On the achievability of blind source separation for high-dimensional nonlinear source mixtures

Takuya Isomura & Taro Toyoizumi
Laboratory for Neural Computation and Adaptation, RIKEN Center for Brain Science
Wako, Saitama 351-0198, Japan
takuya.isomura@riken.jp, taro.toyoizumi@riken.jp

Abstract

For many years, a combination of principal component analysis (PCA) and independent component analysis (ICA) has been used as a blind source separation (BSS) technique to separate hidden sources of natural data. However, it is unclear why these linear methods work well because most real-world data involve nonlinear mixtures of sources. We show that a cascade of PCA and ICA can solve this nonlinear BSS problem accurately as the variety of input signals increases. Specifically, we present two theorems that guarantee asymptotically zero-error BSS when sources are mixed by a feedforward network with two processing layers. Our first theorem analytically quantifies the performance of an optimal linear encoder that reconstructs independent sources. Zero-error is asymptotically reached when the number of sources is large and the numbers of inputs and nonlinear bases are large relative to the number of sources. The next question involves finding an optimal linear encoder without observing the underlying sources. Our second theorem guarantees that PCA can reliably extract all the subspace represented by the optimal linear encoder, so that a subsequent application of ICA can separate all sources. Thereby, for almost all nonlinear generative processes with sufficient variety, the cascade of PCA and ICA performs asymptotically zero-error BSS in an unsupervised manner. We analytically and numerically validate the theorems. These results highlight the utility of linear BSS techniques for accurately recovering nonlinearly mixed sources when observations are sufficiently diverse. We also discuss a possible biological BSS implementation.

Introduction

Blind source separation (BSS) is the problem of separating sensory inputs into hidden sources (or causes) without observing the sources or knowing how they have been mixed [1,2]. While numerous BSS algorithms have been developed, the combination of principal component analysis (PCA) and independent component analysis (ICA) is widely used today [3]. PCA [4–9] finds a low-dimensional compressed representation of sensory inputs, i.e., principal components, that best describes the original high-dimensional inputs. These major principal components constitute the essential features to represent the hidden sources. Conversely, ICA [10–14] finds a representation that makes outputs independent of each other. The sequential application of PCA and ICA is commonly used to find an encoder that removes noise and finds a hidden independent representation. Researchers believe that the brain also uses PCA- and ICA-like learning [15], or more generally Bayesian inference, for sensory processing [16–20]. For example, high-dimensional visual inputs are produced by a superposition of signals from objects, and the visual system performs segmentation and dimensionality reduction to perceive the underlying objects [21].

A classical setup for BSS assumes a linear generative process [11], in which sensory inputs are a linear superposition of independent sources. This linear BSS problem has been extensively studied both analytically and numerically [22–24]. In this case, the cascade of PCA and ICA is well known to provide a linear encoder that separates the hidden sources [28–31]. Conversely, a more general BSS setup involves a nonlinear generative process. Here, we study such a nonlinear BSS problem, where the goal is to learn the inverse of the nonlinear generative process and thereby infer the original independent hidden sources solely based on sensory inputs. There are five requirements for solving this problem. The first two involve the representational capacity of the encoder: (1) The encoder’s parameter space must cover a genuine solution that well approximates the inverse
of the generative process. (2) The encoder’s parameter space should not be too large; otherwise,
a nonlinear BSS problem can have infinitely many spurious solutions, at which all outputs are
dependent but dissimilar to the original hidden sources [25, 26]. Hence, it is nontrivial to constrain
the encoder’s representational capacity to satisfy these two opposing requirements. The remaining
three requirements focus on the unsupervised learning algorithm, whose purpose is to find the optimal
encoder’s parameters: (3) the learning dynamics must have a fixed point at the solution that well
approximates the inverse; (4) the fixed point must be linearly stable so that the learning process can
converge to the solution; and (5) the probability of not converging to this fixed point must be small,
i.e., most realistic initial conditions must be within the basin of attraction of this genuine solution.

The cascade of PCA and ICA has been applied to real-world BSS problems that most likely involve
nonlinear generative components [27]; however, there is no guarantee that this linear method will
solve a nonlinear BSS problem. Generally speaking, a linear encoder cannot represent the inverse
of a nonlinear generative process, and this violates Requirement 1 above. A typical approach for
solving a nonlinear BSS problem is to use a nonlinear BSS method [32–33], in which a nonlinear
encoder such as a multilayer neural network [34–36] is trained to invert the generative process. If the
representation capacity of the encoder is large enough (e.g., involving many neurons in the network),
this approach satisfies Requirement 1, and learning algorithms that satisfy Requirements 3 and 4 are
known [17, 37, 38]. However, it is still nontrivial to reliably solve a nonlinear BSS problem. If the
representation capacity of the encoder is too large (i.e., violates Requirement 2), the encoder can
have infinitely many spurious solutions. For a simple two-dimensional toy problem, it is possible to
limit the representation capacity of the nonlinear encoder so that its parameter space only includes
the true solution and no spurious solutions [25]; however, designing a good model representation
is more difficult in more general cases. To our knowledge, there is no theoretical guarantee for
solving a nonlinear BSS problem, except in some low-dimensional examples. Moreover, even when
Requirement 2 is satisfied, there is no guarantee that a learning algorithm converges to the genuine
source representation. It may be trapped in a local minimum, at which outputs are not independent
of each other. Thus, the core of the problem is the lack of knowledge of how to simplify the parameter
space of the inverse model to remove spurious solutions and local minima (to satisfy Requirements 2
and 5) while retaining the representation capacity (Requirement 1).

Here, we give a theoretical guarantee for the use of a linear encoder in solving a nonlinear BSS
problem in cases in which the sensory input is sufficiently high-dimensional and diverse. If the
number of the effecting dimensions of sensory inputs is much greater than the number of sources, even
a linear superposition of sensory inputs can represent all the hidden sources accurately (satisfying
Requirement 1). The use of the linear encoder is beneficial because it can asymptotically find an
optimal linear encoder solely based on sensory data (Requirements 3,4) while avoiding any spurious
solution or local minimum at which the estimated sources are distinct from the original sources
(Requirements 2,5).

In what follows, we estimate the accuracy of and accessibility to a BSS solution by a linear encoder
and find the conditions under which the nonlinear BSS problem is reliably solvable. We propose two
theorems assuming that sensory inputs are generated by a nonlinear generative process composed
of a two-layer network. As the dimensions of sources increase and the number of units in each
layer increases relative to the number of sources, an optimal linear encoder can progressively more
effectively decompose sensory inputs into hidden sources. Our first theorem guarantees that this
reconstruction error reaches zero asymptotically. Our second theorem states that, despite the high-
dimensional sensory inputs, PCA can pick up the relevant subspace spanned by the hidden sources
in a completely unsupervised manner. Thus, by first performing PCA and then performing ICA,
even a linear neural network can reliably find an optimal linear encoder whose reconstruction error
asymptotically reaches zero. Altogether, these results reveal a mathematical condition in which the
linear BSS technique can invert a nonlinear generative process without being trapped in a spurious
solution or local minimum.

Results

Model inversion by linear neural network

First, let us see how an optimal linear encoder can approximately invert a two-layer nonlinear
generative process to separate hidden sources. Suppose \( s \equiv (s_1, \ldots, s_N)^T \in \mathbb{R}^N \) as hidden sources.
The minimization of the above cost function simply gives
This generative process is universal [39, 40] and can represent an arbitrary mapping
E[\text{f}](s) \equiv \prod s_i \in R^{N_f \times N_s} and
B \in R^{N_s \times N_f} as higher- and lower-layer mixing matrices, respectively, a \in R^{N_f} as a constant vector
of offsets, \text{f}(v) : R \rightarrow R as a nonlinear function, and \text{f} \equiv (f_1, \ldots, f_{N_f})^T \equiv f(As + a) as nonlinear bases (see also Fig. 1). Sensory inputs \text{x} \equiv (x_1, \ldots, x_{N_x})^T \in R^{N_x} are defined by
\text{x} \equiv Bf(As + a) \equiv Bf. (1)
This generative process is universal [39, 40] and can represent an arbitrary mapping \text{x} = F(s) as \text{N}_f increases by adjusting the parameters \text{a}, \text{A}, and \text{B}. The model is also universal if each element of
\text{A} and \text{a} are independently generated from a Gaussian distribution N [0, 1/\text{N}_s] [41, 42] as long as
\text{B} is tuned to minimize the cost function E[|BF - F(s)|^2]. Here, E[\text{f}] describes the average over s.
The scaling of \text{A} is to ensure that the argument of \text{f} is order 1. The offset \text{a} \sim 1/\sqrt{\text{N}_s} is introduced for this model to express any generative process \text{F}(s) but it is negligibly small relative to \text{As} for large \text{N}_s. We show that a robust nonlinear BSS is possible if the generative process \text{F}(s) includes sufficient variety of nonlinear mappings from sources to sensory inputs, which we formulate as a mathematical condition later. In the following, we assume \text{N}_f, \text{N}_x \gg \text{N}_s.
Next, we consider a linear encoder (a single-layer linear neural network) defined by
\text{u} \equiv W(x - E[x]), (2)
where \text{u} \equiv (u_1, \ldots, u_{N_s})^T \in R^{N_s} are neural outputs, and \text{W} \in R^{N_s \times N_x} is a synaptic strength matrix.
In the following, we first consider a synaptic strength matrix that minimizes the cost function E[|x - E[x] - W^+ s|^2] and show that the resulting encoder extracts all hidden sources with asymptotically zero error if the inputs have sufficient variability. Note that both \text{x} and \text{s} are required to compute this encoder but \text{s} is unknown in the BSS setup. In the section after next, we show that it is possible to find approximately the same encoder without knowing \text{s}.
The minimization of the above cost function simply gives \text{W}^+ = E[xs^T]. Note that \text{As} is approximately Gaussian distributed for large \text{N}_s by the central limit theorem and, in this case,
E[\text{xs}^T] = BE[<f(As + a)s>] \approx B\text{diag}(E[f_i^2])A \approx \text{T}^TBA \quad (3)
from the Bussgang theorem [43], where \text{diag}(E[f_i^2]) is a diagonal matrix whose i-th diagonal element is \text{E}[f_i^2]. Note that, in case of random \text{A} and small \text{a}, all basis functions \text{f}_i (i = 1, \ldots, \text{N}_f) have asymptotically identical statistics and their derivatives approach a constant \text{f} \equiv \int dv f'(v) \exp(-v^2/2)/\sqrt{2\pi}. The above approximations become exact if \text{As} is Gaussian. Without losing the universality of the model representation, we can assume \text{T} > 0. Hence, we can compute the solution to be
\text{W}^* = \frac{1}{\text{T}}(\text{A}^T\text{B}^TB\text{A})^{-1}\text{A}^TB^T. (4)
Figure 2: Results of numerical experiments. In all panels, hidden sources $s$ are independently generated by an identical uniform distribution with zero mean and unit variance, elements of $A$, $a$ are independently sampled from $\mathcal{N}(0, N_{a}^{-1})$, and elements of $B$ are independently sampled from $\mathcal{N}(0, N_{f}^{-1})$. (A)-(B)(C)(D) One-by-one mapping between hidden sources and neural outputs. A parameter set of $N \equiv (N_{a}, N_{f}, N_{v}) = (10, 10000, 10000)$ and $f(v) = v^3$ is used. When $W = (N_{a}/N_{f}) A^T B^T$, $u_1$ is proportional to $s_1$ (A) while it is independent of $s_2$ (B). This is summarized in a mutual information matrix (C), where only diagonal elements take large values while non-diagonal elements are almost zero. By contrast, when $W$ is a random matrix, mutual information between $u_i$ and $s_j$ is almost zero for all pairs of $i$ and $j$ (D). (E)(F)(G)(H) Dimensional dependency of error. Different parameter sets $N = (2, 10000, 10000)$, $(10, 10000, 10000)$, $(10, 100, 100)$, and $(10, 10000, 1000)$ are used while $f(v) = v^3$ is fixed. (I)(J)(K) Robustness of this theorem to a choice of nonlinear function $f(v)$. Different nonlinear functions $f(v) = v^5$, $\text{sign}(v)$, $\text{ReLU}(v)$ are used while $N = (10, 10000, 10000)$ is fixed. (L) Conversely, when $f(v)$ is uncorrelated with $v$, no linear (signal) component is extracted. When $f(v) = v^2$, only the noise component that is independent of $s_1$ is extracted, as expected by the theorem. The red lines in (A),(B), and (E)-(L) are the lines predicted by the theorem (i.e., $u_1 = \bar{f}s_1$). The green curve in (E) is the curve predicted by the third-order approximation. MATLAB source code is attached as Supplementary Source Code.
This result is intuitively understandable if \( f \) is a linear function, but it also holds with a general \( f \). In the special case that \( B^T B \) is an identity matrix, the above expression simplifies to

\[
W^* \approx \frac{N_s}{f^T N_f} A^T B^T
\]

(5)

because \( (N_s/N_f)(A^T A)_{ij} \) follows \( \mathcal{N}[\delta_{ij}, (1 + \delta_{ij})/N_f] \) and converges in probability to \( \delta_{ij} \) for large \( N_f \).

If we use the encoder in (5), the resulting neural outputs \( u \) asymptotically converge to hidden sources \( s \) as the dimensions of hidden sources \( N_s \) and ratios \( N_f/N_s \) increase. This was confirmed by numerical calculations, where \( u_1 \) is approximately expressed as a function only of \( s_1 \) (Fig. 2A). This also means that \( u_1 \) becomes almost independent from \( s_2 \) (Fig. 2B) and other sources. This means that, while \( u_i \) is generally a function of \( s_1, \ldots, s_{N_s} \), \( u_i \) is a function only of \( s_i \) asymptotically when the encoder \( W^* \) is used.

Asymptotic linearization theorem

In this section, we analytically quantify errors in expressing the hidden sources using the optimal linear encoder. Before introducing a general theorem, we intuitively show that the output vector \( u \) asymptotically converges to the source vector \( s \). In this section, we assume \( N_s \geq N_f \gg N_s \gg 1 \) and \( B^T B \) as the \( N_f \times N_f \) identity matrix. (We relax these assumptions in later sections.) The output here is simply given by \( u = [N_s/(\overline{f} N_f)] A^T f \) from (5). The difference between the outputs \( u \) and hidden sources \( s \) is quantified by

\[
\langle (u - s)(u - s)^T \rangle = \frac{N_s^2}{f^T N_f^2} \langle A^T \text{Cov} [f] A \rangle - I.
\]

(6)

from (3) and again \( (N_s/N_f)(A^T A)_{ij} \) \( \sim \mathcal{N}[\delta_{ij}, (1 + \delta_{ij})/N_f] \) for a random \( A \), where \( \text{Cov}[f] \equiv \text{E}[f f^T] - \text{E}[f] \text{E}[f]^T \) is the covariance. Here, \( \langle \bullet \rangle \) describes an average over \( s, A, \) and \( a \). If two zero-mean and unit-variance Gaussian variables \( u \) and \( w \) have small correlation \( c = \text{E}[uw] \), we find

\[
\text{Cov}[f(v), f(w)] = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)^2}{n!} c^n/n! \text{ with } f^{(n)} \equiv \text{E}[f^{(n)}(v)] \text{ from the Taylor expansion by } c \text{ (see Lemma 1 in Methods). Applying this expansion, we obtain the leading order } \langle A^T \text{Cov} [f] A \rangle \approx \frac{(\overline{f}^2 - \overline{f}'^2)N_f/N_s + \overline{f}^2 N_f^2/N_s^2 + \overline{f}'^2 N_f^2/(2 N_s^2)}{1} \text{I}. \text{ Altogether, we find}
\]

\[
\langle (u - s)(u - s)^T \rangle \approx \frac{N_s}{f^T N_f^2} \left( \frac{\overline{f}^2 - \overline{f}'^2}{2 N_s \overline{f}^2} \right) \text{I}.
\]

(7)

Because the squared error cannot be negative, this equation shows that the output \( u \) converges to the hidden sources \( s \) as \( N_f/N_s \) and \( N_s \) tend to infinity for almost all \( A \) and \( a \).

In the above expression, we focused on characterizing the distribution of \( u - s \) while treating all nonlinear components such as \( s_i^2 \) for \( i = 1, \ldots, N_s \), as the error. Below, we take such nonlinear terms into account and formally quantify the conditional probability distribution of \( u_i \) given \( s_i \), \( A \), and \( a \), namely, \( p(u_i|s_i, A, a) \equiv \int p(u_i|s_i, A, a)p(s_i|s_{i-1}, s_{i+1}, \ldots, s_{N_s}) \text{d}s_{i-1} \text{d}s_{i+1} \ldots \text{d}s_{N_s} \).

Theorem 1 (asymptotic linearization) Suppose \( N_s \geq N_f \gg N_s \), \( A \in \mathbb{R}^{N_f \times N_s} \) as a random mixing matrix and \( a \in \mathbb{R}^{N_f} \) as a random offset vector whose elements independently follow \( \mathcal{N}[0, \frac{1}{N_s}] \), and \( B \in \mathbb{R}^{N_f \times N_f} \) as a matrix that satisfies \( B^T B = I \). When we use the encoder with the synaptic weight matrix in (5), the conditional probability of the \( i \)-th \( (i = 1, \ldots, N_s) \) neural output asymptotically follows Gaussian distribution \( p(u_i|s_i, A, a) = \mathcal{N}[\mu_i, \sigma_i^2] \), where the conditional mean and conditional variance follow

\[
p(\mu_i | s_i, A, a) = \mathcal{N} \left[ \frac{\sqrt{N_s}}{n} \left \langle w f \left( v + \frac{w s_i}{\sqrt{N_s}} \right) \right \rangle_{v,w}, \frac{2 s_i^2}{N_f} \right]
\]

(8)

\[
p(\sigma_i^2 | s_i, A, a) = \mathcal{N} \left[ \frac{N_s}{f^T N_f^2} \left \langle \text{Var}_v \left \langle w f \left( v + \frac{w s_i}{\sqrt{N_s}} \right) \right \rangle \right \rangle_{v,w} + \frac{1}{2 f^T} \left \langle w f'' \left( v + \frac{w s_i}{\sqrt{N_s}} \right) \right \rangle_{v,w}^2, \frac{N_s}{N_f^2} \right]
\]
When we study a general case involving $W$ and $v$, where $v$ and $w$ are unit random Gaussian variables. Further, the covariance between two different neural outputs $u_i$ and $u_j$ ($i \neq j$) is distributed according to $\mathcal{N}[0, N_s/N_f^2]$. Thus, as $N_f$ increases, the conditional means, conditional variances, and covariances quickly converge to values that are not dependent on $A$ and $a$. See the Methods section for the proof of Theorem 1.

In short, we demonstrated that, when the number of sources $N_s$ and the ratio $N_f/N_s$ of the number of basis functions to the number of sources are both large, we can analytically quantify the accuracy of the linear encoder $W^*$ in inverting the nonlinear generative process to separate hidden sources $s$. The accuracy increases as $N_s$ and $N_f/N_s$ increase, and the output $u$ asymptotically converges to $s$ itself. In this manner, the linear encoder expresses the inverse of the nonlinear generative process, and its outputs express the genuine hidden sources.

Theorem 1 is empirically examined by numerical experiments. Examples are illustrated in Fig. 2. When $f(v) = v^3$, $u_1$ approximates $s_1$ but not $s_2$ (Fig. 2A,B) as expected. Indeed, $u_1, \ldots, u_{N_s}$ approximate $s_1, \ldots, s_{N_s}$, respectively. Whereas, when $W$ is a random synaptic matrix with the same variance as $W^*$, $u_1, \ldots, u_{N_s}$ are approximately independent of all sources and their variances are almost zero (Fig. 2D). Theorem 1 predicts the marginal output distribution $p(u|s, A, a)$ for a wide range of $N_s$, $N_f$, and $N_r$ (Fig. 2E-H) for different nonlinear basis functions (Fig. 2I-K). Note that when $f(v)$ is uncorrelated with unit Gaussian variable $v$, only a noise component is extracted; no linear (signal) component is extracted (Fig. 2L). Then, Theorem 1 is quantitatively verified (Fig. 3). The log-log plot illustrates that when $N_s = N_f$ and $B^TB = I$, the mean square error of $u_i$ from $s_i$ reduces inversely proportional to $N_f/N_s$ but saturates around $N_f = N_s^2$ (Fig. 3A). This shows the existence of the order $1/N_s$ error.

In contrast, when $W$ is a random synaptic matrix, the variance of $u_i$ decreases inversely proportional to $N_f/N_s$ regardless of $N_s$ (Fig. 3B). All these results are predicted by Theorem 1.

**Extracting the optimal encoder without knowing the sources**

Theorem 1 states that with the proposed synaptic strength matrix, a linear neural network can accurately extract hidden sources. However, whether the network can find this linear encoder through unsupervised learning is another problem. If the nonlinear generative process $x = F(s)$ is unknown and only $x$ is observed as in the BSS problem, finding $W^*$ may be difficult especially when $N_f$ is large. Indeed, when $W$ is randomly chosen, rate $\rho$ containing components of optimal linear encoders decreases inversely proportional to $N_f/N_s$ (Fig. 3C). Hence, with large $N_f$, only a fraction of $W$ in the entire space is close to $W^*$ by chance. However, we show below that it is possible to extract $W^*$ within the BSS problem setup.

Thus far, we have considered a special case in which $N_s \geq N_f$ and $B^TB$ is an identity matrix. Here, we study a general case involving $N_s < N_f$ and non-identity matrix $B^TB$. Numerical experiments show that the scale of the uncertainty term depends on both $N_s$ and $N_f$ when $B$ is randomly sampled from a Gaussian distribution. The log-log plot illustrates that the mean square error of $u_i$ from $s_i$ is determined by a smaller one between $N_s$ and $N_f$ (Fig. 3D).

To analytically study the general case, we assume that $f(\bullet)$ is an odd nonlinear function. Although it simplifies mathematical expressions, this assumption does not weaken our claims because the presumed generative process in (1) remains universal.

In the special case of $B = I$ (i.e., $x = f$), we also know from Theorem 1 that the best encoder that minimizes the cost function $E[(f - E[f]) - W^*s]^2]$ is $W^* = E[f s^T] \approx T^A$ for large $N_s$. Hence, the residual must be uncorrelated with the hidden sources, i.e., $E[(f - E[f] - T^As)s^T] = O$. Further, using Lemma 1, the covariance of the residual is $\text{Cov}(f - E[f] - T^As) \approx (T^2 - T^2)I$ to the leading order. Hence, the inputs are described by

$$x - E[x] \approx T^BAs + \sqrt{T^2 - T^2}Bz.$$  (9)

Here we denote the residual as $f - E[f] - T^As \equiv \sqrt{T^2 - T^2}z$. The first term represents the signal that linearly encodes the hidden sources, and the second term represents the noise introduced by the nonlinearity. Hence, the hidden sources can be extracted if the first term is sufficiently larger than the second term.
Applying this coordinate transformation, the inputs are described by

\[ \text{Relationship between } R. \]

The orthogonality condition

\[ \text{To clearly determine which factor in the above equation dominates, let us consider a coordinate} \]

\[ \text{notation, we obtain} \]

\[ \text{variance of} \]

\[ \text{parallel to the matrix} \]

\[ \text{as in Fig. 2. (A) The representation error for an optimal linear encoder. We suppose} \]

\[ \text{fix } B^T B \text{ as the } N_f \times N_f \text{ identity matrix. A synaptic strength matrix is assumed to be} \]

\[ \text{The representation error is defined by the mean square error} \]

\[ \text{show simulation results for } N_s = 10, 100, \text{ and } 1000, \text{ respectively. Dashed gray lines represent} \]

\[ \text{entire behaviors are well matched with curves predicted by the theorem. (C) The rate of a random} \]

\[ \text{synaptic weight matrix is independently generated by} \]

\[ \text{variance of outputs} \]

\[ \text{Error in (A); as the nonlinear basis dimension increases, the error monotonically decreases. These} \]

\[ \text{generated by} \]

\[ \text{is larger than} \]

\[ \text{of outputs from a random encoder. Each element of the synaptic weight matrix is independently} \]

\[ \text{fix} \]

\[ \text{distribution with zero mean and unit variance, and} \]

\[ \text{as in Fig. 2.} \]

\[ \text{Figure 3: Relationship between accuracy and the dimensions of hidden sources, nonlinear bases, and} \]

\[ \text{entire behaviors are well matched with curves predicted by the theorem. (C) The rate of a random} \]

\[ \text{synaptic weight matrix is independently generated by} \]

\[ \text{variance of} \]

\[ \text{as the} \]

\[ \text{as the} \]

\[ \text{false as predicted by the theorem.} \]

To clearly determine which factor in the above equation dominates, let us consider a coordinate transformation. We introduce a rotation matrix

\[ R \equiv (R_{\parallel}, R_{\perp}) \in \mathbb{R}^{N_f \times N_f}, \]

with block dimensions

\[ R_{\parallel} \in \mathbb{R}^{N_f \times N_f} \] and \[ R_{\perp} \in \mathbb{R}^{N_f \times (N_f - N_s)} \], to decompose \( B \) into the two orthogonal subspaces—one parallel to the matrix \( A \), i.e., \[ B_{\parallel} \equiv BR_{\parallel} \] and the other perpendicular to it, i.e., \[ B_{\perp} \equiv BR_{\perp}. \]

The orthogonality condition \( R^T R = I \) imposes that \( R_{\parallel}^T R_{\parallel} \) and \( R_{\perp}^T R_{\perp} \) are identity matrices with the corresponding dimensions; thus, \( R_{\parallel}^T R_{\perp} = O \) and \( R_{\parallel} = A(A^T A)^{-1/2} \) are satisfied. Using this notation, we obtain

\[ B = B_{\parallel} R_{\parallel}^T + B_{\perp} R_{\perp}^T. \]

We assume that the generative process includes a sufficient variety of nonlinearity in the sense that all singular values of \( B \) are of the same order of magnitude. Applying this coordinate transformation, the inputs are described by

\[ x - E[x] \approx T B_{\parallel} (R_{\parallel}^T A)s + \sqrt{T^2 - T_{\parallel}^2 (B_{\parallel} R_{\parallel}^T + B_{\perp} R_{\perp}^T)} z. \] (10)
There is an important $N_f/N_s$ dependency in the above equation, namely, $\sqrt{N_s/N_f} B^T_B A \rightarrow I$ in the limit of $N_f/N_s \rightarrow \infty$. Altogether, we find that the covariance of the inputs is
\[
\text{Cov}[x] \approx \frac{N_f}{N_s} T^2 B^T_B + \left( T^2 - T^2 \right) (B^T_B + B^T_B^T).
\]  
(11)

Because $B_B$ and $B_B^T$ have the same order of magnitude, the first signal term (signal covariance) asymptotically dominates the major principal components for large $N_f/N_s$. Hence, all the $N_s$ directions of $B_B$ must be extracted from Cov[\(x\)] as the major components. Specifically, by applying PCA to Cov[x], the signal term is described by the first to the $N_s$ eigenmodes:

**Theorem 2 (eigenvalue decomposition)** We sort the real eigenvalues of Cov[x] in descending order and express the first to the $N_s$ major eigenvalues by $N_s \times N_s$ diagonal matrix $\Lambda$, and the corresponding $N_s \times N_s$ eigenvector matrix $P$. From the above argument, we asymptotically obtain
\[
PA^TP \approx \frac{N_f}{N_s} T^2 B^T_B
\]  
for large $N_f/N_s$. See the Methods section for a more quantitative condition for finite $N_f/N_s$.

This expression directly provides a key factor to express the optimal encoder, i.e., $T^T B A = \sqrt{N_f/N_s} T^T B_B \approx Q P^{1/2}$, where $Q$ is an arbitrary $N_s \times N_s$ rotation matrix. Indeed, the synaptic strength matrix of the optimal encoder (9) is summarized by
\[
Q W = Q Q^T A^{1/2} P^T P^{1/2} Q |^{-1} Q^T A^{1/2} P^T = \Lambda^{-1/2} P^T
\]  
(13)

This shows that we can compute the optimal encoding weight up to an arbitrary rotation factor $Q$ from the major eigenvalues and eigenvectors of Cov[x], which are available under the BSS setting. In the limit of large $N_s$ and $N_f/N_s$, the outputs of the encoder asymptotically converge to $u \rightarrow Q s$. This means that the outputs are not independent of each other because of the multiplication with $Q$. However, what is remarkable about this approach is that it converts the original nonlinear BSS problem to a linear BSS problem. Namely, $u$ is now a linear mixture of the hidden sources, such that we can extract all the hidden sources by further applying a linear ICA method to the outputs $u$.

As shown in Fig. 4, it is numerically validated that the first to $N_s$-th major principal components of Cov[x] well approximate the signal covariance (Fig. 4A), i.e., $(N_f/N_s) T^T B_B$. Moreover, as $N_s$ increases, the subspace of the major principal components asymptotically matches to the subspace of the signal components (Fig. 4B). This indicates that PCA is a promising method for finding an optimal linear encoder up to a linear random rotation of sources.

**Hebbian-like learning rules can find an asymptotically optimal linear encoder through a cascade of PCA and ICA**

In the previous section, we showed that PCA can reliably extract a subspace spanned by components of the optimal linear encoder as the first to the $N_s$-th major principal components. As the dimensions of sensory inputs and nonlinear bases become large, PCA can more accurately extract the sources. We next explore if a more biologically plausible learning rule for the encoder can also extract the subspace spanned by the optimal linear encoder.

Oja’s subspace rule [6], a type of modified Hebbian plasticity (see also Methods for its update rule), is known to perform PCA and extract a subspace of the first to $N_s$-th major principal components without being trapped by a spurious solution or local minimum [28,29]. Hence, as a corollary of Theorem 2, starting from almost all random initial conditions, Hebbian-like learning rules can reliably find an optimal linear encoder in an unsupervised manner:

Suppose a synaptic strength matrix is given by $W = CPT + CmP^T_m$ with coefficient matrices $C \in \mathbb{R}^{N_s \times N_s}$ and $C_m \in \mathbb{R}^{N_s \times (N_s-N_s)}$. As described above, $P$ is the major eigenvector matrix of Cov[x], while $P_m$ is the remaining minor eigenvector matrix perpendicular to $P$. Together, $(P,P_m) \in \mathbb{R}^{N_s \times N_s}$ is a rotation matrix. From (13), the synaptic strength matrix of the optimal encoder is expressed by the $CP^T$ term. The so-called Hebbian factor is expressed as the product of the outputs and inputs:
\[
E \left[u(x - E[x])^T\right] = WCov[x]
= (CP^T + (C_mA_m)P^T_m)
\]  
(14)
where we write the eigenvalue decomposition of the input covariance as $W = \sum \Lambda_m v_m v_m^T$, and the major principal components of Cov$[x]$, i.e., $PAP^T$, when $N_s = 10$, $N_f = N_x = 10^3$, and $f(v) = \text{sign}(v)$. This comparison empirically verifies Theorem 2. (B) A rate containing components of optimal linear encoders when $N_e = N_f$ and $f(v) = v^3$. PCA achieved by eigenvalue decomposition could reliably find a subspace spanned by components of optimal linear encoders. Simulations were conducted 10 times with different $A, B, a$ for each parameter set. Solid curves represent the means and shaded areas represent areas between maximum and minimum values in 10 simulations.

Figure 4: PCA can find components of optimal linear encoders. Hidden sources $s$ are independently generated by an identical uniform distribution with zero mean and unit variance, and $A, B, a$ are sampled from Gaussian distributions as in Fig. 2. (A) Comparison between the signal covariance, i.e., $(N_f/N_s) v_m^T B_y B_y^T$, and the major principal components of $\text{Cov}[x]$, i.e., $PAP^T$, when $N_s = 10$, $N_f = N_x = 10^3$, and $f(v) = \text{sign}(v)$. This comparison empirically verifies Theorem 2. (B) A rate containing components of optimal linear encoders when $N_e = N_f$ and $f(v) = v^3$. PCA achieved by eigenvalue decomposition could reliably find a subspace spanned by components of optimal linear encoders. Simulations were conducted 10 times with different $A, B, a$ for each parameter set. Solid curves represent the means and shaded areas represent areas between maximum and minimum values in 10 simulations.

Numerical experiments show that the accuracy of Hebbian product in extracting components of $W^*$ increased as the time steps for the Hebbian update increased, and that it saturated at a containing rate of around 95% when $N_s \geq 100$ and $N_f \geq 10^4$ while $W$ was randomly initialized (Fig. 5A,B). The synaptic strength matrix updated by Oja’s subspace rule for PCA [6] also converged to $W^*$ up to the multiplication of the random matrix from the left. We found that Oja’s subspace rule could extract all components of the optimal linear encoders as the first to the $N_s$-th major principal components. The transition of $W$ is shown in Fig. 5C,D. While $W$ started from random initial states, $W$ reliably converged to a matrix that contains approximately 95% $W^*$ components. These results indicate that Hebbian plasticity can reliably and accurately find all components of the optimal linear encoders.

As we have shown, PCA as well as Hebbian plasticity successfully extract the linear components of the hidden sources covered in the nonlinear components. Thereby, the original nonlinear BSS problem has now become a simple linear BSS problem. Thus, the linear ICA (e.g., Amari’s ICA rule [13]; see also Methods for its update rule) can reliably separate all hidden sources from the features extracted by Hebbian plasticity. We numerically confirmed that this is the case. Crucially, those independent components match to the true sources of the nonlinear generative process up to the permutation and sign-flips. We quantify the BSS error, i.e., the difference between those independent components and the genuine hidden sources, by asking how much the mapping from sources to outputs is different from the identity mapping. We found that a nearly zero BSS error is achieved
Figure 5: Hebbian-like learning rules for finding an optimal linear encoder. In all panels, hidden sources $s$ are independently generated by an identical uniform distribution with zero mean and unit variance, and $A, B, a$ are sampled from Gaussian distributions as in Fig. 2. The nonlinearity $f(v) = v^3$ is supposed. Except in (E), $B$ is fixed as the identical matrix. (A)(B) Transitions in rates indicating whether a Hebbian product contains a matrix component parallel to an optimal linear encoder ($\rho$) with $N_s = 10$ (A) or 100 (B) sources. Black, red, green, and blue curves are simulation results with $N_f = N_s, 10N_s, 10^2N_s,$ and $10^3N_s$, respectively. Even at the first step, $\rho$ is around 0.6 and increases as time $t$ increases. (C)(D) The learning process of synaptic strengths. Shown are transitions in rates indicating whether a synaptic matrix contains a matrix component parallel to an optimal linear encoder ($\rho$) with $N_s = 10$ (C) or 100 (D) sources. Synaptic strengths are updated by Oja’s subspace rule for PCA [6]. Black, red, and green curves are simulation results with $N_f = N_s, 10N_s,$ and $10^2N_s$, respectively. While $\rho$ is close to zero at the beginning, it increases as $t$ increases. (E) The PCA-ICA cascade can reliably find an optimal linear encoder. We applied ICA to extracted major principal components (results shown in Fig. 4B) and could reliably find the true sources. The case of $N_s = 10$ was investigated. BSS error was defined as the ratio of first to second maximum absolute values for every row and column of matrix $K = W_{ICA}WA$. A learning rate of $\eta = 2 \times 10^{-5}$ was used, and $W_{ICA}$ was started from an identity matrix. In those panels, simulations were conducted 10 times with different mixing parameters for each dimension. Curves represent the means, and shaded areas represent areas between maximum and minimum values in the 10 simulations.

when $N_x = N_f = 10^3$ and $N_s$ while the error remains when $N_x/N_s \leq 10$ (Fig. 5E). These results highlight that the cascade of PCA and ICA can find an optimal linear encoder with high accuracy when $N_x, N_f \gg N_s \gg 1$.

Discussion

In this study, we theoretically quantified the accuracy of a model inversion performed using an optimal linear encoder when the sensory inputs are generated from a two-layer nonlinear generative process. First, we introduced the asymptotic linearization theorem, which states that as the dimension of hidden sources increases and the dimensions of sensory inputs and nonlinear bases increase relative to the source dimension, an optimal linear encoder can accurately separate sensory inputs into genuine hidden sources (Theorem 1). By applying the optimal linear encoder to sensory inputs, the linearly encodable component of hidden sources is magnified relative to the nonlinear component in proportion to the ratio of the numbers of bases and outputs to the number of sources;
Amari’s ICA rule in this paper, more biologically plausible local Hebbian learning rules are proposed which successfully satisfies Requirements 1-5 mentioned in the introduction. (This property is related to the fact that there are generally infinitely many spurious solutions for encoder, which might be related to the large numbers of sensory cells in humans; e.g., about 100 organisms might have developed high-dimensional sensors to perform nonlinear BSS with a linear inferring hidden sources of a nonlinear generative process. An interesting possibility is that living theorems provide a license to apply standard linear PCA and ICA to natural data for the purpose of nonlinear BSS. The proposed Because most natural data (including biological, chemical, and social data) are generated from nonlinear generative processes, broad applications of nonlinear BSS are considered. The proposed theorems provide a license to apply standard linear PCA and ICA to natural data for the purpose of inferring hidden sources of a nonlinear generative process. An interesting possibility is that living organisms might have developed high-dimensional sensors to perform nonlinear BSS with a linear encoder, which might be related to the large numbers of sensory cells in humans; e.g., about 100.
Applying this expansion, we obtain to the leading order

\[
\mathbf{E} \{ g \mathbf{v} \} = \mathbf{E} \{ v \mathbf{g} \} = \mathbf{E} \{ \mathbf{v} g \} = \mathbf{E} \{ g \mathbf{v} \},
\]

where \( \mathbf{E} \{ \cdot \} \) denotes the expectation of \( \mathbf{X} \) and the above equation for increasing in dimension and variety, sensory inputs can provide greater evidence about the hidden sources, which remove the possibility of finding spurious solutions. The result guarantees that the PCA-ICA cascade can reliably find genuine hidden sources from high-dimensional nonlinear source mixtures.

**Methods**

**Proof of Lemma 1**

Suppose \( v \) and \( w \) are zero-mean and unit-variance Gaussian variables and \( f(v) \) and \( g(w) \) are arbitrary functions. When \( v \) and \( w \) have small correlation \( \mathbf{E} [vw] \), \( w \) satisfies \( w = cv + \sqrt{1-c^2} \xi \), where \( \xi \) is a zero-mean and unit-variance Gaussian variable that is independent of \( v \). When we define \( \phi(c) \equiv \mathbf{Cov}[f(v), g(w)] \) as the covariance between \( f(v) \) and \( g(w) \), its derivative with respect to \( c \) is given by

\[
\phi'(c) = \mathbf{Cov} \left[ f(v), g'(c v + \sqrt{1-c^2} \xi) \left( v - \frac{c}{\sqrt{1-c^2}} \xi \right) \right] = \mathbf{E} \left[ (f(v) - \mathbf{E}[f(v)]) g'(c v + \sqrt{1-c^2} \xi) \left( v - \frac{c}{\sqrt{1-c^2}} \xi \right) \right].
\]

(15)

From the Bussgang theorem [43],

\[
\mathbf{E} \left[ (f(v) - \mathbf{E}[f(v)]) g'(c v + \sqrt{1-c^2} \xi) \right] = \mathbf{E} \left[ f'(v) g'(c v + \sqrt{1-c^2} \xi) + (f(v) - \mathbf{E}[f(v)]) g''(c v + \sqrt{1-c^2} \xi) c \right]
\]

(16)

and

\[
\mathbf{E} \left[ (f(v) - \mathbf{E}[f(v)]) g'(c v + \sqrt{1-c^2} \xi) \right] = \mathbf{E} \left[ f''(v) g'(c v + \sqrt{1-c^2} \xi) \sqrt{1-c^2} \right].
\]

(17)

Thus, \( \phi'(c) \) becomes

\[
\phi'(c) = \mathbf{E} \left[ f'(v) g'(c v + \sqrt{1-c^2} \xi) \right]
\]

(18)

Hence, we find

\[
\phi^{(n)}(c) = \mathbf{E} \left[ f^{(n)}(v) g^{(n)}(c v + \sqrt{1-c^2} \xi) \right].
\]

(19)

From the Taylor expansion by \( c \),

\[
\phi(c) = \sum_{n=1}^{\infty} \phi^{(n)}(0) \frac{c^n}{n!} = \sum_{n=1}^{\infty} \frac{f^{(n)} \mathbf{g}^{(n)}}{n!},
\]

(20)

where \( \overline{f^{(n)}} \equiv \mathbf{E}[f^{(n)}(v)] = \int dv f^{(n)}(v) \exp(-v^2/2)/\sqrt{2\pi} \) indicates the expectation of \( f^{(n)}(v) \) over \( v \).

Applying this expansion, we obtain to the leading order

\[
\langle \mathbf{A}^T \mathbf{Cov}[\mathbf{f}] \mathbf{A} \rangle \approx \mathbf{A}^T \mathbf{Cov}[\mathbf{f}] \mathbf{A} + \sum_{n=1}^{\infty} \frac{f^{(n)}_2}{n!} \left\{ \mathbf{1} - (\mathbf{A}^T)^{\otimes n} - \mathbf{Cov}[\mathbf{f}] \right\} \mathbf{I}
\]

\[
\approx \left( \frac{f_2^2}{N_s} + \frac{f_2^2 N_s^2}{N_f^2} + \frac{f_2^2}{2N_s^2} \right) \mathbf{I},
\]

(21)

where \( \{X^{\otimes n}\} \equiv X_1^n \) describes the element-wise power. Note that the diagonal components of \( \mathbf{Cov}[\mathbf{f}] \) are evaluated separately from the non-diagonal components in the above equation.
Proof of Theorem 1

Here we suppose \(B^TB\) is the identity matrix. Let us separately consider the \(A_{ji} s_i\) term in the argument of \(f_j\) from other terms. We define a new random variable \(y_{ji} \equiv \sum_{k \neq i} A_{jk} s_k + a_j\). The expectation of \(y_{ji}\), over \(A_{j1}, \ldots, A_{jN_s}\) and \(a_j\) follows a Gaussian distribution \(\mathcal{N}[0, \sum_{k \neq i} A_{jk}^2 + a_j^2] \approx \mathcal{N}[0, 1 + \sqrt{2/N_s} \gamma_{ij}]\), where a fluctuation \(\gamma_{ij}\) follows the unit Gaussian distribution \(\mathcal{N}[0, 1]\). When the synaptic strength matrix is optimal (see (5)), from a Taylor expansion of \(f\), we obtain

\[
\begin{align*}
    u_i &= N_s \frac{N_f}{N_f^T} \sum_{j=1}^{N_f} A_{ji} \left\{ f \left( y_{ji} + A_{ji} s_i \right) - E[f(u)] \right\} \\
    &= N_s \frac{N_f}{N_f^T} \left\{ \sum_{n=0}^{\infty} \frac{s_i^n}{n!} \sum_{j=1}^{N_f} f^{(n)}(y_{ji}) A_{ji}^{n+1} - \sum_{j=1}^{N_f} A_{ji} \right\}.
\end{align*}
\]

(22)

The probability distribution of the mean Because \(f^{(n)}(y_{ji})\) and \(A_{ji}^{n+1}\) are independent of each other, the conditional expectation of the mean of \(u_i\) under fixed \(s_i\) and \(A_{j1}, \ldots, A_{jN_s}\) is given by

\[
\begin{align*}
    \mu_i(s_i) &= E[u_i|s_i, A_{j1}, \ldots, A_{jN_s}] \\
    &= N_s \frac{N_f}{N_f^T} \left\{ \sum_{n=0}^{\infty} \frac{s_i^n}{n!} \sum_{j=1}^{N_f} E\left[f^{(n)}(y_{ji})\right] A_{ji}^{n+1} - \sum_{j=1}^{N_f} A_{ji} \right\} \approx N_s \frac{N_f}{N_f^T} \sum_{n=1}^{\infty} \frac{f^{(n)}(s_i^n)}{n!} \sum_{j=1}^{N_f} A_{ji}^{n+1}.
\end{align*}
\]

(23)

The expectation of \(A_{ji}^n\) is given by \(E[A_{ji}^n] = \langle w^n \rangle / \sqrt{N_s^n}\). Hence, the mean of \(\mu_i(s_i)\) is

\[
E[\mu_i|s_i] = N_s \frac{N_f}{N_f^T} \sum_{n=1}^{\infty} \frac{f^{(n)}(s_i^n)}{n!} \frac{N_f \langle w^{n+1} \rangle}{\sqrt{N_s^{n+1}}} = \frac{\sqrt{N_s}}{f} \left\langle w f \left( v + \frac{w s_i}{\sqrt{N_s}} \right) \right\rangle
\]

(24)

and the second-order moment of \(\mu_i(s_i)\) is

\[
E[\mu_i^2|s_i] \approx \left( N_s \frac{N_f}{N_f^T} \right)^2 \left\langle \left( \sum_{j=1}^{N_f} \bar{f}^j s_i A_{ji}^2 \right)^2 \right\rangle = \left( N_s \frac{N_f}{N_f^T} \right)^2 \left\langle \sum_{j=1}^{N_f} \sum_{k=1}^{N_f} \bar{f}^j s_i^2 A_{ji}^2 A_{ki}^2 \right\rangle
\]

\[
= \left( N_s \frac{N_f}{N_f^T} \right)^2 \bar{f}^2 s_i^2 \frac{3N_f + N_f(N_f - 1)}{N_s^2} = \frac{(2 + N_f)s_i^2}{N_f},
\]

(25)

where \(w\) is a new unit random Gaussian variable. Thus, from \(\text{Var}[\mu_i|s_i] = E[\mu_i^2|s_i] - E[\mu_i|s_i]^2\), we find

\[
p(\mu_i|s_i) = \mathcal{N} \left[ \frac{\sqrt{N_s}}{f} \left\langle w f \left( v + \frac{w s_i}{\sqrt{N_s}} \right) \right\rangle, \frac{2s_i^2}{N_f} \right].
\]

(26)

When \(N_f\) is large, the variance of \(\mu_i\) can be ignored.
The probability distribution of the variance  

The conditional expectation of the variance of $u_i$ under fixed $s_i$ and $A_{j1}, \ldots, A_{jN_s}$ is given by

$$
\sigma_i^2(s_i) = \text{Var}_{y_{j1}, \ldots, y_{jN_f}}[u_i|s_i, A_{j1}, \ldots, A_{jN_s}] = E \left[ u_i^2 | s_i, A_{j1}, \ldots, A_{jN_s} \right] - E \left[ u_i | s_i, A_{j1}, \ldots, A_{jN_s} \right]^2
$$

$$
= E \left[ \left( \frac{N_s}{N_f} \right)^2 \left( \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \frac{s_i^n s_i^m}{n! m!} \left( f^{(n)}(y_{j1}) A_{j1}^{n+1} - \bar{f} \sum_{j=1}^{N_f} A_{ji} \right) \right)^2 \right] - \left( \frac{N_s}{N_f} \right)^2 \left( \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \frac{s_i^n s_i^m}{n! m!} \left\{ f^{(n)}(y_{j1}) \right\} A_{j1}^{n+1} \right)^2
$$

$$
= \left( \frac{N_s}{N_f} \right)^2 \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{s_i^n s_i^m}{n! m!} \left( \sum_{j=1}^{N_f} \left( f^{(n)}(y_{j1}) f^{(m)}(y_{j1}) \right) - \left\{ f^{(m)}(y_{j1}) \right\} \left\{ f^{(n)}(y_{j1}) \right\} \right) A_{j1}^{m+1} A_{j1}^{n+1}
$$

$$
= \left( \frac{N_s}{N_f} \right)^2 \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{s_i^n s_i^m}{n! m!} \left( \sum_{j=1}^{N_f} \text{cov} \left( f^{(m)}(y_{j1}), f^{(n)}(y_{j1}) \right) \right) A_{j1}^{m+n+2} + \sum_{j=1}^{N_f} \sum_{k \neq j} \text{cov} \left( f^{(m)}(y_{j1}), f^{(n)}(y_{j1}) \right) A_{j1}^{m+n+1} A_{j1}^{n+1}
$$

In the last line, we used Lemma 1, namely

$$
\sum_{m=1}^{\infty} \left( f^{(m)}(y_{j1}) \right) \left( f^{(m)}(y_{j1}) \right) = \sum_{m=1}^{\infty} \left( f^{(m)}(y_{j1}) \right) \left( f^{(m)}(y_{j1}) \right) = \sum_{m=1}^{\infty} \left( f^{(m)}(y_{j1}) \right) \left( f^{(m)}(y_{j1}) \right) = \frac{1}{2} \frac{(m+2)}{f^{(m+2)}} \left( f^{(m+2)}(y_{j1}) \right) \left( f^{(m+2)}(y_{j1}) \right) + \frac{1}{2} \frac{(m+3)}{f^{(m+3)}} \left( f^{(m+3)}(y_{j1}) \right) \left( f^{(m+3)}(y_{j1}) \right).
$$

Thus, we have

$$
E[\sigma_i^2(s_i)] = \left( \frac{N_s}{N_f} \right)^2 \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{s_i^n s_i^m}{n! m!} \left( \frac{f^{(m)}(y_{j1}) - f^{(m)}(y_{j1})}{\sqrt{N_s^{m+n+2}}} \right)^2 + \frac{1}{2} \frac{f^{(m+2)}}{f^{(m+2)}} \left( f^{(m+2)}(y_{j1}) \right) \left( f^{(m+2)}(y_{j1}) \right) + \frac{1}{2} \frac{f^{(m+3)}}{f^{(m+3)}} \left( f^{(m+3)}(y_{j1}) \right) \left( f^{(m+3)}(y_{j1}) \right)
$$

$$
= \left( \frac{N_s}{N_f} \right)^2 \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{s_i^n s_i^m}{n! m!} \left( \frac{f^{(m)}(y_{j1}) - f^{(m)}(y_{j1})}{\sqrt{N_s^{m+n+2}}} \right)^2 + \frac{1}{2} \frac{f^{(m+2)}}{f^{(m+2)}} \left( f^{(m+2)}(y_{j1}) \right) \left( f^{(m+2)}(y_{j1}) \right) + \frac{1}{2} \frac{f^{(m+3)}}{f^{(m+3)}} \left( f^{(m+3)}(y_{j1}) \right) \left( f^{(m+3)}(y_{j1}) \right)
$$

and

$$
E[(\sigma_i^2)^2(s_i)] = \left( \frac{N_s}{N_f} \right)^4 \left( f^{(2)} - f^{(2)} \right) \sum_{j=1}^{N_f} A_{j1}^2 + \frac{N_f}{N_s} \sum_{j=1}^{N_f} \sum_{k \neq j} A_{j1} A_{k1} + A_{j1} A_{k1} \right)^2
$$

$$
= \left( \frac{N_s}{N_f} \right)^4 \left( f^{(2)} - f^{(2)} \right)^2 \sum_{j=1}^{N_f} A_{j1}^2 + \frac{N_f}{N_s} \sum_{j=1}^{N_f} \sum_{k \neq j} A_{j1} A_{k1} + A_{j1} A_{k1} \right)^2 A_{j1}^2 A_{k1}^2
$$

$$
= \left( \frac{N_s}{N_f} \right)^4 \left( 3N_f + N_f (N_f - 1) \frac{1}{N_s} \right) \left( f^{(2)} - f^{(2)} \right)^2 + \frac{N_f (N_f - 1)}{N_s^3} \frac{1}{N_s}
$$

$$
= \left( \frac{N_s}{N_f} \right)^2 \left( 2 + N_f \frac{1}{N_f} \right) \left( f^{(2)} - f^{(2)} \right)^2 + \frac{1}{N_s} \approx \left( \frac{N_s}{N_f} \right)^2 \left( \frac{f^{(2)} - f^{(2)}}{f^{(4)}} \right)^2 + \frac{1}{N_s}
$$

(29)
Hence, from \( \text{Var}[\sigma_i^2|A,a] = E[(\sigma_i^2)^2|A,a] - E[\sigma_i^2|A,a]^2 \), we find

\[
p(\sigma_i^2|s_i) = \mathcal{N}
\left[
\frac{N_s}{N_f T} \left( \frac{1}{2} \int \text{Var}_v \left( w f \left( v + \frac{w s_i}{\sqrt{N_s}} \right) \right) + \frac{1}{2} f''^2 \left( v + \frac{w s_i}{\sqrt{N_s}} \right)^2 \right) \frac{N_s}{N_f}
\right].
\]

(30)

The probability distribution of the covariance

The covariance between two different outputs is

\[
\text{cov}(u_i, u_{i'}) = \left( \frac{N_s}{N_f T} \right)^2 \sum_{m=0}^\infty \sum_{n=0}^\infty \sum_{j=1}^{N_f} \sum_{k=1}^{N_f} \text{cov} \left( f^{(m)}(y_{ji}), f^{(n)}(y_{ki'}) \right) A_{ji}^{m+1} A_{ki'}^{n+1}
\]

(31)

Thus, the mean is \( E[\text{cov}(u_i, u_{i'})] = 0 \) and the variance is

\[
E[\text{cov}(u_i, u_{i'})^2] = \left( \frac{N_s}{N_f T} \right)^4 \left\{ \left( \sum_{j=1}^{N_f} \sum_{k=1}^{N_f} \left( \sum_{l \neq i, i'} A_{jl} A_{kl} + a_j a_k \right) \right)^2 A_{ji} A_{ki'} \right\}^2
\]

(32)

Hence, we find

\[
p(\text{cov}(u_i, u_{i'})) = \mathcal{N} \left[ 0, \frac{N_s^2}{N_f^2} \right].
\]

(33)

Proof of Theorem 2

Without loss of generality, any \( B \) matrix is decomposed into three matrix components:

\[
B = (U, U_L) \begin{pmatrix}
S & C \\
O & S_L
\end{pmatrix}
\begin{pmatrix}
V^T & O \\
O & V_L^T
\end{pmatrix}
\begin{pmatrix}
R_{\parallel}^T \\
R_{\perp}^T
\end{pmatrix} = \text{USV}^T R_{\parallel}^T + \text{UCV}_L^T R_{\perp}^T + U_L S_{\perp} V_L^T R_{\perp}^T.
\]

(34)

The first term \( \text{USV}^T R_{\parallel}^T \equiv B_{\parallel} R_{\parallel}^T \) expresses a subspace right-perpendicular to the matrix \( A \), where \( R_{\parallel} \equiv A(A^T A)^{-1/2} \), and \( \text{USV}^T \) is the singular value decomposition of \( B_{\parallel} \). The second and third terms \( (UC + U_L S_{\perp}) V_L^T R_{\perp}^T \equiv B_{\perp} R_{\perp}^T \) express a subspace right-orthogonal to \( A \), where \( U_L S_{\perp} V_L^T \) is a component left-parallel to \( B_{\parallel} \), and \( U_L S_{\perp} V_L^T \) is left-orthogonal to it. Here \( C \in \mathbb{R}^{N_f \times (N_f - N_s)} \) is a rectangular matrix and \( U_L S_{\perp} V_L^T \) is the singular value decomposition. Note that \( (R_{\parallel}, R_{\perp}) \in \mathbb{R}^{N_f \times N_f} \) and \( (U, U_L) \in \mathbb{R}^{N_f \times N_r} \) are rotation matrices. In short, \( \text{USV}^T R_{\parallel}^T \) is the signal (linear) component, \( \text{UCV}_L^T R_{\perp}^T \) is the noise (nonlinear) component left-parallel to the signal component, and \( U_L S_{\perp} V_L^T R_{\perp}^T \) is the noise component left-orthogonal to the signal component.

From Theorem 1 and Lemma 1, the covariance of \( (R_{\parallel}, R_{\perp})^T f \) is

\[
\text{Cov}[(R_{\parallel}, R_{\perp})^T f] = \begin{pmatrix}
R_{\parallel}^T \\
R_{\perp}^T
\end{pmatrix}
\begin{pmatrix}
E[f f^T] - E[f] E[f]^T \\
E[f] E[f]^T
\end{pmatrix}
\begin{pmatrix}
R_{\parallel} \\
R_{\perp}
\end{pmatrix}
\approx \begin{pmatrix}
\mathbf{\alpha} I \\
O
\end{pmatrix}
\begin{pmatrix}
O \\
I
\end{pmatrix},
\]

(35)

15
where \( \alpha \equiv (N_f/N_s)T^2 / (T^2 - T^2) \). Thus, we find
\[
\text{Cov}[x] = \left( T^2 - T^2 \right) \left( U, U_\perp \right) \left( \begin{array}{cc}
\alpha S^2 + CC^T & CS_\perp \\
CS_\perp & S_\perp^2
\end{array} \right) \left( \begin{array}{c}
U^T \\
U_\perp^T
\end{array} \right).
\]
(36)

This \( \alpha \) takes a large value if \( N_f \) is much greater than \( N_s \). In this case, even a small rotation of \( \text{Cov}[x] \) by a rotation matrix \((I, E; -E^T, I)\) with \( E \in \mathbb{R}^{N_s \times (N_s - N)} \) would diagonalize it. Hence, to make
\[
\begin{pmatrix}
I & E \\
-E^T & I
\end{pmatrix}
\begin{pmatrix}
\alpha S^2 + CC^T & CS_\perp \\
CS_\perp & S_\perp^2
\end{pmatrix}
\begin{pmatrix}
I \\
E^T \\
-I
\end{pmatrix}
\approx
\begin{pmatrix}
\alpha S^2 + CC^T + ES_\perp^2 C^T + CS_\perp E^T \\
-ES_\perp^2 C^T + CS_\perp E^T
\end{pmatrix}
\begin{pmatrix}
\alpha S^2 + CC^T - ES_\perp^2 C^T + CS_\perp E^T \\
ES_\perp^2 C^T - CS_\perp E^T
\end{pmatrix}
\]
(37)
a diagonal matrix. \(-\alpha S^2 + CC^T\) should be zero. From \((\alpha S^2 + CC^T)E - ES_\perp^2 = CS_\perp\), \(\text{eig}[E]\) is no more than \(|S_{\max}/(\alpha S_{\min}^2 - S_{\max}^2)|\text{eig}[C]\), where \(S_{\min}\) is the smallest singular value of \(S\) and \(S_{\max}\) is the largest singular value of \(S_\perp\). Because this \(E\) is much smaller than the order 1 value as long as \(S_{\max}^2\) is much smaller than \(\alpha S_{\min}^2\) and eigenvalues of \(CC^T\) are in a smaller order than \(\alpha S_{\min}^2\), we approximate \(E \approx (S_{\max}^2/(\alpha S_{\min}^2 - S_{\max}^2))C\).

Using this \(E\), we find the following inequality on the eigenvalues of the major components
\[
\Lambda = \left( T^2 - T^2 \right) \text{eig}[\alpha S^2 + CC^T + ES_\perp^2 C^T + CS_\perp E^T] \\
\leq \left( T^2 - T^2 \right) \text{eig}[\alpha S^2 + \frac{\alpha S_{\min}^2 + S_{\max}^2}{\alpha S_{\min}^2 - S_{\max}^2} CC^T].
\]
(38)

Moreover, the corresponding eigenvector matrix is \(P = U + U_\perp E^T\). From (38), the signal-to-noise (S/N) ratio of the encoder obtained by PCA is
\[
S/N \text{ ratio} \geq \frac{\alpha S_{\min}^2 - S_{\max}^2}{\alpha S_{\min}^2 + S_{\max}^2} \frac{\alpha S_{\min}^2}{S_{\max}^2},
\]
(39)
where \(S_{\max}^2\) is the largest singular value of \(C\). Therefore, when \(S_{\max}^2\) is several times smaller than \(\alpha S_{\min}^2\) and \(S_{\max}^2/S_{\min}^2\) is in an order less than \(N_f/N_s\), the S/N ratio goes to infinity as \(N_f/N_s\) increases. Hence, the major eigenvalues and the corresponding eigenvectors of \(\text{Cov}[x]\) are expressed as shown in (12). The above condition (39) implies that the S/N ratio is more sensitive to \(S_{\max}^2\) than \(S_{\min}^2\).

**Learning rules**

For the dimensionality reduction, Oja’s subspace rule for PCA is considered [6]. Oja’s subspace rule is a modified version of Hebbian plasticity, which is defined by
\[
W \propto E \left[ u(x^T - E|x|T - u^T W) \right],
\]
(40)
where the dot over \(W\) denotes a temporal derivative. By using the \(N_s\)-dimensional neural outputs, this rule can extract a subspace spanned by the first to the \(N_s\)-th principal components. While Oja’s subspace rule does not have a cost function, there is a gradient descent rule for PCA, which is termed as the least mean squared error-based PCA [8]. The cost function of PCA is defined by \(L_X \equiv E[|x - E|x| - W^T u|^2]\). The purpose of PCA is to obtain a representation using a small dimension of output units with the least loss. Indeed, the gradient descent rule of this cost function is the same as Oja’s subspace rule up to an additional term that does not essentially change the behavior of the algorithm. Importantly, our cost function for the optimal encoder \(E[|x - E|x| - W^T s|^2]\) can be seen as one that replaces \(W^T u\) in the PCA cost function with \(W^+ s\).

For BSS of extracted major principal components, Amari’s ICA rule is considered [13]. Suppose \(v \equiv W_{PCA}(x - E|x|) \in \mathbb{R}^{N_s}\) are neural outputs expressing arbitrary rotation of extracted major principal components, and \(u \equiv W_{ICA}v \in \mathbb{R}^{N_s}\) are neural outputs expressing independent components. The cost function of Amari’s ICA rule is defined by the Kullback–Leibler divergence [3] between the posterior distribution \(p(u)\) and the prior distribution \(p_0(u)\), \(L_A \equiv D_{KL}[p(u)||p_0(u)] \equiv E[\log p(u) - \log p(u)]\).
\[ \log p_0(u), \text{ where } p_0(u) = \prod_i p_0(u_i) \text{ is the prior distribution that the sources are supposed to follow.} \]

The natural gradient of \( L_A \) gives Amari’s ICA rule \(^{13}\)

\[
W_{ICA} \propto -\frac{\partial L_A}{\partial W_{ICA}} W_{ICA}^T W_{ICA} = W_{ICA} - E \left[ g(u)u^T \right] W_{ICA},
\]

where \( g(u) \equiv -d \log p_0(u)/du \) is a nonlinear activation function.

**Supplementary MATLAB source code**

```matlab
Ns = 10; % Dimension of hidden sources
Nf = 10000; % Dimension of nonlinear bases
Nx = 10000; % Dimension of sensory inputs
T = 1000; % Number of sample points

%%% Generative process %%%
A = randn(Nf, Ns) / sqrt(Ns); % Second-layer mixing matrix
B = randn(Nx, Nf) / sqrt(Nf); % First-layer mixing matrix
a = randn(Nf, 1) / sqrt(Ns); % Offset vector
s = rand(Ns, T)*2*sqrt(3) - sqrt(3); % Hidden sources
f = (A * s + a * ones(1, T)).^3; % Nonlinear bases
x = B * f; % Sensory inputs

%%% Neural network %%%
W = Ns/Nx * A'; % Synaptic strength matrix
u = W * x; % Neural outputs

%%% Result %%%
plot(s(1,:),u(1,:),'+'); % s_1 v.s. u_1
```

**Acknowledgements**

This work was supported by RIKEN Center for Brain Science (T.I. and T.T.), Brain/MINDS from AMED under Grant Number JP18dm020700 (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 and defined Theorem 1: T.I. Conceived and defined Theorem 2: T.I. and T.T. Performed the analyses: T.I. and T.T. Wrote the paper: T.I. and T.T.

**References**

1. Cichocki A, Zdunek R, Phan AH, Amari SI (2009) Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation (John Wiley & Sons).
2. Comon P, Jutten C (2010) Handbook of Blind Source Separation: Independent Component Analysis and Applications, eds Comon P, Jutten C (Academic Press, Orlando, FL, USA).
3. Bishop CM (2006) Pattern Recognition and Machine Learning. (Springer-Verlag, New York).
4. Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philos Mag 2(11):559-572.
5. Oja E (1982) Simiplified neuron model as a principal component analyzer. J Math Biol 15(3):267-273.
6. Oja E (1989) Neural networks, principal components, and subspaces. Int J Neural Syst 1(10):61-68.
7. Sanger T-D (1989) Optimal unsupervised learning in a single-layer linear feedforward neural network. Neural Netw 2(6):459-473.
8. Xu L (1993) Least mean square error reconstruction principle for self-organizing neural-nets. Neural Netw 6(5):627-648.
[9] Jolliffe IT (2002) Principal Component Analysis (2nd ed.). (Springer).
[10] Comon P (1994) Independent component analysis, a new concept?. Sig Process 36(3):287-314.
[11] Bell A-J, Sejnowski T-J (1995) An information-maximization approach to blind separation and blind deconvolution. Neural Comput 7(6):1129-1159.
[12] Bell A-J, Sejnowski T-J (1997) The “independent components” of natural scenes are edge filters. Vision Res 37(23):3327-3338.
[13] Amari S-I, Cichocki A, Yang H-H (1996) A new learning algorithm for blind signal separation. Adv Neural Inf Proc Sys 8:757-763.
[14] Hyvärinen A, Oja E (1997) A fast fixed-point algorithm for independent component analysis. Neural Comput 9(7):1483-1492.
[15] Brown G-D, Yamada S, Sejnowski T-J (2001) Independent component analysis at the neural cocktail party. Trends Neurosci 24(1):54-63.
[16] von Helmholtz H (1925) Treatise on physiological optics (Vol. 3). The Optical Society of America.
[17] Dayan P, Hinton G-E, Neal R-M, Zemel R-S (1995) The Helmholtz machine. Neural Comput 7(5):889-904.
[18] Knill D-C, Pouget A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci 27(12):712-719.
[19] Friston K, Kilner J, Harrison L (2006) A free energy principle for the brain. J Physiol Paris 100(1):70-87.
[20] Friston K (2010) The free-energy principle: a unified brain theory? Nat Rev Neurosci 11:127-138.
[21] DiCarlo J-J, Zoccolan D, Rust N-C (2012) How does the brain solve visual object recognition? Neuron 73(3):415-434.
[22] Amari S-I, Chen T-P, Cichocki A (1997) Stability analysis of learning algorithms for blind source separation. Neural Netw 10(8):1345-1351.
[23] Oja E, Yuan Z (2006) The FastICA algorithm revisited: Convergence analysis. IEEE Trans Neural Netw 17(6):1370-1381.
[24] Erdogan A-T (2009) On the convergence of ICA algorithms with symmetric orthogonalization. IEEE Trans Signal Process 57(6):2209-2221.
[25] Hyvärinen A, Pajunen P (1999) Nonlinear independent component analysis: Existence and uniqueness results. Neural Netw 12(3):429-439.
[26] Jutten C, Karhunen J (2004) Advances in blind source separation (BSS) and independent component analysis (ICA) for nonlinear mixtures. Int J Neural Syst 14(5):267-292.
[27] Calhoun V-D, Liu J, Adali T (2009) A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. Neuroimage 45(1):S163-S172.
[28] Baldi P, Hornik K (1989) Neural networks and principal component analysis: Learning from examples without local minima. Neural Netw 2(1):53-58.
[29] Chen T, Hua Y, Yan WY (1998) Global convergence of Oja’s subspace algorithm for principal component extraction. IEEE Trans Neural Netw 9(1):58-67.
[30] Papadias CB (2000) Globally convergent blind source separation based on a multilayer kurtosis maximization criterion. IEEE Trans Signal Process 48(12):3508-3519.
[31] Erdogan A-T (2007) Globally convergent deflationary instantaneous blind source separation algorithm for digital communication signals. IEEE Trans Signal Process 55(5):2182-2192.
[32] Lappalainen H, Honkela A (2000) Bayesian non-linear independent component analysis by multi-layer perceptrons. Advances in Independent Component Analysis (Springer, London), pp 93-121.
[33] Karhunen, J. (2001). Nonlinear independent component analysis. Independent Component Analysis: Principles and Practice, eds Roberts S, Everson, R (Cambridge University Press, Cambridge) pp 113-134.
[34] Hinton G-E, Salakhutdinov R-R (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504-507.

[35] Kingma D-P, Welling M (2013) Auto-encoding variational bayes. arXiv:1312.6114

[36] Dinh L, Krueger D, Bengio Y (2014) NICE: Non-linear independent components estimation. arXiv:1410.8516

[37] Friston K (2008) Hierarchical models in the brain. PLoS Comput. Biol 4:e1000211.

[38] Friston K, Trujillo-Barreto N, Daunizeau J (2008) DEM: A variational treatment of dynamic systems. NeuroImage 41(3):849-885.

[39] Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Netw, 2(5):359-366.

[40] Barron A-R (1993) Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans Info Theory 39(3):930-945.

[41] Rahimi A, Recht B (2008) Uniform approximation of functions with random bases. In Proceedings of the 46th Annual Allerton Conference on Communication, Control, and Computing:555-561.

[42] Rahimi A, Recht B (2008) Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. Adv Neural Info Process Sys 21:1313-1320.

[43] Bussgang J-J (1952) Cross-correlation functions of amplitude-distorted Gaussian signals. Res Lab Elec, Mas Inst Technol, Cambridge MA, Tech Rep 216.

[44] Toyoizumi T, Abbott L-F (2011) Beyond the edge of chaos: Amplification and temporal integration by recurrent networks in the chaotic regime. Phys Rev E 84(5):051908.

[45] Arora S, Risteski A (2017) Provable benefits of representation learning. arXiv:1706.04601

[46] Chandler D-M, Field D-J (2007) Estimates of the information content and dimensionality of natural scenes from proximity distributions. J Opt Soc Am A 24(4):922-941.

[47] Kawaguchi K (2016) Deep learning without poor local minima. Adv Neural Inf Proc Sys 29:586-594.

[48] Lu H, Kawaguchi K (2017) Depth creates no bad local minima. arXiv:1702.08580

[49] Nguyen Q, Hein M (2017) The loss surface of deep and wide neural networks. arXiv:1704.08045

[50] Dahl G-E, Yu D, Deng L, Acero A (2012) Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans Audio Speech Lang Proc 20(1):30-42.

[51] Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR (2012) Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580

[52] Wan L, Zeiler M, Zhang S, LeCun Y, Fergus R (2013) Regularization of neural networks using DropConnect. In Proceedings of Machine Learning Research (28):3:1058-1066.

[53] Wu L, Zhu Z, Weinan E (2017) Towards Understanding Generalization of Deep Learning: Perspective of Loss Landscapes. arXiv:1706.10239

[54] Hebb D-O (1949) The Organization of Behavior: A Neuropsychological Theory (Wiley, New York).

[55] Malenka R-C, Bear M-F (2004) LTP and LTD: an embarrassment of riches. Neuron 44(1):5-21.

[56] Dayan P, Abbott L-F (2001) Theoretical neuroscience: computational and mathematical modeling of neural systems, eds Sejnowski T-J, Poggio T-A (MIT Press, London).

[57] Gerstner W, Kistler W-M (2002) Spiking Neuron Models: Single Neurons, Populations, Plasticity, (Cambridge University Press, Cambridge).

[58] Isomura T, Kotani K, Jimbo Y (2015) Cultured cortical neurons can perform blind source separation according to the free-energy principle. PLoS Comput Biol 11(12):e1004643.

[59] Földiák P (1990) Forming sparse representations by local anti-Hebbian learning. Biol Cybern 64(2):165-170.
[60] Linsker R (1997) A local learning rule that enables information maximization for arbitrary input distributions. Neural Comput 9(8):1661-1665.

[61] Isomura T, Toyoizumi T (2016) A local learning rule for independent component analysis. Sci Rep 6:28073.

[62] Pehlevan C, Mohan S, Chklovskii D-B (2017) Blind nonnegative source separation using biological neural networks. Neural Comput 29(11):2925-2954.

[63] Isomura T, Toyoizumi T (2018) Error-gated Hebbian rule: a local learning rule for principal and independent component analysis. Sci Rep 8:1835.

[64] Leugering J, Pipa G (2018) A unifying framework of synaptic and intrinsic plasticity in neural populations. Neural Comput 30(4):945-986.

[65] Kandel E-R, Schwartz J-H, Jessell T-M, Siegelbaum S-A, Hudspeth A-J (2013) Principles of Neural Science 5th edn. eds Kandel E-R, Schwartz J-H, Jessell T-M, Siegelbaum S-A, Hudspeth A-J (McGraw-Hill, New York).