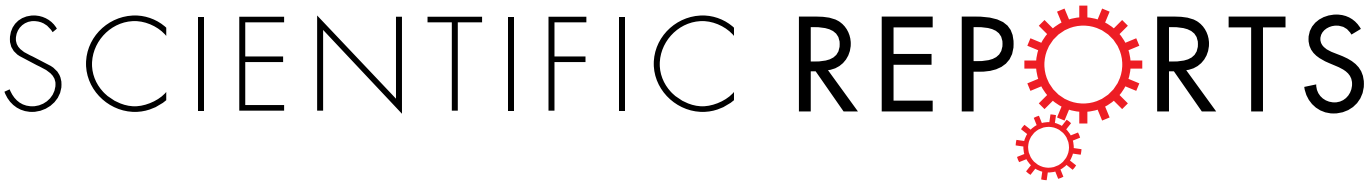

OPEN

A Local Learning Rule for Independent Component Analysis

Takuya Isomura[1,2,3] & Taro Toyoizumi[1]

Received: 16 February 2016
Accepted: 26 May 2016
Published: 21 June 2016

Humans can separately recognize independent sources when they sense their superposition. This decomposition is mathematically formulated as independent component analysis (ICA). While a few biologically plausible learning rules, so-called local learning rules, have been proposed to achieve ICA, their performance varies depending on the parameters characterizing the mixed signals. Here, we propose a new learning rule that is both easy to implement and reliable. Both mathematical and numerical analyses confirm that the proposed rule outperforms other local learning rules over a wide range of parameters. Notably, unlike other rules, the proposed rule can separate independent sources without any preprocessing, even if the number of sources is unknown. The successful performance of the proposed rule is then demonstrated using natural images and movies. We discuss the implications of this finding for our understanding of neuronal information processing and its promising applications to neuromorphic engineering.

One remarkable power of the brain is that it can rapidly identify "objects" from their mixtures. The visual cortex can rapidly identify multiple objects in natural scenes[1], and the auditory cortex can recognize a talker in a noisy social environment, a phenomenon known as the cocktail party effect[2–4]. The problem of separating sensory sources while blind to how they are mixed is termed blind source separation (BSS), which is believed essential for various cognitive tasks[5–8]. Hence, how BSS is performed in the brain can provide a key insight into the way the brain processes sensory information.

Independent component analysis (ICA)[9] is a mathematical model of BSS, where an observer receives linear mixtures of independent sources as inputs and determines the transformation back into their original sources without knowing how they are mixed in the first place. Notably, explicit supervision of which stimulus features belong to what sources is not required to perform ICA. A learner can spontaneously develop the ability to separate independent components only based on stimulus statistics–a concept developed in machine learning as unsupervised learning[10–12]. Several such ICA algorithms (also called learning rules) have been proposed, including those that are based on the information maximization principle[13–16] and the non-Gaussianity of signals[17]. While these learning rules have been successfully used in many engineering applications[18], their neural implementation is not straightforward because each neuron needs to know the information of other unconnected neurons under these rules[16]. Therefore, these learning rules are called *non-local*.

There are a few ICA learning rules that use only the local information available in each neuron and are thus biologically more plausible[19–21]. However, a drawback is that they do not always converge to a desirable solution. A biologically plausible learning rule with reliable performance remains under open investigation.

Here, we proposed a new biologically plausible local learning rule for ICA. First, we propose an extended Hebbian learning rule[22], where changes in synaptic strength are gated by a global error signal summed over a local neural population. We show that this rule can be derived as a gradient descent rule of a cost function that approximates the mutual information between output neurons. Second, we theoretically analyze the stability and uniqueness of its solution. Third, we compare it with other ICA rules and demonstrate that, unlike conventional local rules, the proposed rule reliably converges to an ideal solution over a wide range of mixing matrices, source distributions, and source time scales. Finally, we indicate its promising applications in neuromorphic engineering using natural images and movies.

[1]RIKEN Brain Science Institute, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan. [2]Department of Human and Engineered Environmental Studies, Graduate School of Frontier Sciences, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan. [3]Research Fellow of Japan Society for the Promotion of Science (JSPS), 5-3-1 Kojimachi, Chiyoda-ku, Tokyo 102-0083, Japan. Correspondence and requests for materials should be addressed to T.I. (email: isomura@neuron.t.u-tokyo.ac.jp) or T.T. (email: taro.toyoizumi@brain.riken.jp)

Results

A novel local ICA learning rule. The basic problem is to recover a vector of unobserved independent sources $\mathbf{s}$ from its linear mixture $\mathbf{x} = A\mathbf{s}$ without knowing the mixing matrix A. We assume that the independent sources are distributed according to an identical probability distribution, i.e., $\mathrm{Prob}(\mathbf{s}) = \prod_i p_0(s_i)$. Note that sources must follow a non-Gaussian distribution for ICA to be successful (in Section 7.5 in ref. 9). In this work, we consider a network of N neurons that learn to separate independent sources (Fig. 1A). The output of these neurons is computed by $\mathbf{u} = W\mathbf{x}$, where $\mathbf{x}$ is the activity of the input neurons and W is a matrix of synaptic strengths from the input neurons. Note that each element of $\mathbf{s}$, $\mathbf{x}$, and $\mathbf{u}$ may take a positive or negative value here, as they represent a relative rather than absolute activity level. The goal is to find a synaptic strength matrix that produces independent output. One solution is $W = A^{-1}$, whereby $\mathbf{u} = \mathbf{s}$ is achieved, but any additional permutations and signflips of the output elements also give a solution. We collectively call them ICA solutions. Except when we consider the undercomplete condition later, we assume $\mathbf{s}$, $\mathbf{x}$, and $\mathbf{u}$ are N-dimensional column vectors.

Conventional ICA rules often modify synaptic strength depending on a product of pre- and post-synaptic activity. We called them Hebbian[11,22] rules in a broad sense. Neurons tend to receive a correlated group of inputs under a Hebbian rule. In addition, previous local ICA rules use lateral inhibition to decorrelate the activities of output neurons[19–21]. These mechanisms help neurons to represent separate independent sources. However, modification of neural activity by lateral inhibition is not the only way a neuron influences synaptic plasticity of other neurons. A number of experimental studies have reported that a third-factor, apart from pre- and post-synaptic activity, can play an essential role in modulating the outcome of Hebbian plasticity. For example, GABA[23,24], dopamine[25–27], noradrenalin[28,29], and D-serin[30] are known to modulate Hebbian plasticity. Therefore, local ICA computation may be possible without the need for direct lateral inhibition, if such a third factor monitors the overall state of the neurons and adequately modulates Hebbian plasticity.

In this study, we proposed a novel local learning rule for ICA that extends a Hebbian learning rule by a time varying learning rate, based on a global error signal. We call this the error-gated Hebbian rule (EGHR), expressed as follows:

EGHR

$$\tau_W \dot{W} = \langle (E_0 - E(\mathbf{u}))g(\mathbf{u})\mathbf{x}^T \rangle. \tag{1}$$

Here, the $g(u_i)x_j$ term is a standard Hebbian term, commonly included in many ICA rules (c.f., Equations 2–6), where $g(u_i) = -\mathrm{d}\log p_0(u_i)/\mathrm{d}u_i$ describes a postsynaptic factor of neuron i and x_j describes a presynaptic factor of neuron j. Further, $\langle \bullet \rangle$ describes an expectation over the ensemble of $\mathbf{x}$, the dot over W denotes a temporal derivative, and τ_W is a learning time-constant. In this rule, this Hebbian term is gated by a global error signal $E_0 - E(\mathbf{u})$ composed of constant E_0 and $E(\mathbf{u}) = -\sum_i \log p_0(u_i)$. The term $E(\mathbf{u})$ describes the surprise[31] of observing output $\mathbf{u}$ under the assumption that the distribution of output is $\prod_i p_0(u_i)$, which is achieved after successful learning. Accordingly, the EGHR operation switches with the error signal; the EGHR facilitates a current activity pattern by inducing Hebbian change if $E_0 > E(\mathbf{u})$. In contrast, it suppresses the pattern by inducing anti-Hebbian change if $E_0 < E(\mathbf{u})$. In this way, the EGHR maintains the error $E_0 - E(\mathbf{u})$ close to zero.

Indeed, it is easy to demonstrate that the EGHR is a gradient descent rule that minimizes a cost function $L = \langle (E(\mathbf{u}) - E_0)^2/2 \rangle$ (see Methods). Hence, the basic strategy behind the EGHR is to reduce the fluctuations of $E(\mathbf{u})$ while maintaining its average close to E_0. Independence between outputs (as opposed to highly correlated outputs) helps to keep the fluctuations of $E(\mathbf{u})$ small. In addition, E_0 ($> -N\log p_0(0)$) prevents W from converging to zero, avoiding the trivial solution of $\mathbf{u} = \mathbf{0}$. It turns out that L approximates the common cost function of the Bell-Sejnowski and Amari rules[13–15] if W is near an ICA solution (see Methods and S2.2). Despite this similarity, the EGHR is more biologically plausible than the Bell-Sejnowski and Amari rules because its synaptic changes are based on the local information[16] available at each synapse (see the next section). Note that the computational complexity required by the EGHR is of an N^2-order, $O(N^2)$, in a serial implementation, but is $O(N)$ in a parallel implementation such as using a neuromorphic hardware.

The simple EGHR can straightforwardly perform ICA, as we illustrate using an example with two independent sources obeying a Laplace distribution (Fig. 1B). Figure 1B left shows a typical outcome of the EGHR, where initially non-independent outputs become independent along a gradient descent path of the cost function L (Fig. 1B right; see Methods). We show in separate simulations that the outcome of this rule is robust to the number of independent sources, distribution from which the sources are generated, and deviations of $E(\mathbf{u})$ and $g(\mathbf{u})$ from the above definitions due to the unknown form of the source distribution p_0 (Supplementary Movie 1 and Supplementary Figs S1–3; see also Supplementary Movie 2 for an example with natural scenes). In addition, employing detailed theoretical analyses, we show (Methods and Supplementary information S2.2–4): (1) a mathematical condition under which the ICA solutions are stable fixed points of the EGHR, (2) that the ICA solutions are unique solutions of the EGHR if the source distribution is nearly Gaussian, and (3) the robustness of EGHR solutions to the choice of E_0.

Comparison of ICA rules. In this section, we compare the EGHR with five conventional ICA rules:

1. EGHR:

$$\tau_W \dot{W} = \langle (E_0 - E(\mathbf{u}))g(\mathbf{u})\mathbf{x}^T \rangle.$$

2. Bell-Sejnowski[13,14]:

$$\tau_W \dot{W} = W^{-T} - \langle g(\mathbf{u})\mathbf{x}^T \rangle \tag{2}$$

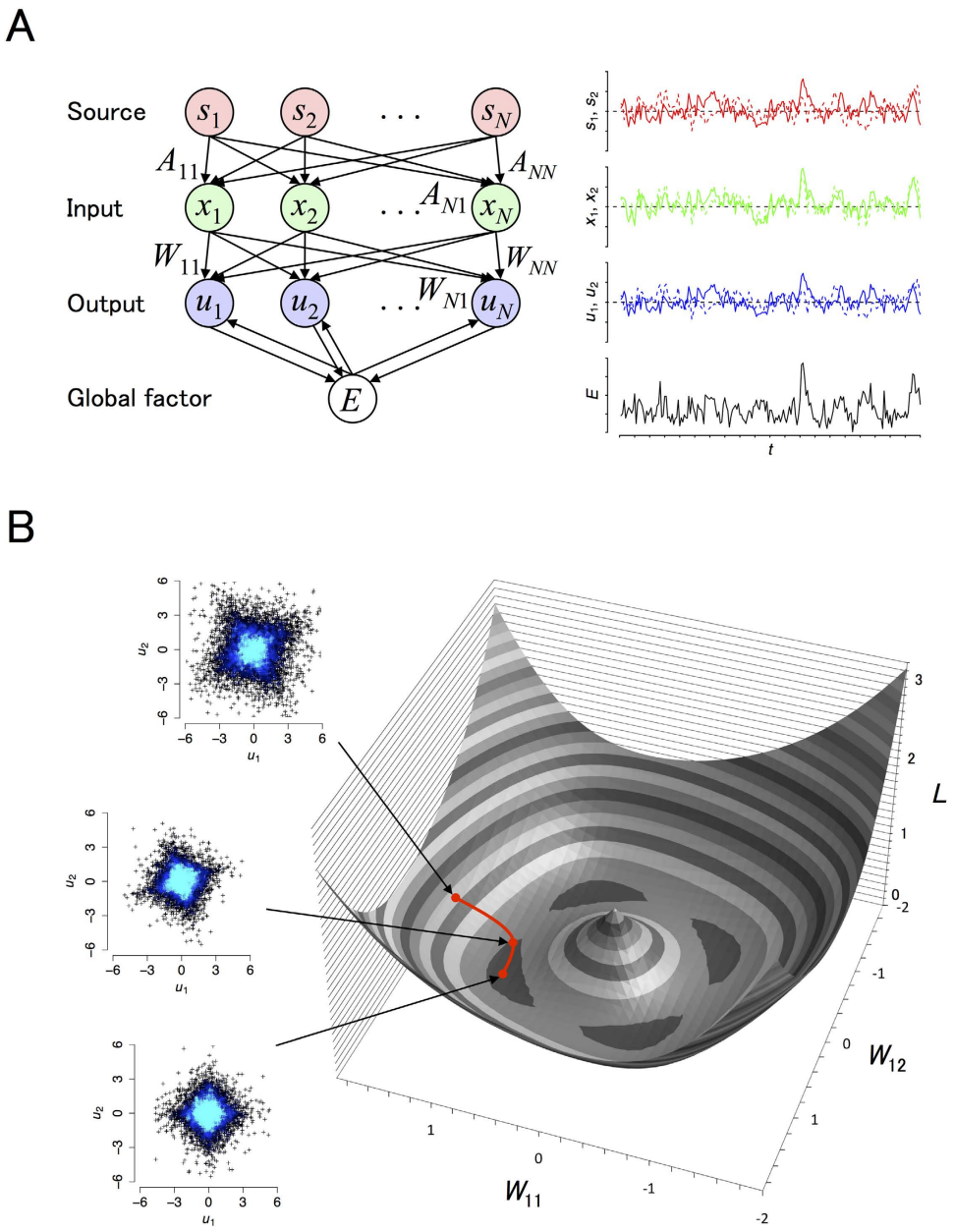

Figure 1. Schematic image of the model setup and results of the proposed learning rule. (A) Left: Schematic image of the model. The input **x** to the neural network is a linear mixture of independent sources **s**, i.e., $\mathbf{x} = A\mathbf{s}$, where A is a mixing matrix. The neural network linearly sums the input and produces the output $\mathbf{u} = W\mathbf{x}$, where W is a synaptic strength matrix. The goal is to learn the $W \propto A^{-1}$ (or its row permutations and signflips) for which the outputs become independent. To this end, a global signal E is computed based on the outputs of individual neurons and gate activity-dependent changes in W during learning. Right: Time traces of **s**, **x**, **u**, and E. **(B)** A dynamic trajectory of the synaptic strength matrix while the network learns to separate independent sources. The learning rule is formulated as a gradient descent algorithm of a cost function L, whose landscape is depicted as a function of synaptic strength parameters (W_{11}, W_{12}). Note that in order to graphically illustrate the results in this three-dimensional plot, we used a two-dimensional rotation matrix with angle $6/\pi$ as the mixing matrix A and restricted W as a rotation and scaling matrix $(W_{11}, W_{12}; -W_{12}, W_{11})$. The red trajectory displays how the gradient descent algorithm reduces the cost function L by adjusting (W_{11}, W_{12}). The three inset panels display the distributions of the network output (u_1, u_2) during the course of the learning. Each point represents sampled outputs and the brightness of the blue color represents probability density. Top: The outputs are not independent at the initial condition $(W_{11}, W_{12}) = (1.5, 0)$. Middle: The distribution of the outputs is rotated during learning. Bottom: The network outputs become independent at the final state $(W_{11}, W_{12}) = (\cos 6/\pi, \sin 6/\pi)$. The two sources are drawn independently from the same Laplace distribution (see Methods). Note that a MATLAB source code of the EGHR is appended as Supplementary Source Code 1.

Name	Symbol	Value
Time resolution of source	dt	1
Simulation time	T	2×10^6
Time constant of source	τ_s	50
Time resolution of algorithm	Δt	100 for EGHR, Amari, Cichocki 10 for Linsker 1 for Foldiak
Time constant of W	τ_W	10^3 for EGHR, Amari, Cichocki 10^4 for Linsker 10^6 for Foldiak
Time constant of Q and $\mathbf{h}$	τ_Q	$\tau_W/10$ for Linsker, Foldiak
Time constant of $\mathbf{h}$	τ_h	$\tau_W/10$ for Linsker, Foldiak
Time constant of $\mathbf{v}$	τ_v	10 for Linsker, Foldiak
Amplification factor	a	1 for Linsker 1.1 for Foldiak
Mean of $\mathbf{v}$	b	$\langle f_F(s_i)\rangle_{p0(si)}$ for Foldiak

Table 1. Model parameters.

3. Amari[15]:

$$\tau_W \dot{W} = (I - \langle g(\mathbf{u})\mathbf{u}^T\rangle)W \tag{3}$$

4. Cichocki[21]:

$$\tau_W \dot{W} = I - \langle g(\mathbf{u})\mathbf{u}^T\rangle \tag{4}$$

5. Linsker[20]:

$$\begin{aligned} \tau_W \dot{W} &= \langle a\mathbf{v}\mathbf{x}^T - g(\mathbf{u})\mathbf{x}^T\rangle, \\ \tau_v \dot{\mathbf{v}} &= -\mathbf{v} + \mathbf{u} + Q\mathbf{v}, \\ \tau_Q \dot{Q} &= -Q + I - a\mathbf{u}\mathbf{u}^T \end{aligned} \tag{5}$$

6. Foldiak[19]:

$$\begin{aligned} \tau_W \dot{W} &= \langle a\mathbf{v}\mathbf{x}^T - \mathrm{Diag}\,[\mathbf{v}/b]\,W\rangle, \\ \tau_v \dot{\mathbf{v}} &= -\mathbf{v} + f_F(\mathbf{u} + Q\mathbf{v} - \mathbf{h}), \\ \tau_Q \dot{Q} &= -\mathbf{v}\mathbf{v}^T + b^2\mathbf{1}\,\mathbf{1}^T, \ (If\ i = j\ or\ Q_{ij} > 0,\ then\ Q_{ij} \leftarrow 0), \\ \tau_h \dot{\mathbf{h}} &= \mathbf{v} - b\mathbf{1} \end{aligned} \tag{6}$$

In these rules, again $\mathbf{x} = A\mathbf{s}$ is the input to each model and $\mathbf{u} = W\mathbf{x}$ is its output. The $g(\mathbf{u})\mathbf{x}^T$ (or $g(\mathbf{u})\mathbf{u}^T W$) term is common across many ICA rules. In addition to the dynamics of W, the Linsker and Foldiak rules assume dynamic updates of neural state $\mathbf{v}$ and lateral connections Q. The Foldiak rule additionally assumes an adaptive threshold $\mathbf{h}$. Note that $\tau_\bullet$ describes the time constant of dynamical variable $\bullet$, I is the $N \times N$ identity matrix, $\mathbf{1}$ is an N-dimensional vector of ones, a and b are constant parameters that we vary in the following, and $f_F(\bullet)$ is a nonlinear function. While $a = 1$ in the original Foldiak rule, tuning a is important in some cases, as we describe below (see Table 1 for parameter values).

Importantly, the Bell-Sejnowki and Amari rules are so called non-local learning rules[16] because updating synaptic strength W_{ij} requires the information of remote synapses such as W_{kl}, where $i \neq k$ or $j \neq l$. On the other hand, the Cichocki, Linsker, and Foldiak rules are so called local learning rules because each synapse is updated based on quantities available there. Note that while the Cichocki rule does not require lateral connections to modify neural activity, they may be required to signal the activity of one neuron to another to achieve a local implementation of the learning rule.

Notably, the Linsker and Foldiak rules are more involved than the others because they need to learn a few sets of dynamical variables in addition to synaptic weight matrix W. In order for these rules to work, the time constants of $\mathbf{s}$, $\mathbf{v}$, Q, and W must satisfy $\tau_v < \tau_s \ll \tau_Q < \tau_W$. This means that the neuronal time constant must be faster than the input time constant and learning needs to be a couple of orders of magnitude slower than the neuronal time constant for these rules. Hence, for these learning rules, we need to introduce slowly time-varying sources. Specifically, we model dynamical sources ($i = 1, \ldots, N$) according to

$$\tau_s \dot{s}_i(t) = -U'(s_i(t)) + \sqrt{2\tau_s}\,\xi_i(t) \tag{7}$$

where $U'(s_i)$ is a derivative of a potential function, $\xi_i(t)$ is a white Gaussian random variable of unit variance and τ_s is the time constant of the sources. The marginal distribution of each source is given in terms of the potential function by $p_0(s_i(t)) \propto \exp(-U(s_i(t)))$. Hence, each source is distributed according to a Laplace distribution with zero mean and a variance of one if the potential function is $U_L(s_i) = \sqrt{2}|s_i|$ and a uniform distribution with zero mean and a variance of one if the potential function is $U_U(s_i) = 0$ for $|s_i| \leq \sqrt{3}$ and a large positive constant for $|s_i| > \sqrt{3}$.

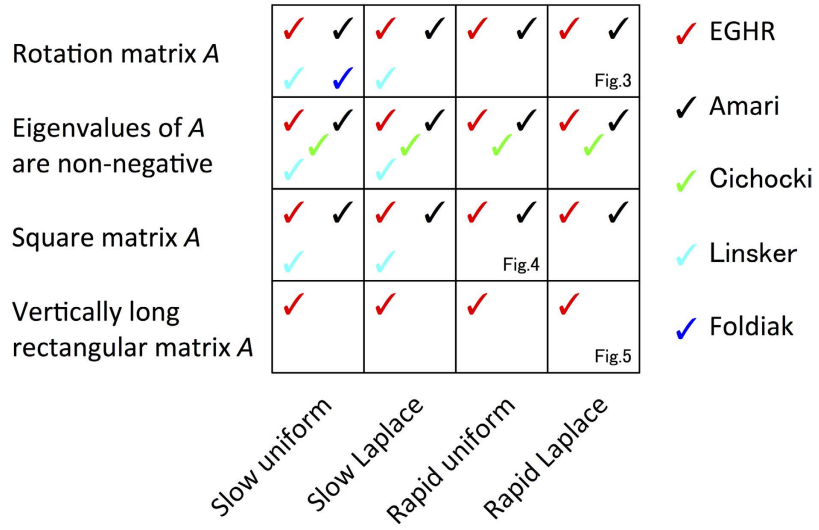

Figure 2. Comparison of ICA rules. Rows and columns describe the features of mixing matrix A and sources $\mathbf{s}$, respectively. Each tick mark indicates that the rule can always perform ICA under the corresponding condition according to our analytical (see Methods and Supplementary Note S2) and numerical (see Figs 3–5 and S1–3) analyses. The EGHR can perform ICA under all listed conditions (see Methods, S2.2–4, Figs 3–5, and S1–3). Previous rules all fail to perform ICA under the undercomplete condition with vertically long rectangular matrix A (see S2.8 and Fig. 5). Under the Linsker rule, the output can converge to zero or diverge to infinity if the sources fluctuate more rapidly than the neuronal time constant (see S2.6.2). The Cichocki rule fails if all eigenvalues of A are negative and can fail depending on the initial synaptic strengths if some eigenvalues are negative (see S2.7.2 and Figs 3 and 4). The Foldiak rule works only if A is proportional to a rotation matrix (see S2.6.3 and Fig. 4). In addition, the Foldiak rule has multiple spurious stable solutions if the sources obey a Laplace distribution (see Fig. 3). Note that the properties of the Bell-Sejnowski rule are the same as those of the Amari rule.

All these rules are capable of performing ICA at least under restricted conditions (see Methods and S2.6). However, the stability of the ICA solutions as well as the existence of spurious solutions depends on the properties of the mixing matrix, source distribution, and how rapidly source signals change relative to the neuronal time scale. Figure 2 summarizes the performance of these learning rules, which we demonstrate in the following simulations and analyses. A tick mark indicates that a learning rule always finds an ICA solution under the corresponding condition according to our analytical (see Methods and Supplementary Note S2) and numerical (see Figs 3–5 and S1–3) analyses. Figure 2 clearly indicates that the EGHR exhibits the most reliable performance over a range of conditions.

First, we analytically investigate the linear stability of ICA solutions (Fig. 2). Because the EGHR is a gradient descent rule of a cost function L, a fixed point of the EGHR is stable if its second derivative d^2L is non-negative in all parameter directions. We found that the EGHR has stable ICA solutions for a wide range of source distributions, including Laplace and uniform distributions, regardless of the form of mixing matrix A (see Methods and S2.2). This property is similar to the Amari rule[32] as well as the Bell-Sejnowski rule because the two share the same cost function (see also S2.7.1). While the Linsker rule approximates the Bell-Sejinowski rule if the input changes more slowly than the neural time constant, its ICA solutions disappear if the input changes too fast (see S2.6.2). Moreover, under the Cichocki rule, some of the ICA solutions are unstable if A has at least one negative eigenvalue, and all ICA solutions are unstable if A is negative definite (see S2.7.2). Additionally, no ICA solutions exist under the Foldiak rule if A is a non-rotation matrix (see S2.6.3). Even if A is a rotation matrix, their stability depends on the detailed relationship between the source distribution and the shape of $f_F(\bullet)$ (see the next section and S2.7.3).

Numerical simulations with a rotation mixing matrix and Laplace source distribution. Here, we numerically investigate the performance of each rule. For visualization purposes, we consider again two neurons to separate two independent sources. We set the mixing matrix to rotation matrix $A = (\cos\theta, -\sin\theta; \sin\theta, \cos\theta)$ with $\theta = \pi/6$. The sources are generated from a Laplace distribution using the potential function U_L as described above. Other parameter values are summarized in Table 1.

Figure 3A depicts the initial (top left) and final (other panels) distributions of the output (u_1, u_2). The two output variables should become independent if each rule is successful. The EGHR and the two non-local learning rules, the Bell-Sejnowski and Amari rules, can successfully separate independent sources. Because the Bell-Sejnowski and Amari rules perform similarly, we only plotted the result of the Amari rule. In contrast, the outcome of other local learning rules, i.e., the Linsker, Cichocki, and Foldiak rules, depends on the initial condition of the synaptic strength matrix (an unsuccessful case is shown in the main panels and a successful case is shown in the inset panels).

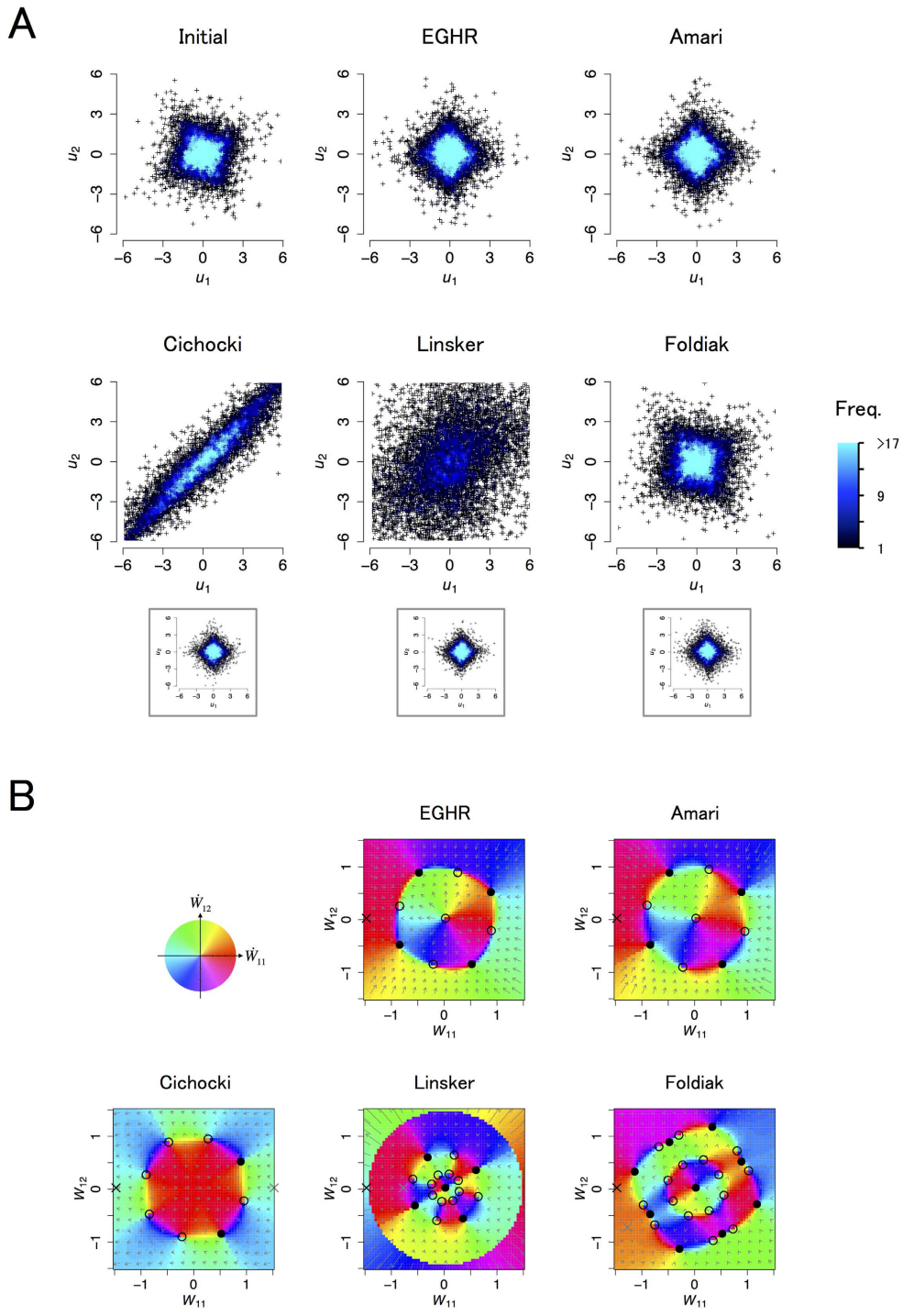

Figure 3. Results of ICA with a rotation mixing matrix and Laplace source distribution. (A) Initial and final distribution of $\mathbf{u} = (u_1, u_2)$ for each rule. Top left: Initial distribution of outputs common to all rules. Other panels: Final distribution of outputs for each rule. Horizontal and vertical axes respectively represent u_1 and u_2. Panels show samples of output signals pooled over the first or last 10^4 steps. **(B)** Velocity map of each ICA rule. Horizontal and vertical axes respectively indicate W_{11} and W_{12}, where synaptic strength matrix $W = (W_{11}, W_{12}; -W_{12}, W_{11})$. The direction of the arrow and color at each location represent the direction of the change of synaptic strengths, (W_{11}, W_{12}), and the length of the arrow represents the magnitude of the change. ICA solutions are located at $(W_{11}, W_{12}) = (\cos\theta, \sin\theta)$ for $\theta = \pi/6, 2\pi/3, 7\pi/6$, and $5\pi/3$. Top left: The color scale. Other panels: Velocity maps for the EGHR, Amari, Linsker, Cichocki, and Foldiak rules, respectively. The time constant of the sources was $\tau_s = 50$. The neuronal time constant for the Linsker and Foldiak rules was $\tau_v = 10$. The filled and open circles respectively indicate stable and unstable equilibrium points. The black and gray cross marks respectively indicate the common initial condition used in the main panels of (**A**) and the specific initial conditions used in the inset panels of (**A**). Note that the ICA result and the map of the Bell-Sejnowski rule are similar to those of the Amari rule.

To better understand the results, we next explore a velocity map that characterizes the dynamics of the synaptic weight matrix (Fig. 3B). Because A is a rotation matrix in this simulation, the synaptic weight matrix $W = (W_{11}, W_{12}; -W_{12}, W_{11})$ can be characterized by only two parameters, W_{11} and W_{12}. This is because W remains a rotation matrix during the entire learning phase, as long as it is initially set so. On each velocity map (see Methods for computational procedures), the directions of changes in the synaptic weight matrix are indicated by color for different values of W_{11} and W_{12}. Under the EGHR and Amari rule, all ICA solutions are stable and no spurious solutions exist (there are four ICA solutions at $(W_{11}, W_{12}) = (\cos\theta, \sin\theta)$ for $\theta = \pi/6$, $2\pi/3$, $7\pi/6$, and $5\pi/3$). In contrast, the three conventional local rules, the Linsker, Cichocki, and Foldiak rules, have basins of attraction for spurious solutions. The Linsker rule can approximate the Bell-Sejnowki rule if the time scales of dynamical variables are set appropriately and the time bin Δt is set small enough. Otherwise, it has spurious stable solutions at $W = 0$ and at infinity. This means that W converges to zero (or infinity) if an initial W is started too small (or too big). The basins of attraction for spurious solutions expand as the neuronal time constant (τ_y) becomes slow relative to the sources (τ_s) (see Fig. 3B for $\tau_y/\tau_s = 0.2$). They eventually remove all ICA solutions if $\tau_y \gg \tau_s$ (see S2.5.1 and S2.6.2 for analyses). The Foldiak rule also has a similar spurious stable solution at $W = 0$ even if τ_y is small, and has four additional spurious solutions near the diagonal lines of the plot, indicating that it fails if W is initially small or at near diagonal lines. The Cichocki rule can also fail depending on the initial conditions because one of the eigenvalues of A is negative in this example. In the current case, two of the ICA solutions, $(W_{11}, W_{12}) = (\cos\theta, \sin\theta)$ for $\theta = 2\pi/3$ or $7\pi/6$ are unstable, causing the synaptic strength matrix to diverge for a range of initial W.

Numerical simulations with a non-rotation mixing matrix and uniform source distribution. We next numerically explore another example using a non-rotation mixing matrix $A = (1, 0.5; 0.5; 1)$. We consider again two neurons to separate two independent sources for visualization purposes. Sources are generated from a uniform distribution, using the potential function U_U as explained above. Other parameter values are summarized in Table 1.

Figure 4A depicts the initial (top left) and final (other panels) distributions of the output (u_1, u_2). The two output variables should become independent if each rule is successful. Similar to the previous example, the EGHR and Amari rule successfully separate the independent sources, while the other local learning rules fail depending on the initial conditions. The Linsker rule does not work for the same reason as in the previous example. The Cichocki rule can fail regardless of the source distribution if the mixing matrix has a negative eigenvalue. The Foldiak rule generally cannot perform ICA if the mixing matrix is non-rotational (see S2.6.3).

Unlike the case with a rotation mixing matrix, we cannot easily visualize a velocity map with a non-rotation mixing matrix because W is characterized by more than two parameters. Instead, we monitor the time course of learning using the mutual information of outputs, defined by $I(\mathbf{u}) = \int d\mathbf{u} \mathrm{Prob}(\mathbf{u}) \log[\mathrm{Prob}(\mathbf{u})/\prod_i \mathrm{Prob}(u_i)]$[12] (Fig. 4B). This mutual information initially takes a finite value and then may converge to zero if all sources are successfully separated. Consistent with the results of Fig. 4A, the mutual information for the Linsker and Cichocki rules does not decrease if the initial synaptic strength matrix is not set appropriately. Even in the successful cases, the number of computational steps required for the Linsker and Foldiak rules is more than 10 times greater than that required for the EGHR because they need to update variable $\mathbf{v}$ in high time resolution before sources significantly change. In contrast, other learning rules, including the EGHR, require only sparse sampling of the input and yet can reach a solution within a similar physical time. (The bin size Δt is set to 10 for the Linsker rule, 1 for the Foldiak rule, and 100 for other rules in Fig. 4A,B.) Thus, the EGHR's tolerance to sparse sampling of input and the lack of a need to update extra variables (i.e., Q and $\mathbf{h}$) should be highly beneficial for hardware implementation–a slow clock time for a digital device or slow dynamics for an analog device with respect to a signal of interest would be sufficient for ICA.

In sum, while all conventional local learning rules have problems even with simple examples, the EGHR can reliably perform ICA similarly to the powerful non-local learning rules. Indeed, extensive numerical simulations demonstrate that the EGHR always converges to an ICA solution for a wide range of source dimensions and randomly sampled mixing matrices (Figs S1–3). Taken together, the simplicity and robust performance of the EGHR are highly advantageous for parallel and biological computation of ICA.

Undercomplete condition. In visual information processing, the number of neurons is usually much larger than that of the relevant sources; this case is called the undercomplete condition[16]. More generally, because a number of sources is unknown a priori and can change dynamically, it may be a good strategy to prepare a sufficient number of neurons in case they are required. Thus, if the brain performs ICA, the learning rule should be robust to the undercomplete condition. Here, we investigate whether the EGHR and conventional ICA rules can handle this condition.

We consider 32 neurons to separate two-dimensional sources. A mixing matrix A (32×2) is defined as a stack of 2×2 rotation matrices (see Methods for detail). To visualize the learning outcome, we define two-dimensional column vectors $\mathbf{k}_1, \ldots, \mathbf{k}_{32}$ according to $K = (\mathbf{k}_1, \ldots, \mathbf{k}_{32})^T = WA$, where these vectors characterize the relation between the outputs of individual neurons and the two sources by $u_i = \mathbf{k}_i^T \mathbf{s}$. In order for the model to perform ICA, a subset of neurons must encode the first source with $\mathbf{k}_i^T \propto (1, 0)$ and another subset must encode the second source with $\mathbf{k}_i^T \propto (0, 1)$.

For all rules, the $\mathbf{k}_i$ are initially distributed randomly on a unit circle as we defined (Fig. 5 top left). We found that only the EGHR successfully found an optimal representation of sources, where $\mathbf{k}_1, \ldots, \mathbf{k}_{32}$ were either along the horizontal or vertical axis (Fig. 5 top center). This result indicates that every neuron became specialized to one of two sources. In contrast, other learning rules were not successful in the undercomplete setting. Even with the non-local Amari rule, all the $\mathbf{k}_i$ kept mixing the two sources, thus failing to achieve ICA (Fig. 5 top right).

Indeed, we can mathematically show that the EGHR has a stable solution characterized by $\mathbf{k}_i^T \propto (1, 0)$ for some neurons and $\mathbf{k}_i^T \propto (0, 1)$ for others in a general undercomplete case (see Methods). The other rules cannot

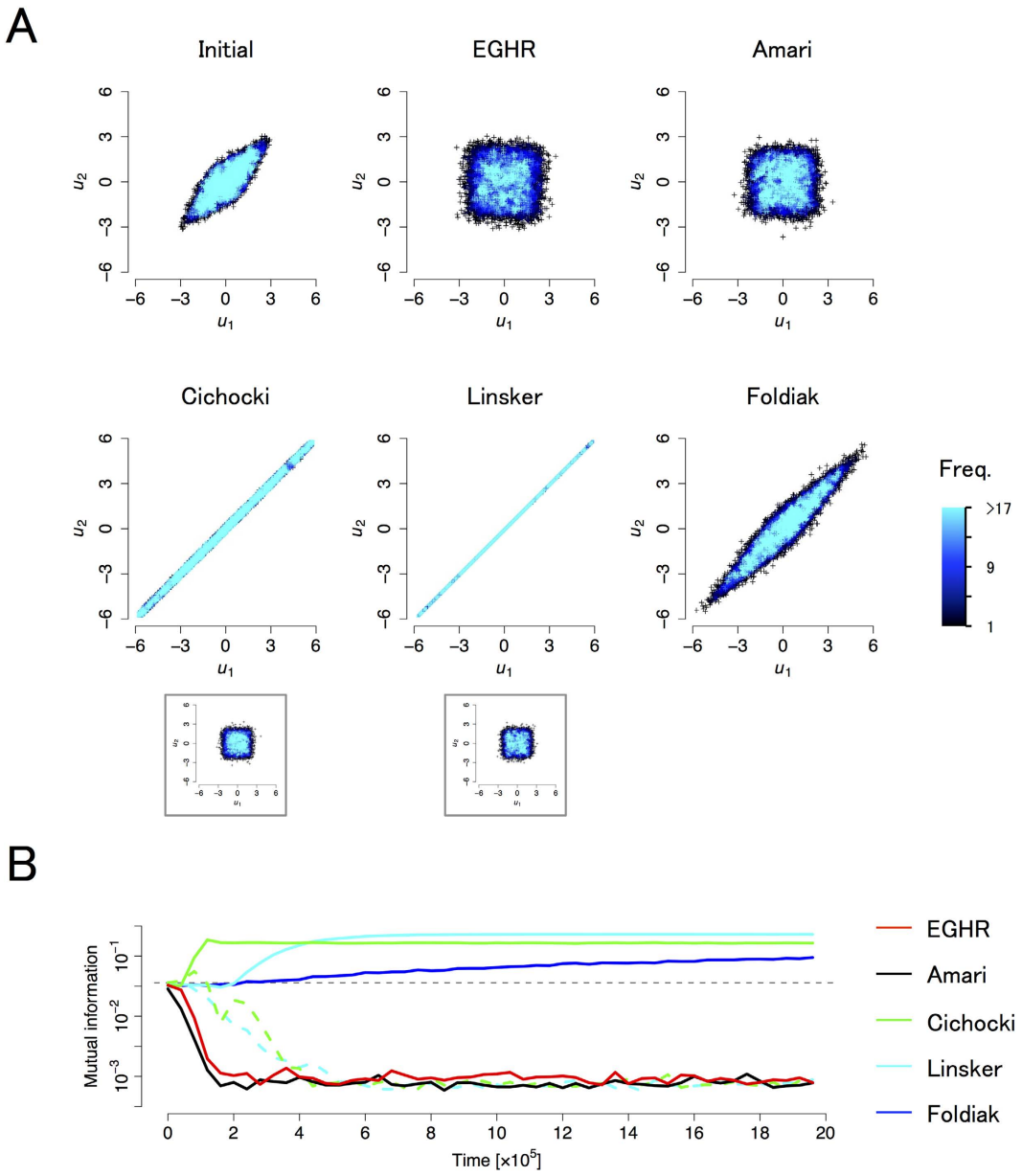

Figure 4. Results of ICA with a non-rotation mixing matrix and uniform source distribution. (A) Initial and final distributions of the outputs for each rule. Conventions are as in Fig. 3A. (B) The learning time course of each method assessed by mutual information of the outputs $I(\mathbf{u})$. The EGHR and Amari rule successfully perform ICA, whereas the Linsker and Cichocki rules fail depending on the initial synaptic strength matrix, and the Foldiak rule cannot handle a non-rotation mixing matrix. Time (x-axis) is defined by $k \times \Delta t$, where k is the number of computing steps and Δt is the time bin. Because the Linsker and Foldiak rules need a smaller time bin ($\Delta t = 10$ for Linsker, $= 1$ for Foldiak, and $= 100$ for others), they required more computational steps than other rules to reach a solution. The color of each learning rule is shown in the legend. Solid curves indicate the time courses of learning when started from a common initial condition $W = (-2.2, 0; 0, -2.2)$. Dashed curves are the time courses for the Cichocki and Linsker rules when started from a good initial condition, that is, $W = (-0.8, 0; 0, -0.8)$. A gray dashed line indicates $I(\mathbf{u})$ at the beginning of learning. Other parameters are summarized in Table 1.

generically find such representation (see S2.8). Furthermore, the EGHR successfully separates sources even if the number of sources (more than two) dynamically changes (Supplementary Movie 1). Taken together, only the EGHR can reliably separate and extract all independent sources in the undercomplete condition.

Application to separate natural images. Finally, we conducted computer simulations to demonstrate a promising application of the EGHR for BSS based on natural scenes. Figure 6 displays the result of BSS using the EGHR. Three pictures of a distinct felidae animal and one white noise image were used as sources (Fig. 6 top). The color intensities of the individual pixels were processed to gray scale pixels and then converted to real numbers

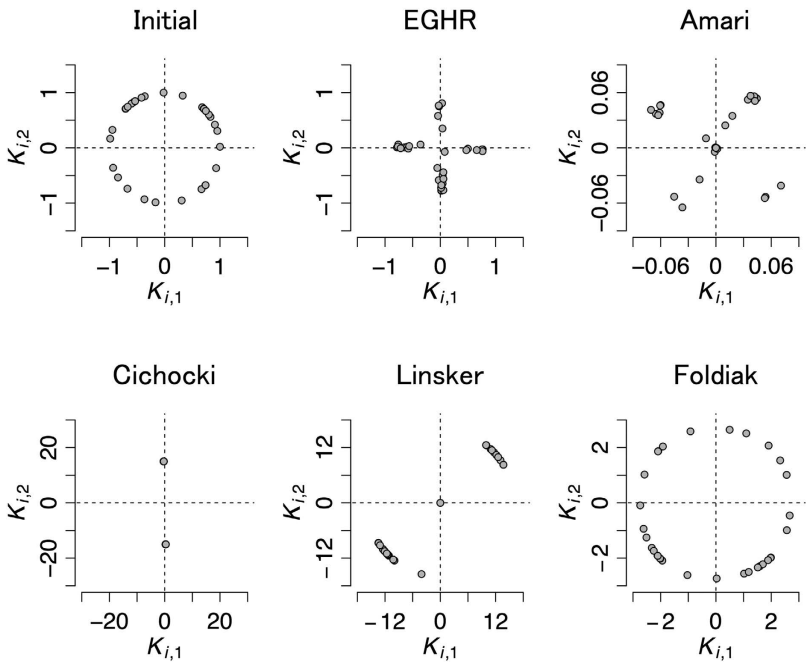

Figure 5. Results of ICA in the undercomplete condition. Thirty-two neurons were used to separate two sources. In each panel, the horizontal and vertical axes respectively represent the first and second elements of two-dimensional vector $\mathbf{k}_i$ ($i = 1, \ldots, 32$), which respectively represents the responsiveness of neuron i to the two sources. ICA is successful if $\mathbf{k}_i{}^T \propto (1, 0)$ for some i and $\mathbf{k}_i{}^T \propto (0, 1)$ for others. Initially, vectors $\mathbf{k}_1, \ldots, \mathbf{k}_{32}$ are randomly sampled on a unit circle (top left). The EGHR successfully performed ICA as indicated by $\mathbf{k}_1, \ldots, \mathbf{k}_{32}$ directed either along the horizontal or vertical axis (top center). On the other hand, the other learning rules did not achieve ICA (other panels). See Methods for other simulation details.

following[21] (see also Methods). The four images were randomly superposed to produce four mixed images using a 4×4 mixing matrix (Fig. 6 middle). These four mixed images were simultaneously sampled one pixel at a time (from an identical position) and fed into four model neurons as input. The model then learned synaptic strength matrix W according to the EGHR. Although grayscale images are used during training, the final results are obtained by providing color images as input to the learned network.

One issue is that, while the rule needs to assume a specific distribution of sources to compute the updates of the synaptic strengths in Equation 1, this distribution is unknown in practice. Based on our observation that the EGHR is robust to the detailed shape of a source distribution (Supplementary Fig. S3), we tested Laplace and uniform distributions for p_0 because only the difference of super- vs. sub-Gaussian is important. We found that the uniform distribution worked better with these images. Indeed, a posthoc analysis of the original source images confirmed that the true sources tended to obey a sub-Gaussian distribution with negative kurtosis (see inset panels in Fig. 6 top). Specifically, we ran the neural network for 2×10^7 steps using a uniform distribution for p_0 with learning time constant $\tau_W = 2 \times 10^4$, and found that a series of output $\mathbf{u}$ calculated after learning successfully achieved BSS by reconstructing natural images close to the originals (Fig. 6 bottom).

To further show the wide applicability of the EGHR, we next applied the learning rule to movies. This application was straightforward and the outcome was successful (Fig. 7; Supplementary Movie 2). This suggests the EGHR's potential for a wide range of applications. One minor difference in this example compared to the previous one is that, while the distribution of sources remained mostly sub-Gaussian, it sporadically turned super-Gaussian. Because of this transition, a small fraction of elements of the synaptic strength matrix did not converge and kept fluctuating. Nonetheless, BSS was overall successful using $g(\mathbf{u})$ and $E(\mathbf{u})$ functions designed based on a uniform source distribution.

Discussion

In this work, we proposed a new ICA rule, the EGHR, that requires only local information at each synapse for learning. We also showed that, in comparison to other ICA rules, the EGHR is the only local ICA rule that reliably works with various source statistics, mixing matrices, and number of sources.

Although we have focused on extracting independent sources in the external world in this paper, the method has a more general benefit–to adaptively organize the output of neurons nearly independent, regardless of the nature of the input. Generally, the "curse of dimensionality" makes it difficult to find a way to extract relevant information from multiple neurons[33] unless they use a particularly simple representation[1]. The situation is known to be much easier if each neuron codes information independently[34]. Thus, the EGHR could be a general computational principle in the brain to avoid the curse of dimensionality for decoding by self-organizing nearby neurons to acquire a nearly independent information coding scheme.

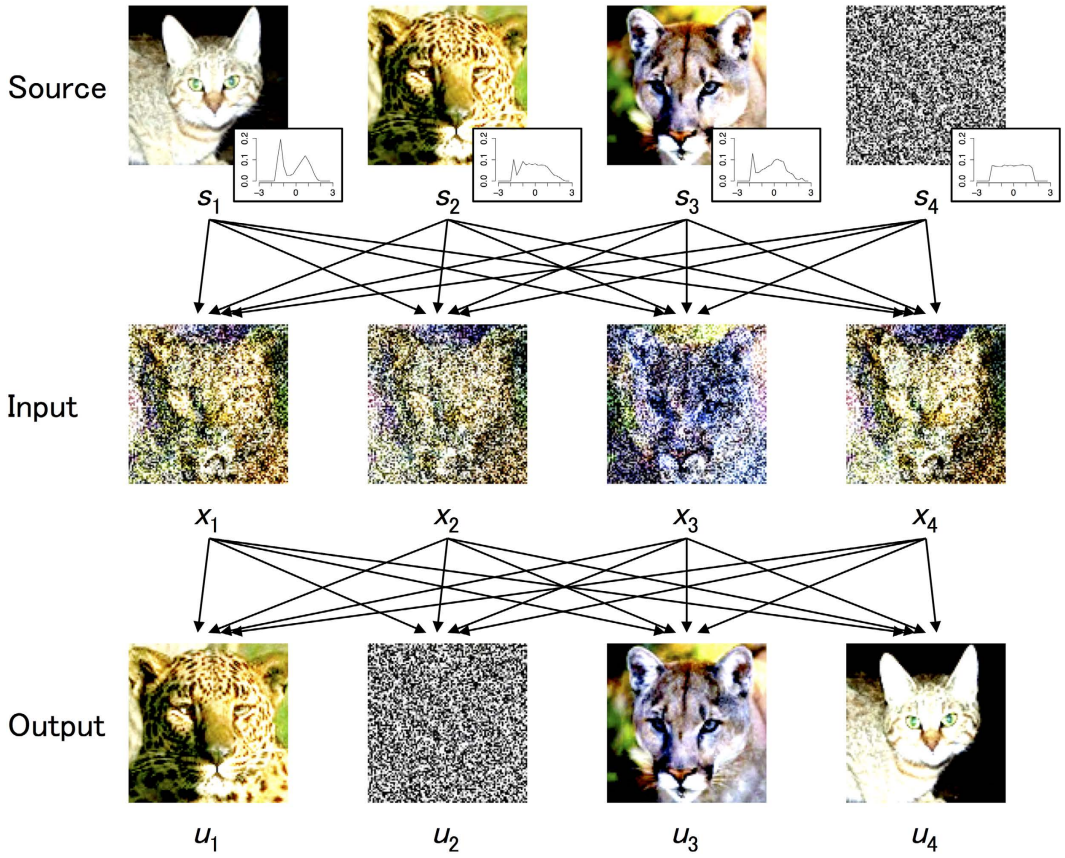

Figure 6. BSS of natural images. Top: Four original images as hidden signal sources. Middle: Four superposed images provided as input to the model. Bottom: Final states of the outputs of the neural network reconstructed original images. We retrieved these pictures from the Caltech101 dataset[50] (http://www.vision.caltech.edu/Image_Datasets/Caltech101/) and processed them accordingly.

In this study, we compared the EGHR with previously proposed learning rules, including local[19–21] and non-local[13–15] rules. We found that the EGHR is one of the most reliable learning rules among the previously proposed ICA rules in terms of the stability of genuine solutions and the absence of spurious solutions. In particular, all but the EGHR failed to extract independent sources under the undercomplete setting. Notably, the resulting stimulus representation by the EGHR utilizing all neurons, as opposed to a minimal number of neurons, is optimal according to the "infomax" principle for accurately representing sources in the presence of noise[16,35]. Although a non-local algorithm has been proposed to achieve ICA under the undercomplete setting[36], to our knowledge, the EGHR is unique in achieving such an undercomplete representation by only using local computations. In the real world, the number of independent sources is unknown and may differ from one condition to the next. Hence, it is natural to prepare enough neurons in case they are needed. In this view, the EGHR's capacity to handle undercomplete conditions is extremely beneficial in biological settings as well as in engineering applications.

Moreover, the EGHR automatically can whiten the inputs and rotate pre-whitened inputs to extract independent outputs (see S2.3–4 for additional analyses). In contrast, kurtosis-based methods such as Fast ICA[17] assume that inputs are already whitened. This is not ideal for parallel and biological implementation of ICA because signals are typically correlated in biological systems, e.g., in the cocktail party effect. Especially, inputs are inevitably correlated in undercomplete condition. Therefore, it is a big advantage of the EGHR to be able to perform decorrelation (i.e., whitening) and ICA (i.e., increasing non-Gaussianity) simultaneously.

To date, all ICA algorithms based on neuron-like units require extensive information sharing among output neurons. However, the communication is much simpler for the EGHR than it is for other rules. While specific communication for each pair of neurons is required for other ICA algorithms, a single global signal is sufficient for the EGHR. Moreover, the frequency of communication required is also advantageous for the EGHR over the Linsker and Foldiak rules. While the Linsker and Foldiak rules require virtually continuous updating of the neural activity during learning for stability, the EGHR requires only one forward and backward information passing before stimulus significantly changes with its time constant τ_s. Hence, the EGHR can successfully separate rapidly changing sources while requiring only minimal communication and processing by neurons.

In the brain, the strength of each synapse changes according to the co-activation of the pre- and post-synaptic factors, as described by Hebbian plasticity[22,37] or by its variants such as spike-timing dependent plasticity (STDP)[38,39]. It is critical for the EGHR that the outcome of Hebbian plasticity is modulated by a global signal.

Figure 7. Snapshots of BSS results using movies. Top: Four original images as hidden signal sources. Middle: Four superposed images provided as input to the model. Bottom: The final states of the outputs of the neural network reconstructed the original movies well (Supplementary Movie 2). We retrieved these movies from MotionElements (https://www.motionelements.com) and processed them accordingly.

Consistent with our proposal, recent experimental studies have reported the essential role of a third-factor such as GABA[23,24], neuromodulators[25–29], or glial signaling[30] in directly modulating Hebbian plasticity. While the importance of three-factor learning has been shown in many computational models[40–44], this is to our knowledge the first demonstration that it can play an essential role in ICA. A simple summation of surprise signal from output neurons is sufficient to compute the global error signal used for the EGHR. The surprise signal in each neuron is large if neural activity takes an unexpectedly high (or low) value, which happens more often in an unfamiliar rather than familiar environment. In this sense, the property of the surprise signal has noticeable commonality with the sensory saliency signal[45,46].

In addition, a simple circuit architecture of the EGHR provides an ease for neuromorphic computation. Conventional local ICA rules, including a recently proposed one[47], use lateral inhibition to decorrelate output neurons. This requires $N \times N$ mutual connections among the output neurons, and they need to be tuned in an activity-dependent manner (Fig. 8 bottom). Biologically, recurrent connections are rather sparse (~10% or less connectivity[48]), and this limitation can reduce the performance of conventional learning rules. Exactly the same set of problems arises if conventional learning rules are implemented in neuromorphic chips. The necessity of dense, specific, and dynamic recurrent connections can easily complicate the circuit architecture. In contrast, the EGHR robustly and efficiently performs ICA, only needing the interaction between output neurons and a global factor–thus, only $2N$ non-plastic recurrent connections suffice (Fig. 8 top). This feature makes the EGHR an excellent candidate for implementation using neuromorphic technology[49].

In summary, we developed a new local ICA rule based on a global error signal. The proposed rule performs an extremely robust ICA computation using only locally available information and a minimum number of operations. The broad applicability and easy implementation of the present rule could further advance neuromorphic computation and may reveal the principle underlying BSS computation in the brain.

Methods

Derivation of the proposed learning rule. The generative model of input signals is represented as $\mathbf{x} = A\mathbf{s}$, where $\mathbf{s} = (s_1, \ldots, s_M)^T$ and $\mathbf{x} = (x_1, \ldots, x_N)^T$ are M- and N-dimensional column vectors of independent sources and merged inputs, respectively, and A is an $N \times M$ transform matrix from sources to inputs (see Fig. 1A). The true probability distribution of $s_1, \ldots, s_M$ is represented as $p(s_i)$, i.e., we assume $s_1, \ldots, s_M$ independently obey an identical distribution. However, the true distribution is usually unknown to the observer, so we set the prior distribution of s_i as $p_0(s_i)$. The prior of $\mathbf{s}$ is represented as $p_0(\mathbf{s}) = \prod_i p_0(s_i)$. In addition, the distribution of $\mathbf{x}$ is represented as $p(\mathbf{x})$. The neural network model with linear firing rate is defined as $\mathbf{u} = W\mathbf{x}$, where $\mathbf{u} = (u_1, \ldots, u_N)^T$ is

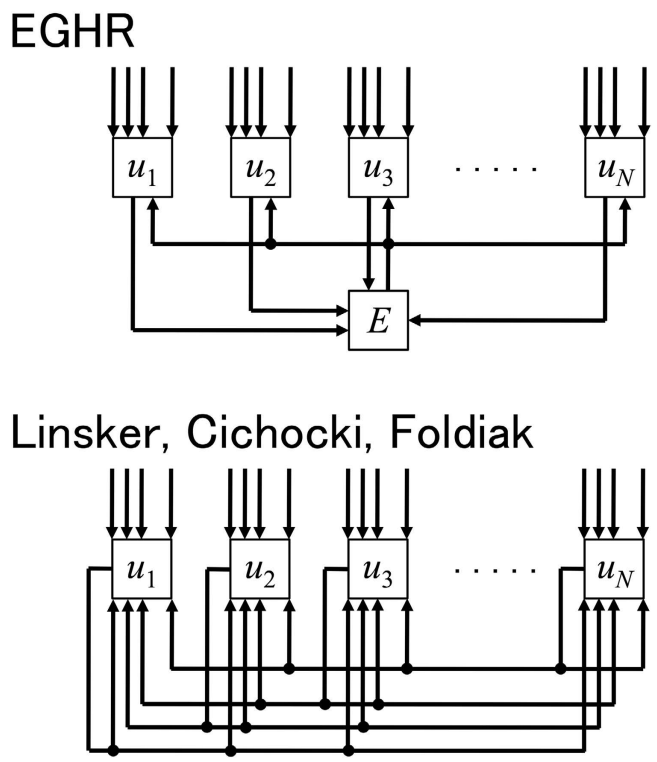

Figure 8. Diagrams of the proposed scheme and conventional scheme. Top: Diagram of a circuit implementing the EGHR. The EGHR requires recurrent connections, but they are only between the output neurons and the unique global factor E; therefore, the total number of recurrent connections is only $2N$. Bottom: Diagram of a conventional circuit implementing local ICA rules. Conventional local ICA rules use $N \times N$ recurrent connections between all output neurons (i.e., the Linsker, Cichocki, and Foldiak rules) and their strengths need to be learned (for the Linsker and Foldiak rules). Thus, the small number of fixed recurrent connections is a significant advantage of the EGHR with respect to possible applications in neuromorphic engineering.

an N-dimensional column vector of outputs, and W is an $N \times N$ transform matrix from the inputs to outputs. We define the probability distribution of $\mathbf{u}$ as $q(\mathbf{u})$, which is a posterior that the neural network recognizes. Unless specifically mentioned, we assume $M = N$. To perform infomax learning, as in the Bell-Sejnowski and Amari rules, $q(\mathbf{u})$ should become the same shape as $p_0(\mathbf{u})$ because $u_1, \ldots, u_N$ become independent of each other if and only if $q(\mathbf{u}) = p_0(\mathbf{u})$[16]. Hence, the Bell-Sejnowki and Amari rules minimize the Kullback-Leibler divergence[12] between $q(\mathbf{u})$ and $p_0(\mathbf{u})$, computed by $L_A = D_{KL}[q(\mathbf{u})||p_0(\mathbf{u})]$, to evaluate the distance of two distributions[13–15]. The idea of the proposed method is to use another cost function that is more tractable for the neural network. The proposed method (EGHR) is derived from a cost function L, which is a functional of a prediction error $E(\mathbf{u})$ (also known as a prior energy function). First, we define $E(\mathbf{u})$ by

$$E(\mathbf{u}) = -\sum_m \log p_0(u_m) = \sum_m z(u_m), \tag{8}$$

where $z(u_m) = -\log p_0(u_m)$ for all $m = 1, \ldots, N$. Next, we defined cost function L by

$$L = \frac{1}{2}\langle (E(\mathbf{u}) - E_0)^2 \rangle_{p(\mathbf{x})}, \tag{9}$$

where E_0 is a positive constant value defined depending on the shape of $p_0(\mathbf{s})$. The bracket $\langle \bullet \rangle_{p(\mathbf{x})}$ represents an expectation over input distribution $p(\mathbf{x})$, that is, $\langle \bullet \rangle_{p(\mathbf{x})} = \int \bullet\, p(\mathbf{x})\, d\mathbf{x}$. For simplification, we also write $\langle \bullet \rangle_{p(\mathbf{x})}$ as $\langle \bullet \rangle$. Learning in the EGHR occurs with the change in W and its goal is to minimize L, so that the dynamics of W are defined as the first order derivative of L, which is calculated as

$$\begin{aligned}\dot{W} &\propto -\frac{\partial L}{\partial W}\\ &= -\langle \partial/\partial W\{1/2\cdot(E(\mathbf{u})-E_0)^2\}\rangle\\ &= -\langle \partial\{1/2\cdot(E(\mathbf{u})-E_0)^2\}/\partial E\cdot\partial E/\partial\mathbf{u}(\partial\mathbf{u}/\partial W)^T\rangle\\ &= -\langle (E(\mathbf{u})-E_0)g(\mathbf{u})\mathbf{x}^T\rangle, \end{aligned} \tag{10}$$

where $g(u) = -\mathrm{d}\log p_0(u)/\mathrm{d}u$. Accordingly, Equation 1 is derived from Equation 9, although the learning time constant τ_W needs to be defined separately.

Equation 1 indicates that W is proportional to the expectation of the multiplication of global factor $(E_0 - E(\mathbf{u}))$ (a scalar) by the Hebbian term $g(\mathbf{u})\mathbf{x}^T$ (a matrix). The former can be regarded as a learning efficacy depending on $\mathbf{u}$ that is common for all neurons. Because we do not assume that $g(s_i)$ is a monotone increasing function of s_i, the EGHR potentially can be applied to sources with multimodal distributions if the distribution is within the linear-stability condition (see the following sections). Specifically, when a source distribution is $p_0(s_i) = 1/\sqrt{2}\cdot\exp(-\sqrt{2}\,|s_i|)$ (normal Laplace distribution), $g(s_i)$ becomes $g_L(s_i) = \sqrt{2}\,\mathrm{sgn}(s_i)$ and is approximated as $\sqrt{2}\,\tanh(\gamma\, s_i)$ for numerical calculations, where γ is a large positive constant. Similarly, when a source distribution is $p_0(s_i) = 1/2\sqrt{3}$ for $|s_i| < \sqrt{3}$ or 0 for otherwise (normal uniform distribution), $g(s_i)$ is approximated as $g_U(s_i) = -\gamma\tanh(-\gamma(s_i+\sqrt{3})) + \gamma\tanh(\gamma(s_i-\sqrt{3}))$ using large positive constant γ.

Although $E(\mathbf{u})$ in the EGHR is tractable, $H[q(\mathbf{u})]$ for infomax rules is more difficult to calculate for both biological neurons and computers because handling of non-Gaussian distribution $q(\mathbf{u})$ is required. This leads to the known difficulty of calculating the partial differential of $H[q(\mathbf{u})]$ by W, i.e., $\partial H[q(\mathbf{u})]/\partial W = W^{-T}$, in the Bell-Sejnowski equation (see Equation 2)[13,14]. The EGHR instead calculates $E(\mathbf{u})^2$, so that its partial differential $2E(\mathbf{u})g(\mathbf{u})\mathbf{x}^T$ is more tractable than W^{-T} for neurons.

Equilibrium point of the EGHR. We show $W = A^{-1}$ is an equilibrium point of Equation 1. Again, we write $z(u_i) = -\log p_0(u_i)$ and $E(\mathbf{u}) = \sum_m z(u_m)$. When $W = A^{-1}$, Equation 1 becomes $W \propto \langle (E_0 - \sum_m z(s_m))g(\mathbf{s})\mathbf{s}^T\rangle A^T$ since the relationship of $\mathbf{u} = A^{-1}A\mathbf{s} = \mathbf{s}$ holds. For simplification, we assume that $\langle \mathbf{s}\rangle = \mathbf{0}$. In this case, $\langle g(\mathbf{s})\rangle = -\int \mathrm{d}\mathbf{s}\ \mathrm{d}p_0(\mathbf{s})/\mathrm{d}\mathbf{s} = \mathbf{0}$. As $s_1, \ldots, s_N$ independently obey an identical distribution, we obtain $\langle g(s_i)s_j\rangle = \langle g(s_i)\rangle\langle s_j\rangle = 0$ for $i \neq j$. In addition, we obtain $\langle \sum_m z(s_m)g(s_i)s_j\rangle = 0$ for all m when $i \neq j$. On the other hand, when $i = j$, $\langle g(s_i)s_i\rangle$ and $\langle z(s_m)g(s_i)s_i\rangle$ have non-zero values. Using the relationships of $\langle g(s_i)s_i\rangle = 1$, $\langle z(s_m)g(s_i)s_i\rangle = \langle z(s_m)\rangle$ for $m \neq i$, and $\langle z(s_i)g(s_i)s_i\rangle = \langle z(s_i)\rangle + 1$ (see S2.1.1), we find that $\langle g(s_i)\, s_j\rangle = \delta_{ij}$ and $\langle \sum_m z(s_m)g(s_i)s_j\rangle = (N\langle z(s_i)\rangle + 1)\delta_{ij}$, where δ_{ij} is Kronecker's delta such that $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ for $i \neq j$. Let us derive the condition such that $\dot{W} = 0$ holds when $W = A^{-1}$. Using these relationships, Equation 1 is further calculated as $\dot{W} = (E_0 - N\langle z(s_i)\rangle - 1)A^T = 0$. Therefore, if and only if

$$E_0 = N\langle z(s_i)\rangle + 1 = \langle E(\mathbf{s})\rangle + 1 \tag{11}$$

holds, $W = A^{-1}$ is an equilibrium point of Equation 1. In this case, E_0 is a constant that only depends on the shape of $p_0(s_i)$ and the dimensions of $\mathbf{u}$. Notably, $W = 0$ is another equilibrium point of Equation 1 if we assume $g(\mathbf{0}) = \mathbf{0}$. However, this turns out to be an unstable equilibrium point (see S2.2).

With an unknown source distribution. In practical cases, however, the shape of the true distribution $p(s_i)$ is usually unknown. This means that the optimal choices for E_0 and g, i.e., $E_0 = -N\langle\log p(s_i)\rangle + 1$ and $g(u) = -\mathrm{d}\log p(u)/\mathrm{d}u$, are also unknown. Here, we show that the EGHR finds ICA solutions even if we choose E_0 to be an arbitrary positive scalar. (While we assume the optimal g in this section, we show in Fig. S3 that the performance of the EGHR is also robust to the choice of g.) Let us consider the situation where W is proportional to A^{-1}, that is, $W = cA^{-1}$, where c is a positive scalar. We assume that $E(c\mathbf{s})$ is an even function of $\mathbf{s}$ and $s_1, \ldots, s_N$ obey independently an identical distribution $p(s_i)$. In this condition, when $W = cA^{-1}$, Equation 1 becomes

$$\begin{aligned}\dot{W} &\propto \langle (E_0 - E(c\mathbf{s}))g(c\mathbf{s})\mathbf{s}^T A^T)\rangle\\ &= \mathrm{Diag}[E_0\langle g(cs_i)s_i\rangle - \langle E(c\mathbf{s})g(cs_i)s_i\rangle]A^T, \end{aligned} \tag{12}$$

where $\mathrm{Diag}[x_i]$ is a diagonal matrix in which the (i, i) elements are x_i and the non-diagonal elements are zero. Thus, if and only if the relationship of

$$E_0 = \frac{\langle E(c\mathbf{s})g(cs_i)s_i\rangle}{\langle g(cs_i)s_i\rangle} = \frac{\langle z(cs_i)g(cs_i)s_i\rangle}{\langle g(cs_i)s_i\rangle} + (N-1)\langle z(cs_i)\rangle \tag{13}$$

holds, $W = cA^{-1}$ becomes an equilibrium state of Equation 1. The existence of c that satisfies Equation 13 is guaranteed, if we assume that $g(\mathbf{0}) = \mathbf{0}$, $z(cs_i)$ is a convex function, and $g(cs_i)$ is a monotonically increasing function. In this case, the right hand side of Equation 13 is a monotonically increasing function of c that takes 0 at $c = 0$ and tends to be ∞ as c approaches ∞. Therefore, for any $E_0 > 0$, there is a positive c that gives the equilibrium point of the EGHR. For example, if we assume that sources obey $p_0(s_i) \propto \exp(-\beta\,|s_i|^\alpha)$ $(\alpha > 0, \beta > 0)$, then $z(cs_i)$ and $g(cs_i)$ are written as $z(cs_i) = \beta|cs_i|^\alpha = c^\alpha z(s_i)$ and $g(cs_i) = c^\alpha g(s_i)$. Therefore, $W = cA^{-1}$ is a equilibrium point of the EGHR if and only if $E_0 = c^\alpha(N\langle z(s_i)\rangle + 1)$ in this example.

Linear stability. We investigated the necessary and sufficient conditions for linear stability. In this and the following sections, we assume that the prior $p_0(s_i)$ is the same as the true distribution of the source, $W=A^{-1}$ is a solution of the EGHR according to Equation 11, and that $p_0(s_i)$ is an even function of s_i. Let us set $\rho=\mathrm{cov}(z(s_i), g'(s_i)s_i^2)$ and $\omega=\mathrm{cov}(z(s_i), g'(s_i))\langle s_i^2\rangle+\mathrm{cov}(z(s_i), s_i^2)\langle g'(s_i)\rangle$, where $\mathrm{cov}(x, y)$ indicates the covariance between x and y. We calculate d^2L, the second order differential form of L, at $W=A^{-1}$ as

$$d^2L=\sum_{i=1}^{N}(\rho+1)dK_{ii}{}^2+\left(\sum_{i=1}^{N}dK_{ii}\right)^2+\frac{1}{2}\sum_{i\neq j}\left(\omega dK_{ij}{}^2+2dK_{ij}dK_{ji}+\omega dK_{ji}{}^2\right), \tag{14}$$

Notably, K_{ij} is an element of matrix $K=WA$ and dK_{ij} is its differential form. We confirm that d^2L at $W=A^{-1}$ is definitely non-negative if and only if $\rho>-1$ and $\omega>1$ hold because a discriminant of a quadratic equation in the third term would be negative definite. Under this condition, $W=A^{-1}$ is a stable equilibrium point and gives the minimum value of Equation 9 (see S2.2 for details).

For example, if the sources obey $p_0(s_i)\propto\exp(-\beta|s_i|^\alpha)$ $(\alpha>0, \beta>0)$, we obtain $\rho=\alpha-1$ and $\omega=\langle s_i^2\rangle\langle(g(s_i))^2\rangle$. Therefore, d^2L is further calculated as $d^2L=\sum_i\alpha dK_{ii}{}^2+(\sum_i dK_{ii})^2+1/2\cdot\sum_{i\neq j}(\langle s_i^2\rangle\langle(z'(s_i))^2\rangle dK_{ij}{}^2+2dK_{ij}dK_{ji}+\langle s_i^2\rangle\langle(z'(s_i))^2\rangle dK_{ji}{}^2)$, which is definitely non-negative as long as $\alpha>0$ and $\langle s_i^2\rangle\langle(z'(s_i))^2\rangle>1$. Notably, numerical simulations suggested that $\langle s_i^2\rangle\langle(z'(s_i))^2\rangle$ is no less than one when $\alpha>0$. The above second order differential form is the same as that of the Amari rule[32] except for the extra $(\sum_i dK_{ii})^2$ term, where this positive term provides additional stability for the EGHR compared to the Amari rule.

The absence of spurious solutions and relaxation time. We analytically and numerically evaluated the absence of spurious solutions and relaxation time of the EGHR if there are more than two sources. We analytically showed that, if the source distribution is close to Gaussian, $W=A^{-1}$ and its permutation and sign-flips are the only stable equilibrium points of the EGHR (see S2.3 and S2.4 for details).

We then numerically confirmed that there was no local minimum found when the source obeyed either a Laplace or uniform distribution by calculating the relaxation time of W to a fixed point. Figs S1–3 graph the relaxation time with a variety of transform matrices (Fig. S1), source dimensions (Fig. S2), and presumed source distribution shapes (Fig. S3).

Undercomplete condition. We investigate the dynamics in the case where the output dimension $\dim(\mathbf{u})=N$ is larger than that of sources $\dim(\mathbf{s})=M$, that is, $N>M$. The input dimension $\dim(\mathbf{x})$ is the same as $\dim(\mathbf{u})=N$. As a special case, let us assume that $\mathbf{x}$ and $\mathbf{u}$ are $2M$-dimensional column vectors $(N=2M)$. Then W is a $2M\times 2M$ square matrix, and $A=({A_1}^T, {A_2}^T)^T$ is a $2M\times M$ block matrix. Similar to the previous section, we define a $2M\times M$ vertically long matrix $K=WA$. From the infomax viewpoint, the optimal solutions comprise $K=(\Lambda_1, \Lambda_2)^T$ and its permutations and sign-flips, where Λ_1 and Λ_2 are non-zero diagonal matrices. This is because the representation using two neurons per source $(\mathbf{u}=K\mathbf{s}=(\mathbf{s}^T\Lambda_1, \mathbf{s}^T\Lambda_2)^T)$ can more accurately convey the information of the sources than using a single neuron per source $(\mathbf{u}=(\mathbf{s}^T\Lambda_1, \mathbf{0}^T)^T)$ if there is a small amount of background noise. Therefore, when $K=(\Lambda_1, \Lambda_2)^T$, Equation 1 becomes

$$\dot{W}\propto\left\langle\left(E_0-E\begin{pmatrix}\Lambda_1\mathbf{s}\\ \Lambda_2\mathbf{s}\end{pmatrix}\right)\begin{pmatrix}g(\Lambda_1\mathbf{s})\\ g(\Lambda_2\mathbf{s})\end{pmatrix}\mathbf{s}^T\right\rangle A^T. \tag{15}$$

If we assume $K=(I, I)^T$, since the elements of $\mathbf{s}$ are independent of each other, Equation 15 is further calculated as

$$\begin{aligned}\dot{W}&\propto\left\langle\left(E_0-E\begin{pmatrix}\mathbf{s}\\ \mathbf{s}\end{pmatrix}\right)\begin{pmatrix}g(\mathbf{s})\\ g(\mathbf{s})\end{pmatrix}\mathbf{s}^T\right\rangle A^T\\ &=\left\langle\left(E_0-E\begin{pmatrix}\mathbf{s}\\ \mathbf{s}\end{pmatrix}\right)g(s_i)s_i\right\rangle\begin{pmatrix}I\\ I\end{pmatrix}A^T.\end{aligned} \tag{16}$$

Similar to the case where $\dim(\mathbf{s})=\dim(\mathbf{u})$, Equation 16 is in the equilibrium point if E_0 satisfies $\langle(E_0-E((\mathbf{s}^T, \mathbf{s}^T)^T))g(s_i)s_i\rangle=0$. Therefore, $K=(I, I)^T$ is an ICA solution of the EGHR. The same explanation can be applied to any case where $\dim(\mathbf{u})>\dim(\mathbf{s})$. Linear stability in the undercomplete condition also can be shown in a similar way.

For Fig. 1. We used a two-dimensional colored (Fig. 1A) and white (Fig. 1B) noises obeying a Laplace distribution. A transform matrix A was defined as $A=(1, 0.5; 0.5, 1)$ (Fig. 1A) or $A=(\cos 6/\pi, -\sin 6/\pi; \sin 6/\pi, \cos 6/\pi)$ (Fig. 1B). The initial state of connection strength matrix W was set to $W=(1.5, 0; 0, 1.5)$, i.e., the initial u_1 and u_2 were not independent. A learning time constant of $\tau_W=10^3$ and a time resolution of $\Delta t=100$ were used. Simulations were conducted over $T=2\times 10^6$ steps.

For Fig. 3A. A two-dimensional colored noise obeying a Laplace distribution with zero mean and a variance of one, generated by Equation 7 with $U_L(s)=\sqrt{2}|s_i|$, was used. The mixing matrix was set to rotation matrix $A=(\cos\theta, -\sin\theta; \sin\theta, \cos\theta)$ with $\theta=\pi/6$. The synaptic strength matrix was initially started from $W=(-1.5, 0; 0, -1.5)$ in the main panels. In the insets, final distributions with the desired initial conditions were used, namely $W=(1.5, 0; 0, 1.5)$ initially for the Cichocki rule, $W=(-0.8, 0; 0, -0.8)$ initially for the Linsker rule, and $W=(1.5\cos\pi/6, -1.5\sin\pi/6; 1.5\sin\pi/6, 1.5\cos\pi/6)$ initially for the Foldiak rule. A common learning time constant $\tau_W=10^3$ was used for the EGHR, Amari, and Cichocki rules. For the Linsker and Foldiak rules,

$\tau_W = 10^4$ and 10^6 were used, respectively. The time resolutions for each rule were $\Delta t = 100$ for the EGHR, Amari, and Cichocki rules, $\Delta t = 10$ for the Linsker rule, and $\Delta t = 1$ for the Foldiak rule. Simulations continued for $T = 2 \times 10^6$ steps. For the Foldiak rule, $f_F(u_i) = 1/(1 + \exp(-\sqrt{2}u_i^3))/0.225$ was used. To prevent the divergence of W, whenever $\sum_j W_{ij}^2$ exceeded 4^2, $(W_{i1}, \ldots, W_{iN})$ was rescaled to $(W_{i1}, \ldots, W_{iN}) \cdot 4/\sqrt{\sum_j W_{ij}^2}$. See also Table 1 for parameter details.

For Fig. 3B. A numerical integration along a probability distribution of source $p_0(\mathbf{s})$ was used instead of the Monte Carlo sampling method to calculate the expectations. A spatial resolution of $ds = 0.1$ and a range of $-20 \le s_i \le 20$ were used for all i. Parameters W_{11}, and W_{12} were moved within $-1.5 \le W_{11} < 1.5$, and $-1.5 \le W_{12} < 1.5$ in increments of 0.05 steps. For the Foldiak rule, $f_F(u_i) = 1/(1 + \exp(-\sqrt{2}u_i^3))/0.225$ was used. For the numerical calculation, we analytically simplified the Linsker rule as

$$\tau_W \dot{W} = a\Delta t/\tau_v \sum_{k=0}^{\infty} (I - aKK^T \Delta t/\tau_v)^k \rho((k+1)\Delta t) KK^T W^{-T} - \langle g(\mathbf{u})\mathbf{x}^T \rangle \tag{17}$$

and the Foldiak rule as

$$\tau_W \dot{W} = a\langle f(\mathbf{u})\mathbf{x}^T \rangle - W. \tag{18}$$

See S2.5.1 and S2.5.2 for derivation details. Note that $\rho(t)$ is the auto-correlation of a signal train generated from Equation 7. We define Equations 17 and 18 to be the reduced Linsker (R-Linsker) and the reduced Foldiak (R-Foldiak) rules, respectively. The numerical calculation in Fig. 3B is based on this R-Lisnker and R-Foldiak rules.

For Fig. 4A. Source signals were independently drawn from a two-dimensional colored uniform distribution with zero mean and a variance of one, generated by Equation 7 with $U_U(s) = 1/(2\sqrt{3})$ for $|s_i| \le \sqrt{3}$ and $U_U(s_i) = 0$ for $|s_i| > \sqrt{3}$. A non-rotation transform matrix $A = (1, 0.5; 0.5, 1)$ was used. In the main panels, the initial and final distributions with $W = (-2.2, 0; 0, -2.2)$ initially are shown. In the insets, the final distributions with desired initial conditions, namely $W = (-0.8, 0; 0, -0.8)$ initially for the Cichocki and Linsker rules, were used. For the Foldiak rule, $f_F(u_i) = 1/(1 + \exp(-100u_i))$ was used. Parameters other than $U(s)$, $p_0(\mathbf{s})$, A, initial W, and $f_F(u_i)$ are the same as in Fig. 3A.

For Fig. 4B. Mutual information between u_1 and u_2, $I(\mathbf{u}) = \langle \log q(\mathbf{u}) - \log q(u_1) - \log q(u_2) \rangle_{q(\mathbf{u})}$, was used for evaluation, where $q(\mathbf{u})$, $q(u_1)$, and $q(u_2)$ were calculated using a histogram method. The parameters are the same as in Fig. 4A.

For Fig. 5 and Supplementary Movie 1. We used two-dimensional colored noise obeying a Laplace distribution. A transform matrix A (32×2 vertically elongated rectangular matrix) was defined as a stack of 2×2 rotation matrices (16 of them vertically aligned) characterized by randomly selected angles. Sets of 32 inputs ($\mathbf{x}$) and 32 output neurons ($\mathbf{u}$) were prepared. Synaptic strength matrix was initially started from $W = I$ (the 32×32 identity matrix) for all learning rules. A common time constant $\tau_W = 2 \times 10^6$ was used for the EGHR, Amari, and Cichocki rules. For the Linsker rule, $\tau_W = 2 \times 10^7$. For the Foldiak rule, $\tau_W = 2 \times 10^9$. Time resolutions for each rule are the same as in Figs 3 and 4 (see Table 1). Simulations continued for $T = 4 \times 10^6$ steps. To prevent the divergence of W, whenever $\sum_j W_{ij}^2$ exceeded 4^2, $(W_{i1}, \ldots, W_{iN})$ was rescaled to $(W_{i1}, \ldots, W_{iN}) \cdot 4/\sqrt{\sum_j W_{ij}^2}$. For the Linsker and Foldiak rules, to prevent the divergence of Q and $\mathbf{v}$, whenever $\sum_j Q_{ij}^2$ exceeded 4^2, $(Q_{i1}, \ldots, Q_{iN})$ was rescaled to $(Q_{i1}, \ldots, Q_{iN}) \cdot 4/\sqrt{\sum_j Q_{ij}^2}$, and $(v_1, \ldots, v_N)$ was restricted within $-20 \le v_i \le 20$. Other parameters are the same as in Fig. 3A.

Supplementary Movie 1 shows the sources (left), inputs (center), and outputs (right). The dimension of the sources is changed once every 10^6 steps. Sets of 32 inputs ($\mathbf{x}$) and 32 output neurons ($\mathbf{u}$) were prepared. We calculated I_{ij}, the mutual information between u_i and u_j for all i, j. The positions of $u_1, \ldots, u_N$ were moved by a repelling force anti-proportional to I_{ij}. A learning time constant of $\tau_W = 10^6$ was used.

For Figs 6 and 7 and Supplementary Movie 2. The natural image data was processed according to a modified version of Cichocki's scheme[21]. We prepared four images with 100×100 pixels, three natural images (animal faces) and a noise image generated from a uniform distribution (Fig. 6 top). We retrieved these pictures from the Caltech101 dataset[50] (http://www.vision.caltech.edu/Image_Datasets/Caltech101/) and processed them accordingly. We used images transformed to grayscale as sources. Signal trains were defined as trains constructed by the values of each pixel that were randomly extracted from the entire area of the image to construct the $T = 2 \times 10^7$ length train. The signal trains were transformed to zero-mean signals and their variances were normalized before the start of the learning procedure. As the sources were four images, we obtained four-dimensional source trains. Input trains $\mathbf{x}$ were prepared as the multiplication $\mathbf{x} = A\mathbf{s}$ of a transform matrix $A = (0.4, 0.65, -0.4, -0.8; 0.4, 0.4, -0.4, 0.9; -0.4, -0.4, 0.6, 0.8; 0.7, 0.5, -0.5, -0.8)$ and a vector $\mathbf{s}$, where $\mathbf{x}$ could also be images that were a mixture of the original images (Fig. 6 middle). The EGHR with a learning time constant of $\tau_W = 2 \times 10^6$ was used to perform ICA (Fig. 6C bottom). The results of BSS are displayed using the color images as input while the grayscale images were used during training.

For BSS of movies (Fig. 7 and Supplementary Movie 2), data was randomly sampled in the same way. We retrieved these movies from MotionElements (https://www.motionelements.com) and processed them accordingly. We prepared four movies with 200×200 pixels, three natural movies and a noise movie generated from a uniform distribution. Other parameters are common with Fig. 6.

References

1. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73,** 415–434 (2012).
2. Bronkhorst, A. W. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica United with Acustica* **86,** 117–128 (2000).
3. Haykin, S. & Chen, Z. The cocktail party problem. *Neural Comput.* **17,** 1875−1902 (2005).
4. Golumbic, E. M. Z. *et al.* Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron* **77,** 980–991 (2013).
5. Belouchrani, A., Abed-Meraim, K., Cardoso, J. F. & Moulines, E. A blind source separation technique using second-order statistics. *Signal Processing IEEE Trans.* **45,** 434–444 (1997).
6. Choi, S., Cichocki, A., Park, H. M. & Lee, S. Y. Blind source separation and independent component analysis: A review. *Neural Inf. Proc. Lett. Rev.* **6,** 1–57 (2005).
7. Cichocki, A., Zdunek, R., Phan, A. H. & Amari, S. I. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. (John Wiley & Sons, 2009).
8. Comon, P. & Jutten, C. (Eds.) *Handbook of Blind Source Separation: Independent Component Analysis and Applications.* (Academic Press, 2010).
9. Hyvärinen, A., Karhunen, J. & Oja, E. *Independent Component Analysis* (John Wiley & Sons, 2004).
10. Dayan, P. & Abbott, L. F. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (MIT Press, London, 2001).
11. Gerstner, W. & Kistler, W. M. *Spiking Neuron Models: Single Neurons, Populations, Plasticity* (Cambridge University Press, Cambridge, 2002).
12. Bishop, C. M. & Nasrabadi, N. M. *Pattern Recognition and Machine Learning* (Springer, New York, 2006).
13. Bell, A. J. & Sejnowski, T. J. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* **7,** 1129–1159 (1995).
14. Bell, A. J. & Sejnowski, T. J. The "independent components" of natural scenes are edge filters. *Vision Res.* **37,** 3327–3338 (1997).
15. Amari, S. I., Cichocki, A. & Yang, H. H. A new learning algorithm for blind signal separation. *Adv. Neural. Inf. Proc. Sys.* **8,** 757–763 (1996).
16. Lee, T. W., Girolami, M., Bell, A. J. & Sejnowski, T. J. A unifying information-theoretic framework for independent component analysis. *Comput. Math. Appl.* **39,** 1–21 (2000).
17. Hyvärinen, A. & Oja, E. A fast fixed-point algorithm for independent component analysis. *Neural Comput.* **9,** 1483–1492 (1997).
18. Hyvärinen, A. & Oja, E. Independent component analysis: Algorithms and applications. *Neural Net.* **13,** 411–430 (2000).
19. Foldiak, P. Forming sparse representations by local anti-Hebbian learning. *Biol. Cybern.* **64,** 165–170 (1990).
20. Linsker, R. A local learning rule that enables information maximization for arbitrary input distributions. *Neural Comput.* **9,** 1661–1665 (1997).
21. Cichocki, A., Karhunen, J., Kasprzak, W. & Vigario, R. Neural networks for blind separation with unknown number of sources. *Neurocomputing* **24,** 55–93 (1999).
22. Hebb, D. O. *The Organization of Behavior* (Wiley, New York, 1949).
23. Hayama, T. *et al.* GABA promotes the competitive selection of dendritic spines by controlling local Ca^{2+} signaling. *Nat. Neurosci.* **16,** 1409–1416 (2013).
24. Paille, V. *et al.* GABAergic Circuits Control Spike-Timing-Dependent Plasticity. *J. Neurosci.* **33,** 9353–9363 (2013).
25. Reynolds, J. N., Hyland, B. I. & Wickens, J. R. A cellular mechanism of reward-related learning. *Nature* **413,** 67–70 (2001).
26. Zhang, J. C., Lau, P. M. & Bi, G. Q. Gain in sensitivity and loss in temporal contrast of STDP by dopaminergic modulation at hippocampal synapses. *Proc. Natl. Acad. Sci. USA* **106,** 13028–13033 (2009).
27. Yagishita, S. *et al.* A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science* **345,** 1616–1620 (2014).
28. Salgado, H., Köhr, G. & Treviño, M. Noradrenergic 'tone' determines dichotomous control of cortical spike-timing-dependent plasticity. *Sci. Rep.* **2,** 417 (2012).
29. Johansen, J. P. *et al.* Hebbian and neuromodulatory mechanisms interact to trigger associative memory formation. *Proc. Natl. Acad. Sci. USA* **111,** E5584–E5592 (2014).
30. Henneberger, C., Papouin, T., Oliet, S. H. & Rusakov, D. A. Long-term potentiation depends on release of D-serine from astrocytes. *Nature* **463,** 232–236 (2010).
31. Cover, T. M. & Thomas, J. A. *Elements of information theory* (John Wiley & Sons, 2012).
32. Amari, S. I., Chen, T. P. & Cichocki, A. Stability analysis of learning algorithms for blind source separation. *Neural. Net.* **10,** 1345–1351 (1997).
33. Latham, P. E. & Nirenberg, S. Synergy, redundancy, and independence in population codes, revisited. *J. Neurosci.* **25,** 5195–5206 (2005).
34. Wu, S., Nakahara, H. & Amari, S. I. Population coding with correlation and an unfaithful model. *Neural Comput.* **13,** 775–797 (2001).
35. Linsker, R. Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Comput.* **4,** 691–702 (1992).
36. Amari, S. I., Chen, T. & Cichocki, A. Nonholonomic orthogonal learning algorithms for blind source separation. *Neural Comput.* **12,** 1463–1484 (2000).
37. Bliss, T. V. & Lømo, T. Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *J. Physiol.* **232,** 331–356 (1973).
38. Markram, H., Lübke, J., Frotscher, M. & Sakmann, B. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* **275,** 213–215 (1997).
39. Bi, G. Q. & Poo, M. M. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing synaptic strength and postsynaptic cell type. *J. Neurosci.* **18,** 10464–10472 (1998).
40. Izhikevich, E. M. Solving the distal reward problem through linkage of STDP and dopamine signaling *Cereb. Cortex* **17,** 2443–2452 (2007).
41. Florian, R. V. Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. *Neural Comput.* **19,** 1468–1502 (2007).
42. Legenstein, R., Pecevski, D. & Maass, W. A learning theory for reward-modulated spike-timing-dependent plasticity with application to biofeedback. *PLos Comput. Biol.* **4,** e1000180 (2008).
43. Urbanczik, R. & Senn, W. Reinforcement learning in populations of spiking neurons. *Nat. Neurosci.* **12,** 250–252 (2009).
44. Frémaux, N., Sprekeler, H. & Gerstner, W. Functional requirements for reward-modulated spike-timing-dependent plasticity. *J. Neurosci.* **30,** 13326–13337 (2010).
45. Itti, L., Koch, C. & Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pat. Anal. Mach. Intel.* **20,** 1254–1259 (1998).
46. Li, Z. A saliency map in primary visual cortex. *Trends. Cogn. Sci.* **6,** 9–16 (2002).
47. Brito, C. S. & Gerstner, W. Nonlinear Hebbian learning as a unifying principle in receptive field formation. *arXiv preprint* arXiv:1601.00701 (2016).

48. Song, S. *et al.* Highly nonrandom features of synaptic connectivity in local cortical circuits *PLos Biol.* **3,** e68 (2005).
49. Chicca, E., Stefanini, F., Bartolozzi, C. & Indiveri, G. Neuromorphic electronic circuits for building autonomous cognitive systems. *Proc. IEEE* **102,** 1367–1388 (2014).
50. Fei-Fei, L., Fergus, R. & Perona, P. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *IEEE CVPR 2004*, *Workshop on Generative-Model Based Vision 200*4. 178 (2003).

Acknowledgements

We are grateful to Andrzej Cichocki and Shun-ichi Amari for helpful discussions. This work was supported by RIKEN Brain Science Institute (TT), Brain/MINDS from AMED (TT), and the Japan Society for the Promotion of Science (https://www.jsps.go.jp/english/) through Grant-in-Aid for JSPS Fellows, Grant 26-8435 (TI). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Conceived and designed the experiments: T.I. and T.T. Performed the experiments: T.I. and T.T. Analyzed the data: T.I. and T.T. Wrote the paper: T.I. and T.T.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Isomura, T. and Toyoizumi, T. A Local Learning Rule for Independent Component Analysis. *Sci. Rep.* **6**, 28073; doi: 10.1038/srep28073 (2016).