SCIENTIFIC REPORTS

OPEN

Multi-context blind source separation by error-gated Hebbian rule

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

Received: 27 November 2018
Accepted: 23 April 2019
Published online: 09 May 2019

Animals need to adjust their inferences according to the context they are in. This is required for the multi-context blind source separation (BSS) task, where an agent needs to infer hidden sources from their context-dependent mixtures. The agent is expected to invert this mixing process for all contexts. Here, we show that a neural network that implements the *error-gated Hebbian rule* (EGHR) with sufficiently redundant sensory inputs can successfully learn this task. After training, the network can perform the multi-context BSS without further updating synapses, by retaining memories of all experienced contexts. This demonstrates an attractive use of the EGHR for dimensionality reduction by extracting low-dimensional sources across contexts. Finally, if there is a common feature shared across contexts, the EGHR can extract it and generalize the task to even inexperienced contexts. The results highlight the utility of the EGHR as a model for perceptual adaptation in animals.

Inference of the causes of a sensory input is one of the most essential abilities of animals[1–3] — a famous example is the cocktail party effect, i.e., the ability of a partygoer to distinguish a particular speaker's voice against a background of crowd noise[4,5]. This ability has been modelled by blind source separation (BSS) algorithms[6,7], by considering that several hidden sources (speakers) independently generate signal trains (voices), while an agent receives mixtures of signals as sensory inputs. A neural network, possibly inside the brain, can invert this mixing process and separate these sensory inputs into hidden sources using a BSS algorithm. Independent component analysis (ICA), achieves BSS by minimizing the dependency between output units[8,9]. Numerous ICA algorithms have been proposed for both rate-coding[10–13] and spiking neural networks[14].

Previously, we developed a biologically plausible ICA algorithm, referred to as the *error-gated Hebbian rule (EGHR)*[15]. This learning rule can robustly estimate the hidden sources that generate sensory data without supervised signals. Importantly, it can reliably perform ICA in undercomplete conditions[16], where the number of inputs is greater than that of outputs. A simple extension of the EGHR can separate sources while removing noise within a single-layer neural network[17], by simultaneously performing principal component analysis (PCA)[18,19] and ICA. The EGHR is expressed as a product of pre- and post-synaptic neuronal activities and a third modulatory factor, each of which can be computed locally (i.e., local learning rule[16]). In this sense, the EGHR is more biologically plausible than non-local engineering ICA algorithms[10–12]. Because of these desirable properties, the EGHR is considered as a candidate mechanism for neurobiological BSS[20–22], as well as a next-generation neuromorphic implementation[23,24] for energy efficient BSS.

The optimal inference and behavior often depend on context. Indeed, our perception and decisions reflect this context dependency, i.e., cognitive flexibility[25]. Studies in primates have suggested that a contextual-cue-dependent dynamic process in the prefrontal cortex controls this behavior[26–28], and several computational studies have modeled it[29–32]. Likewise, context dependence of auditory perceptual inference has been modeled[33]. In addition to experimental evidence, recent progress in machine learning has also addressed this multi-context problem, in an attempt to create artificial general intelligence[34–36]. By implementing (task-specific) synaptic consolidation, a neural network can learn a new environment, while retaining past memories, by protecting synaptic strengths that are important to memorizing past environments. Those findings indicate the importance of multi-context processes for cognitive flexibility.

Unlike the above-mentioned tasks, BSS in several different contexts has some difficulty. Conventional ICA algorithms assume the same number of input and output neurons[10–12,37,38] and cannot straightforwardly perform a multi-context BSS. After learning, the synaptic strength matrix of these algorithms converges to the inverse of the mixing matrix of the current context (or its permutation or sign-flip), which is generally different from that in the previous context. Hence, when the network subsequently encounters a previously learnt context, it needs to

[1]Laboratory for Neural Computation and Adaptation, RIKEN Center for Brain Science, Wako, Saitama, 351-0198, Japan. [2]RIKEN CBS-OMRON Collaboration Center, Wako, Saitama, 351-0198, Japan. Correspondence and requests for materials should be addressed to T.I. (email: takuya.isomura@riken.jp) or T.T. (email: taro.toyoizumi@riken.jp)

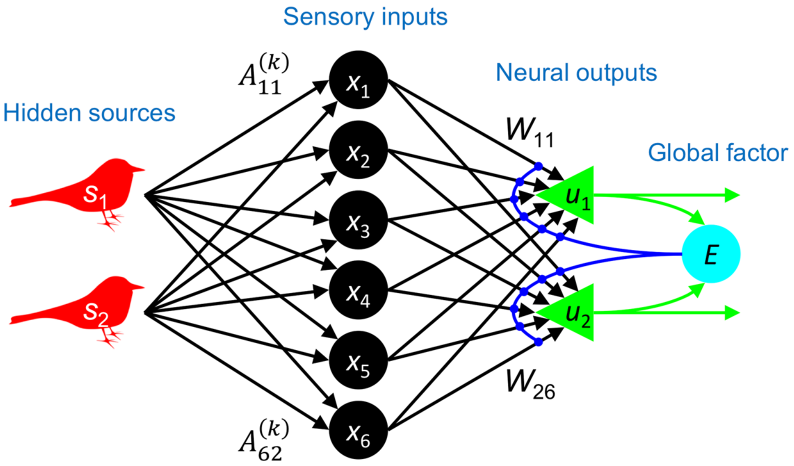

Figure 1. Model setup for multi-context BSS task. In this model, $s_1, \ldots, s_{N_s}$ are hidden sources (e.g., birdsongs); $x_1, \ldots, x_{N_x}$ are sensory inputs that an agent receives; $u_1, \ldots, u_{N_u}$ are neural outputs; $A_{11}^{(k)}, \ldots, A_{1N_s}^{(k)}, A_{21}^{(k)}, \ldots, A_{N_xN_s}^{(k)}$ are elements of the k-th-context mixing matrix; $W_{11}, \ldots, W_{1N_x}, W_{21}, \ldots, W_{N_uN_x}$ are synaptic strengths; and E is a scalar global factor that mediates synaptic plasticity. Synaptic strengths are adjusted to perform multi-context BSS by the EGHR.

relearn the synaptic strengths from the very beginning. More involved engineering ICA algorithms, such as the non-holonomic ICA algorithm[39] and the ICA mixture algorithm[40,41], are expected to perform the multi-context BSS. However, a biological implementation of these non-local learning rules is unclear. Further, as we show below, they cannot learn to compress redundant inputs by extracting the underlying low-dimensional hidden sources.

Here we show that the EGHR can perform multi-context BSS when a neural network receives redundant sensory inputs. It can retain memories of previously experienced contexts and process the BSS right after contextual switching to a previously learnt context. This suggests that the EGHR can also be used as a powerful data compression method[42], since it extracts low-dimensional hidden sources across contexts, despite the proportional increase in data dimensions to the number of contexts. Moreover, when a common feature is shared across contexts, the EGHR can extract it to perform BSS, while filtering out features that vary among contexts. Once the learning is achieved, the network can perform BSS even in an inexperienced context, indicating some generalization capability or transfer learning. We demonstrate that the EGHR with sufficiently redundant sensory inputs learns to distinguish birdsongs from their superpositions and retains this ability even after learning different sets of birdsongs. The rule finds a general representation that is capable of separating an unheard set of birdsongs. Finally, possible neurobiological implementations of the EGHR are discussed.

Results

Error-gated Hebbian rule (EGHR).

In a BSS task, several hidden sources (s) independently generate signal traces, while our agent receives their mixtures as sensory inputs (x). In this study, we considered a multi-context BSS task, in which a set of contexts with different mixing weights was used. Sensory inputs were randomly generated from one of these contexts for a period of time, with k ($=1, \ldots, C$) being an index of context. Our experimental setup consisted of an N_s-dimensional vector of hidden sources $s \equiv (s_1, \ldots, s_{N_s})^T$ whose elements s_i independently follow a non-Gaussian distribution $p(s_i)$, an N_x-dimensional vector of sensory inputs $x \equiv (x_1, \ldots, x_{N_x})^T$, and an N_u-dimensional vector of neural outputs $u \equiv (u_1, \ldots, u_{N_u})^T$ (Fig. 1). The sensory inputs in the k-th condition were generated by transforming the hidden sources, i.e., the so-called generative process:

$$\textbf{Sensory inputs} \quad x = A^{(k)} s. \tag{1}$$

Here $A^{(k)}$ is the $N_x \times N_s$ mixing matrix for the k-th context that defines the magnitude of inputs, when each source generates a signal. To ensure that each $A^{(k)}$ represents a different context and that each context has an ICA solution, column vectors of a block matrix $(A^{(1)}, A^{(2)}, \ldots, A^{(C)})$ are supposed to be linearly independent of each other. We designed the task such that these contexts appear sequentially or randomly. The neural outputs were expressed as sums of inputs weighted by an $N_u \times N_x$ synaptic strength matrix W, and calculated by:

$$\textbf{Neural outputs} \quad u = Wx. \tag{2}$$

It is well known that when a presynaptic neuron (x_j) and a postsynaptic neuron (u_i) fire together, Hebbian plasticity occurs, and the synaptic connection from x_j to u_i, denoted by W_{ij}, is strengthened[43,44]. Because this constitutes associative learning, correlations between $x_1, \ldots, x_{N_x}$ and u_i are usually enhanced; thereby, correlations among neural outputs also increase. This process is distinct from separation of signals (i.e., BSS) for which each neural output is expected to encode a specific source. To separate signals, we introduced a global scalar factor (i.e., a third factor) given by the sum of nonlinearly-transformed output units[15]:

$$\textbf{Global factor}\quad E(u)\equiv-\log p_0(u)=-\sum_{i=1}^{N_u}\log p_0(u_i). \tag{3}$$

Here $p_0(u)$ is the prior distribution that the agent expects the hidden sources to follow; e.g., when $p_0(u)$ is a Laplace distribution of mean zero and unit variance, then $E(u)=\sqrt{2}\left(|u_1|+\ldots+|u_{N_u}|\right)+\text{const}$. We supposed that this global factor modulates Hebbian plasticity. Recent experimental studies have reported that synaptic plasticity can be modulated by various neuromodulators[45–49], GABAergic inputs[50,51], or glial factors[52]. Possible neurobiological implementations of the global factor are further discussed in the Discussion section. Overall, the synaptic strength matrix W is updated by the EGHR in the following way:

$$\textbf{Synaptic plasticity (EGHR)}\quad \dot{W}\propto\langle\underbrace{(E_0-E(u))}_{\text{global factor}}\underbrace{g(u)}_{\text{post}}\underbrace{x^T}_{\text{pre}}\rangle, \tag{4}$$

where $\dot{W}$ with respect to time, $\langle\cdot\rangle$ is the expectation over the input distribution, and $g(u)\equiv \mathrm{d}E(u)/\mathrm{d}u$ is a non-linear function usually associated with a nonlinear activation function. A constant E_0 scales the neural outputs; the output scale becomes equivalent to the source scale when $E_0=\langle-\log p_0(s)\rangle+1$. In short, the EGHR constitutes a Hebbian learning rule when the global factor is smaller than the threshold ($E(u)<E_0$); otherwise ($E(u)>E_0$), it becomes an anti-Hebbian rule. This mechanism makes output neurons independent from each other. The detailed derivation and theoretical proofs of the EGHR have been described in our previous reports[15,17]. Briefly, the EGHR is derived as the gradient descent of the cost function $L\equiv\langle(E(u)-E_0)^2\rangle/2$. This is the cost for having dependency among outputs, designed for measuring the nonlinear correlation among elements of u. Hence, the minimization of L makes the elements of u independent of each other. The formal relationship between the EGHR and ICA algorithm based on the infomax principle is described in[17].

Memory capacity of the EGHR. First, we analytically show the memory capacity of a neural network established by the EGHR. As the number of contexts increases, larger dimensions of inputs are needed to retain information pertaining to past contexts in the neural network. For simplicity, we supposed that $N_u=N_s$. Because the network represents a linear inverse model of the generative processes, the goal of the multi-context BSS is generally given by:

$$W(A^{(1)},\ \ldots,\ A^{(C)})=(\Omega^{(1)},\ \ldots,\ \Omega^{(C)}), \tag{5}$$

where $\Omega^{(k)}$ is an $N_u\times N_s$ matrix equivalent to the identity matrix, up to permutations and sign-flips. This is because the success of BSS is defined by one-by-one mapping from sources to outputs. Thus, the multi-context BSS is successful if and only if a set of mixing matrices $(A^{(1)},\ldots,A^{(C)})$ expresses a full-column-rank matrix (see Methods for the derivation). Hence, we found that the following conditions are necessary to achieve the multi-context BSS for a generic $(A^{(1)},\ldots,A^{(C)})$: (1) the input dimension needs to be equal to or larger than the number of contexts times the number of sources, $N_x\geq CN_s$; and (2) the output dimension needs to be equal to or larger than the source dimension, $N_u\geq N_s$. Note that the neural network learns the information representation that compresses the sensory inputs, because we considered the input dimensions that are much greater than the output dimensions.

The memory capacity of the EGHR was empirically confirmed by numerical simulations (Fig. 2). Here, we supposed that two contexts generated inputs alternately. In each context, six-dimensional inputs were generated from two-dimensional sources with different mixing weights, as denoted by $A^{(1)}$ and $A^{(2)}$ (see top and middle rows in Fig. 2A). A neural network consisting of six input and two output neurons received the inputs and changed its synaptic strengths through the EGHR (i.e., training). After training, each neural output came to selectively respond to (i.e., encode) one of the two sources (bottom row in Fig. 2A). Thus, the network achieved separation of the sensory inputs into their sources without being taught the mixing weights (i.e., BSS).

Crucially, the neural network was able to retain the information learnt for all past contexts if provided with sufficiently redundant sensory inputs. This property is illustrated by the trajectories of the BSS error and EGHR cost function in Fig. 2B. We defined the BSS error for context k as the ratio of first to second maximum absolute values averaged for every row and column of matrix $K^{(k)}\equiv WA^{(k)}$ (see Methods for the mathematical definition of the BSS error). Here $K^{(k)}$ expresses the mapping from sources to outputs, which is equivalent to the covariance matrix between hidden sources and neural outputs $K^{(k)}=\mathrm{Cov}(u,s)$. This definition of the BSS error was made to ensure that the value was zero if and only if one source mapped onto one output, and *vice versa*; otherwise the value was positive and less than one. Moreover, the cost function of the EGHR was defined as the expectation of the square of the global factor: $L=\langle(E_0-E(u))^2\rangle/2$. Context 1 (red in Fig. 2A) was provided in the first session. Since synaptic strengths started from a random initial state, the BSS error at the beginning of the first session was large; then, the network learned an optimal set of synaptic strengths, and the error became zero, which was achieved by minimizing the cost function through gradient descent updates. When context 2 (blue in Fig. 2A) was provided for the first time in the second session, the EGHR cost function transiently increased, as it needed to learn the new mixing matrix. An important point was revealed at the first step of the third session, in which context 1 was provided again. The BSS error was significantly smaller than that in the first session and close to zero from the beginning of this session, indicating that the network retained synaptic strengths that were optimized for context 1 even after learning context 2. After several iterations, the BSS error for both contexts converged to zero. The success of learning was also confirmed by the trajectory of the EGHR cost function that also converged to the minimum value (Fig. 2B bottom).

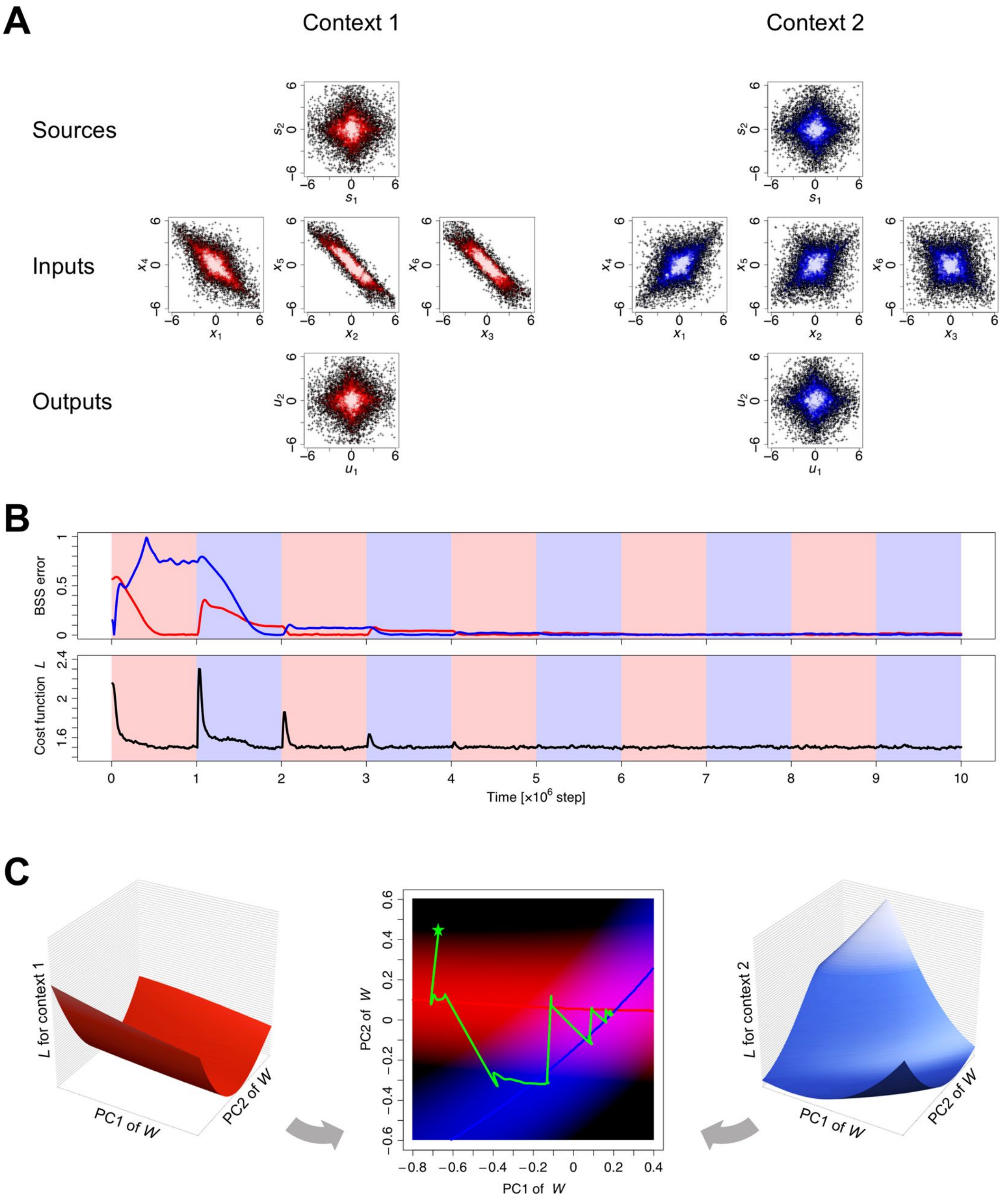

Figure 2. Results of multi-context BSS. (**A**) Distributions of sources, inputs, and outputs for context 1 and 2. (**B**) Trajectories of BSS error in the two contexts (top, red: context 1, blue: context 2) and cost function (bottom). (**C**) Visualization of null spaces. The panel illustrates the shapes of the cost function under each context (left and right) and the trajectory of synaptic strengths (W) projected in a subspace spanned by the first (PC1) and second (PC2) principal components (center). The trajectory is determined by the gradient of either cost function, depending on the context. On the PC1-PC2 plane, null spaces are illustrated as nullclines; red and blue curves are nullclines for contexts 1 and 2, respectively. Low cost areas (i.e., valleys of the cost functions) are highlighted by red or blue shading. The synaptic strength matrix starts from a random initial state (star mark), shifts to the nullcline of context 1 or 2, and eventually converges to the cross point of the two nullclines, where the synaptic strengths perform the BSS for both contexts. Each source was randomly generated by the unit Laplace distribution. A learning rate of $\eta = 4 \times 10^{-6}$ was used. The MATLAB source code for this simulation is appended as Supplementary Source Codes.

These results show that an undercomplete EGHR increased the speed of re-adaptation to previously experienced contexts, suggesting that memory of past experiences was preserved within the network. Moreover, the network learned the optimal set of synaptic strengths that entertained both contexts after several iterations. A key feature for this ability is the "null space" in the synaptic strength matrix. While only four (2×2) dimensions were required to express a mapping from two-dimensional sources to two-dimensional outputs in one context, the synaptic strength matrix still comprised eight ($2 \times 6 - 2 \times 2$)-dimensional degrees of freedom. This freedom

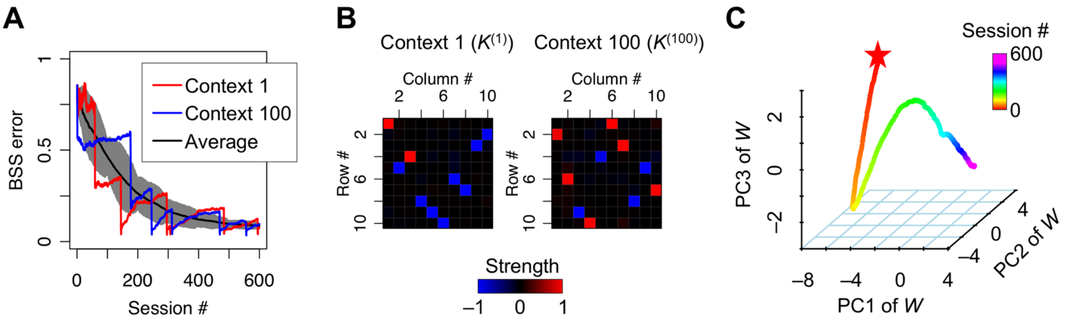

Figure 3. BSS with large number of contexts. One of 100 different contexts was randomly selected for each session. Each session contained $T = 10^5$ time steps, and the training continued for 600 sessions. In each session, the 2000-dimensional sensory inputs (x) were generated from ten-dimensional hidden sources (s), which independently followed the unit Laplace distribution, through a context-dependent random mixing matrix $A^{(k)}$. The neural network consisting of ten-dimensional neural outputs (u) was trained with a learning rate of $\eta = 10^{-5}$. (**A**) Trajectories of BSS error for context 1 and 100 and the average BSS error over contexts 1 to 100. The shaded area shows the standard deviation. (**B**) Mappings from ten sources to ten outputs in contexts 1 and 100 after training. Elements of matrix $K^{(k)} = W A^{(k)}$ with $k = 1$ and 100 are illustrated by the heat map. Only one element in each row and column takes ±1, indicating the one-to-one mapping from sources to outputs, i.e., the success of multi-context BSS. (**C**) The dynamics of synaptic strength matrix W projected in the three-dimensional space spanned by the first to third principal components (PC1 to PC3). The matrix starts from a random initial point (star mark) and converges to the null space, in which synaptic strengths are optimized for all trained contexts. The C code for this simulation is appended as Supplementary Source Codes.

spanned a null space in which synaptic strengths were equally optimized with zero BSS error. Similarly, when two different contexts were considered, four-dimensional degrees of freedom remained, as an overlap between the two eight-dimensional null spaces. To visualize such a null space, we projected synaptic strengths onto a subspace spanned by the first (PC1) and second (PC2) principal components of the trajectory of synaptic strengths (Fig. 2C). On this PC1-PC2 plane, a null space was illustrated as a nullcline. Since the dynamics of synaptic strengths were determined to go down the slope of a cost function for either context 1 or 2, synaptic strengths were started from a random initial state and reached the nullcline of either context 1 or 2, in turn. Crucially, this trajectory converged to the cross point of the two nullclines, where the synaptic strengths entertained both contexts. Because of this, the BSS error reached zero after iterative training; i.e., the network solved ICA for both contexts.

Furthermore, we examined the multi-context BSS by the EGHR using a large number of contexts (Fig. 3). Our agent received redundant (2000-dimensional) sensory inputs, comprising 100 sets (contexts) of mixtures of ten hidden sources (1000 sources in total), that were generated as products of the context-dependent mixing matrix and sources. Ten outputs neurons learned to infer each source from their mixtures by updating synaptic strengths through the EGHR. After training, we found that they successfully represented the ten sources for every context, without further updating synaptic strengths, as illustrated by the reduction of the BSS error for all 100 contexts (Fig. 3A) and the convergence of the covariance between sources and outputs to a diagonal matrix (up to permutations and sign-flips) (Fig. 3B). This was because synaptic strengths had sufficient capacity and were formed to express the inverse of the concatenated mixing matrices from all contexts, which was further confirmed by the convergence of the synaptic strength matrix in the null space (Fig. 3C).

BSS in constantly time-varying environments. In the previous section, we described a general condition for the neural network to achieve the multi-context BSS. In special cases, where the mixing matrices in each context have common features, the neural network can perform the multi-context BSS beyond the maximum number of contexts described above. Here, we show that when contexts are generated from a low-dimensional subspace of mixing matrices and, therefore, are dependent on each other, the EGHR can find the common features and use them to perform the multi-context BSS.

As a corollary of the property of the EGHR when provided with redundant inputs, the EGHR can perform the BSS even when the mixing matrix changes constantly as a function of time (Fig. 4A). Without loss of generality, a time-dependent mixing matrix is expressed by the sum of time-invariant and time-variant components, as follows:

$$A(t) \equiv \underbrace{A^{(0)}}_{\text{time-invariant}} + \underbrace{A^{(1)}R(t)}_{\text{time-variant}}, \tag{6}$$

where $A^{(0)}$ is a full-column-rank constant matrix with the same size as $A(t)$, $A^{(1)}$ is a full-column-rank constant vertically-long rectangular (or square) matrix, and $R(t)$ is a matrix composed of either smoothly or discontinuously changing functions in time. Each component of $R(t)$ is supposed to on average slowly change, i.e., their time-derivatives are typically much smaller in magnitude than those of $s(t)$. This condition is required to distinguish whether changes in inputs are caused by changes in the mixing matrix $A(t)$ or the hidden sources $s(t)$. Formally, $A(t)$ expresses infinite contexts along the trajectory of $R(t)$. This is a more complicated setup than the standard BSS in the sense that both sources and the mixing matrix change in time. Nonetheless, the EGHR can achieve BSS for all contexts if a solution of the synaptic strength matrix that satisfies $W(A^{(0)}, A^{(1)}) = (\Omega, O)$ exists.

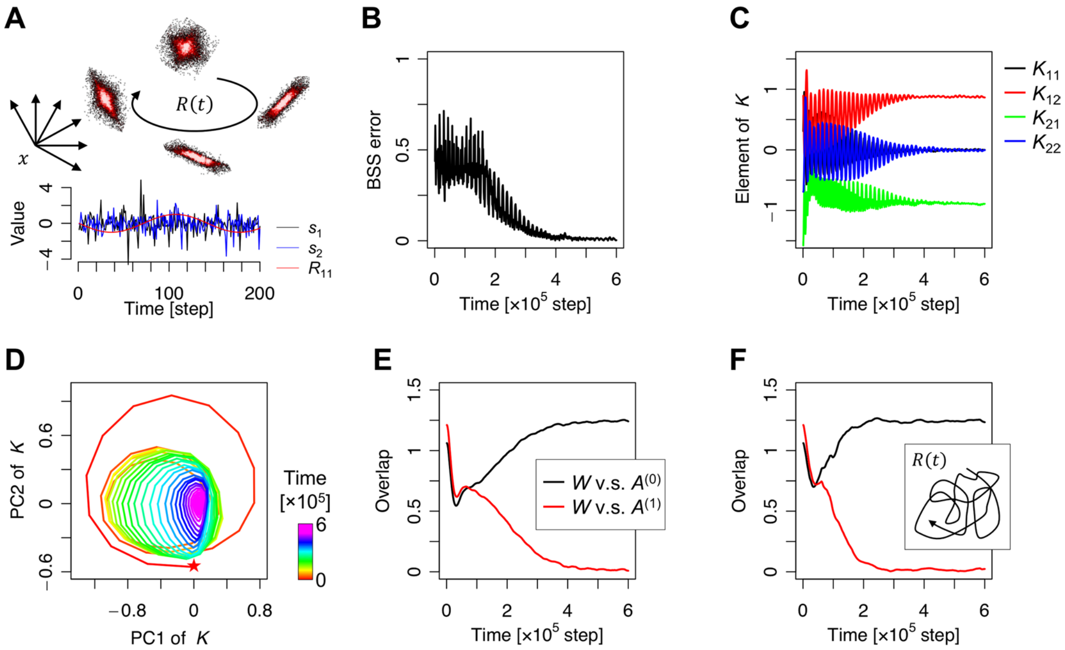

Figure 4. BSS with time-varying mixing matrix. (**A**) Top: Schematic image of sensory inputs generated from two sources through time-varying mixing matrix $A(t)$. The mixing matrix is controlled by the low-dimensional rotation matrix $R(t)$. Bottom: Trajectories of hidden sources and an element of $R(t)$, showing the difference in their time courses. (**B**) Trajectory of BSS error. (**C**) Trajectories of mapping weights from sources to outputs, i.e., matrix $K = WA(t)$. (**D**) Dynamics of matrix K projected on the first two-dimensional PCA subspace of K's trajectory over training. The matrix starts from a random initial point (star mark) and follows a spiral trajectory as it converges to a subspace in which synaptic matrix W is perpendicular to the time-varying component $A^{(1)}$. (**E**) Overlap of synaptic matrix W with time-invariant component $A^{(0)}$ and time-variant component $A^{(1)}$. The overlap between two matrices was defined by the Frobenius norm of their product, i.e., $|WA^{(k)}|_F \equiv \sqrt{\sum_{ij}(WA^{(k)})_{ij}^2}$. (**F**) Overlap of synaptic matrix W with $A^{(0)}$ and $A^{(1)}$ when elements of $R(t)$ were modeled as OU processes. Each source was randomly generated by the unit Laplace distribution. The learning rate of $\eta = 10^{-5}$ was used. The C code for this simulation is appended as Supplementary Source Codes.

Here, Ω represents the identity matrix up to permutations and sign-flips and O represents a matrix with zero elements. Such a solution generally exists if and only if $(A^{(0)}, A^{(1)})$ is a full-column-rank matrix (see Methods for the derivation). The above condition means that the network performs BSS based on the time-invariant features $A^{(0)}$ of the mixing matrix, while neglecting the time-varying features $A^{(1)}R(t)$. This can be viewed as a way to compress high-dimensional data. This is distinct from the standard dimensionality reduction approach by PCA, which would preferentially extract the time-variant features due to their extra variances. Moreover, the ability to perform dimensionality reduction is an important advantage of the EGHR over conventional ICA algorithms, such as the infomax-based ICA[10,11], natural gradient[12] and nonholonomic[39] algorithms, and the ICA mixture model[40], because these learning algorithms do not learn effective dimensionality reduction in the multi-context BSS setup due to their construction (see Methods for mathematical explanations).

In the simulation, we supposed $R(t)$ to be a two-dimensional rotation matrix, $R(t) = (\cos\omega t, -\sin\omega t; \sin\omega t, \cos\omega t)$, with an angular frequency of $\omega = \sqrt{2}\pi/100$. The simulation showed a reduction in the BSS error (Fig. 4B). At the same time, $K = WA$ converged to the identity matrix up to permutations and sign-flips, $K \to (0,1; -1,0)$ in this case, although A continuously changed in time (Fig. 4C,D). As illustrated in Fig. 4E, the synaptic matrix W became perpendicular to the time-varying features $A^{(1)}$ (i.e., $WA^{(1)} = O$), by a monotonic reduction of the overlap between W and $A^{(1)}$ (defined by the Frobenius norm of their product). After training, the overlap converged to zero. Hence, synaptic strengths were optimized regardless of $R(t)$ at this solution, which enabled the network to perform BSS with a virtually infinite number of contexts. In addition, the neural network that implements the EGHR could learn W perpendicular to $A^{(1)}$ in another simulation setup, where $R(t)$ was a 2 × 2 matrix and its elements were modeled as Ornstein-Uhlenbeck (OU) processes with time constant $\tau = 10^{-3}$ (Fig. 4F). These results indicate that the EGHR can perform the multi-context BSS with a wide range of time-varying mixing matrices. Indeed, a mathematical analysis shows that multi-context BSS is possible for a general time-varying matrix $R(t)$ as long as it changes slowly enough (see Methods).

Next, we demonstrated the utility of the EGHR, when supplied with redundant inputs, by using natural birdsongs and a time-variant mixing matrix that expressed a natural contextual change. Figure 5 illustrates the BSS task of two birdsongs when birds moved around the agent; thereby, the mixing matrix changed in time according to the positions of the birds (see, the entire movie at http://toyoizumilab.brain.riken.jp/dataset/Isomura2019/Isomura_Toyoizumi_SciRep2019_SupplementaryMovieS1.mp4). To obtain time-independent features, we

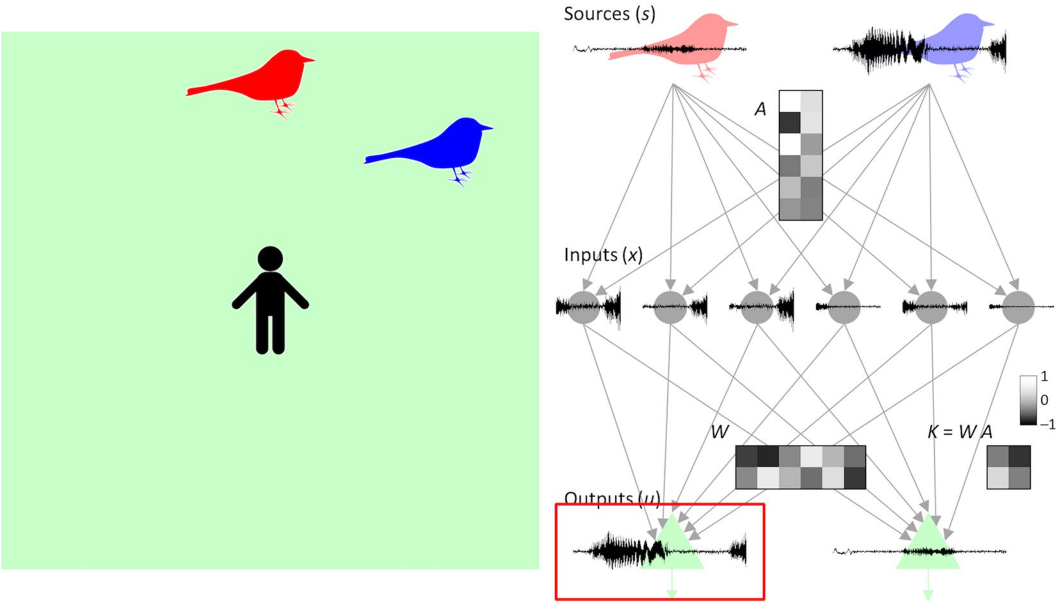

Figure 5. BSS of birdsongs when two birds move around the agent. A snapshot of the simulation overview movie after training is shown. Songs (or sources) generated by two birds s_1, s_2 (right top) are mixed with a time-varying mixing matrix A, resulting in six-dimensional sensory inputs $x_1, \ldots, x_6$ (right middle). The mixed signals correspond to the recording through six microphones with different preferences. The neural network converts the six inputs into two neural outputs, u_1 and u_2 (right bottom), using synaptic strength matrix W. The synaptic updates by the EGHR enable the outputs to encode each birdsong. Matrix $K = W\,A$ represents the mapping from sources to outputs. See Methods for the detailed simulation setup.

assumed that the two birds moved around in non-overlapping areas. For simplicity, we also assumed that the two birds moved around at different heights. The agent received mixtures of the two birdsongs through six microphones with different direction preferences. In the current context, the z-axis of the birds was time-invariant and the x- and y-axes of the birds were time-variant, although the observer was not informed about this. By tuning synaptic strengths by the EGHR, neural outputs were established to infer each birdsong, while the mixing matrix changed continuously. Crucially, after training, the mapping from the sources to the outputs ($K = W\,A$) became constant with time, although matrix A was time-dependent. More precisely, the EGHR found a representation where W satisfied $W(A^{(0)}, A^{(1)}) = (\Omega, O)$. Hence, neural outputs could separate the two birdsongs, although the amplitudes of the songs recorded by the microphones continuously changed depending on the positions of birds.

Generalization for inexperienced environments.

Finally, we examined the generalization capability of the multi-context BSS by the EGHR using natural birdsongs. For the sake of simplicity, we reduced Eq. (6) by considering $R(t)$ that changes discontinuously at the beginning of each session but otherwise is constant. Specifically, we considered the mixing matrix

$$A(v) = \underbrace{A^{(0)}}_{\text{context-independent}} + \underbrace{A^{(1)}v_1 + \cdots + A^{(n)}v_n}_{\text{context-dependent}}, \tag{7}$$

written using context-independent matrix $A^{(0)}$, context-dependent matrices $\{A^{(1)}, \ldots, A^{(n)}\}$, and context vector $v \equiv (v_1, \ldots, v_n)$ that discontinuously changes at the beginning of a new session. The first term in the right-hand side of Eq. (7) corresponds to the context-independent (i.e., constant) component, which should be a full-column-rank matrix to provide an ICA solution. Similarly to the case with the continuously time-varying mixing matrix, the EGHR can establish synaptic matrix W that expresses the pseudo inverse of $A^{(0)}$ up to permutations and sign-flips, while keeping W perpendicular to $A^{(1)}, \ldots, A^{(n)}$, i.e., $W(A^{(0)}, A^{(1)}, \ldots, A^{(n)}) = (\Omega, O, \ldots, O)$. Notably, the EGHR can establish such W by using only a handful samples of v out of combinatorially many possibilities. This is because the mappings from sources to inputs are restricted to be a linear transformation, and thereby, observations with the polynomial (probably quadratic) order number of contexts can identify the mapping for all contexts. This property is particularly useful when v is high dimensional.

In this demonstration, ten sets (contexts) of mixtures of ten birdsongs were introduced to our agent, with redundant sensory inputs composed of 100 mixed sound waves (Fig. 6). Those contexts were defined by random mixing matrices $A^{(0)}, A^{(1)}, \ldots, A^{(4)}$. We trained the network using only 10 contexts: $v = (1,0,0,0)$, $(½,½,0,0)$, $(0,1,0,0)$, $(0,½,½,0)$, $(0,0,1,0)$, $(0,0,½,½)$, $(0,0,0,1)$, $(½,0,0,½)$, $(½,0,½,0)$, $(0,½,0,½)$. At the beginning of each session, v was randomly selected from the above ten vectors, which provided a discrete random transition among 10 contexts. Ten output neurons learned to infer each birdsong from their mixtures, by updating synaptic strengths through the EGHR. After training, they successfully represented the ten birdsongs without further updating

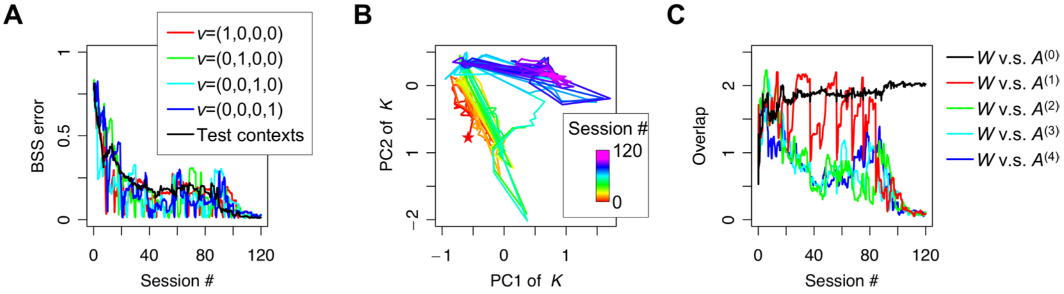

Figure 6. Generalization of multi-context BSS. (**A**) Trajectories of the BSS error with four trained contexts and the average BSS error over 20 inexperienced test contexts, created using randomly sampled v. (**B**) Dynamics of matrix K projected on the first two-dimensional PCA subspace. The matrix starts from a random initial point (star mark) and converges to a fixed point, at which the synaptic matrix entertains every trained context. (**C**) Overlap of synaptic matrix W with context-independent component $A^{(0)}$ and context-dependent components $A^{(1)}, \ldots, A^{(4)}$. Overlap between two matrices is defined by $|WA^{(k)}|_F$, as described in Fig. 4. See Methods for the detailed simulation setup.

synaptic strengths. Crucially, the network could perform BSS even in an inexperienced context (for example, in $v = (¼,¼,¼,¼)$). This speaks to the generalization of the multi-context BSS for unseen test contexts.

We quantitatively showed that, as learning progresses, the BSS error for test contexts (defined using 20 randomly sampled v that were inexperienced in the training), as well as for trained contexts, decreased (Fig. 6A). The trajectory of the first two principal components (PC1 and PC2) of K exhibits their convergence to a fixed point at later sessions (Fig. 6B). Here PC1 and PC2 together captured 63.4% of the total variance. Regardless of the given context, matrix K converged to a constant matrix that was the same as the identity matrix up to permutations and sign-flips. The convergence of W to this fixed point was validated by plotting the trajectories of the overlaps between W and A components (Fig. 6C). While the overlap between W and $A^{(0)}$ increased as learning progressed, the overlap with context-dependent components ($A^{(1)}, \ldots, A^{(4)}$) decreased and converged to zero, showing that W became perpendicular to $A^{(1)}, \ldots, A^{(4)}$ by the EGHR. We conducted a series of simulations with different initial conditions and confirmed the reliability of convergence, although the convergence speed depended on the initial relative position of W compared to $A^{(0)}, A^{(1)}, \ldots, A^{(4)}$. Hence, this learnt network could perform BSS with $A(v)$ determined by arbitrary v in the four-dimensional space, without further synaptic updating or transient error, while the network was trained only with 10 contexts. Those results highlight the significant generalization capability of the neural network established by the EGHR and the robustness against inexperienced environments for performing BSS.

Discussion

While a real environment comprises several different contexts, humans and animals retain the experience of past contexts to perform well when they find themselves in the same context in the future. This ability is known as conservation of learning or cognitive flexibility[25]. Although analogous learning is likely to happen during BSS, the conventional biological BSS algorithms[37,38] must forget the memory of past contexts to learn a new one. Thereby, when the agent subsequently encounters a previously experienced context, it needs to relearn it from the very beginning. We overcame this limitation by using the described algorithm, the EGHR. The crucial property of the EGHR is that when the number of inputs is larger than the number of sources, the synaptic matrix contains a null space in which synaptic strengths are equally optimized for performing BSS. Hence, with sufficiently redundant inputs, the EGHR can make the synaptic matrix optimal for every experienced context. This is an ability that the conventional biologically plausible BSS algorithms do not have, due to the constraint that the number of inputs and outputs must be equal[15]; however, we argue that this ability is crucial for animals to perceive and adapt to dynamically changing multi-context environments. It is also crucial for animals to generalize past learning to inexperienced contexts. We also found that, if there is a common feature shared across the training contexts, the EGHR can extract it and generalize the BSS result to inexperienced test contexts. This speaks to a generalization capability and transfer learning, implying the prevention of overfitting to a specific context; alternatively, one might see this as an extraction of a general concept across contexts. Therefore, we argue that the EGHR is a good candidate model for describing the neural mechanism of conservation of learning or cognitive flexibility for BSS.

Moreover, the process of extracting hidden sources in a multi-context BSS setup can be seen as a novel concept of dimensionality reduction[42]. If the dimensions of input are greater than the product of the number of sources and the number of contexts, the EGHR can extract the low-dimensional sources (up to context-dependent permutations and sign-flips), while filtering out a large number of context-dependent signals induced by changes in the mixing matrix. ICA algorithms for multi-context BSS[39–41] and undercomplete ICA for compressing data dimensionality[15,17,53] have been separately developed. Nevertheless, conventional ICA algorithms for multi-context BSS cannot learn efficient dimensionality reduction, and thus, to our knowledge, our study is the first to attempt dimensionality reduction in the multi-context BSS setup. This method is particularly powerful when a common feature is shared across the contexts, because the EGHR can make each neuron encode an identical source across all contexts. Our results are different from those obtained using standard dimensionality reduction approaches by PCA[18,19], because PCA is used for extracting subspaces of high-variance principal components and hence would preferentially extract the context-dependent varying features, given that each source has the same variance. Therefore, our study proposes an attractive use of the EGHR for dimensionality reduction.

It is worth noting that the application of standard ICA algorithms to high-pass filtered inputs cannot solve the multi-context BSS problem. This is because context-dependent changes in the mixing matrix not only change the means of the inputs, which can be removed by high-pass filtering, but also change the gain of how fluctuations of each source are propagated to input fluctuations. Hence, the difference in contexts cannot be expressed as a linear ICA problem after high-pass input filtering. Therefore, selective extraction of context-invariant features is an advantage of the EGHR. Moreover, if provided with redundant input, the EGHR can solve multi-context BSS even if the context changes continuously in time, as we demonstrated in Figs. 4, 5.

We demonstrated that a neural network learns to distinguish individual birdsongs from their superposition. Young songbirds learn songs by mimicking adult birds' songs[54–57]. A study reported that neurons in songbirds' higher auditory cortex exhibit a teacher specific activity[58]. One can imagine those neurons correspond to the expectation of hidden sources (u), as considered in this study. Importantly, the natural environment that young songbirds encounter is dynamic, as we considered in Fig. 5. Therefore, the conventional BSS setup, which assumes a static environment or context, is not suitable for explaining this problem. It is interesting to consider that young songbirds might employ some computational mechanism similar to the EGHR to distinguish a teacher's song from other songs in a dynamically changing environment.

Biological neural networks implement an EGHR-like learning rule. The main ingredients of the EGHR are Hebbian plasticity and the third scalar factor that modulates it. Hebbian plasticity occurs in the brain depending on the activity level[44,59], spike timings[60–63], or burst timings[64] of pre- and post-synaptic neurons. In contrast, the third scalar factor can modify the learning rate and even invert Hebbian to anti-Hebbian plasticity[50], similarly to what we propose for the EGHR. In general, such a modulation forms the basis of a three-factor learning rule, a concept that has recently received attention (see[20,65,66] for reviews), and is supported by experiments on various neuromodulators and neurotransmitters, such as dopamine[45–47], noradrenaline[48,49], muscarine[67], and GABA[50,51], as well as glial factors[52]. (These factors may encode reward[68–72], likelihood[73], novelty/surprise[74], or error from a prior belief[15,17] to achieve various types of learning, implying the existence of a unified three-factor learning framework.) Importantly, the EGHR only requires such a signal that conveys global information to neurons to achieve learning. Furthermore, a study using *in vitro* neural networks suggested that neurons perform simple BSS using a plasticity rule that is different from the most basic form of Hebbian plasticity, by which synaptic strengths are updated purely as a product of pre- and postsynaptic activity[75,76]. A candidate implementation of the EGHR can be made for cortical pyramidal cells and inhibitory neurons; the former constituting the EGHR output neurons and encoding the expectations of hidden sources, and the latter constituting the third scalar factor and calculating the nonlinear sum of activity in surrounding pyramidal cells. This view is consistent with the circuit structure reported for the visual cortex[77,78]. These empirical evidences support the biological plausibility of the EGHR as a candidate model of neuronal BSS.

A local computation of the EGHR is highly desirable for neuromorphic engineering[23,24,79,80]. The EGHR updates synapses by a simple product of pre- and postsynaptic neurons' activity and a global scalar factor. Because of this, less information transfer between neurons is required, compared to conventional ICA methods that require non-local information[10–12], all-to-all plastic lateral inhibition between output neurons[37,38], or an additional processing step for decorrelation[13]. The simplicity of the EGHR is a great advantage when implemented in a neuromorphic chip because it can reduce the space for wiring and the energy consumption. Furthermore, unlike the conventional ICA algorithms that assume an equal number of input and output neurons, a neuromorphic chip that employs the EGHR with redundant inputs would perform BSS in multiple contexts, as allowed by the network memory capacity, without requiring readaptation. The generalization capability of the EGHR, as demonstrated in Fig. 6, is an additional benefit, as the EGHR captures the common features shared across training contexts to perform BSS in inexperienced test contexts.

Notably, although we considered a linear BSS problem in this study, multi-context BSS can be extended to non-linear BSS, in which the inputs are generated through a non-linear mixture of sources[81,82]. To solve this problem, a promising approach would be to use a linear neural network. A recent study showed that when the ratio of input-to-source dimensions and source number are large, a linear neural network can find an optimal linear encoder that separates the true sources through PCA and ICA, thus asymptotically achieving zero BSS error[83]. Because both the asymptotic linearization and multi-context BSS by the EGHR are based on high-dimensional sensory inputs, combining these two might be a useful approach to solve the multi-context and non-linear BSS problem.

In summary, we demonstrated that the EGHR can retain memories of past contexts and, once the learning is achieved for every context, it can perform multi-context BSS without further updating synapses. Moreover, the EGHR can find common features shared across contexts, if present, and uses them to generalize the learning result to inexperienced contexts. Therefore, the EGHR will be useful for understanding the neural mechanisms of flexible inference and sensory representation under dynamically changing environments, and for creating brain-inspired artificial general intelligence.

Methods

Model and learning rule. The neural network model and used learning rule (the EGHR) are described in the Results section.

Definition of BSS error. We calculated the maximum and second maximum rows as $i' = \text{argmax}_i \left| K_{ij}^{(k)} \right|$ and $i'' = \text{argmax}_{i \neq i'} \left| K_{ij}^{(k)} \right|$ and defined the BSS error of column j by the ratio of the values in the two rows:

$\varepsilon_j^c = \left|K_{i''j}^{(k)}\right| / \left|K_{i'j}^{(k)}\right|$. Similarly, the BSS error of row i: $\varepsilon_i^r = \left|K_{ij''}^{(k)}\right| / \left|K_{ij'}^{(k)}\right|$ was obtained from the ratio of the maximum and second maximum columns, where $j' = \mathrm{argmax}_j \left|K_{ij}^{(k)}\right|$ and $j'' = \mathrm{argmax}_{j\neq j'} \left|K_{ij}^{(k)}\right|$. The BSS error (for the whole K) was defined as the average of them: $BSS\ error \equiv (\varepsilon_1^c + \ldots + \varepsilon_{N_s}^c)/2N_s + (\varepsilon_1^r + \ldots + \varepsilon_{N_u}^r)/2N_u$.

Analysis of BSS solution: existence and linear stability.

Supposing that $N_u = N_s$, we defined the transform matrix $K^{(k)}$ by

$$K^{(k)} \equiv WA^{(k)}. \tag{8}$$

For $N_x \geq N_s$, the ICA for context k is achieved when $K^{(k)}$ is the identical matrix up to permutations and sign-flips. Hence, when amd only when column vectors of a block matrix $(A^{(1)}, \ldots, A^{(C)})$ are linearly independent of each other, i.e., if and only if $(A^{(1)}, \ldots, A^{(C)})$ is a full-column-rank matrix, an ICA solution that separates all sources for context 1, …, C exists. Namely, W achieves the multi-context BSS when it satisfies

$$(K^{(1)}, \ldots, K^{(C)}) = W(A^{(1)}, \ldots, A^{(C)}) = (\Omega^{(1)}, \ldots, \Omega^{(C)}), \tag{9}$$

where $\Omega^{(k)}$ is an $N_u \times N_s$ matrix equivalent to the identity matrix up to permutations and sign-flips. Regarding the i-th row of matrix $K^{(k)}$, as denoted by a row vector $\left(K_{i1}^{(k)}, \ldots, K_{iN_s}^{(k)}\right)$, the achievement of ICA is justified when one element is one and the others are zero. Thus, there are many candidate sets of $\left(W_{i1}, \ldots, W_{iN_x}\right)$ that can achieve ICA, because N_x is larger than N_s. Our numerical analyses showed that among these potential solutions, the one that is the nearest to the solution for the previous context is likely to be chosen. This can be understood as follows: when the network finds an ICA solution for all contexts, the error (i.e., cost function of the EGHR), including transient periods between two contexts, is minimized; hence, according to the gradient descent, synaptic strengths converge to such a solution as training progresses. Owing to this mechanism, the initial errors converge to zero when previously experienced environments are provided as stimuli.

We showed that W that satisfies $K = WA = \Omega$ gives a fixed point for the EGHR cost function, $\dot{W} = -\partial L/\partial W = (E_0 - E(u))g(u)x^T = O$, and thus gives an ICA solution, where A is a vertically long or square full-rank mixing matrix[15,17]. Regarding BSS with a time-varying mixing matrix, from $A = A^{(0)} + A^{(1)}R$, the time differential of K yields $\dot{K} = \dot{W}(A^{(0)} + A^{(1)}R) + WA^{(1)}\dot{R} = O$. Here, we assume that $A^{(0)}$ and $A^{(1)}$ are full column-rank matrices and R is a general $N_R \times N_s$ time-varying matrix. Because $\dot{W} = O$ holds for the fixed point, W gives an ICA solution if and only if $WA^{(1)}\dot{R} = O$. Thus, W needs to satisfy $W(A^{(0)}, A^{(1)}) = (\Omega, O)$ to give a multi-context ICA solution. The condition for such an ICA solution to exist was obtained as follows: we considered this as a BSS problem such that

$$x = (A^{(0)}, A^{(1)})\begin{pmatrix} s \\ Rs \end{pmatrix}. \tag{10}$$

The singular value decomposition is given by $(A^{(0)}, A^{(1)}) = US(V_0^T, V_1^T)$, where $U \in \mathbb{R}^{N_x \times (N_s + N_R)}$, $V_0 \in \mathbb{R}^{N_s \times (N_s + N_R)}$, and $V_1 \in \mathbb{R}^{N_R \times (N_s + N_R)}$ with $V_0 V_1^T = O$ are orthogonal matrices and $S \in \mathbb{R}^{(N_s + N_R) \times (N_s + N_R)}$ is a diagonal matrix of singular values. From this, $W = \Omega V_0 S^{-1} X$ should hold to ensure $W(A^{(0)}, A^{(1)}) = (\Omega, O)$, where X is an orthogonal matrix satisfying $XU = I$. Hence, ICA solutions exist when and only when column vectors of $(A^{(0)}, A^{(1)})$ are linearly independent of each other.

Moreover, we analyzed a sufficient condition on the time constant of $R(t)$ for the stability of the ICA solution. From our previous analysis, the linear stability for fixed points is determined by the following second differential form[15,17]:

$$d^2L = \left(\sum_{i=1}^{N_u} dK_{ii}\right)^2 + \sum_{i=1}^{N_u} (1 + \Phi_{ii})\, dK_{ii}^2 + \frac{1}{2}\sum_{i=1}^{N_u}\sum_{i\neq j} (\Phi_{ij} dK_{ij}^2 + 2dK_{ij}dK_{ji} + \Phi_{ji} dK_{ji}^2), \tag{11}$$

where $\Phi_{ii} \equiv \mathrm{cov}[-\log p_0(s_i), g'(s_i)s_i^2]$ and $\Phi_{ij} \equiv \mathrm{cov}[-\log p_0(s_i), g'(s_i)]s_j^2 + \mathrm{cov}\left[-\log p_0(s_j), s_j^2\right]g'(s_i)$ for $i \neq j$ (note that cov[,] is the covariance). The magnitude of dW is assumed to be negligible due to a small learning rate. The solution is linearly stable when and only when $\Phi_{ii} > -1$ and $\Phi_{ij}\Phi_{ji} > 1$. When change in $R(t)$ is sufficiently slower than that of $s(t)$ on average, i.e., when $dK = dW(A^{(0)} + A^{(1)}R) + WA^{(1)}\, dR$ is sufficiently small, the above linear stability condition determines the stability of the fixed point. However, when $R(t)$ changes faster than or as fa*st* as $s(t)$, dK is no longer a small fluctuation, because of large dR, and therefore K may leave from the neighborhood of the fixed point to the region where the second order approximation is no longer accurate. Therefore, as long as the time constan*t* of $R(t)$ is chosen to ensure the averaged fluctuation is small and thus K is within the neighborhood of the fixed point, the EGHR with a time-varying mixing matrix has the same linear stability condition as the conventional EGHR without context switching.

Analysis of conventional ICA algorithms.

Here we show that, unlike the multi-context EGHR, conventional ICA algorithms cannot be used for the dimensionality reduction purpose. Some of the ICA algorithms in consideration are written as $\dot{W} \propto F(u(t), x(t))W$ or, equivalently,

$$W(t+1) = [I + \eta F(u(t), x(t))]W(t) \tag{12}$$

in each discrete time step t ($t = 1, 2, \ldots, T$) with learning rate η. The functional F specifies an individual learning rule, namely, the natural gradient algorithm takes $F(u, x) = I - \langle g(u)u^T \rangle$[12] and the non-holonomic algorithm takes $F(u, x) = \langle \mathrm{diag}[g(u) \odot u] - g(u)u^T \rangle$[39], where $\odot$ expresses the element-wise product of two vectors and diag[·] indicates a diagonal matrix comprising a vector. This class of ICA algorithms cannot perform dimensionality reduction. Following Eq. (12), the synaptic strength matrix after training (i.e., at time T) is expressed as

$$W = \left[\prod_{t=1}^{T} (I + \eta F(u(t), x(t)))\right] W_0, \tag{13}$$

where W_0 is the initial synaptic matrix. In dimensionality reduction, we are interested in horizontally long $N_u \times N_x$ matrices W and W_0, which compress N_x-dimensional signal x to N_u-dimensional output u with $N_u < N_x$. However, $\prod_{t=1}^{T} (I + \eta F(u(t), x(t)))$ changes the strength only within the $N_u \times N_u$ degree of freedom, so that this is equivalent to the ICA of N_u-dimensional signals $W_0 x$ that is already compressed by the non-optimal matrix W_0. Hence, this class of ICA algorithm can be used for separating already (sub-optimally) compressed signals $W_0 x$ but not for reducing signal dimensions. The infomax-based ICA algorithm[10,11] has the same fixed point and linear stability conditions as the natural gradient algorithm; thus, again, it does not perform dimensionality reduction. Next, the ICA mixture model was proposed, which is a combination of ICA and a mixture model, to perform multi-context ICA by assigning one of the multiple ICA models to each context[40]. In this model, the pseudo inverse of the synaptic matrix W^k for the ^{k}-th model is updated instead of W^k by $\mathrm{d}(W^k)^+/\mathrm{d}t \propto z^k(t)(W^k) + (I - g(u(t))u(t)^T)$, where $z_k(t) \in [0, 1]$ is the probability of the k-th model being selected. Similar to Eq. (13), the pseudo inverse of the synaptic strength matrix after training is expressed as

$$(W^k)^+ = (W_0^k)^+ \left[\prod_{t=1}^{T} (I + \eta z_k(t)(I - g(u(t))u(t)^T))\right], \tag{14}$$

which indicates again that the compression is determined by $W_0^k x$. Therefore, the ICA mixture model does not perform dimensionality reduction, either. Hence, the use of multi-context ICA for dimensionality reduction is our novel contribution to the literature, which is beyond the original proposal of the EGHR or conventional multi-context ICA algorithms.

Simulation protocols. *For figure 5:* Two birdsongs were downloaded from Xeno-canto (https://www.xeno-canto.org/132149, https://www.xeno-canto.org/133054). Two hidden sources were created by trimming the first 60 s of these songs (with 4410-Hz time resolution) and normalizing them, to ensure each source sequence had zero mean and unit variance. During the training, the song sequences were repeated. To add stochasticity, a hidden source was defined by the sum of a song and a white-noise sequence generated by a Laplace distribution. The mixing matrix was defined by 6×2 random matrices, $A^{(0)}$, $A^{(1)}$, and a rotation matrix, $R(t) \equiv (\cos\omega t, -\sin\omega t;\ \sin\omega t, \cos\omega t)$. The angular frequency ω was randomly set as -0.1π, 0, or 0.1π [rad/s], by following Markov process with a transition probability of 1/8820. The training time and learning rate were defined by $T = 4410 \times 6000$ [step] and $\eta = 10^{-7}$.

For figure 6: Ten birdsongs were downloaded from Xeno-canto (the URLs are https://www.xeno-canto.org/****** where ****** was replaced with the following numbers: 27060, 64735, 67307, 110303, 121326, 121691, 126481, 132149, 133054, 133862). Ten hidden sources were created in the same manner as described above. The mixing matrix was defined by 100×10 random matrices $A^{(0)}, A^{(1)}, A^{(2)}, A^{(3)}, A^{(4)}$, where $\hat{A}^{(1)}, \ldots, \hat{A}^{(4)}$ were randomly generated and $A^{(0)}, \ldots, A^{(4)}$ were defined by $A^{(0)} = (\hat{A}^{(1)} + \hat{A}^{(2)} + \hat{A}^{(3)} + \hat{A}^{(4)})/4$ and $A^{(k)} = \hat{A}^{(k)} - A^{(0)}$, for $k = 1, \ldots, 4$. This treatment was served to ensure that $A^{(1)}, \ldots, A^{(4)}$ do not involve common features across contexts. The training comprised 120 sessions, with each session continued for $T = 4410 \times 600$ [step]. The context vector v randomly chose one of the following ten vectors, $v = (1,0,0,0)$, (½,½,0,0), (0,1,0,0), (0,½,½,0), (0,0,1,0), (0,0,½,½), (0,0,0,1), (½,0,0,½), (½,0,½,0), (0,½,0,½), at the beginning of each session and maintained the value during the session. The learning rate was defined by $\eta = 2 \times 10^{-7}$. For the test, 20 randomly generated vectors were used, and their elements were randomly sampled from [0,1] and then normalized to satisfy $v_1 + v_2 + v_3 + v_4 = 1$.

References

1. Helmholtz, H. *Treatise on physiological optics* Vol. III (Dover Publications, 1925).
2. Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719 (2004).
3. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
4. Brown, G. D., Yamada, S. & Sejnowski, T. J. Independent component analysis at the neural cocktail party. *Trends Neurosci.* **24**, 54–63 (2001).
5. Mesgarani, N. & Chang, E. F. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* **485**, 233–236 (2012).
6. Belouchrani, A., Abed-Meraim, K., Cardoso, J. F. & Moulines, E. A blind source separation technique using second-order statistics. *IEEE Trans. Signal Process.* **45**, 434–444 (1997).
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, West Sussex, UK, 2009).
8. Comon, P. Independent component analysis, a new concept? *Signal Process.* **36**, 287–314 (1994).
9. Comon, P. & Jutten, C. In Comon, P. & Jutten, C. (Eds), *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. (Orlando, FL: Academic Press, 2010).

10. Bell, A. J. & Sejnowski, T. J. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* **7**, 1129–1159 (1995).
11. Bell, A. J. & Sejnowski, T. J. The "independent components" of natural scenes are edge filters. *Vision Res.* **37**, 3327–3338 (1997).
12. Amari, S. I., Cichocki, A. & Yang, H. H. A new learning algorithm for blind signal separation. *Adv. Neural Inf. Process. Syst.* **8**, 757–763 (1996).
13. Hyvärinen, A. & Oja, E. A fast fixed-point algorithm for independent component analysis. *Neural Comput.* **9**, 1483–1492 (1997).
14. Savin, C., Joshi, P. & Triesch, J. Independent component analysis in spiking neurons. *PLoS Comput. Biol.* **6**, e1000757 (2010).
15. Isomura, T. & Toyoizumi, T. A local learning rule for independent component analysis. *Sci. Rep.* **6**, 28073 (2016).
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. Isomura, T. & Toyoizumi, T. Error-gated Hebbian rule: A local learning rule for principal and independent component analysis. *Sci. Rep.* **8**, 1835 (2018).
18. Pearson, K. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **2**, 559–572 (1901).
19. Oja, E. Neural networks, principal components, and subspaces. *Int. J. Neural Syst.* **1**, 61–68 (1989).
20. Kuśmierz, Ł., Isomura, T. & Toyoizumi, T. Learning with three factors: modulating Hebbian plasticity with errors. *Curr. Opin. Neurobiol.* **46**, 170–177 (2017).
21. Avitan, L. & Goodhill, G. J. Code under construction: neural coding over development. *Trends Neurosci.* **41**, 599–609 (2018).
22. Goodhill, G. J. Theoretical models of neural development. *iScience* **8**, 183–199 (2018).
23. Neftci, E. Data and power efficient intelligence with neuromorphic learning machines. *iScience* **5**, 52–68 (2018).
24. Fouda, M., Neftci, E., Eltawil, A. M. & Kurdahi, F. Independent component analysis using RRAMs. *IEEE Trans. Nanotech.*; https://doi.org/10.1109/TNANO.2018.2880734 (2018).
25. Dajani, D. R. & Uddin, L. Q. Demystifying cognitive flexibility: Implications for clinical and developmental neuroscience. *Trends Neurosci.* **38**, 571–578 (2015).
26. Dehaene, S. & Changeux, J. P. The Wisconsin Card Sorting Test: Theoretical analysis and modeling in a neuronal network. *Cereb. Cortex* **1**, 62–79 (1991).
27. Gilbert, C. D. & Sigman, M. Brain states: top-down influences in sensory processing. *Neuron* **54**, 677–696 (2007).
28. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
29. Song, H. F., Yang, G. R. & Wang, X.-J. Training excitatory-inhibitory recurrent neural networks for cognitive tasks: A simple and flexible framework. *PLoS Comput. Biol.* **12**, e1004792 (2016).
30. Song, H. F., Yang, G. R. & Wang, X.-J. Reward-based training of recurrent neural networks for cognitive and value-based tasks. *eLife* **6**, 679–684 (2017).
31. Miconi, T. Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *eLife* **6**, 229–256 (2017).
32. Chaisangmongkon, W., Swaminathan, S. K., Freedman, D. J. & Wang, X. J. Computing by robust transience: how the fronto-parietal network performs sequential, category-based decisions. *Neuron* **93**, 1504–1517 (2017).
33. Ahrens, M. B., Linden, J. F. & Sahani, M. Nonlinearities and contextual influences in auditory cortical responses modeled with multilinear spectrotemporal methods. *J. Neurosci.* **28**, 1929–1942 (2008).
34. Yu, D., Deng, L. & Dahl, G. Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition. In *Proceeding of NIPS Workshop on Deep Learning and Unsupervised Feature Learning*. (2010).
35. Kirkpatrick, J. *et al*. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. USA* **114**, 3521–3526 (2017).
36. Zenke, F., Poole, B. & Ganguli, S. Continual learning through synaptic intelligence. *International Conference on Machine Learning*, 3987–3995; https://arxiv.org/abs/1703.04200 (2017).
37. Földiák, P. Forming sparse representations by local anti-Hebbian learning. *Biol. Cybern.* **64**, 165–170 (1990).
38. Linsker, R. A local learning rule that enables information maximization for arbitrary input distributions. *Neural Comput.* **9**, 1661–1665 (1997).
39. Amari, S. I., Chen, T. & Cichocki, A. Nonholonomic orthogonal learning algorithms for blind source separation. *Neural Comput.* **12**, 1463–1484 (2000).
40. Lee, T. W., Lewicki, M. S. & Sejnowski, T. J. ICA mixture models for unsupervised classification of non-Gaussian classes and automatic context switching in blind signal separation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 1078–1089 (2000).
41. Hirayama, J. I., Ogawa, T. & Hyvärinen, A. Unifying blind separation and clustering for resting-state EEG/MEG functional connectivity analysis. *Neural Comput.* **27**, 1373–1404 (2015).
42. Cunningham, J. P. & Ghahramani, Z. Linear dimensionality reduction: Survey, insights, and generalizations. *J. Mach. Learn. Res.* **16**, 2859–2900 (2015).
43. Hebb, D. O. *The Organization of Behavior: A Neuropsychological Theory*. (Wiley, New York, 1949).
44. 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).
45. Reynolds, J. N. J., Hyland, B. I. & Wickens, J. R. A cellular mechanism of reward-related learning. *Nature* **413**, 67–70 (2001).
46. 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).
47. Salgado, H., Köhr, G. & Treviño, M. Noradrenergic "tone" determines dichotomous control of cortical spike-timing-dependent plasticity. *Sci. Rep.* **2**, 417 (2012).
48. Yagishita, S. *et al*. A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science* **345**, 1616–1620 (2014).
49. Johansen, J. P. *et al*. Hebbian and neuromodulatory mechanisms interact to trigger associative memory formation. *Proc. Natl. Acad. Sci. USA* **111**, E5584–92 (2014).
50. Paille, V. *et al*. GABAergic circuits control spike-timing-dependent plasticity. *J. Neurosci.* **33**, 9353–9363 (2013).
51. Hayama, T. *et al*. GABA promotes the competitive selection of dendritic spines by controlling local Ca^{2+} signaling. *Nat. Neurosci.* **16**, 1409–1416 (2013).
52. Ben Achour, S. & Pascual, O. Glia: the many ways to modulate synaptic plasticity. *Neurochem. Int.* **57**, 440–445 (2010).
53. Porrill, J. & Stone, J. V. Undercomplete independent component analysis for signal separation and dimension reduction. *Technical report*, *University of Sheffield, Department of Psychology*. (1998).
54. Tchernichovski, O., Mitra, P. P., Lints, T. & Nottebohm, F. Dynamics of the vocal imitation process: how a zebra finch learns its song. *Science* **291**, 2564–2569 (2001).
55. Woolley, S. Early experience shapes vocal neural coding and perception in songbirds. *Dev. Psychobiol.* **54**, 612–631 (2012).
56. Lipkind, D. *et al*. Stepwise acquisition of vocal combinatorial capacity in songbirds and human infants. *Nature* **498**, 104–108 (2013).
57. Lipkind, D. *et al*. Song-birds work around computational complexity by learning song vocabulary independently of sequence. *Nat. Commun.* **8**, 1247 (2017).
58. Yanagihara, S. & Yazaki-Sugiyama, Y. Auditory experience-dependent cortical circuit shaping for memory formation in bird song learning. *Nat. Commun.* **7**, 11946 (2016).

59. Dudek, S. M. & Bear, M. F. Homosynaptic long-term depression in area CA1 of hippocampus and effects of N-methyl-D-aspartate receptor blockade. *Proc. Natl. Acad. Sci. USA* **89**, 4363–4367 (1992).
60. Markram, H., Lübke, J., Frotscher, M. & Sakmann, B. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* **275**, 213–215 (1997).
61. 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).
62. Zhang, L. I., Tao, H. W., Holt, C. E., Harris, W. A. & Poo, M. M. A critical window for cooperation and competition among developing retinotectal synapses. *Nature* **395**, 37–44 (1998).
63. Feldman, D. E. The spike-timing dependence of plasticity. *Neuron* **75**, 556–571 (2012).
64. Butts, D. A., Kanold, P. O. & Shatz, C. J. A burst-based "Hebbian" learning rule at retinogeniculate synapses links retinal waves to activity-dependent refinement. *PLoS Biol.* **5**, e61 (2007).
65. Pawlak, V., Wickens, J. R., Kirkwood, A. & Kerr, J. N. Timing is not everything: neuromodulation opens the STDP gate. *Front. Synaptic Neurosci.* **2**, 146 (2010).
66. Frémaux, N. & Gerstner, W. Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Front. Neural Circuits* **9**, 85 (2016).
67. Seol, G. H. *et al.* Neuromodulators control the polarity of spike-timing-dependent synaptic plasticity. *Neuron* **55**, 919–929 (2007).
68. Izhikevich, E. M. Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cereb. Cortex* **17**, 2443–2452 (2007).
69. Florian, R. V. Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. *Neural Comput.* **19**, 1468–1502 (2007).
70. 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).
71. Urbanczik, R. & Senn, W. Reinforcement learning in populations of spiking neurons. *Nat. Neurosci.* **12**, 250–252 (2009).
72. Frémaux, N., Sprekeler, H. & Gerstner, W. Functional requirements for reward-modulated spike-timing-dependent plasticity. *J. Neurosci.* **30**, 13326–13337 (2010).
73. Brea, J., Senn, W. & Pfister, J. P. Matching recall and storage in sequence learning with spiking neural networks. *J. Neurosci.* **33**, 9565–9575 (2013).
74. Rezende, D. J. & Gerstner, W. Stochastic variational learning in recurrent spiking networks. *Front. Comput. Neurosci.* **8**, 38 (2014).
75. Isomura, T., Kotani, K. & Jimbo, Y. Cultured cortical neurons can perform blind source separation according to the free-energy principle. *PLoS Comput. Biol.* **11**, e1004643 (2015).
76. Isomura, T. & Friston, K. *In vitro* neural networks minimise variational free energy. *Sci. Rep.* **8**, 16926 (2018).
77. Harris, K. D. & Mrsic-Flogel, T. D. Cortical connectivity and sensory coding. *Nature* **503**, 51–58 (2013).
78. Hofer, S. B. *et al.* Differential connectivity and response dynamics of excitatory and inhibitory neurons in visual cortex. *Nat. Neurosci.* **14**, 1045–1052 (2011).
79. Merolla, P. A. *et al.* A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **345**, 668–673 (2014).
80. Chicca, E., Stefanini, F., Bartolozzi, C. & Indiveri, G. Neuromorphic electronic circuits for building autonomous cognitive systems. *Proc. IEEE* **102**, 1367–1388 (2014).
81. Lappalainen, H. & Honkela, A. Bayesian non-linear independent component analysis by multi-layer perceptrons. In *Advances in independent component analysis* (pp. 93–121) (London, UK: Springer, 2000).
82. Karhunen, J. Nonlinear independent component analysis. In Roberts, S. & Everson, R. (Eds), *Independent component analysis: principles and practice* (pp. 113–134) (Cambridge, UK: Cambridge University Press, 2001).
83. Isomura, T. & Toyoizumi, T. On the achievability of blind source separation for high-dimensional nonlinear source mixtures. Preprint at, https://arxiv.org/abs/1808.00668 (2018).

Acknowledgements

This work was supported by RIKEN Center for Brain Science (T.I. and T.T.), Brain/MINDS from AMED under Grant Number JP19dm020700 (T.T.), and JSPS KAKENHI Grant Number JP18H05432 (T.T.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Conceived the idea: T.I. and T.T. Performed the analyses: T.I. and T.T. Wrote the paper: T.I. and T.T.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-43423-z.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.