Independent Component Analysis Using Potts Models

Jiann-Ming Wu and Shih-Jang Chiu

***Abstract*—In this work, we explore the extending application of Potts encoding to the task of independent component analysis, which primarily deals with the problem of minimizing the Kullback–Leibler (KL) divergence between the joint distribution and the product of all marginal distributions of output components. The competitive mechanism of Potts neurons is used to encode the overlapping projections from observations to output components. Based on these projections, the marginal distributions and the entropy of output components are made tractable for computation and the adaptation of the demixing matrix toward independent output components is obtained. The Potts model for ICA is well formulated by an objective function subject to a set of constraints, which leads to a novel energy function. A hybrid of the mean field annealing and the gradient descent method is applied to the energy function so that a set of dynamics and mean field equations can be obtained for the evolution of the continuous geometrical and discrete combinatorial neural variables. Our approach to independent component analysis presents a new criterion for ICA which is distinct from the cumulative-expansion based criterion of Comon and the edgeworth-expansion-based entropy estimation of Amari. The performance of the Potts model for ICA given by our numerical simulations is encouraging.**

***Index Terms*—Entropy, independent component analysis (ICA), mean field annealing, Potts model, unsupervised learning.**

I. Introduction

INDEPENDENT component analysis (ICA) has received much attention from the field of neural computation due to its potential application to the process of array signals, such as speech [19], natural images [5], [13], [14], event related potential [23] and functional MRI [24]. ICA algorithms have been considered to be information-theory-based unsupervised learning rules [1]–[4], [6], [17], [20], [21], [26]. Given a set of multidimensional observations, which are assumed to be linear mixtures of unknown independent sources through an unknown mixing structure, an ICA algorithm performs a search for the demixing matrix by which the observations can be linearly translated to form independent output components. This process has been termed blind source separation [10], [11].

Encoding of criteria for the independence of output components affects the derivation of an ICA algorithm. A set of output components or random variables are independent if their joint distribution coincides with the product of all individual marginal distributions. The independence of output components can be quantitatively measured by the Kullback–Leibler (KL) divergence [1], [10] which is an expected value of the log ratio of the joint distribution to the product of all individual marginal distributions. To facilitate computational tractability for the minimization of the KL divergence, the individual marginal distributions of output components are estimated from observations in an ultra-precise form by which the demixing matrix can be optimized. Existing methods for the estimation primarily include the cumulative-expansion-based approach [10], the edgeworth-expansion-based estimation [31] and the truncated Gram–Charlier expansion based estimation [32]. Using the competitive mechanism of Potts neural variables [25], this work presents a novel encoding for the marginal distributions of output components and explores the resulting KL divergence and the ICA algorithm.

The Potts encoding possesses flexibility for effective internal representations and reliable capability in collective decisions. These properties are essential for designing neural networks suitable for fundamental complex tasks, such as combinatorial optimizations [25], self-organization [22], [29], [30], classification and regression [27]. The employment of multistate Potts neurons, generalized from two-state spin neurons, can significantly reduce the search complexity for feasible configurations, thereby facilitating modeling of the problem. To avoid the trap of tremendous local minima in a circumspect energy function, the evolution of the mean configuration of Potts neurons is controlled by an annealing process an analogous to physical annealing, which is a process of gradually and carefully scaling the temperature from a sufficiently large value to a small one. When this process is used, the probability of a Potts neuron in an individual state, denoted by one mean activation of a Potts neural variable, is increasingly influenced by the injected mean field which measures the weighted sum of the mean activations provided by the other neural variables through interconnections. At the beginning, mean activations are independent of the injected mean fields and the system acts following the principle of maximal entropy. Under these circumstances, a Potts neuron has the same probability for each individual state. As the process progresses, the system always arrives at a stationary configuration which is a tradeoff between the principle of maximum entropy and that of minimal mean energy. Toward the end of the process, the determination of the mean activations is thoroughly dominated by the force of minimal mean energy. That is essentially equivalent to the winner-take-all principle. The Potts encoding has been shown to be suitable for modeling collective decisions in parallel and distributed computations [22], [25]. Its applicability to the task of independent component analysis is explored in this work.

The ICA algorithm developed in this work iteratively invokes the estimation of the individual marginal distributions of output components and the adaptation of the demixing matrix for the minimization of the KL divergence. The key to the coordina-

Manuscript received October 5, 1999; revised June 14, 2000. This work was supported by the National Science Council under Project NSC 88-2218-E-259-004.

The authors are with the Department of Applied Mathematics, National Donghwa University, Hualien,Taiwan, R.O.C. (e-mail: jmwu@server.am.ndhu.edu.tw).

Publisher Item Identifier S 1045-9227(01)02052-5.

tion of the two subtasks depends on the method used to map observations to output components. Due to uncertainty during the intermediate process, the mapping cannot be straightforward but must go through modulation from overlapping projection to nonoverlapping projection. The Potts encoding and the annealing process plays central role in fine-tuning the mapping during modulation. Assuming that the response of each output component is within a set of finite disjointed states or bins, the nonoverlapping projection maps an observation to one and only one state of every output component, while, in contrast, the overlapping projection activates all states to each observation, with each state having its own projected probability attached. This being the case, the nonoverlapping projection is a special case of the overlapping projection in which the projected probability of the only active state is one and the others are zero. The projected probabilities of every output component to an observation are related to the mean activations of a Potts neuron. The competitive mechanism of Potts neurons then takes over the natural modulation of the mapping from overlapping projection to nonoverlapping projection. The projected probabilities compensate for the uncertainty of the demixing matrix and the marginal distributions during intermediate process. By tracing all observations, we can sketch a normalized histogram for the estimation of the marginal distribution of each output component. Based on this, the intermediate demixing matrix can be optimized, since the KL divergence is already tractable. The whole idea is realized by the minimization of an objective function subject to a set of constraints. The key is the encoding of the projected probabilities and the demixing matrix respectively into discrete Potts neural variables and continuous receptive fields of mixed linear and integer programming. Related dynamics are further obtained by applying the mean field annealing and the gradient descent method to an energy function which is derived from the mathematical framework. The two sets of dynamics interactively evolve throughout the annealing process toward the global or near global minimum of the energy function. The novel Potts model minimizes the KL divergence and reduces redundancy among raw signals without any assumption about the form of prior distributions of independent sources.

In the next section, we state the ICA problem and our assumptions, and we introduce our new algorithm for ICA. In the final section, we examine the simulation results and discuss our new algorithm.

II. Potts Models for ICA

A. The Problem

Assume that the M unknown mutually independent sources are of zero mean and are denoted by a random vector $\boldsymbol{s} = [s_1, \ldots, s_M]'$; that the observations are samples from the linear transformation of these independent sources via an unknown mixing matrix, such as

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} \tag{1}$$

where $\boldsymbol{A}$ is an $N \times M$ scalar matrix and $\boldsymbol{x} = [x_1, \ldots, x_N]'$. Independent component analysis aims to recover original sources through a set of output components of which the joint distribution is as close as possible to the product of their marginal distributions

$$p(\mathbf{y}) = \prod_{i=1}^{M} p_i(y_i) \tag{2}$$

where $p_i(y_i)$ denotes the marginal distribution of the ith recovered source or output component y_i. The estimation $\mathbf{y}$ of independent sources is the linear transformation of the observations via a demixing matrix W as follows:

$$\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{s}. \tag{3}$$

When $\mathbf{W}$ is identical to the inverse of $\mathbf{A}$ (i.e., $\mathbf{W} = \mathbf{A}^{-1}$) or the product $\mathbf{WA}$ is an identity matrix, the estimation $\mathbf{y}$ recovers sources $\mathbf{s}$ exactly. However, since the condition of independence (2) does not limit the source signals to be recovered to an exact order or scale, a valid demixing matrix can have a form of

$$\mathbf{W} = \mathbf{\Lambda}\mathbf{P}\mathbf{A}^{-1} \tag{4}$$

where $\mathbf{P}$ is a permutation matrix and $\mathbf{\Lambda}$ is a nonsingular diagonal matrix for arbitrary scaling.

B. The Kullback–Leibler Divergence

The dependency among output components $\mathbf{y}$ is quantitatively measured by the KL divergence which is the expected value of the log ratio of the joint distribution to the product of the marginal distributions. The KL divergence is defined by

$$D(\mathbf{y}) = \int p(\mathbf{y}) \log \frac{p(\mathbf{y})}{\prod_{i=1}^{N} p_i(y_i)} d\mathbf{y}. \tag{5}$$

The minimization of the KL divergence has produced many well-known ICA algorithms, including information maximization [4], negentropy maximization [1], and higher order moments and cumulants [9]. The KL divergence can be separated into two terms

$$D(\mathbf{y}) = -H(\mathbf{y}) + \sum_{i=1}^{N} H_i(y_i) \tag{6}$$

where

$$H(\mathbf{y}) = -\int p(\mathbf{y}) \log p(\mathbf{y}) d\mathbf{y}$$

denotes the joint entropy and

$$H_i(y_i) = -\int p_i(y_i) \log p_i(y_i) dy_i \tag{7}$$

denotes the marginal entropy. Since $\mathbf{y} = \mathbf{W}\mathbf{x}$, we have $H(\mathbf{y}) = H(\mathbf{x}) + \log|\det(\mathbf{W})|$ and then

$$D(\mathbf{y}) = -H(\mathbf{x}) - \log|\det(\mathbf{W})| + \sum_{i=1}^{N} H_i(y_i). \tag{8}$$

Since it is independent of the demixing matrix, the first term is negligible. The tractability of the last term for optimizing the KL divergence in (8) is resolved by the Potts encoding in the next section.

C. Potts Modeling

To approximate the marginal distribution of an output component y_i by a normalized histogram, we first quantize the range of y into a set of discrete states or bins by a symmetrical partition $\{h_{i1} < h_{i2} < \ldots < h_{iK}\}$, where h_{i1} is equal to $-h_{iK}$ and $h_{ik} - h_{ik-1}$ coincides with d for all $2 \leqslant k \leqslant K$. The parameters K and d, respectively, control the number of states and the length of the middle $K-1$ intervals. The states are numbered from one to K. Using the method of nonoverlapping projection, when given a sample $y_i(t) = \mathbf{W}_i\mathbf{x}(t)$, with $\mathbf{W}_i$ as the ith row of $\mathbf{W}$, the ith output component responds a state numbered $k^* = \arg\min_k \|y_i(t) - h_{ik}\|^2$. It follows $\|y_i(t) - h_{ik^*}\|^2 = \min_k \|y_i(t) - h_{ik}\|^2$. The normalized histogram for the occurrences of the states of y_i can be estimated by tracing all samples. But this form of the marginal distribution is not differentiable with respect to $\mathbf{W}_i$ and does not compensate for uncertainty, so in order to achieve a tractable KL divergence, we consider the overlapping projection.

Let the unitary vector $\boldsymbol{\delta}_{it} = [\delta_{it1}, \ldots, \delta_{itK}]'$ denote the membership vector for indicating the response state of the ith output component when the tth sample is given, where $\delta_{itk} \in \{0, 1\}$ for $1 \leqslant i \leqslant N$, $1 \leqslant t \leqslant T$ and $1 \leqslant k \leqslant K$. The only active bit within $\boldsymbol{\delta}_{it}$, for example, the αth bit, indicates that the tth sample $y_i(t) = W_i\mathbf{x}(t)$ is mapped to the αth state of the ith output component. Consider the following objective function:

$$L_1 = \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K}\delta_{itk}\,\|y_i(t) - h_{ik}\|^2 \tag{9}$$

subject to

$$\begin{aligned}&\sum_{k=1}^{K}\delta_{itk} = 1, \text{ for all } i,\, t \\ &\quad\delta_{itk} \in \{0, 1\}, \text{ for all } i,\, t,\, k.\end{aligned} \tag{10}$$

If the only active bit in every membership vector $\boldsymbol{\delta}_{it}$ has an index identical to $\arg\min_k \|y_i(t) - h_{ik}\|$, L_1 is minimized. The minimizer $\{\boldsymbol{\delta}_{it}\}$ is a map of the nonoverlapping projection. For a fixed $\mathbf{W}$, the quantity $-1/2\,\|y_i(t) - h_{ik}\|^2$, denoted by E_{itk}, is a constant. The objective function L_1 with constraints (10) exactly describes an interactive neural system, as every membership vector $\boldsymbol{\delta}_{it}$ is associated with a Potts neuron. In an analogy with statistical mechanism under thermal equilibrium, the mean field theory states that the probability of δ_{itk} being active, or the expected value of δ_{itk} is proportional to the following Boltzmann distribution

$$\begin{aligned}\Pr\,(\delta_{itk} = 1) &= \langle\delta_{itk}\rangle \\ &\propto \exp\,(\beta E_{itk})\end{aligned} \tag{11}$$

where β denotes the inverse of an artificial temperature and E_{itk} is the local mean field. According to the unitary constraint in (10), the probability $\Pr(\delta_{itk} = 1)$ has the following normalized form:

$$\Pr\,(\delta_{itk} = 1) = \frac{\exp\,(\beta E_{itk})}{\sum_{l=1}^{K}\exp\,(\beta E_{itl})}. \tag{12}$$

Let $\Pr\,(\delta_{itk} = 1)$ be the projected probability of the response of the kth state to the sample $y_i(t)$. The overlapping projection stochastically activates each state of an output component to a sample according to its own projection probability. Then a discrete form of the individual marginal distribution can be approximated by a normalized histogram of the state occurrence of an output component y_i, which sums the normalized projection probabilities of all samples

$$p_{ik} = \frac{1}{T}\sum_{t=1}^{T}\Pr\,(\delta_{itk} = 1)\,. \tag{13}$$

Furthermore, the marginal entropy can be estimated by follows:

$$\begin{aligned}H_i(y_i) &= -\int p_i(y_i)\,\log\,p_i(y_i)dy_i \\ &\approx -\sum_{k=1}^{K}p_{ik}\,\log\,p_{ik}.\end{aligned} \tag{14}$$

The β parameter in (12) is gradually scaled by the annealing process from a sufficiently low value to a large one. At the beginning of the process, the probability $\Pr\,(\delta_{itk} = 1)$ is not affected by the local mean field E_{itk} and approaches $1/K$. $y_i(t)$ is then assigned to each state with an almost equal projection probability. As the process progresses, the local mean field increasingly contributes to the mean activation. At an extremely large β, the rule of (12) obeys the principle of winner-take-all, and the system behaves in conformity with nonoverlapping projection. The β parameter clearly modulates the degree of overlapping projection. With Potts encoding, the adaptation of $\mathbf{W}$ toward the minimum value of marginal entropy $H_i(y_i)$ in (14) can be directly realized by the gradient descent method.

The new ICA algorithm iteratively executes the adaptation of an intermediate mixing matrix and the estimation of all marginal entropies as the annealing process is carried out. Returning to the KL divergence in (8) one can derive the overall algorithm. By neglecting the first term, replacing the last term with the discrete marginal entropy in (14) and inserting the objective function L_1 with corresponding constraints, we model the task of the ICA as a mathematical program which minimizes

$$\begin{aligned}L' = &\frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K}\delta_{itk}\,\|\mathbf{W}_i\mathbf{x}(t) - h_{ik}\|^2 \\ &+ C_1(-\log\,|\det\,(\mathbf{W})|) \\ &+ C_2\left(-\sum_{i=1}^{N}\sum_{k=1}^{K}p_{ik}\,\log\,p_{ik}\right)\end{aligned} \tag{15}$$

subject to

$$\sum_{k=1}^{K}\delta_{itk} = 1,\; 1 \leqslant i \leqslant N,\; 1 \leqslant t \leqslant T \tag{16}$$

$$\delta_{itk} \in \{0, 1\}, \text{ for all } i\; t,\, k \tag{17}$$

$$p_{ik} = \frac{1}{T}\sum_{t=1}^{T}\delta_{itk},\; 1 \leqslant i \leqslant N,\; 1 \leqslant k \leqslant K \tag{18}$$

where C_1 and C_2 are weighting constants. The task of ICA now turns to find $\{\boldsymbol{\delta}_{it}\}$ and $\mathbf{W}$ which satisfy the constraints and minimize the objective L'. The mathematical framework consists of a mixed linear and integer program. The variables $\{\boldsymbol{\delta}_{it}\}$ are discrete nonordered Potts neural variables for the combinatorial projection, and the demixing matrix $\mathbf{W}$ represents continuous geometrical features. To develop our algorithm, we apply a hybrid of the mean field annealing and the gradient descent method to optimize these two kinds of variables.

When we treat all membership vectors $\{\boldsymbol{\delta}_{it}\}$ as Potts neural variables, the first two constraints (16) and (17) are naturally taken over by the property of the Potts neural activation function as in (12). By further substituting the constraints (18) into all p_{ik} in the objective (15) and following the optimization procedure through mean field annealing, we can rewrite the programming (15)–(18) as minimizing the following energy function:

$$\begin{aligned} L(\boldsymbol{\delta}, \mathbf{W}) =& \tfrac{1}{2} \sum_i \sum_t \sum_k \delta_{itk} \left\|\mathbf{W}_i \mathbf{x}(t) - h_{ik}\right\|^2 \\ &+ C_1(-\log|\det(\mathbf{W})|) \\ &+ \frac{C_2}{T}\left(-\sum_i \sum_t \sum_k \delta_{itk} \log\left(\frac{1}{T}\sum_t \delta_{itk}\right)\right) \end{aligned} \tag{19}$$

where $\boldsymbol{\delta}$ denotes the collection of all $\boldsymbol{\delta}_{it}$. By fixing $\mathbf{W}$, at each temperature, the mean field annealing seeks the mean configuration $\langle\boldsymbol{\delta}\rangle$ under thermal equilibrium, where the probability of the system configuration is proportional to the following Boltzmann distribution

$$\Pr(\boldsymbol{\delta}) \propto \exp(-\beta L(\boldsymbol{\delta})) \tag{20}$$

where $L(\boldsymbol{\delta})$ with only one argument denotes that the other argument $\mathbf{W}$ has been considered constant. At a sufficiently large β value, the Boltzmann distribution leads to an optimal configuration

$$\lim_{\beta\to\infty} \Pr(\boldsymbol{\delta}^*) = 1$$

where

$$L(\boldsymbol{\delta}^*) = \min_{\boldsymbol{\delta}} L(\boldsymbol{\delta}).$$

To approximate the optimal configuration, the mean field annealing tracks the mean configuration along the annealing process, which gradually increases the β parameter from a sufficiently low value to a large one. At each β value, the mean field equations iteratively run to reach a fixed point which approximates the mean configuration. The mean configuration obtained at each β value is consequently used as an initial mean configuration for evolution to the subsequent β value. The mean field equations can be derived from the following free energy function which have been proposed by Peterson and Söderberg [25]

$$\begin{aligned} \psi(\mathbf{u}, \mathbf{v}, \mathbf{W}, \beta) =& L(\mathbf{v}, \mathbf{W}) + \sum_i \sum_t \mathbf{v'}_{it}\mathbf{u}_{it} \\ &- \frac{1}{\beta}\sum_i \sum_t \ln z(\mathbf{u}_{it}, \beta) \\ z(\mathbf{u}_{it}, \beta) =& \sum_\alpha \exp(\beta u_{it\alpha}) \end{aligned} \tag{21}$$

where $\mathbf{v}_{it}$ denotes the mean of $\boldsymbol{\delta}_{it}$, $\mathbf{v}$ denotes a collection of all $\mathbf{v}_{it}$, $\mathbf{u}_{it}$ is an auxiliary vector, and $\mathbf{v}'_{it}$ denotes the transposition of $\mathbf{v}_{it}$. The stationary point of the free energy function embodies the following mean field equations:

$$\frac{\partial\psi}{\partial\mathbf{v}_{it}} = 0 \Rightarrow \mathbf{u}_{it} = \frac{-\partial L(\mathbf{v}, \mathbf{W})}{\partial\mathbf{v}_{it}} \tag{22}$$

$$\frac{\partial\psi}{\partial\mathbf{u}_{it}} = 0 \Rightarrow \mathbf{v}_{it} = \left[\frac{\exp(\beta u_{it1})}{\sum_l \exp(\beta u_{itl})} \cdots \frac{\exp(\beta u_{itK})}{\sum_l \exp(\beta u_{itl})}\right]' \tag{23}$$

of which the detailed form is

$$\begin{aligned} u_{itk} =& -\frac{1}{2}\left\|\mathbf{W}_i\mathbf{x}(t) - h_{ik}\right\|^2 \\ &+ \frac{C_2}{T}\log\left(\frac{1}{T}\sum_{t=1}^{T} v_{itk}\right) \end{aligned} \tag{24}$$

$$v_{itk} = \frac{\exp(\beta u_{itk})}{\sum_l \exp(\beta u_{itl})} \tag{25}$$

where the constant terms in (24) have been neglected.

During the stage in which the mean configuration of Potts neural variables at each β value is evaluated, the demixing matrix W is considered constant. Then the mean configuration feeds back to the adaptation of the demixing matrix. By applying the gradient descent method to the free energy, we have the following updating rule for each element $\mathbf{W}_{mn}$ in the matrix $\mathbf{W}$:

$$\Delta\mathbf{W}_{mn} \equiv -\eta\frac{\partial\psi}{\partial\mathbf{W}_{mn}} = -\eta\frac{\partial L(\mathbf{v}, \mathbf{W})}{\partial\mathbf{W}_{\mathbf{mn}}} \tag{26}$$

$$\begin{aligned} =& -\eta\left[C_1(\mathbf{W}')^{-1}_{mn} + \sum_t \sum_k v_{mtk}\right. \\ &\left.\cdot(\mathbf{W}_m\mathbf{x}(t) - h_{mk})\,\mathbf{x}_n(t)\right]. \end{aligned} \tag{27}$$

The mean field equations (24) and (25) and the updating rule (27) constitute the following ICA algorithm.

```
1.  Initialize $\beta$ as a sufficiently low value, $\mathbf{W}$ as an identity matrix and every $v_{itk}$ as a value near $1/K$.
2.  Update every $u_{itk}$ and $v_{itk}$ by (24) and (25) iteratively to a stationary point.
3.  Update $\mathbf{W}$ by (27).
4.  If the value $\sum_{itk} v_{itk}^2$ is less than a halting threshold, increase the $\beta$ value by an annealing schedule and then go to step 2; otherwise halt.
```

The convergence of the algorithm is shown in the Appendix.

The above approach is an energy-function-oriented neural network, which is composed of two sets of interactive dynamics, (24), (25), and (27), for the evolution of Potts neural activations and the demixing matrix, respectively. The two sets of interactive dynamics are iteratively executed toward the minimum of the energy function (19) along the annealing process. The energy function is essentially equivalent to the KL divergence,

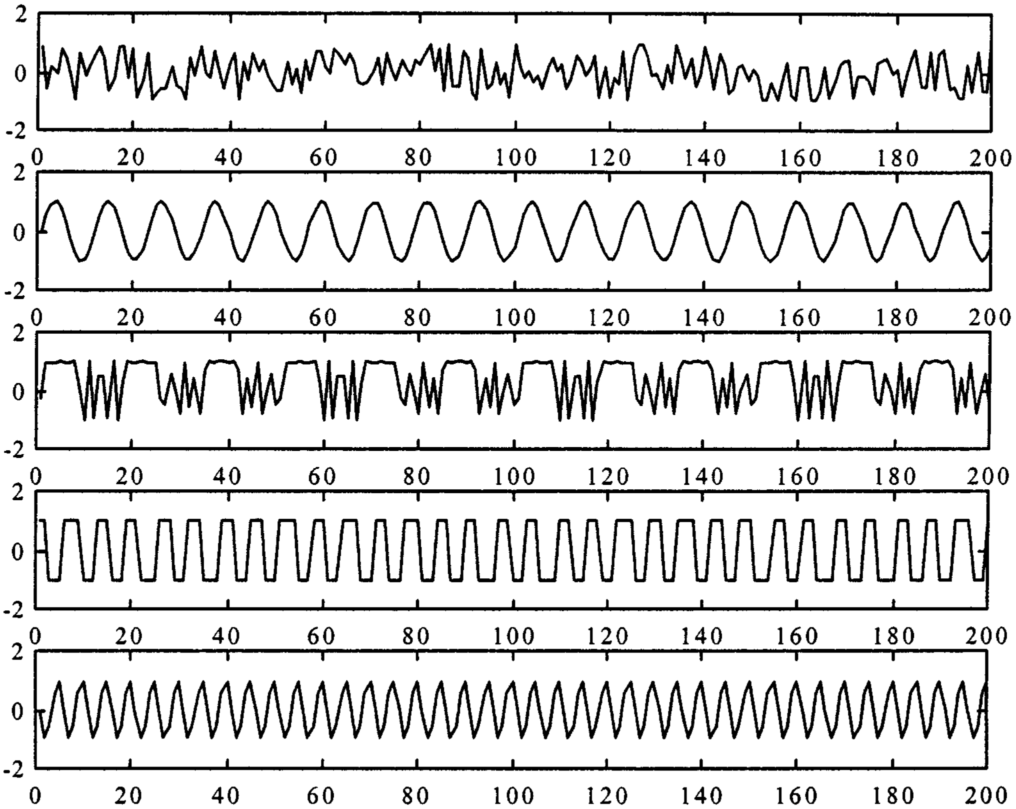

Fig. 1. The five source signals in the first test.

which has been encoded with the criteria for independent output components without any assumption about the prior distributions of the unknown sources.

All means $\{v_{itk}\}$ obtained at step 2 denote the result of overlapping projections of all observations into output components via current $\mathbf{W}$. Comparison of the local mean field u_{itk} in (24) with E_{itk} in (12). u_{itk} depends on all means $v_{it'k}$, $1 \leqslant t' \leqslant T$, but E_{itk} is constant. u_{itk} is influenced by the projection feature as well as the term oriented by the minimization of the marginal entropy $H_i(y_i)$. The minimization of the KL divergence in (8) has been decomposed into the maximization of the joint entropy $H(y)$ and the minimization of all marginal entropies $H_i(y_i)$. The update rule in (27) is responsible for the maximization of the joint entropy. The two subtasks at steps 2 and 3 are joined by the first term of the energy function L in (19), which allows the first term in (24) and the second term in (27) to produce reliable projections. The competitive mechanism in (25) and the annealing process at step 4 subsequently modulate the degree of overlapping projections in an attempt to minimize the energy function L.

III. Numerical Simulations and Conclusion

In the following simulations, we use the performance measure proposed by Amari *et al.* [1].

$$E = \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \frac{|q_{ij}|}{\max\limits_{k} |q_{ik}|} - 1 \right) + \sum_{j=1}^{N} \left(\sum_{i=1}^{N} \frac{|q_{ij}|}{\max\limits_{k} |q_{kj}|} - 1 \right) \tag{28}$$

where q_{ij} denotes the joint element of the ith row and the jth column of the product of the mixing matrix and the demixing matrix. We explore the performance and stability of the new algorithm by comparison with the fast fixed point algorithm (FastICA) [16], [15] and the JadeICA approach [6], [7].

The following sources have been used by Amari *et al.* in [1], $s(t) = [\text{sign}(\cos(2\pi 155t)),\ \sin(2\pi 800t), \sin(2\pi 300t + 6\cos(2\pi 60t)), \sin(2\pi 90t), r(t)]'$, where the first four components of $s(t)$ are modulating the data signals and $r(t)$ is a noise uniformly distributed in $[-1,\ 1]$. Assume that the five sources are unknown to the algorithms and are mixed by a mixing matrix A ,of which the diagonal entries are randomly generated by $0.8+(z-0.5)*0.3$ and the off-diagonal entries are generated by $0.1+(z-0.5)*0.3$, where z is of a uniform distribution in $[0,\ 1]$. The source signals are shown in Fig. 1 and their normalized histograms are shown in Fig. 2. The mixed signals are sampled at a sampling rate of 10K Hz. We feed the first 200 samples of the mixed signals in Fig. 3 to the three algorithms. In the simulation of the PottsICA, the parameter setting includes the learning rate $\eta = 0.002$ in (27), and the weights $C1 = 8,\ C2 = 2$ in (19). For the annealing process, the artificial temperature or the inverse of the β parameter has an initial value 2.5 and a decreasing factor 0.95. The second step in the procedure is executed no more than 20 times and the third step is executed ten times. The halting condition in the last step is $\sum v_{itk}^2 < 0.95NT$, for this case $N = 5$ and $T = 200$. The input to the log function in (24) is automatically added by a constant 10^{-5} to avoid vanishing to zero. The partition parameter h_{ik} is $-1 + (2k-1)/K$ and $K = 20$. The PottsICA was coded in MATLAB and executed in

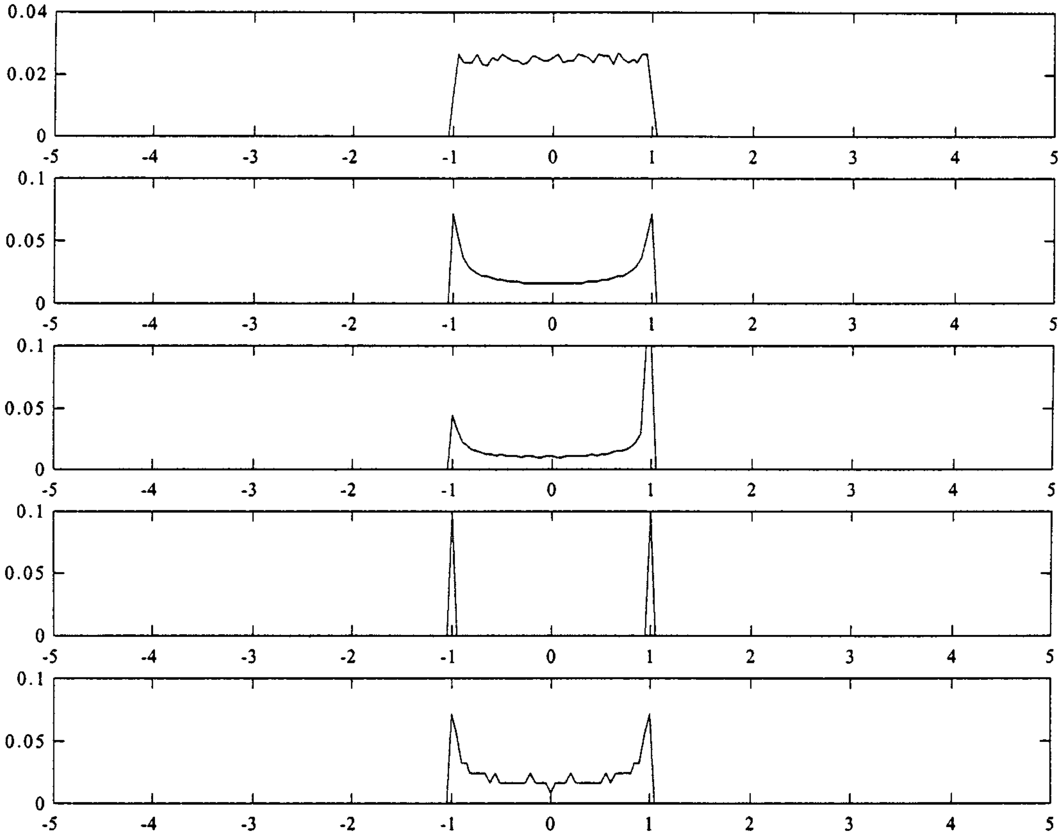

Fig. 2. The normalized histograms of the five sources in Fig. 1.

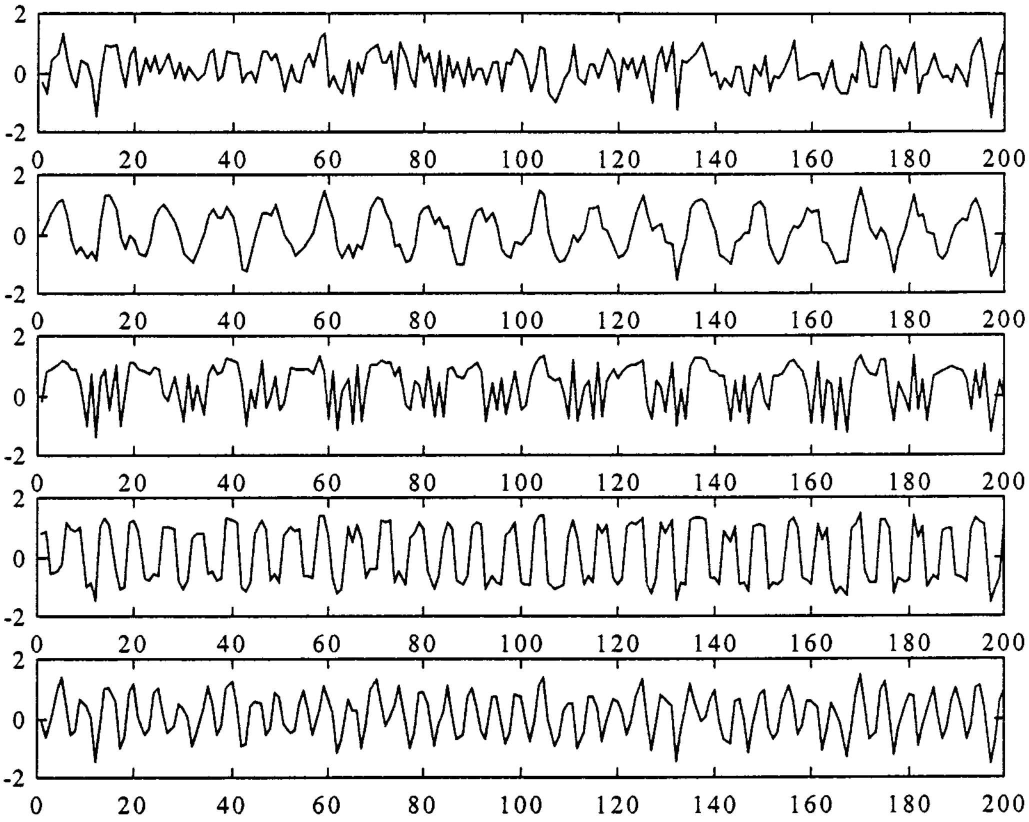

Fig. 3. The mixed signals of the source signals in Fig. 1 by a randomly generated mixing matrix.

Pentium III. The MATLAB codes for the FastICA and JadeICA were downloaded from the homepages provided by the authors of [15], [16] and [6], [7]. Fig. 4 shows the signals separated by the PottsICA. The same experiment was repeated ten times. For

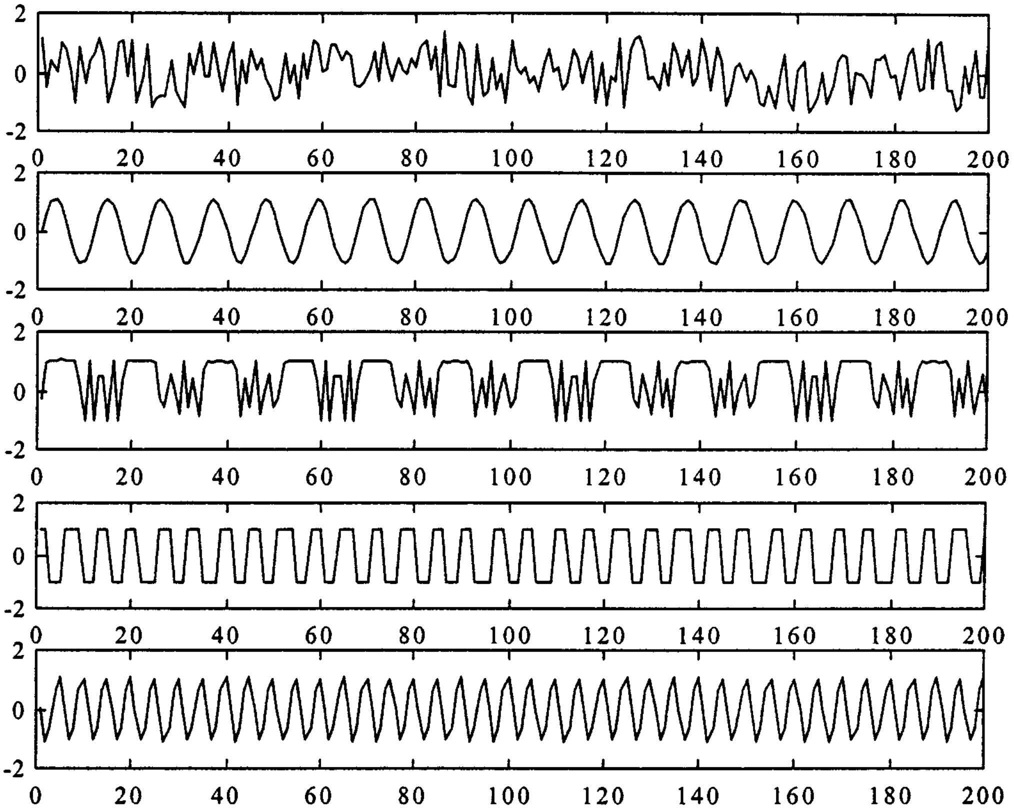

Fig. 4. The recovered signals by the PottsICA from the mixed signals in Fig. 3.

each repetition the mixing matrix and the source signal were renewed. The mean of the performance measure in (28) over the ten experiments for each of the three algorithms is listed in first row of Table I.

The second test adds a Gaussian noise with mean zero and variance 0.25 to the sources in the last test as the sixth source, and the resulting performance of the three algorithms is listed in the second row of Table I. The third row of the same table lists the result for another test with eight sources, for which the two additional sources are in a sub-Gaussian distribution and a super-Gaussian distribution, respectively. The sub-Gaussian is of the distribution $(\mathbf{N}(\mu, \sigma^2) + \mathbf{N}(-\mu, \sigma^2))/2$, where $\mathbf{N}(\mu, \sigma^2)$ is the normal density with mean $\mu = 0.5$ and variance $\sigma^2 = 0.25$. The super-Gaussian source is generated by $n/2 + 1/4 * \sinh(n)$, where a random variable n is in a Gaussian density with mean zero and variance 0.25. The normalized histograms of the eight sources are shown in Fig. 5. For all three tests, the PottsICA is better than the other two algorithms in performance, but the other two algorithms are superior to the PottsICA in speed. The last column of Table I lists the average CPU-time of the PottsICA. This result was achieved using a sequential machine, but PottsICA may be executed in a parallel machine or a delicate machine to improve computational speed.

TABLE I
THE PERFORMANCE OF THE THREE ALGORITHMS FOR THE TESTS

mean E	PottsICA	JadeICA	FastICA	cpu-time(PottsICA)
example 1(N=5)	0.28	0.60	0.75	314 secs
example 2(N=6)	1.28	3.02	1.97	400 secs
example 3(N=8)	4.40	15.30	11.07	566 secs

The following test explores the dependence of the performance on problem size. The source number is increased one by one from $N = 2$ to $N = 20$. Table II lists the result of separating uniform distributions by the three algorithms. Every uniform distribution is in interval $[-0.5, 0.5]$. The result of a similar test is listed in Table III, in which the sources include one Gaussian source with mean $\mu = 0.5$ and variance $\sigma^2 = 0.25$ and $N - 1$ uniform sources in interval $[-0.5, 0.5]$. For cases in which problem size is larger than 15, the FastICA failed to recover all sources due to divergence of some components and the corresponding performance is replaced by a mark "$*$." The performance in Tables II and III shows that the PottsICA is better than the other two algorithms. This superiority becomes more significant as the problem size becomes large.

The extension of the Potts encoding to the task of ICA has been demonstrated in this work. We have employed mean activations of Potts neural variables to realize overlapping projections from observations to output components. The interactions between the estimation of the individual marginal entropy for a tractable KL divergence and the adaptation of the intermediate demixing matrix is well developed by the optimization structure of mean field annealing. For independent component analysis, we have derived a novel energy function, to which a hybrid of the mean field annealing and the gradient descent method has

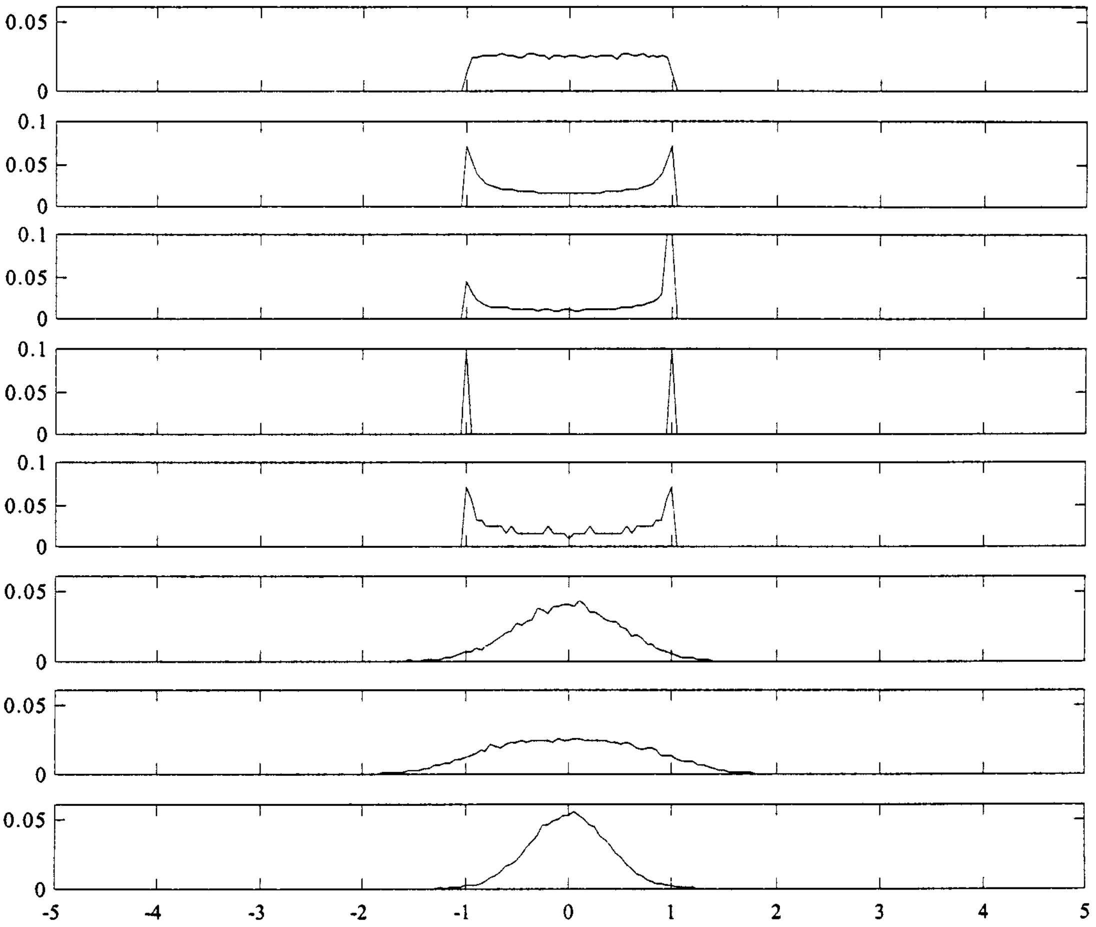

Fig. 5. The normalized histograms of eight sources for the third test in Table I.

been applied. The two sets of interactive dynamics, (24), (25), and (27), are shown to be effective for ICA by our numerical simulations. In comparison with the JadeICA and the FastICA, the new algorithm does a better job of handling the sources whether uniform, sub-Gaussian or super-Gaussian in distribution, and the new algorithm is still reliable for cases involving a larger number of sources. For all of our tests, the new algorithm uses the same set of parameters. Since there is no prior assumption concerning the distributions of the sources in the derivation, there is no need to use different objectives for different distributions of sources. The result of our numerical simulations are encouraging.

Like the other complex tasks, such as constrained optimizations [25], and unsupervised learning [12], [22], the ICA problem is solved by an energy-function-oriented neural network. It is notable that the two sets of interactive dynamics, (24), (25), and (27), are similar to those of the elastic net for the task of self-organization [12], [22]. The key is the Potts encoding, by which the KL divergence for the ICA is connected to the proposed energy function in this work. The usage of Potts neural variables for resolving the individual marginal entropy directly addresses the core issue involved in reducing the KL divergence for the ICA. The new algorithm possesses the following two computational advantages of the energy function oriented neural networks as in [25], [22]. First, the solution obtained by the collective decision of the PottsICA is reliable in quality, which has been shown by our numerical simulations. Second, the PottsICA is suitable for a parallel and distributed process and can be significantly speeded up by a parallel machine or a delicate hardware. The PottsICA has potential for real applications. In addition, we are interested in extending the current work to include overcomplete representations and nonlinear ICA in future work.

Appendix

That steps 2 and 3 in the ICA algorithm converge can be proved by follows. Rewrite the mean field equations in the context as the following continuous form:

$$\frac{d\mathbf{u}_{ij}}{dt} = -\frac{\partial\psi}{\partial\mathbf{v}_{ij}} = \frac{-\partial L(\mathbf{v},\mathbf{W})}{\partial\mathbf{v}_{ij}} \tag{29}$$

$$\mathbf{v}_{ij} = \left[\frac{\exp(\beta u_{ij1})}{\sum_l \exp(\beta u_{ijl})} \cdots \frac{\exp(\beta u_{ijK})}{\sum_l \exp(\beta u_{ijl})}\right]' \tag{30}$$

$$= \sum_k \frac{\exp(\beta u_{ijk})}{\sum_l \exp(\beta u_{ijl})} e_k \tag{31}$$

TABLE II
TEST PERFORMANCE OF THE THREE ALGORITHMS FOR DIFFERENT PROBLEM SIZES WITH ALL SOURCES IN UNIFORM DISTRIBUTIONS

mean E	PottsICA	JadeICA	FastICA	cpu-time (secs) of PottsICA
N=2	0.13	0.17	0.18	153
N=3	0.38	0.48	0.64	224
N=4	0.84	0.91	1.13	317
N=5	1.48	1.66	2.45	386
N=6	2.00	2.73	3.77	460
N=7	3.19	3.64	5.50	556
N=8	4.88	6.54	7.35	619
N=9	6.41	13.20	11.26	716
N=10	7.98	33.85	16.91	782
N=11	9.64	77.96	24.04	861
N=12	13.37	115.26	38.82	950
N=13	15.74	141.02	40.64	1040
N=14	18.00	166.29	55.52	1123
N=15	21.82	183.66	76.51	1199
N=16	25.81	208.93	*	1322
N=17	29.24	233.51	*	1363
N=18	34.67	260.71	*	1437
N=19	39.53	292.50	*	1525
N=20	44.63	323.31	*	1614

where vector e_k is a standard unit vector of which the kth element is one. Then rewrite the updating rule as the following dynamics:

$$\frac{d\mathbf{W}_{mn}}{dt} \equiv -\eta \frac{\partial \psi}{\partial \mathbf{W}_{mn}} = -\eta \frac{\partial L(\mathbf{v}, \mathbf{W})}{\partial \mathbf{W}_{mn}}.$$

Then the convergence of the free energy ψ along the trace of two sets of dynamics can be shown

$$\begin{aligned} \frac{d\psi}{dt} &= \sum_{ij} \left(\frac{\partial \psi}{\partial v_{ij}}\right)' \frac{dv_{ij}}{dt} + \sum_{mn} \left(\frac{\partial \psi}{\partial W_{mn}}\right)' \frac{dW_{mn}}{dt} \\ &= -\sum_{ij} \left(\frac{du_{ij}}{dt}\right)' \left(\Lambda \frac{du_{ij}}{dt}\right) \\ &\quad -\eta \sum_{mn} \left(\frac{dW_{mn}}{dt}\right) \left(\frac{dW_{mn}}{dt}\right) \leqslant 0 \end{aligned} \tag{32}$$

TABLE III
TEST PERFORMANCE OF THE THREE ALGORITHMS FOR DIFFERENT PROBLEM SIZES WITH ONE GAUSSIAN SOURCE AND $N - 1$ UNIFORM SOURCES

mean E	PottsICA	JadeICA	FastICA	cpu-time(secs) of PottsICA
N=2	0.31	0.27	0.31	145
N=3	0.77	0.73	0.92	216
N=4	1.35	1.37	1.38	298
N=5	2.21	2.69	2.66	381
N=6	3.21	5.16	4.26	439
N=7	3.95	5.09	5.68	531
N=8	5.48	17.96	8.87	605
N=9	7.28	19.75	15.20	688
N=10	9.02	48.03	20.67	763
N=11	11.18	74.30	22.30	853
N=12	14.27	110.57	36.10	949
N=13	17.49	133.84	55.21	1024
N=14	20.72	155.62	72.72	1107
N=15	23.86	173.66	76.04	1188
N=16	27.38	201.70	*	1309
N=17	32.80	235.56	*	1364
N=18	36.86	255.16	*	1422
N=19	44.28	285.16	*	1534
N=20	47.52	313.09	*	1625

where Λ is the Hessian of $\ln z(u_{ij}, \beta)$

$$\Lambda = \frac{\sum_{[\sigma_k]} \exp(\beta v'_{ij} \sigma_k)[\sigma_k - v_{ij}][\sigma_k - v_{ij}]'}{\sum_{[\sigma_k]} \exp(\beta v'_{ij} \sigma_k)}$$

$[\sigma_k]$ runs over $\{e_1, \ldots, e_K\}$. Since Λ is positive definite

$$\left(\frac{du_{ij}}{dt}\right)' \left(\Lambda \frac{du_{ij}}{dt}\right) > 0$$

$d\psi/dt \leqslant 0$ is shown.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for fruitful comments.

REFERENCES

[1] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," *Advances Neural Inform. Processing Syst.*, vol. 8, pp. 757–763, 1996.

[2] H. Attias, “Independent factor analysis,” *Neural Comput.*, vol. 11, pp. 803–851, 1999.

[3] J. Basak and S. Amari, “Blind separation of a mixture of uniformly distributed source signals: A novel approach,” *Neural Comput.*, vol. 11, pp. 1011–1034, 1999.

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

[5] ——, “Edges are the independent components of natural scenes,” *Advances Neural Inform. Processing Syst.*, pp. 831–837, 1996.

[6] J. Cardoso, “High-order contrasts for independent component analysis,” *Neural Comput.*, vol. 11, pp. 157–192, 1999.

[7] ——, “Blind signal separation:statistical principles,” *Proc. IEEE*, vol. 9, no. 10, pp. 2009–2025, Oct. 1998.

[8] A. Cichocki, S. C. Douglas, and S. Amari, “Robust techniques for independent component analysis (ICA) with noisy data,” *Neurocomput. 22*, vol. 1–3, pp. 113–129, Nov. 1998.

[9] G. Deco and D. Obradovic, *An Information-Theoretic Approach to Neural Computing*. New York: Springer-Verlag, 1996.

[10] P. Comon, “Independent component analysis, a new concept,” *Signal Processing*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[11] ——, “Contrasts for multichannel blind deconvolution,” *IEEE Signal Processing Lett.*, vol. 3, pp. 209–211, July 1996.

[12] R. Durbin and D. Willshaw, “An analog approach to the travelling salesman problem using an elastic net method,” *Nature*, vol. 326, no. 6114, pp. 689–691, Apr. 16, 1987.

[13] J. H. van Hateren and D. L. Ruderman, “Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex,” *Proc. Roy. Soc. Lond. B. Bio.*, vol. 265, no. 1412, pp. 2315–2320, Dec. 7, 1998.

[14] J. H. van Hateren and A. van der Schaaf, “Independent component filters of natural images compared with simple cells in primary visual cortex,” *Proc. Roy. Soc. Lond. B. Bio.*, vol. 265, no. 1394, pp. 359–366, Mar. 7, 1998.

[15] A. Hyvärinen and E. Oja, “A fast fixed-point algorithm for independent component analysis,” *Neural Comput.*, vol. 9, pp. 1483–1492, 1997.

[16] A. Hyvarinen, “Fast and robust fixed-point slgorithms for independent component analysis,” *IEEE Trans. Neural Networks*, vol. 10, pp. 626–634, May 1999.

[17] J. Karhunen, E. Oja, L. Wang, R. Vigario, and J. Joutsensalo, “A class of neural networks for independent component analysis,” *IEEE Trans. Neural Networks*, vol. 8, pp. 487–504, 1997.

[18] T. Kohonen, *Self-Organizing Maps*. New York: Springer-Verlag, 1995.

[19] T. Lee, M. Girolami, A. J. Bell, and T. J. Sejnowski, “A unifying information-theoretic framework for independent component analysis,” *Int. J. Comput. Math. Applicat.*, 1999.

[20] J. K. Lin, D. G. Grier, and J. D. Cowan, “Faithful representation of separable distributions,” *Neural Comput.*, vol. 9, pp. 1305–1320, 1997.

[21] R. Linsker, “A local learning rule that enables information maximization for arbitrary input distributions,” *Neural Comput.*, vol. 9, no. 8, pp. 1661–1665, Nov. 15, 1997.

[22] C. Y. Liou and J. M. Wu, “Self-organization using potts models,” *Neural Networks*, vol. 9, no. 4, pp. 671–684, 1996.

[23] S. Makeig, T. P. Jung, and A. J. Bell *et al.*, “Blind separation of auditory event-related brain responses into independent components,” *P. Nat. Academy Sci. USA*, vol. 94, no. 20, pp. 10 979–10 984, Sep 30, 1997.

[24] M. J. McKeown, S. Makeig, G. G. Brown, T.-P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski, “Analysis of fMRI data by blind separation into independent components,” *Human Brain Mapping*, vol. 6, pp. 1–31, 1998.

[25] C. Peterson and B. Söderberg, “A new method for mapping optimization problems onto neural network,” *Int. J. Neural Syst.*, 1989.

[26] E. Oja, “The nonlinear PCA learning rule in independent component analysis,” *Neurocomput.*, vol. 17, pp. 25–45, 1997.

[27] A. V. Rao, D. J. Miller, K. Rose, and A. Gersho, “A deterministic annealing approach for parsimonious design of piecewise regression models,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, Feb. 1999.

[28] M. Rattray, D. Saad, and S. Amari, “Natural gradient descent for on-line learning,” *Phys. Rev. Lett.*, vol. 81, no. 24, pp. 5461–5464, Dec. 14, 1998.

[29] K. Rose, E. Gurewitz, and G. C. Fox, “Statistical mechanics and phase transitions in clustering,” *Phys. Rev. Lett.*, vol. 65, no. 8, pp. 945–948, 1990.

[30] ——, “Constrained clustering as an optimization method,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, pp. 785–794, 1993.

[31] A. Stuart and J. K. Ord, *Kendall’s Advanced Theory of Statistics*. New York: Edward Arnold, 1994.

[32] H. H. Yang and S. Amari, “Adaptive on-line learning algorithms for blind separation—Maximum entropy and minimum mutual information,” *Neural Comput.*, vol. 9, pp. 1457–1482, 1997.

Jiann-Ming Wu was born in Taiwan, R.O.C., on November 22, 1966. He received the B.S. degree in computer science in 1988 from National Chiao Tung University, the M.S. degree and the Ph.D. degree in computer science and information engineering from National Taiwan University in 1990 and 1994, respectively.

In 1996, he joined the faculty as an Assistant Professor at the Department of Applied Mathematics in National DongHwa University, where he is currently an Associate Professor. His current research interests include neural networks and signal processing.

Shih-Jang Chiu was born in Taiwan, on January 23, 1975. He received the B.S. degree in mathematics in 1997 from Chung Yaun Christian University and the M.S. degree in applied mathematics in 1999 from National DongHwa University.

His research interest is primarily in neural computation.