A Normative and Biologically Plausible Algorithm for Independent Component Analysis

Anonymous Author(s)
Affiliation
Address
email

Abstract

The brain effortlessly solves blind source separation (BSS) problems, but the algorithm it uses remains elusive. Independent Component Analysis (ICA) is a popular way to solve a class of BSS problems in signal processing. The implementation of ICA with a neural circuit needs to satisfy the following minimal biological constraints: 1. The algorithm must operate in the online setting where data samples are streamed one at a time. In other words, the neural network (NN) computes the sources on the fly without storing any significant fraction of the data in memory. 2. The synaptic weight updates are local, i.e., they depend only on the variables available to the synapse locally. Here, we propose a novel objective function for ICA from which we derive a biologically plausible NN, including both the neural architecture and the synaptic learning rules. Interestingly, our algorithm relies on modulating synaptic plasticity by the global activity of the output neurons. In the brain, this could be accomplished by neuromodulators, extracellular calcium, local field potential, or nitric oxide.

1 Introduction

Various brain areas, such as the early visual, auditory, and olfactory systems, are known to identify sources from their mixtures effortlessly [20, 85, 5, 52], an unsupervised task also known as blind source separation (BSS) [19]. BSS is often addressed by Independent Component Analysis (ICA) [18, 32], which assumes a generative model, wherein the observed stimuli are linear combinations of independent sources. ICA algorithms determine the transformation back from the observed stimuli into their original sources without knowing how they were mixed in the first place. Due to the ubiquity of BSS problems, it is believed that the development of biologically plausible neural circuits capable of performing ICA may provide critical insight into general neural computation principles.

A variety of ICA algorithms have been proposed over the years. Notable work include that of [6, 2, 72, 42, 43] who proposed algorithms based on the information-theoretic framework. A set of competing models are based instead on bio-inspired neural networks (NN) [40, 1, 31, 14, 32] designed to operate online. However, when applied to the extraction of more than one component, these NNs rely on non-local learning rules, i.e., a synapse needs to "know" about the specific activity of neurons other than the two it connects. A few ICA learning rules use only the local information available in each neuron and are more biologically plausible but are often hand-engineered and lack good theoretical guarantees, or are limited to one sign of kurtotic distributions [22, 45, 14]. Since analyzing the outcome of NN with arbitrary learning rules is difficult, the alternative approach is often called normative, in which an optimization problem with known offline solutions is formulated and then mapped onto a NN. In the limited settings of nonnegative ICA [69, 47], or bounded component analysis [78, 21] progress towards developing more biologically plausible NN from a normative approach has been made.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

In this work we focus on kurtosis-based ICA methods [10, 32, 53]. The popularity of kurtosis-based methods originates from an early interest in datasets (such as speech signals) where the non-Gaussianity of random processes mainly comes from kurtosis and have naturally symmetric distributions [55]. Within this setup, we search for a solution that can be implemented using a biologically plausible NN. For a NN to be biologically plausible, it must satisfy the following two minimal biological requirements. First, it must operate in the online (or streaming) setting, namely, the input dataset is not available as a whole but is streamed one data vector at a time, and the corresponding output must be computed before the next data vector arrives. Second, the weights of synapses in a NN must be updated using local learning rules.

Contributions: We propose a novel similarity-preserving objective for ICA, with a precise geometric interpretation. We rephrase the objective as a min-max optimization problem and solve it online by stochastic gradient optimization. We demonstrate the performance of our algorithm using synthetic datasets, audio signals, and natural images.

Finally, we show that our online algorithm maps onto a single-layer NN that can separate independent sources without pre-processing. The synaptic weights in our NN are updated using local learning rules, extending more conventional Hebbian learning rules by a time-varying modulating factor, which is a function of the total output activity.

Our result suggests the importance of addressing the extracellular context in which axons and dendrites are present. Modulation of the plasticity rules by overall output intensity agrees with several experimental studies that have reported that a third factor, in addition to pre- and post-synaptic activities, can play a crucial role in modulating the outcome of Hebbian plasticity. This could be accomplished by neuromodulators, extracellular calcium, local field potential, or nitric oxide.

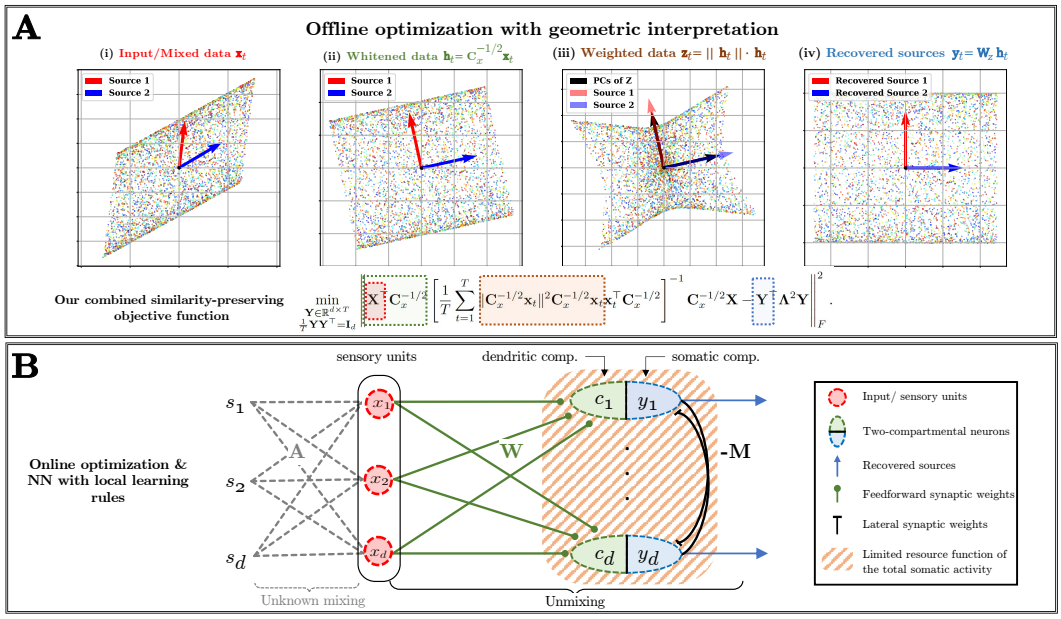

Figure 1: **[A]** illustrates the insights from the FOBI procedure used in our approach using scatterplots of 2D observations, described in Sec. 2.1. We show in **(i)** the observed signal, $\mathbf{x}_t$, in **(ii)** the whitened data, $\mathbf{h}_t$, in **(iii)** their weighted transformation, $\mathbf{z}_t$, and in **(iv)** the recovered sources, $\mathbf{y}_t$. The $\mathbf{y}_t$s are obtained by projection of $\mathbf{h}_t$ onto the principal directions of $\mathbf{z}_t$. The red and blue arrows show the axes of the independent sources s_i, and in **(iii)** the black arrows represent the principal directions of $\mathbf{z}_t$. The principal directions of $\mathbf{z}_t$ overlap with the sources providing a reasonable estimate for the sources s_i. Our objective function shown in **[A]** combines those three components into one single objective function as shown by the color-coding. **[B]** shows the NN we derived. It is composed of two-compartment neurons and trained with local learning rules, described in Sec. 5. The mixed inputs a represented in the sensory units (in red), whereas the dendritic compartment of the output neurons performs and stores the whitened data. The somatic compartment instead reconstructs the sources by rotating the whitened data. In orange, we represent a limited resource used for the plasticity of the feedforward connections, $\mathbf{W}$, and accounts for the nonlinearity introduced by the weighted data.

2 Problem statement and inspiration

The problem of BSS consists of recovering a set of unobservable source signals from observed mixtures. ICA is a natural tool for a class of BSS problems in linear combinations as it decomposes an observed random vector into statistically independent variables.

Mathematically, ICA assumes the following generative model. There are d sources recorded T times forming the columns of $\mathbf{S} := [\mathbf{s}_1, \dots, \mathbf{s}_T] \in \mathbb{R}^{d \times T}$ whose components $s_t^1, \dots, s_t^d$ are assumed non-Gaussian and independent. Without loss of generality, we assume that each source has zero-mean, unit variance, and finite and distinct kurtosis, a common assumption among kurtosis-based ICA methods [53]. The kurtosis of a random variable v is defined as $\mathrm{kurt}[v] = \mathbb{E}\left[(v - \mathbb{E}(v))^4\right] / \left(\mathbb{E}\left[(v - \mathbb{E}(v))^2\right]\right)^2$. Finally, sources are assumed to be mixed through a linear system, i.e., there exists a full rank mixing matrix, $\mathbf{A} \in \mathbb{R}^{d \times d}$, producing the d-dimensional mixture, $\mathbf{x}_t$, expressed as

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t \qquad \forall t \in \{1, \dots, T\} \ . \tag{1}$$

The goal of the various ICA algorithms is then to determine a signal, $\mathbf{y}_t$, obtained from a fixed linear transformation of the observed signal, $\mathbf{x}_t$, i.e., $\exists \mathbf{W}_{ICA} \in \mathbb{R}^{d \times d}$, such that

$$\mathbf{y}_t = \mathbf{\Xi}\mathbf{\Pi}\mathbf{s}_t \quad \text{and} \quad \mathbf{y}_t := \mathbf{W}_{ICA}\mathbf{x}_t, \quad \forall t \in \{1, \dots, T\} \ , \tag{2}$$

where $\mathbf{\Xi}$ is a diagonal matrix with ± 1's on the diagonal, and $\mathbf{\Pi}$ a permutation matrix. As a result, $\mathbf{y}_t$ represents the ideally recovered unknown sources.

2.1 The FOBI procedure.

A method used to learn orthogonal projections is Principal Component Analysis (PCA) [64, 37]. Although PCA and ICA are often contrasted, the authors of [10, 89] noticed that they are connected. Indeed, ICA can be performed in three steps, known as fourth-order blind identification (FOBI). We illustrate the FOBI procedure in Fig. 1**A** and recall the proof from [10, 89] in Appendix A.

Description of the procedure. First of all, the data must be whitened, i.e., all components become decorrelated and of variance one, as shown in Fig. 1**A(ii)**. The use of the whitened data is common in many ICA algorithms because recovering the sources after whitening corresponds to finding a rotation matrix. The way to choose the rotation is what distinguishes the various ICA methods. The whitening step can be performed as follows, with $\mathbf{C}_x = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t^\top$,

$$\textbf{Step 1:} \text{ whiten} \quad \mathbf{h}_t = \mathbf{C}_x^{-1/2}\mathbf{x}_t \ .$$

The key insight of FOBI is that the information about the original sources can be recovered from higher moments of the whitened data, $\mathbf{h}_t$. If we scale the whitened data vector by its norm, denoted by $\mathbf{z}_t$, we can distinguish the direction of distinct kurtosis in $\mathbf{h}_t$ and $\mathbf{s}_t$ in a simple manner. We do so by finding the eigenvectors of the sample covariance matrix $\frac{1}{T}\sum_{t=1}^{T} \mathbf{z}_t \mathbf{z}_t^\top$, as shown in Fig. 1**A(iii)**.

$$\textbf{Step 2a:} \text{ transform } \mathbf{z}_t = \|\mathbf{h}_t\| \cdot \mathbf{h}_t \ ; \ \textbf{Step 2b:} \text{ optimize } \mathbf{W}_z := \underset{\substack{\mathbf{W} \in \mathbb{R}^{d \times d} \\ \mathbf{W}\mathbf{W}^\top = \mathbf{I}_d}}{\arg\max} \ \frac{1}{T} \operatorname{Tr}\left(\mathbf{W} \sum_{t=1}^{T} \mathbf{z}_t \mathbf{z}_t^\top \mathbf{W}^\top\right).$$

Now that we have found the directions corresponding to the sources as the columns of $\mathbf{W}_z$, we simply project the whitened data $\mathbf{h}_t$ onto $\mathbf{W}_z$, shown in Fig. 1**A(iv)** to recover the sources as

$$\textbf{Step 3:} \text{ project} \quad \mathbf{y}_t = \mathbf{W}_z \mathbf{h}_t \ .$$

The problem of implementing FOBI via a biologically plausible NN. Although, no NN was proposed for FOBI, the procedure above can be carried out using usual Oja-based NNs [56]. For example, **Step 2b**, a form of PCA, can be solved by a stochastic gradient ascent algorithm, with $\eta > 0$ the learning rate, as

$$\Delta \mathbf{W}_z \approx \eta \left(\mathbf{u}_t \mathbf{z}_t^\top - \mathbf{u}_t \mathbf{u}_t^\top \mathbf{W}_z\right) \ . \tag{3}$$

Note that (3) is an Oja-like rule and makes sense in a network of neurons with pre-synaptic activity $\mathbf{z}_t$, with synapses of which the strength encodes $\mathbf{W}_z$, and with neurons outputting $\mathbf{u}_t := \mathbf{W}_z \mathbf{z}_t$. However, according to **step 3**, the true output must be $\mathbf{y}_t$, obtained by applying $\mathbf{W}_z$ to the whitened inputs $\mathbf{h}_t$, before the scaling by the norm. To achieve such a goal, one would need some form of weight transport. Alternatively, avoiding weight transport requires a non-local update rule for $\mathbf{W}_z$.

2.2 Similarity matching for principal subspace analysis

To find a biologically realistic implementation of ICA, we need a different formulation of the objective function for ICA that still utilizes the insights from the FOBI procedure. To understand the context of our approach, we need a brief recapitulation of the Similarity Matching (SM) [50, 66, 70, 54] route to an NN performing principal subspace analysis (PSP) [81], a variant of PCA. The authors proposed to perform PSP by solving the following optimization problem

$$\min_{\mathbf{Y}\in\mathbb{R}^{m\times T}} \left\|\mathbf{X}^\top\mathbf{X} - \mathbf{Y}^\top\mathbf{Y}\right\|_F^2 \quad , \tag{4}$$

where $\mathbf{X} := [\mathbf{x}_1, \dots, \mathbf{x}_T]$ is the data matrix, $\mathbf{Y} := [\mathbf{y}_1, \dots, \mathbf{y}_T]$ is the output matrix, and $\|\cdot\|_F$ the Frobenius norm. Notably, the authors demonstrated that the optimization of objective (4) could be done online by networks of linear neurons trained with local learning rules.

Based on our discussion of the FOBI approach, it is easy to understand that naively applying the SM method, a multichannel generalization of the Oja algorithm, to the transformed variables $\mathbf{z}_t$ does not extract the original sources. It is nonetheless an important starting point; it had already proven capable of solving eigenvalue problems with a biologically plausible NN. The SM approach has also already been applied to other computational tasks such as clustering [67], learning manifolds [77] and transformations [4], or canonical correlation analysis [71, 46] and others [74]. In the context of BSS, so far, it has only been used in the limited settings of nonnegative ICA [69, 47], or bounded component analysis [78, 21].

3 A similarity-preserving objective for ICA

To derive a single-layer NN for ICA, which can be trained with local learning rules, we adopt a normative approach and propose a novel objective function. Here we combine the insights of the FOBI procedure with the SM approach by proposing the following generalized nonlinearly weighted similarity matching objective using the notations, $\mathbf{H} := [\mathbf{h}_1, \dots, \mathbf{h}_T]$ and $\mathbf{Z} := [\mathbf{z}_1, \dots, \mathbf{z}_T]$ as defined earlier in Sec. 2.1 and illustrated in Fig. 1**A**,

$$\min_{\mathbf{Y}\in\mathbb{R}^{d\times T}} \left\|\mathbf{H}^\top\left[\frac{1}{T}\mathbf{Z}\mathbf{Z}^\top\right]^{-1}\mathbf{H} - \mathbf{Y}^\top\mathbf{\Lambda}^2\mathbf{Y}\right\|_F^2 \quad , \quad \text{s.t.} \quad \frac{1}{T}\mathbf{Y}\mathbf{Y}^\top = \mathbf{I}_d \ , \tag{5}$$

with $\mathbf{\Lambda}^2 = \mathrm{diag}(\lambda_1^2, \dots, \lambda_d^2)$ any diagonal matrix with distinct finite positive entries. However, for the sake of the derivation of our NN, we explicit our objective (5) and use the full formulation

$$\min_{\substack{\mathbf{Y}\in\mathbb{R}^{d\times T}\\ \frac{1}{T}\mathbf{Y}\mathbf{Y}^\top=\mathbf{I}_d}} \left\|\mathbf{X}^\top\mathbf{C}_x^{-1/2}\left[\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{C}_x^{-1/2}\mathbf{x}_t\|^2\mathbf{C}_x^{-1/2}\mathbf{x}_t\mathbf{x}_t^\top\mathbf{C}_x^{-1/2}\right]^{-1}\mathbf{C}_x^{-1/2}\mathbf{X} - \mathbf{Y}^\top\mathbf{\Lambda}^2\mathbf{Y}\right\|_F^2 . \tag{6}$$

The theorem below states one of our main results: the global minima of our objective (6) perfectly recovers the sources.

Theorem 1. *Given that the sources are independent, centered, have unit variance, and distinct kurtosis (c.f. Sec 2), then the global optimal solution for our objective* (6)*, denoted by* $\mathbf{Y}^*$ *satisfies*

$$\mathbf{Y}^* = \mathbf{\Xi}\mathbf{\Pi}\mathbf{S} \tag{7}$$

where $\mathbf{\Xi}$ *and* $\mathbf{\Pi}$*, defined in* (2)*, represents the sign and permutation ambiguity of the solution, and is thus a solution to the ICA problem.*

Proof. We give the detailed proof of the theorem in Appendix A. □

4 Algorithm derivation

While our objective (6) can be minimized by taking gradient descent steps with respect to $\mathbf{Y}$, this would not lead to an online algorithm because such computation requires combining data from different time steps. Instead, we introduce auxiliary matrix variables, namely $\mathbf{W}$ and $\mathbf{M}$, which store sufficient statistics allowing for the ICA computation, using solely instantaneous inputs.

The introduction of the auxiliary variables, $\mathbf{W}$, and $\mathbf{M}$, is essential to the mapping of our algorithm into a NN. We refer the reader to Section 5 for the neural implementation of our algorithm where we identify $\mathbf{W}$ and $\mathbf{M}$ with synaptic weights and the gradient-descent ascent steps with plasticity rules.

4.1 Min-max formulation

In the following, we rewrite the minimization problem (6) as a min-max problem. For notational convenience, we introduce $\mathbf{\Gamma}_x := \mathbf{C}_x^{-1/2}\left[\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{C}_x^{-1/2}\mathbf{x}_t\|^2\mathbf{C}_x^{-1/2}\mathbf{x}_t\mathbf{x}_t^\top\mathbf{C}_x^{-1/2}\right]^{-1}\mathbf{C}_x^{-1/2}$. We expand the square in Eq. (6), normalizing by T^2, and dropping terms that do not depend on $\mathbf{Y}$ yields:

$$\min_{\mathbf{Y}\in\mathbb{R}^{d\times T}} \frac{1}{T^2}\operatorname{Tr}\left(-2\mathbf{X}^\top\mathbf{\Gamma}_x\mathbf{X}\mathbf{Y}^\top\mathbf{\Lambda}^2\mathbf{Y}+\mathbf{Y}^\top\mathbf{\Lambda}^2\mathbf{Y}\mathbf{Y}^\top\mathbf{\Lambda}^2\mathbf{Y}\right) \quad \text{s.t.} \quad \frac{1}{T}\mathbf{Y}\mathbf{Y}^\top=\mathbf{I}_d \ . \tag{8}$$

The quartic term in $\mathbf{Y}$ in (8) is a constant under the decorrelation constraint and can be dropped from the optimization. We now introduce auxiliary matrix variables $\mathbf{W}$ and $\mathbf{M}$, which results in:

$$\min_{\mathbf{Y}\in\mathbb{R}^{d\times T}}\ \min_{\mathbf{W}\in\mathbb{R}^{d\times d}}\ \max_{\mathbf{M}\in\mathbb{R}^{d\times d}} \mathcal{L}(\mathbf{W},\mathbf{M},\mathbf{Y}) \ , \tag{9}$$

where $\quad \mathcal{L}(\mathbf{W},\mathbf{M},\mathbf{Y}) := \frac{1}{T}\operatorname{Tr}\left(-2\mathbf{X}^\top\mathbf{W}^\top\mathbf{Y}+\mathbf{Y}^\top\mathbf{M}\mathbf{Y}\right)+\operatorname{Tr}\left(\mathbf{W}\mathbf{\Gamma}_x^{-1}\mathbf{W}^\top\mathbf{\Lambda}^{-2}-\mathbf{M}\right) \ .$

The equivalence between the minimization problem (8) and the min-max problem (9) can be seen by taking partial derivatives of $\mathcal{L}(\mathbf{W},\mathbf{M},\mathbf{Y})$ with respect to $\mathbf{W}$ (resp. $\mathbf{M}$) and noting the minimum (resp. maximum) is achieved when $\mathbf{W}=\frac{1}{T}\mathbf{\Lambda}^2\mathbf{Y}\mathbf{X}^\top\mathbf{\Gamma}_x$ (resp. $\frac{1}{T}\mathbf{Y}\mathbf{Y}^\top=\mathbf{I}_d$). Interchanging the order of minimization with respect to $\mathbf{Y}$, with the optimization with respect to $\mathbf{W}$ and $\mathbf{M}$, yields

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times d}}\ \max_{\mathbf{M}\in\mathbb{R}^{d\times d}}\ \min_{\mathbf{Y}\in\mathbb{R}^{d\times T}} \mathcal{L}(\mathbf{W},\mathbf{M},\mathbf{Y}). \tag{10}$$

The interchange is justified by saddle point property of $\mathcal{L}(\mathbf{W},\mathbf{M},\mathbf{Y})$ with respect to $\mathbf{Y}$ and $\mathbf{M}$ [70].

4.2 Gradient optimization in the offline setting

We first optimize the objective (6) in the offline setting, where the entire data matrix $\mathbf{X}$ is accessible. In this case, we solve the min-max problem (9) by alternating optimization steps. For fixed $\mathbf{W}$ and $\mathbf{M}$, we minimize the objective function $\mathcal{L}(\mathbf{W},\mathbf{M},\mathbf{Y})$ over $\mathbf{Y}$, which yields the relation

$$\mathbf{Y} := \arg\min_{\mathbf{Y}\in\mathbb{R}^{d\times T}} \mathcal{L}(\mathbf{W},\mathbf{M},\mathbf{Y}) \ = \ \mathbf{M}^{-1}\mathbf{W}\mathbf{X} \ . \tag{11}$$

Before applying gradient optimization steps of the objective function $\mathcal{L}(\mathbf{W},\mathbf{M},\mathbf{Y})$ with respect to $\mathbf{W}$ and $\mathbf{M}$, we first investigate the term $\mathbf{\Gamma}_x^{-1}$ appearing in (10), as $\operatorname{Tr}\left(\mathbf{W}\mathbf{\Gamma}_x^{-1}\mathbf{W}^\top\mathbf{\Lambda}^{-2}\right)$,

$$\mathbf{\Gamma}_x^{-1} = \mathbf{C}_x^{1/2}\left[\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{C}_x^{-1/2}\mathbf{x}_t\|^2\mathbf{C}_x^{-1/2}\mathbf{x}_t\mathbf{x}_t^\top\mathbf{C}_x^{-1/2}\right]\mathbf{C}_x^{1/2} \ = \ \frac{1}{T}\sum_{t}^{T}\alpha_t\mathbf{x}_t\mathbf{x}_t^\top \ , \tag{12}$$

where $\alpha_t = \|\mathbf{C}_x^{-1/2}\mathbf{x}_t\|^2$. The α_t term corresponds to the norm-squared of the whitened data; more importantly it is also the norm of the sources and of the outputs yielding

$$\alpha_t = \|\mathbf{C}_x^{-1/2}\mathbf{x}_t\|^2 = \|\mathbf{s}_t\|^2 = \|\mathbf{y}_t\|^2 \ . \tag{13}$$

The identification of α_t with $\|\mathbf{y}_t\|^2$ in (13) is a direct consequence of the learning of a rotation matrix, as in **Step 2b** of the FOBI approach (Sec 2.1), which by definition preserves the Euclidean norm. We then use (13) to rewrite $\mathbf{\Gamma}_x^{-1}$ as $\frac{1}{T}\mathbf{X}\operatorname{ddiag}(\mathbf{Y}^\top\mathbf{Y})\mathbf{X}^\top = \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{y}_t\|^2\mathbf{x}_t\mathbf{x}_t^\top$, where $\operatorname{ddiag}(\cdot)$ represents the diagonal matrix whose diagonal elements are those of the matrix in the argument. We now obtain the update rules for $\mathbf{W}$ and $\mathbf{M}$ by gradient-descent ascent and by replacing $\mathbf{\Gamma}_x^{-1}$

$$\mathbf{W}\leftarrow \mathbf{W}+2\eta\left(\tfrac{1}{T}\mathbf{Y}\mathbf{X}^\top-\mathbf{\Lambda}^{-2}\mathbf{W}\mathbf{\Gamma}_x^{-1}\right) = \mathbf{W}+\frac{2\eta}{T}\left(\mathbf{Y}\mathbf{X}^\top-\mathbf{\Lambda}^{-2}\mathbf{W}\mathbf{X}\operatorname{ddiag}(\mathbf{Y}^\top\mathbf{Y})\mathbf{X}^\top\right) \ , \tag{14}$$

$$\mathbf{M}\leftarrow \mathbf{M}+\frac{\eta}{\tau}\left(\frac{1}{T}\mathbf{Y}\mathbf{Y}^\top-\mathbf{I}_d\right) \ . \tag{15}$$

Here $\tau>0$ is the ratio between the learning rates for $\mathbf{W}$ and $\mathbf{M}$, and $\eta\in(0,\tau)$ is the learning rate for $\mathbf{W}$, ensuring that $\mathbf{M}$ remains positive definite given a positive definite initialization.

4.3 Online algorithm

We now solve the min-max objective (9) in the online setting. At each time step t, we minimize over the output $\mathbf{y}_t$ by repeating the following gradient descent steps until convergence:

$$\mathbf{y}_t \leftarrow \mathbf{y}_t + \gamma(\mathbf{c}_t - \mathbf{M}\mathbf{y}_t) \ , \tag{16}$$

where γ is a small step size, and we have defined the projection $\mathbf{c}_t := \mathbf{W}\mathbf{x}_t$, with biological interpretation described in Sec 5. We then take stochastic gradient descent-ascent steps in $\mathbf{W}$ and $\mathbf{M}$. We thus replace the averages in Eqs. (14)-(15) with their online approximations

$$\frac{1}{T}\mathbf{Y}\mathbf{X}^\top = \mathbf{y}_t\mathbf{x}_t^\top \ ; \ \frac{1}{T}\mathbf{Y}\mathbf{Y}^\top = \mathbf{y}_t\mathbf{y}_t^\top \ ; \ \frac{1}{T}\mathbf{\Lambda}^{-2}\mathbf{W}\mathbf{X}\,\mathrm{ddiag}(\mathbf{Y}^\top\mathbf{Y})\mathbf{X}^\top = \|\mathbf{y}_t\|^2\mathbf{\Lambda}^{-2}\mathbf{c}_t\mathbf{x}_t^\top \ .$$

This yields our online algorithm for ICA detailed in Algorithm 1 and biologically plausible NN detailed in Section 5.

Algorithm 1 A similarity-preserving algorithm for Independent Component Analysis.

input data $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$; dimension d
output $\{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$; dimension d ▷ estimated sources.
initialize the matrix $\mathbf{W}$, and positive definite matrix $\mathbf{M}$.
for $t = 1, 2, \ldots, T$ **do**
  $\mathbf{c}_t \leftarrow \mathbf{W}\mathbf{x}_t$; ▷ projection of inputs
  run the following until convergence:
    $\frac{d\mathbf{y}_t(\gamma)}{d\gamma} = \mathbf{c}_t - \mathbf{M}\mathbf{y}_t(\gamma)$; ▷ neural dynamics
  $\mathbf{W} \leftarrow \mathbf{W} + 2\eta(\mathbf{y}_t - \|\mathbf{y}_t\|^2\mathbf{\Lambda}^{-2}\mathbf{c}_t)\mathbf{x}_t^\top$; $\mathbf{M} \leftarrow \mathbf{M} + \frac{\eta}{\tau}(\mathbf{y}_t\mathbf{y}_t^\top - \mathbf{I}_d)$; ▷ synaptic updates
end for

5 Biological interpretation and neural implementation

We now show that our online ICA algorithm (Alg.1) maps onto a NN with local, activity-dependent synaptic update rules, which emulate aspects of synaptic plasticity observed experimentally.

5.1 Neural architecture and dynamics

Our algorithm can be implemented by a biologically plausible NN presented in Fig.1**B**. The network consists of an input layer of d neurons, representing the input data to be separated into independent components, and an output layer of d neurons, with separate dendritic and somatic compartments estimating the unknown sources. It also comprises two fully connected pathways, a feedforward pathway between inputs and dendrites and a recurrent lateral pathway between the somas.

At each time step t, the mixture $\mathbf{x}_t$, represented in the input neurons, is multiplied by the weight matrix $\mathbf{W}$ encoded by the feedforward synapses connecting the input neurons to the output neurons. This yields the projection $\mathbf{c}_t = \mathbf{W}\mathbf{x}_t$ computed in the dendritic compartments of the output neurons and then propagated to their somatic compartments. This is followed by the fast recurrent neural dynamics in the somatic compartments described by the dynamics Eq. (16).

The matrix $\mathbf{M}$ is encoded by the lateral synapses connecting the somas. The equilibrium value of the dynamic (16) corresponds to the assignment for $\mathbf{y}_t$ in Alg. 1. The d-dimensional output signal $\mathbf{y}_t$ is represented as somatic activity in the output neurons and corresponds to the estimated sources.

5.2 Synaptic plasticity rules

We now highlight the locality of our learning by rewriting the element-wise synaptic updates for $\mathbf{W}$ and $\mathbf{M}$ in Algorithm 1 as

$$W_{ij} \leftarrow W_{ij} + 2\eta\left(y_t^i x_t^j - \|\mathbf{y}_t\|^2\frac{c_t^i}{\lambda_i^2}x_t^j\right) \ ; \ M_{ij} \leftarrow M_{ij} + \frac{\eta}{\tau}\left(y_t^i y_t^j - \delta_{ij}\right), \ 1 \le i, j \le d \ . \tag{17}$$

In Eqs. (17), the y_t^i, x_t^j, and c_t^j terms describe respectively the post-synaptic factor of the ith neuron, the pre-synaptic factor and the dendritic current of the jth neuron at time t. Furthermore, the influence of the local dendritic currents c_j^t on a synapse's strength is modulated by the term $\|\mathbf{y}_t\|^2$, representing the overall activity of the output neurons. We argue in the following how this information gets accessed locally via the reduction of the abundance of a diffusible plasticity modulating factor.

Near equilibrium and optimality the model has a clear interpretation. Indeed, the dendritic current can then be identified with $\mathbf{c}_t = \mathbf{W}\mathbf{x}_t = \mathbf{M}\mathbf{y}_t$, from Eq. (16), with $\mathbf{M} \approx \mathbf{I}_d$. The learning rule of the feedforward weights $\mathbf{W}$, Eq.(17), then becomes

$$\Delta\mathbf{W} = \eta_t \left(\mathbf{I} - \|\mathbf{y}_t\|^2 \mathbf{\Lambda}^{-2}\mathbf{M}\right) \mathbf{y}_t\mathbf{x}_t^\top \approx \eta_t \left(\mathbf{I} - \|\mathbf{y}_t\|^2 \mathbf{\Lambda}^{-2}\right) \mathbf{y}_t\mathbf{x}_t^\top \quad . \tag{18}$$

It now becomes apparent that the type of plasticity switches depending on the total somatic activity and can be considered a thresholded nonlinearly modulated Hebbian rule. For global sub-threshold total somatic activity, i.e., $\|\mathbf{y}_t\|^2 < \lambda_i^2$, the abundance of the modulating factor facilitates a current activity pattern by inducing Hebbian change. In contrast, for global super-threshold output, i.e., $\|\mathbf{y}_t\|^2 > \lambda_i^2$, its depletion induces anti-Hebbian change. In this way, the modulating factor maintains the difference between the somatic output and the scaled dendritic current close to zero. In short, the rule induces long-term potentiation (LTP) for global sub-threshold responses and long-term depression (LTD) for global super-threshold responses. We contrast this global activity-dependent modulation of plasticity with Bienenstock, Cooper, and Munro (BCM) rule in subsection 5.3.

Existence of a global activity-dependent plasticity modulating factor. In the learning rule for feedforward synaptic weights (18), the update $\Delta\mathbf{W}$, depends on the overall neural activity of the output layer, $\|\mathbf{y}_t\|^2$. How could overall neural activity be signaled to each feedforward synapse in the brain? It is natural to suggest that a diffusible molecule whose concentration depends on overall neural activity may affect synaptic plasticity. One candidate proposed for this role in the cerebellum is extracellular calcium [16]. A number of neuromodulators could modulate synaptic plasticity in an activity-dependent way: GABA[26, 61], dopamine [75, 88, 87], noradrenaline [76, 36] or D-Serin [27] as also argued in [34, 35]. Nitric oxide (NO) is another chemical implicated in synaptic plasticity, although its range of action is contested [29, 28, 24]. Finally, local field potential can also affect synaptic plasticity [83].

Whereas the importance of three-factor learning has been shown in other computational models, primarily for reward-based models [80, 51, 84, 41], our model is the first to propose such a factor as part of a fully normative approach for ICA.

5.3 Comparison with existing rules

To understand the distinctive features of our model against existing approaches, we contrast it with three important works: 1. Oja-based models, 2. the famous BCM rule [7], and 3. error-gated Hebbian rule (EGHR) [34].

1. Oja-based models. The pioneering work of the authors of [57, 31], known as nonlinear Oja, further generalized the standard Hebbian and Oja's rule, Eq. (3), with a component-wise nonlinear function $g(\cdot)$ as $\Delta\mathbf{W} = g(\mathbf{y}_t)\mathbf{x}_t^\top - g(\mathbf{y}_t)g(\mathbf{y}_t)^\top\mathbf{W}$ (see [57, 31] for details). However, this model and follow-up work inherited the main drawbacks of the standard Oja's rule, i.e., they require pre-whitening of the data and rely on non-local learning rules. Indeed, the last term of the learning rules of nonlinear Oja implies that updating the weight of a synapse requires precise knowledge of output activities of all other neurons which are not available to the synapse (c.f. [68] for details on standard PCA rules and networks of nonlinear neurons [60]).

2. Contrast with the BCM rule. We showed in Eq. (18) that our learning is approximately a nonlinear thresholded Hebbian rule and is evocative to one of the most influential theories of synaptic plasticity with similar properties: the BCM rule. It was initially postulated and later connected to an objective function [33, 13] characterizing the deviation from Gaussian distribution but mainly focusing on skewness rather than kurtosis as in our model. When looking at a single neuron, the BCM rule induces LTD for local sub-threshold responses and LTP local for super-threshold responses, with a threshold function of average output activity. At the population level, output neurons [33, 13] tend to respond to the same dominant feature, producing an incomplete, highly redundant code for visual information. In contrast, the lateral inhibitory connections in our network are updated via anti-Hebbian rule Eq. (17). In our online algorithm, it is the learning of the inhibitory connections that leads to the globally optimal solution of our objective leading to the exhaustive recovery of the

sources. Although experimental evidence has validated BCM-like plasticity in parts of the visual cortex and hippocampus, other brain areas have yet to show similar behavior. Interestingly, an "inverse" BCM rule, like our own has been suggested in the cerebellum [17, 38, 82].

3. Modulated Hebbian rules. The recent set of work in [34, 41, 35] also introduced a set of modulating Hebbian rules for ICA as $\Delta\mathbf{W} = (E_0 - E(\mathbf{y}_t))g(\mathbf{y}_t)\mathbf{x}_t^\top$, with E_0 a constant, $E(\cdot)$ a nonlinear function of the total activity, and $g(\cdot)$ a component-wise nonlinear function. This learning rule shares many similarities with our own. The term E_0 is a constant characterizing the source distributions, which could identify with our λ_i terms, and the function $E(\cdot)$ resembles our $\|\mathbf{y}_t\|^2$ but is model dependent in their approach. This is where the similarities end as their objective function is inspired by the information-theoretic framework of [6, 32] and ours to the insight from the FOBI procedure [10, 62] and spectral methods from the SM method [70].

Their model can be considered partly normative since the neural architecture is predetermined and uses a hand-designed error-computing neuron to determine the global modulating factor rather than being derived from an optimization problem. Interestingly, their model does not use lateral connections for output decorrelation resulting in a model without direct interaction between outputs.

6 Numerical simulations

In this section, we verify by numerical experiments our theoretical results. We use our model to perform ICA on both synthetic and real world datasets. We detail the parameters used and propose further experimental results in Appendix C. More precisely, three sets of experiments were designed to illustrate the performance of our approach. In the first set, shown in Fig. 2**A** we used artificially generated signals, in the second set natural speech signals, Fig. 2**B**, and in the third set natural scene images, Fig. 2**C**, as sources, $\mathbf{s}_t$. We then used a random full rank square mixing matrices, $\mathbf{A}$, to generate the observed mixed signals, $\mathbf{x}_t$ (with the generative model defined in Eq. (1)) from which we aimed at recovering the original sources. We show that our model recovers sources indiscriminately of the sign of the kurtosis of the distributions considered, which is essential in natural datasets.

Synthetic data. We first evaluate our algorithm on a synthetic dataset generated by independent and identically distributed samples. The data are generated from meaningful signals, i.e., square-periodic, sine-wave, saw-tooth, and Laplace random noise. The data were chosen with the purpose of including both super- and sub-Gaussian distribution known respectively as leptokurtic ("spiky", e.g., the Laplace distribution) and platykurtic ("flat-topped", the three other source signal). We show in Fig. 2**A** the mixed signals in black, on the left plots. We show on the right plot the recovered sources, in red, overlapped with the original sources, in blue, and the residual in green. We also show the histogram of each signal on the right side of each plot. Results are shown for 300 samples. We observe that the recovered and true sources nearly perfectly overlap, explaining the low value of the residual, which shows the almost perfect reconstruct performed by our algorithm.

Real world data - Speech signals. For the audio separation task, we used speech recordings from the freely available TSP data set [39][1] , recorded at 16kHz. The first source we use was obtained from a male speaker (MA02 04.wav), the second source from a female speaker (FA01 03.wav), and the third source synthetically generated from a uniform noise, as was previously used in the literature [9]. We show our results in Fig.2**B**. We show the mixtures, the true sources, the recovered sources, and the residual. It is clear from the figure that our algorithm's outputs recover the true sources similarly to the synthetic dataset.

Real world data - Natural scene images. We finally applied our algorithm to the task of recovering images from their mixtures, on data already used for BSS tasks [32, 30, 47] [2], as shown in Fig.2**C**. Here, we show separately the original sources, top images, the mixtures, middle images, and the recovered sources, bottom images of Fig.2**C**. We considered three grayscale images of size 256×512 pixels (shifted and scaled to have zero-mean and unit variance), such that each image is treated as one source, with the pixel intensities representing the samples. We again observe in Fig.2**C** that the recovered sources are nearly identical to the original sources. We can also see that the histograms of the recovered sources nearly match the histograms of the original sources, up to their sign. However, it is not perfect, which could be explained by the partial dependence of natural images.

[1] Freely available at http://www.mmsp.ece.mcgill.ca/Documents/Data/ (Accessed May 24th 2021).

[2] Freely available at https://research.ics.aalto.fi/ica/data/images/ (Accessed May 24th 2021).

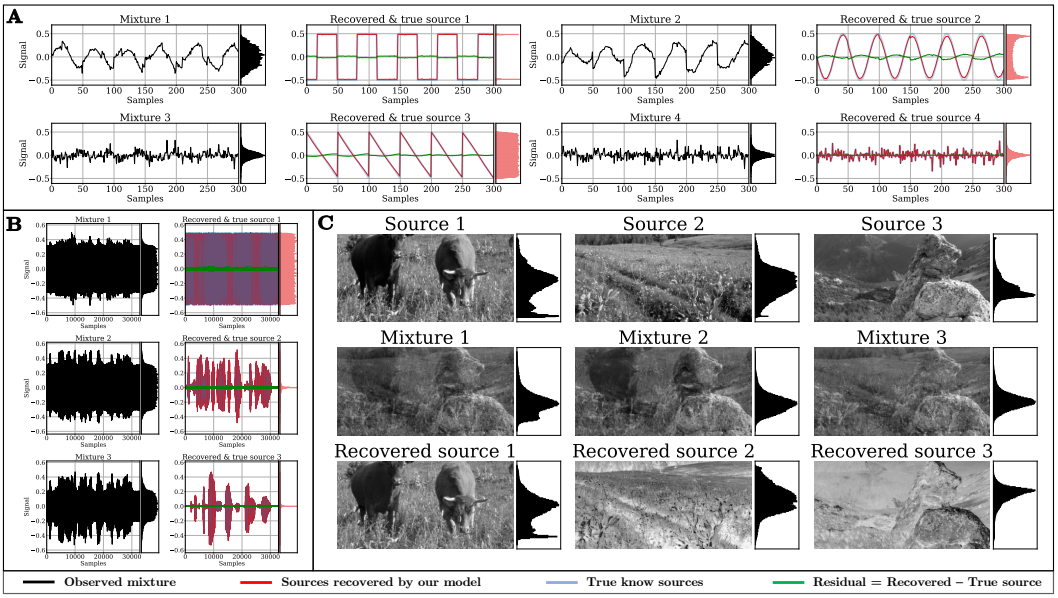

Figure 2: Illustration of the performance of our algorithm on various datasets. Our recovers the sources from mixed signals on various datasets. **[A]** shows the results obtained for synthetic data, **[B]** the speech separation and **[C]** the image separation task. In **[A-B]** we show in black the mixed signals and in red (resp. blue) the recovered (resp. true) sources, and in green their difference called Residual. We also show the histogram of the associated distributions. We show separately in **[C]** the sources, mixture, and recovered sources, on top, middle and bottom images respectively.

7 Discussion

We proposed a new single-layer NN with biologically plausible local learning rules for ICA. The normative nature of our approach makes the biologically realistic features of our NNs readily interpretable. In particular, our NN uses neurons with two separate compartments and is trained with extended Hebbian learning rules. The changes in synaptic strength are gated by a global signal summed over a local neural population, equivalent to performing gradient optimization of our objective function. We demonstrated that the proposed rule reliably converges to an ideal solution over a wide range of mixing matrices, synthetic, and natural datasets. The broad applicability and easy implementation of our NN and learning rules could further advance neuromorphic computation [73, 23, 65] and may reveal the principle underlying BSS computation in the brain.

Recent work on canonical correlation analysis [12, 25, 63] and slow feature analysis [86] have led to biologically plausible NNs [48, 46], which rely on two-compartment neurons. These NNs could, in principle, be used for popular BSS tasks known as second-order blind identification [79, 49, 8, 15] or in the context of kernel ICA [3, 44]. This observation suggests the existence of a single model of two-compartment neurons and non-trivial local learning rules for BSS. In future work, we aim at proposing such a model, including high-order statistics, temporal correlation, and diversity of views.

One limitation of our approach is the inability of the model to separate sources with the same kurtosis. As long as even some even order moments are different for the sources, we can alter our scaling rule and separate the sources [10]. Another limitation is the well-known sensitivity of kurtosis to outliers. A change of the scaling could again address this, this time as a sublinear function of the total activity [89]. These changes do not affect the locality of the learning rules nor the neural architecture.

Clarifying the limitations of our model leads us to ask various follow-up questions left for future work. How can we further generalize the solution beyond the choice of nonlinearity and beyond the task of linear ICA? We could envision considering more than two covariance matrices as in the famous JADE algorithm [11, 58, 59], which effectively performs joint-diagonalization of arbitrarily many matrices. A neural solution was proposed in [90] but again relies on non-local Oja-based rules.

References

[1] S.-I. Amari and A. Cichocki. Adaptive blind signal processing-neural network approaches. *Proceedings of the IEEE*, 86(10):2026–2048, 1998.

[2] S.-I. Amari, A. Cichocki, and H. Yang. A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems*, 8:757–763, 1995.

[3] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3(Jul):1–48, 2002.

[4] Y. Bahroun, D. Chklovskii, and A. Sengupta. A similarity-preserving network trained on transformed images recapitulates salient features of the fly motion detection circuit. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[5] M. A. Bee and C. Micheyl. The cocktail party problem: what is it? how can it be solved? and why should animal behaviorists study it? *Journal of comparative psychology*, 122(3):235, 2008.

[6] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[7] E. L. Bienenstock, L. N. Cooper, and P. W. Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, 2(1):32–48, 1982.

[8] T. Blaschke, P. Berkes, and L. Wiskott. What is the relation between slow feature analysis and independent component analysis? *Neural Computation*, 18(10):2495–2508, 2006.

[9] P. Brakel and Y. Bengio. Learning independent features with adversarial nets for non-linear ica. *arXiv preprint arXiv:1710.05050*, 2017.

[10] J.-F. Cardoso. Source separation using higher order moments. In *ICASSP*, pages 2109–2112. IEEE, 1989.

[11] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-gaussian signals. *IEE Proceedings F (Radar and Signal Processing)*, 140(6):362–370, 1993.

[12] J. D. Carroll. Generalization of canonical correlation analysis to three or more sets of variables. *Proc. APA*, pages 227–228, 1968.

[13] G. Castellani, N. Intrator, H. Shouval, and L. Cooper. Solutions of the bcm learning rule in a network of lateral interacting nonlinear neurons. *Network: Computation in Neural Systems*, 10(2):111–121, 1999.

[14] A. Cichocki, J. Karhunen, W. Kasprzak, and R. Vigario. Neural networks for blind separation with unknown number of sources. *Neurocomputing*, 24(1-3):55–93, 1999.

[15] C. Clopath, A. Longtin, and W. Gerstner. An online Hebbian learning rule that performs independent component analysis. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

[16] O. Coenen, D. Eagleman, V. Mitsner, T. Bartol, A. Bell, and T. Sejnowski. Cerebellar glomeruli: Does limited extracellular calcium direct a new kind of plasticity. In *Society for Neuroscience Abstracts*, volume 27, 2001.

[17] M. Coesmans, J. T. Weber, C. I. De Zeeuw, and C. Hansel. Bidirectional parallel fiber plasticity in the cerebellum under climbing fiber control. *Neuron*, 44(4):691–700, 2004.

[18] P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[19] P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[20] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1):193–222, 1995.

[21] A. T. Erdogan and C. Pehlevan. Blind bounded source separation using neural networks with local learning rules. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3812–3816. IEEE, 2020.

[22] P. Földiak. Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics*, 64(2):165–170, 1990.

[23] M. E. Fouda, E. Neftci, A. Eltawil, and F. Kurdahi. Independent component analysis using rrams. *IEEE Transactions on Nanotechnology*, 18:611–615, 2018.

[24] N. Hardingham, J. Dachtler, and K. Fox. The role of nitric oxide in pre-synaptic plasticity and homeostasis. *Frontiers in cellular neuroscience*, 7:190, 2013.

[25] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[26] T. Hayama, J. Noguchi, S. Watanabe, N. Takahashi, A. Hayashi-Takagi, G. C. Ellis-Davies, M. Matsuzaki, and H. Kasai. Gaba promotes the competitive selection of dendritic spines by controlling local ca 2+ signaling. *Nature neuroscience*, 16(10):1409, 2013.

[27] C. Henneberger, T. Papouin, S. H. Oliet, and D. A. Rusakov. Long-term potentiation depends on release of d-serine from astrocytes. *Nature*, 463(7278):232–236, 2010.

[28] C. Hölscher. Nitric oxide, the enigmatic neuronal messenger: its role in synaptic plasticity. *Trends in neurosciences*, 20(7):298–303, 1997.

[29] E. P. Huang. Synaptic plasticity: a role for nitric oxide in ltp. *Current Biology*, 7(3):R141–R143, 1997.

[30] A. Hyvärinen and P. Hoyer. Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural computation*, 12(7):1705–1720, 2000.

[31] A. Hyvärinen and E. Oja. Independent component analysis by general nonlinear Hebbian-like learning rules. *Signal Processing*, 64(3):301–313, 1998.

[32] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

[33] N. Intrator and L. N. Cooper. Objective function formulation of the bcm theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks*, 5(1):3–17, 1992.

[34] T. Isomura and T. Toyoizumi. A local learning rule for independent component analysis. *Scientific Reports*, 6(1):1–17, 2016.

[35] T. Isomura and T. Toyoizumi. Error-gated Hebbian rule: A local learning rule for principal and independent component analysis. *Scientific Reports*, 8(1):1–11, 2018.

[36] J. P. Johansen, L. Diaz-Mataix, H. Hamanaka, T. Ozawa, E. Ycu, J. Koivumaa, A. Kumar, M. Hou, K. Deisseroth, E. S. Boyden, et al. Hebbian and neuromodulatory mechanisms interact to trigger associative memory formation. *Proceedings of the National Academy of Sciences*, 111(51):E5584–E5592, 2014.

[37] I. T. Jolliffe. *Principal Component Analysis and Factor Analysis*, volume 1. Springer, 1986.

[38] H. Jörntell and C. Hansel. Synaptic memories upside down: bidirectional plasticity at cerebellar parallel fiber-purkinje cell synapses. *Neuron*, 52(2):227–238, 2006.

[39] P. Kabal. TSP speech database. *McGill University, Database Version*, 1(0):09–02, 2002.

[40] J. Karhunen, E. Oja, L. Wang, R. Vigario, and J. Joutsensalo. A class of neural networks for independent component analysis. *IEEE Transactions on Neural Networks*, 8(3):486–504, 1997.

[41] L. Kuśmierz, T. Isomura, and T. Toyoizumi. Learning with three factors: modulating Hebbian plasticity with errors. *Current Opinion in Neurobiology*, 46:170–177, 2017.

[42] T.-W. Lee. Independent component analysis. In *Independent component analysis*, pages 27–66. Springer, 1998.

[43] T.-W. Lee, M. Girolami, A. J. Bell, and T. J. Sejnowski. A unifying information-theoretic framework for independent component analysis. *Computers & Mathematics with Applications*, 39(11):1–21, 2000.

[44] Y.-O. Li, T. Adali, W. Wang, and V. D. Calhoun. Joint blind source separation by multiset canonical correlation analysis. *IEEE Transactions on Signal Processing*, 57(10):3918–3929, 2009.

[45] R. Linsker. A local learning rule that enables information maximization for arbitrary input distributions. *Neural Computation*, 9(8):1661–1665, 1997.

[46] D. Lipshutz, Y. Bahroun, S. Golkar, A. M. Sengupta, and D. B. Chklovskii. A biologically plausible neural network for multi-channel canonical correlation analysis. *arXiv preprint arXiv:2010.00525*, 2020.

[47] D. Lipshutz and D. B. Chklovskii. Bio-NICA: A biologically inspired single-layer network for nonnegative independent component analysis. *arXiv preprint arXiv:2010.12632*, 2020.

[48] D. Lipshutz, C. Windolf, S. Golkar, and D. Chklovskii. A biologically plausible neural network for slow feature analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 14986–14996, 2020.

[49] H. Liu and Y. Cheung. A learning framework for blind source separation using generalized eigenvalues. In *International Symposium on Neural Networks*, pages 472–477. Springer, 2005.

[50] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis (probability and mathematical statistics)*. Academic Press London, 1980.

[51] P. Mazzoni, R. A. Andersen, and M. I. Jordan. A more biologically plausible learning rule for neural networks. *Proceedings of the National Academy of Sciences*, 88(10):4433–4437, 1991.

[52] J. H. McDermott. The cocktail party problem. *Current Biology*, 19(22):R1024–R1027, 2009.

[53] J. Miettinen, S. Taskinen, K. Nordhausen, and H. Oja. Fourth moments and independent component analysis. *Statistical Science*, 30(3):372–390, 2015.

[54] V. Minden, C. Pehlevan, and D. B. Chklovskii. Biologically plausible online principal component analysis without recurrent neural dynamics. In *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, pages 104–111. IEEE, 2018.

[55] C. L. Nikias and J. M. Mendel. Signal processing with higher-order spectra. *IEEE Signal Processing Magazine*, 10(3):10–37, 1993.

[56] E. Oja. Principal components, minor components, and linear neural networks. *Neural Networks*, 5(6):927–935, 1992.

[57] E. Oja. The nonlinear pca learning rule in independent component analysis. *Neurocomputing*, 17(1):25–45, 1997.

[58] H. Oja, S. Sirkiä, and J. Eriksson. Scatter matrices and independent component analysis. *Austrian Journal of Statistics*, 35(2&3):175–189, 2006.

[59] E. Ollila, H. Oja, and V. Koivunen. Complex-valued ICA based on a pair of generalized covariance matrices. *Computational Statistics & Data Analysis*, 52(7):3789–3805, 2008.

[60] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[61] V. Paille, E. Fino, K. Du, T. Morera-Herreras, S. Perez, J. H. Kotaleski, and L. Venance. Gabaergic circuits control spike-timing-dependent plasticity. *Journal of Neuroscience*, 33(22):9353–9363, 2013.

[62] L. Parra and P. Sajda. Blind source separation via generalized eigenvalue decomposition. *Journal of Machine Learning Research*, 4:1261–1269, 2003.

[63] L. C. Parra. Multiset canonical correlation analysis simply explained. arXiv preprint arXiv:1802.03759, 2018.

[64] K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[65] C. Pehlevan. A spiking neural network with local learning rules derived from nonnegative similarity matching. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7958–7962. IEEE, 2019.

[66] C. Pehlevan and D. Chklovskii. A normative theory of adaptive dimensionality reduction in neural networks. In *Advances in Neural Information Processing Systems*, pages 2269–2277, 2015.

[67] C. Pehlevan and D. B. Chklovskii. A Hebbian/anti-Hebbian network derived from online non-negative matrix factorization can cluster and discover sparse features. In *2014 48th Asilomar Conference on Signals, Systems and Computers*, pages 769–775. IEEE, 2014.

[68] C. Pehlevan and D. B. Chklovskii. Neuroscience-inspired online unsupervised learning algorithms: Artificial neural networks. *IEEE Signal Processing Magazine*, 36(6):88–96, 2019.

[69] C. Pehlevan, S. Mohan, and D. B. Chklovskii. Blind nonnegative source separation using biological neural networks. *Neural Computation*, 29(11):2925–2954, 2017.

[70] C. Pehlevan, A. M. Sengupta, and D. B. Chklovskii. Why do similarity matching objectives lead to Hebbian/anti-Hebbian networks? *Neural Computation*, 30(1):84–124, 2017.

[71] C. Pehlevan, X. Zhao, A. M. Sengupta, and D. Chklovskii. Neurons as canonical correlation analyzers. *Frontiers in Computational Neuroscience*, 14:55, 2020.

[72] D. T. Pham and P. Garat. Blind separation of mixture of independent sources through a maximum likelihood approach. In *In proc. eusipco*. Citeseer, 1997.

[73] J. H. Poikonen and M. Laiho. Online linear subspace learning in an analog array computing architecture. In *CNNA 2016; 15th International Workshop on Cellular Nanoscale Networks and their Applications*, pages 1–2. VDE, 2016.

[74] S. Qin, N. Mudur, and C. Pehlevan. Contrastive similarity matching for supervised learning. *Neural Computation*, 33(5):1300–1328, 2021.

[75] J. N. Reynolds, B. I. Hyland, and J. R. Wickens. A cellular mechanism of reward-related learning. *Nature*, 413(6851):67–70, 2001.

[76] H. Salgado, G. Köhr, and M. Trevino. Noradrenergic 'tone'determines dichotomous control of cortical spike-timing-dependent plasticity. *Scientific reports*, 2(1):1–7, 2012.

[77] A. Sengupta, C. Pehlevan, M. Tepper, A. Genkin, and D. Chklovskii. Manifold-tiling localized receptive fields are optimal in similarity-preserving neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[78] B. Simsek and A. T. Erdogan. Online bounded component analysis: A simple recurrent neural network with local update rule for unsupervised separation of dependent and independent sources. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pages 1639–1643. IEEE, 2019.

[79] J. V. Stone. Blind source separation using temporal predictability. *Neural Computation*, 13(7):1559–1574, 2001.

[80] R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

[81] A.-J. Van Der Veen, E. F. Deprettere, and A. L. Swindlehurst. Subspace-based signal analysis using singular value decomposition. *Proceedings of the IEEE*, 81(9):1277–1308, 1993.

[82] K. E. Vogt and M. Canepari. On the induction of postsynaptic granule cell–purkinje neuron ltp and ltd. *The Cerebellum*, 9(3):284–290, 2010.

[83] M. Weckström and S. Laughlin. Extracellular potentials modify the transfer of information at photoreceptor output synapses in the blowfly compound eye. *Journal of Neuroscience*, 30(28):9557–9566, 2010.

[84] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[85] R. I. Wilson and Z. F. Mainen. Early events in olfactory processing. *Annual Review of Neuroscience*, 29:163–201, 2006.

[86] L. Wiskott and T. J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002.

[87] S. Yagishita, A. Hayashi-Takagi, G. C. Ellis-Davies, H. Urakubo, S. Ishii, and H. Kasai. A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, 345(6204):1616–1620, 2014.

[88] J.-C. Zhang, P.-M. Lau, and G.-Q. Bi. Gain in sensitivity and loss in temporal contrast of stdp by dopaminergic modulation at hippocampal synapses. *Proceedings of the National Academy of Sciences*, 106(31):13028–13033, 2009.

[89] K. Zhang and L.-W. Chan. ICA by PCA approach: relating higher-order statistics to second-order moments. In *International Conference on Independent Component Analysis and Signal Separation*, pages 311–318. Springer, 2006.

[90] C. Ziegaus and E. W. Lang. A neural implementation of the jade algorithm (nJADE) using higher-order neurons. *Neurocomputing*, 56:79–100, 2004.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section 6
- Did you include the license to the code and datasets? [No] The code and the data will be made publicly available but for now are included with the submission.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Yes, we clearly state the setup in which we work,i.e., kurtosis-based linear ICA with sources with finite distinct kurtosis and search for a biologically plausible neural network.
   (b) Did you describe the limitations of your work? [Yes] Yes, we describe extensively the limitations of our model in Sec 7. We also mention possible future research direction
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We do not forsee any negative impact rather positive one, it can have wide-ranging benefits for helping to manage the adverse effects of neurological disorders.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] Yes.

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Yes.
   (b) Did you include complete proofs of all theoretical results? [Yes] We left the proofs in the supplementary materials but gave some intuitions illustrated with toy examples as in Fig. 1.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We include the code to the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Yes, we explain how data are generated, and where the datasets used are available as well as training parameters.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] For Figs. in the supplementary material for comparison with existing non-biologically plausible models.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We only run on a basic computer as the model is online and is designed to not require much computing power and memory.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] In Sec 6
   (b) Did you mention the license of the assets? [No] All data used either synthetic data or publicly available.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] No new assets aside from our code of the new algorithm.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]