. There, a scale factor is introduced. Section 3 gives gradient descent for the convex divergence and ICA. Section 4 shows methods of software implementations. Use of the past and/or future is considered. Speed evaluations by simulations are given. Applications to real world data such as brain fMRI are also addressed. Section 5 gives concluding remarks such that the α-divergence essentially "spans" the methods of the f-divergence if the convex function is twice continuous differentiable.

2. CONVEX DIVERGENCE AND ITS PROPERTIES

2.1. Definitions and Basic Properties

The convex divergence between two probability densities p and q is defined by the following equations [7].

$$D_f(p\|q) = \int_{\mathcal{Y}} q(y) f(p(y)/q(y)) dy \quad (1)$$

$$= \int_{\mathcal{Y}} p(y) g(q(y)/p(y)) dy \quad (2)$$

$$= D_g(q\|p) \geq g(1) = f(1). \quad (3)$$

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

This work was partially supported by Grant-in-Aid for Scientific Research, and Waseda University's Research Projects on High Technologies and New Technologies.

dual function $g(r)$ satisfies

$$g(r) = rf(1/r) \tag{4}$$

so that it is also convex. The inequality (3) is the equality if and only if $p(y) = q(y)$, y-a.e. Since the normalization of $f(1)$ is arbitrary, we can choose

$$f(1) = g(1) = 0. \tag{5}$$

Then, the convex divergence can be regarded as a directed distance between p and q.

2.2. Convex Functions with Twice Continuous Differentiability

In the definitions of the convex divergences D_f and D_g, differentiability of $f(r)$ and $g(r)$ is not necessarily required. But, we are interested in the case that these functions are twice continuously differentiable. This is because we will derive learning algorithms based upon gradients. Then, for

$$C \stackrel{\text{def}}{=} \frac{f''(1)}{f'(1)} = -\frac{g''(1)}{g'(1)} \in (-\infty, \infty), \tag{6}$$

we have the following equalities around $r = 1$.

$$\frac{f(r)}{f''(1)} = \frac{1}{C(1-C)}(r - r^C) + o(1) \tag{7}$$

$$\frac{g(r)}{g''(1)} = \frac{-1}{C(1-C)}(r^{1-C} - 1) + o(1) \tag{8}$$

Here, $o(1)$ is a higher order term. It is important to note that

$$\begin{aligned} \frac{1}{C(1-C)}(r - r^C) &= \left\{\frac{1}{C}r^C\right\}\left\{\frac{1}{1-C}(r^{1-C} - 1)\right\} && (9) \\ &\stackrel{\text{def}}{=} U^{(C)}(r)L^{(C)}(r) && (10) \end{aligned}$$

In the above expression,

$$L^{(C)}(r) = \frac{1}{1-C}(r^{1-C} - 1) \tag{11}$$

is a compelling function. This is a parameterized class of monotone functions whose convexity is controlled by the parameter C from the ultimate concavity to the ultimate convexity. It is important to note that

$$L^{(1)}(r) = \log r. \tag{12}$$

Thus, $L^{(C)}(r)$ can be regarded as a wide-sense logarithm. We can call this function the C-logarithm. If the argument r is replaced by a probability density p, $L^{(C)}(p)$ can be interpreted as a generalized score function.

2.3. Information Matrix and Cramér-Rao Bound

By using the C-logarithm, we have the following equality.

$$\begin{aligned} M^{(C)}(\varphi) &\stackrel{\text{def}}{=} E_p\left[Cp^{-2(1-C)}\left(\frac{\partial L_C}{\partial \varphi}\right)\left(\frac{\partial L_C}{\partial \varphi^T}\right)\right] && (13) \\ &= -E_p\left[p^{-(1-C)}\left(\frac{\partial^2 L_C}{\partial \varphi \partial \varphi^T}\right)\right] && (14) \end{aligned}$$

This equality can be regarded as a generalization of the Fisher information matrix. In fact, it holds that

$$M^{(C)}(\varphi) = CM^{(1)}(\varphi) = CF(\varphi). \tag{15}$$

Here, $F(\varphi)$ is the Fisher information matrix. The constant C can be regarded as a scale factor.

The information matrix $M^{(C)}(\varphi)$ is related to the Cramér-Rao bound. Let $h(\varphi)$ be an unknown vector function of a vector variable φ for a statistical model $p_{Y|\varphi}(y|\varphi)$. Let $\hat{h}(Y)$ be an unbiased estimate for $h(\varphi)$. Let

$$V(\hat{h}(Y)) \stackrel{\text{def}}{=} \left[\text{Cov}\left(\hat{h}_i(Y), \hat{h}_j(Y)\right)\right] \tag{16}$$

and

$$\Omega(\varphi) \stackrel{\text{def}}{=} \frac{\partial h(\varphi)}{\partial \varphi^T}. \tag{17}$$

Then, the following inequality holds.

$$\begin{aligned} V(\hat{h}(Y)) &\geq C\Omega(\varphi)\{M^{(C)}(\varphi)\}^{-1}\Omega(\varphi)^T \\ &= \Omega(\varphi)\{M^{(1)}(\varphi)\}^{-1}\Omega(\varphi)^T && (18) \end{aligned}$$

This corresponds to the Cramér-Rao inequality. Thus, the bound is not degraded by the choice of C.

3. CONVEX DIVERGENCE ICA

3.1. Gradient of the Convex Divergence

In the problem of ICA, we are given a set of N vector random variables.

$$x(n) = \text{col}[x_1(n), \cdots, x_K(n)], \quad (n = 1, \cdots, N). \tag{19}$$

Each $x(n)$ is a mixture by an unknown matrix A such that

$$As(n) = x(n). \tag{20}$$

Here, the vector

$$s(n) = \text{col}[s_1(n), \cdots, s_K(n)] \tag{21}$$

is also unknown except that its components are independent each other. Then, we want to find a demixing matrix $W = \Lambda\Pi A^{-1}$ so that the components of

$$Wx(n) \stackrel{\text{def}}{=} y(n) = \text{col}[y_1(n), \cdots, y_K(n)] \tag{22}$$

are independent each other for every n. Here, Λ is a nonsingular diagonal matrix and Π is a permutation matrix. Both matrices are also unknown.

Let

$$p(y) = p(y_1, \cdots, y_K) \tag{23}$$

be a joint probability density and

$$q(y) = \prod_{i=1}^K q_i(y_i) \tag{24}$$

be a product probability density. Then, we have

$$\begin{aligned} I_f(\bigwedge_{i=1}^K Y_i) &\stackrel{\text{def}}{=} D_f\left(p(y_1, \cdots, y_K) \| \prod_{i=1}^K q_i(y_i)\right) \\ &\stackrel{\text{def}}{=} D_f(p(y)\|q(y)) \\ &= D_g(q(y)\|p(y)) \\ &= I_g(\bigwedge_{i=1}^K Y_i) \\ &= \int_{\mathcal{X}} p(x) g\left(\frac{|W|q(y)}{p(x)}\right) dx. \end{aligned} \tag{25}$$

Here, we used

$$dy = |W| dx \tag{26}$$

as well as

$$p(y)dy = p(x)dx. \tag{27}$$

The symbol "$\bigwedge$" is used instead of "$;$" which appears in standard references [6]. It is important to observe that the determinant $|W|$ appears at only one place in the last expression of (25).

The negative gradient is obtained as follows.

$$\begin{aligned} -\nabla I_g(\bigwedge_{i=1}^K Y_i) &\stackrel{\text{def}}{=} -\frac{\partial I_g(\bigwedge_{i=1}^K Y_i)}{\partial W} \\ &= \int_{\mathcal{X}} |W| q(y) g'\left(\frac{|W|q(y)}{p(x)}\right) \left\{W^{-T} - \varphi(y) x^T\right\} dx \\ &= -\nabla I_f(\bigwedge_{i=1}^K Y_i) \end{aligned} \tag{28}$$

Here,

$$g'(r) = \frac{d}{dr} g(r). \tag{29}$$

and

$$-\varphi(y) = \mathrm{col}\left[\frac{q_1'(y_1)}{q_1(y_1)}, \ldots, \frac{q_K'(y_K)}{q_K(y_K)}\right]. \tag{30}$$

Then, a simple update equation is

$$W(t+1) = W(t) + \Delta_f W(t) \tag{31}$$

with

$$\begin{aligned} \Delta_f W(t) &= \rho_t \left\{-\nabla I_f(\bigwedge_{i=1}^K Y_i)\right\}_{W=W(t)} \\ &= \rho_t \left\{-\nabla I_g(\bigwedge_{i=1}^K Y_i)\right\}_{W=W(t)} \\ &= \Delta_g W(t). \end{aligned} \tag{32}$$

Here, t is the index for iteration and ρ_t is a learning rate.

3.2. Removal of the Inverse Matrix

In Equation (28), a matrix inverse and transpose W^{-T} appears. The matrix inversion and transposition can be removed by using a natural or relative gradient [1], [4]. By considering (15), we multiply CW^TW. Then, we have

$$\begin{aligned} -\tilde{\nabla} I_g(\bigwedge_{i=1}^K Y_i) &\stackrel{\text{def}}{=} -\frac{\partial I_g(\bigwedge_{i=1}^K Y_i)}{\partial W}(CW^TW) \\ &= -C \int_{\mathcal{X}} q(y) g'\left(\frac{|W|q(y)}{p(x)}\right) \left\{I - \varphi(y) x^T W^T\right\} |W| dx\, W \\ &= -C \int_{\mathcal{Y}} q(y) g'\left(\frac{q(y)}{p(y)}\right) \left\{I - \varphi(y) y^T\right\} dy\, W. \end{aligned} \tag{33}$$

An important next step is how to evaluate the core of the integrand of (33). It is a key to observe

$$q g'(q/p) = -g''(1)p + \{g'(1) + g''(1)\}q + o(1) \tag{34}$$

around $p \approx q$. Then, by virtue of (6), we have

$$\begin{aligned} q(y) g'(\tfrac{q(y)}{p(y)}) &= -g''(1)p(y)\left[1 + \frac{1 + \frac{g''(1)}{g'(1)}}{-\frac{g''(1)}{g'(1)}} \frac{q(y)}{p(y)}\right] + o(1) \\ &= -g''(1)p(y)\left[1 + \frac{1-C}{C}\frac{q(y)}{p(y)}\right] + o(1). \end{aligned} \tag{35}$$

Therefore, we have the following equation.

$$\begin{aligned} &-\frac{\partial I_f}{\partial W}(CW^TW) \\ &= -\frac{\partial I_g}{\partial W}(CW^TW) \\ &= f''(1)\Big[C\left\{I - E_{p(y)}[\varphi(y)y^T]\right\} W \\ &\quad + (1-C)\left\{I - E_{q(y)}[\varphi(y)y^T]\right\} W\Big] + o(1). \end{aligned} \tag{36}$$

Therefore,

$$0 < C \leq 1 \tag{37}$$

is a region of faster convergence with the ratio of $1 + (\frac{1-C}{C})\frac{q}{p}$. Note that $C = 1$ is the case of the minimum mutual information ICA because of (12).

3.3. Special Classes of the f-ICA

A useful class of convex functions satisfies the following equality.

$$f(xy) = k f(x) f(y) \tag{38}$$

The following function satisfies this equality.

$$f(r) = \frac{r^\beta}{k(\beta)} \tag{39}$$

Here, β and $k(\beta)$ should have a relationship so that $f(r)$ be a convex function. If we choose $f(1) = g(1) = 0$ and $f''(1) = g''(1) = 1$, then

$$f^{(\alpha)}(r) = \frac{4}{1-\alpha^2}(r - r^{\frac{1-\alpha}{2}}), \tag{40}$$

and

$$g^{(\alpha)}(r) = \frac{4}{1-\alpha^2}(1 - r^{\frac{1+\alpha}{2}}) \tag{41}$$

are such convex functions for $\alpha \in (-\infty, \infty)$. In this case,

$$C = \frac{f''(1)}{f'(1)} = -\frac{g''(1)}{g'(1)} = \frac{1-\alpha}{2} \tag{42}$$

and

$$1 - C = 1 - \frac{f''(1)}{f'(1)} = 1 + \frac{g''(1)}{g'(1)} = \frac{1+\alpha}{2}. \tag{43}$$

Note that (37) corresponds to

$$-1 \leq \alpha < 1. \tag{44}$$

Thus, the α-divergence which uses $f^{(\alpha)}(r)$ and $g^{(\alpha)}(r)$ inherits the convexity control ability of the f-divergence through the parameter α instead of the parameter C.

4. IMPLEMENTATION OF THE CONVEX DIVERGENCE ICA

4.1. Non-Anticipatory Realization as the Momentum f-ICA

First, we observe that $q(y)$ is the target function of $p(y)$ such that

$$q(y) = \lim_{t\to\infty} p^{(t)}(y) \tag{45}$$

under an appropriate convergence criterion. Here, t is the index for the iteration. Then, there is a non-anticipatory approximation at the t-th iteration such that

$$q(y) \Leftarrow p^{(t)}(y) \quad \text{and} \quad p(y) \Leftarrow p^{(t-\tau)}(y). \tag{46}$$

By this approximation, we have the following sample-based learning algorithm.

[Momentum f-ICA]

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

$$\begin{aligned}
&\tilde{\Delta}_f W(t) = \tilde{\Delta} W(t) + \mu_f \tilde{\Delta} W(t-\tau) \\
&= \rho_t \Big[\{I - \varphi(y(t))y(t)^T\} W(t) \\
&\quad + \mu_f \{I - \varphi(y(t-\tau))y(t-\tau)^T\} W(t-\tau) \Big]
\end{aligned} \tag{47}$$

Here,

$$\mu_f = \frac{C}{1-C} \tag{48}$$

Thus, we added a momentum term $\tilde{\Delta} W(t-\tau)$ weighted by μ_f. Figure 1 illustrates a flow of data and updates. Note that the case of $\mu_f = \frac{1-\alpha}{1+\alpha}$ corresponds to the α-ICA [9], [12]. Further special case of $\alpha = 1$, i.e., $\mu_f = 0$ is

$$\tilde{\Delta}_f W(t) = \tilde{\Delta} W(t) \tag{49}$$

which is the plain minimum mutual information method of [15].

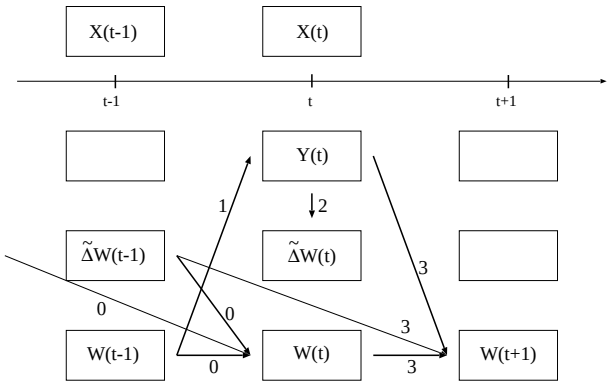

Figure 1: Diagram of the Momentum f-ICA.

4.2. Anticipatory Realization as the Turbo f-ICA

There is an anticipatory approximation at the t-th iteration such that

$$q(y) \Leftarrow p^{(t+\tau)}(y) \quad \text{and} \quad p(y) \Leftarrow p^{(t)}(y). \tag{50}$$

This is more natural than the momentum f-ICA since $p(y)$ has the present iteration index. Then, we have the following sample-based learning algorithm.

[Turbo (Look-ahead) f-ICA]

$$\begin{aligned}
&\tilde{\Delta}_f W(t) = \tilde{\Delta} W(t) + \nu_f \tilde{\Delta} W(t+\tau) \\
&= \rho_t \Big[\{I - \varphi(y(t))y(t)^T\} W(t) \\
&\quad + \nu_f \{I - \varphi(\hat{y}(t+\tau))\hat{y}(t+\tau)^T\} \hat{W}(t+\tau) \Big]
\end{aligned} \tag{51}$$

Here,

$$\nu_f = \frac{1}{\mu_f} = \frac{1-C}{C} \tag{52}$$

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 ν_f. Figure 2 illustrates the flow of data and update terms. We comment here that there is a duality between Equations (47) and (51). We also note in advance that $\tau = 1$ works effectively enough for both causal and non-causal methods

4.3. Orthogonal f-ICA

Amari et al. [2] introduced an orthogonal ICA which is expected to suppress zero-power fake signals. The idea is to find an update term, say $\tilde{\Delta}^+ W$, which is orthogonal to $\tilde{\Delta} W$ so that

$$< \tilde{\Delta} W, \tilde{\Delta}^+ W >_W = 0. \tag{53}$$

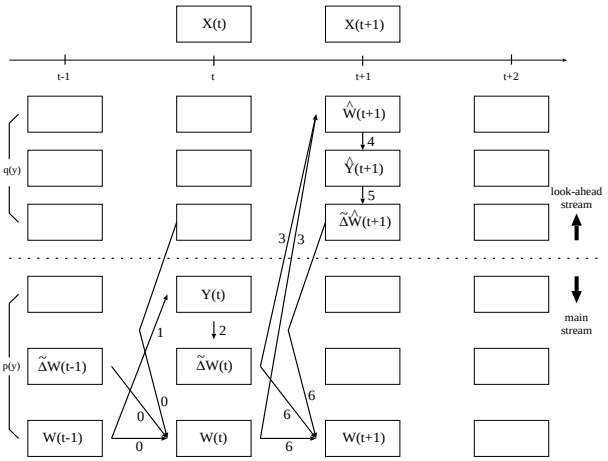

Figure 2: Diagram of the Momentum f-ICA.

Such an update term $\tilde{\Delta}^{+}W$ is obtained as follows. Let

$$\Lambda = \mathrm{diag}\,[\lambda_i]_{i=1}^{K} \tag{54}$$

be a non-singular diagonal matrix. Let

$$W + \tilde{\Delta}W = (I + d\Lambda)W. \tag{55}$$

Then, it holds that

$$\tilde{\Delta}^{+}W = \rho\{\Lambda - \varphi(y)y^{T}\}W, \tag{56}$$

where

$$\Lambda = \mathrm{diag}\,[\varphi_i(y_i)y_i]_{i=1}^{K}. \tag{57}$$

We can obtain four types of orthogonal f-ICA algorithms as is given in [10].

4.4. Combination of Momentum and Turbo f-ICA's

It is possible to use both momentum and turbo effects.

$$\tilde{\Delta}_f W(t) = \tilde{\Delta}W(t) + \mu_t\tilde{\Delta}W(t-\tau) + \nu_t\tilde{\Delta}\hat{W}(t+\tau) \tag{58}$$

We can give reasoning to use this update equation from the definitions of the convex divergence. Definition (1) means that the current environment is considered to be $q(y)$. On the other hand, definition (2) takes $p(y)$ as the current environment. Thus,

$$\begin{aligned} D(p\|q) &= D_{f_1}(p\|q) + D_{f_2}(q\|p) \\ &= D_{f_1}(p\|q) + D_{g_2}(p\|q) \end{aligned} \tag{59}$$

gives the joint momentum and turbo f-ICA.

4.5. Experiments

4.5.1. Experimental Evaluation

Since we are given $\{x(n)\}_{n=1}^{N}$ as a set of mixture source vectors, the expectation $E[\cdot]$ is approximated by $\frac{1}{T}\sum_{i=1}^{T}[\cdot]$ where T is the number of samples in a selected window. The case of $T = N$ is the full batch mode. If we use $T < N$ as a window, it becomes a semi-batch mode. If $T = 1$, the case is an incremental learning. It is possible to choose a window size smaller than N for the look-ahead part so that computation is alleviated.

We chose mixtures of five time series as benchmarking problems. The non-linearity of $\varphi(y) = y^3$ [8] was used. The convergence speed was measured by the cross-talking error [15] which checks the closeness of the matrix WA to $\Lambda\Pi$.

Our first experiment is, (i) to obtain a limit large ρ for the plain MMI. $\rho = 0.50$ was found to be the limit after many trial runs. Following to this step, (ii) the f-ICA was applied by using this ρ. Table I shows the speed of convergence.

Table I Iteration counts for ICA's with $\rho = 0.50$.

plain MMI	momentum	turbo	m+t
23	18	9	7

Thus, the f-ICA strategies are effective.

Next, we try experiments from a different angle. We have a rule-of-thumb; say, $\rho = 0.1$. Table II compares this case.

Table II Iteration counts for ICA's with $\rho = 0.1$.

plain MMI	momentum	turbo	m+t
115	39	16	14

Recommended figures are $C = 0.7$ for the momentum f-ICA, and $1 - C = 0.85$ for the turbo f-ICA .

4.5.2. Applications

After [14], we have tried processing of brain fMRI data. We applied the f-ICA to find active areas when a tested person watches moving images. Figure 3 shows an active area at the rear of the right hemisphere (male). Because of the f-ICA, a conventional personal computer was enough.

5. CONCLUDING REMARKS

The convex divergence, or the f-divergence, is an intriguing quantity which measures a directed distance of two probability densities. If the kernel function is convex with twice continuous differentiability, we can find an accompanied function which can be regarded as a generalized logarithm In this context, there are

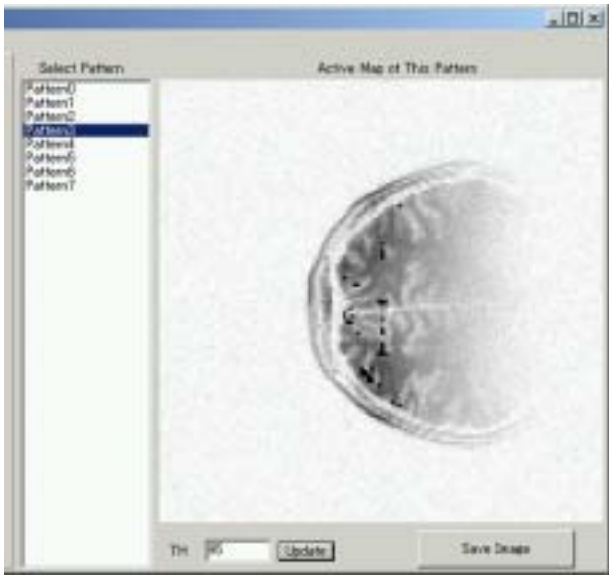

Figure 3: Activation pattern for "*ssmmssmm*" and its associated map.

generalizations of the score function and the information matrix. The Cramér-Rao bound remains non-deteriorated.

In this paper, we focused on the derivation of ICA algorithms using the f-divergence as a surrogate function for independence. It is worth noting that the EM algorithm was also extended by using a divergence measure [13]. The present paper revealed that the α-divergence used in [13] essentially "spans" the methods of the f-divergence.

ACKNOWLEDGEMENT

The authors are quite thankful to Dr. R. Allen Waggoner, Dr. Keiji Tanaka and Dr. Hiroshige Takeichi of RIKEN BRI for permitting them to try out the test data set.

REFERENCES

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

[2] S. Amari T-P. Chen and A.J. Cichocki, Non-holonomic constraints in learning blind source separation, Proc. ICONIP'97, vol. 1, pp. 633-636, 1997.

[3] 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.

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

[5] P. Comon, Independent component analysis, A new concept?, Signal Processing, vol. 36, pp. 287-314, 1994.

[6] T.M. Cover and J.A. Thomas, *Elements of Information Theory,* Wiley, 1991.

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

[8] 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.

[9] Y. Matsuyama and S. Imahara, The α-ICA algorithm and brain map distillation from fMRI images, Proc. ICONIP2000, vol. 2, pp. 708-713, 2000.

[10] Y. Matsuyama and S. Imahara, Independent component analysis by convex divergence minimization: Applications to brain fMRI analysis, Proc. IJCNN2001, vol. x, pp. y-z, 2001.

[11] Y. Matsuyama, N. Katsumata and S. Imahara, Independent component analysis using convex divergence, Proc. ICONIP2001, vol. x, pp. y-z, 2001.

[12] Y. Matuyama, N. Katsumata, Y. Suzuki and S. Imahara, The α-ICA algorithm, Proc. ICA2000, pp. 297-302, 2000.

[13] Y. Matsuyama, T. Niimoto, N. Katsumata, Y. Suzuki and S. Furukawa, α-EM algorithm and α-ICA learning based upon extended logarithmic information measures, Proc. IJCNN2000, vol. III, pp. 351-356, 2000.

[14] M.J. McKeown, T-P. Jung, S. Makeig, G. Brown, S.S. Kindermann, T-W. Lee and T.J. Sejnowski, Spacially independent activity patterns in functional MRI data during the Stroop color-naming task, Proc. Natl. Acad. Sci. USA, vol. 95, pp. 803-810, 1998.

[15] 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.