 doi:10.20944/preprints201711.0020.v2

Article

A Measure of Information Available for Inference

Takuya Isomura [1]*

[1] Laboratory for Neural Computation and Adaptation, RIKEN Center for Brain Science, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan.
* Correspondence: takuya.isomura@riken.jp; Tel.: +81-48-467-9644

Academic Editor: name
Version May 11, 2018 submitted to Entropy

Abstract: The mutual information between the state of a neural network and the state of the external world represents the amount of information stored in the neural network that is associated with the external world. In contrast, the surprise of the sensory input indicates the unpredictability of the current input. In other words, this is a measure of inference ability, and an upper bound of the surprise is known as the variational free energy. According to the free-energy principle (FEP), a neural network continuously minimizes the free energy to perceive the external world. For the survival of animals, inference ability is considered to be more important than simply memorized information. In this study, the free energy is shown to represent the gap between the amount of information stored in the neural network and that available for inference. This concept involves both the FEP and the infomax principle, and will be a useful measure for quantifying the amount of information available for inference.

Keywords: free-energy principle; internal model hypothesis; unconscious inference; infomax principle; independent component analysis; principal component analysis

1. Introduction

Sensory perception comprises complex responses of the brain to sensory inputs. For example, the visual cortex can distinguish objects from their background [1], while the auditory cortex can recognize a certain sound in a noisy place with high sensitivity, a phenomenon known as the cocktail party effect [2–7]. The brain (i.e., a neural network) has acquired these perceptual abilities without supervision, which is referred to as unsupervised learning [8–10]. Unsupervised learning, or implicit learning, is defined as the learning that happens in the absence of a teacher or supervisor; it is achieved through adaptation to past environments, which is necessary for higher brain functions. An understanding of the physiological mechanisms that mediate unsupervised learning is fundamental to augmenting our knowledge of information processing in the brain.

One of the consequent benefits of unsupervised learning is inference, which is the action of guessing unknown matters based on known facts or certain observations; i.e., it is the process of drawing conclusions through reasoning and estimation. While inference is thought to be an act of the conscious mind in the ordinary sense of the word, where consciousness often represents a state of self-awareness, it can occur even in the unconscious mind. Hermann von Helmholtz, a 19th-century physicist/physiologist, realized that perception often requires inference by the unconscious mind and coined the word *unconscious inference* [11]. According to Helmholtz, conscious inference and unconscious inference can be distinguished based on whether conscious knowledge is involved in the process. For example, when an astronomer computes the positions or distances of stars in space based on images taken at various times from different parts of the orbit of the Earth, he or she performs conscious inference. This is because the process is "based on a conscious knowledge of the laws of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2018 doi:10.20944/preprints201711.0020.v2

optics"; by contrast, "in the ordinary acts of vision, this knowledge of optics is lacking" [11]. Thus, the latter process is performed by the unconscious mind. Nevertheless, the results of conscious and unconscious inference are clearly similar. Similar to conscious inference, unconscious inference is crucial so that cognitive processes in the unconscious mind can estimate the overall picture from partial observations.

In the field of theoretical and computational neuroscience, unconscious inference has been translated as the successive inference of the generative process of the external world (in terms of Bayesian inference) that animals perform in order to achieve perception. One hypothesis, the so-called internal model hypothesis [12–19], states that animals reconstruct a model of the external world in their brain through past experiences. This internal model helps animals infer hidden causes and predict future inputs automatically; in other words, this inference process happens unconsciously. This is also known as the predictive coding hypothesis [20,21]. In the past decade, a mathematical foundation for unconscious inference, called the free-energy principle (FEP), has been proposed [13–17], and is a candidate unified theory of higher brain functions. Briefly, this principle hypothesizes that parameters of the generative model are learned through unsupervised learning, while hidden variables are inferred in the subsequent inference step. The FEP provides a unified framework for higher brain functions including perceptual learning [14], reinforcement learning [23], motor learning [22,23], communication [24,25], emotion, mental disorders [26,27], and evolution. However, the difference between the FEP and a related theory, namely the information maximization (infomax) principle [28–31], is still not fully understood.

In this study, the relationship between the FEP and the infomax principle is investigated. As one of most simple and important examples, the study focuses on blind source separation (BSS), which is the task of separating sensory inputs into hidden sources (or causes) [32–35]. BSS is shown to be a subset of the inference problem considered in the FEP, and variational free energy is demonstrated to represent the difference between the information stored in the neural network (which is the measure of the infomax principle [28]) and the information available for inferring current sensory inputs.

2. Methods

2.1. Definition of a system

Let us suppose $s \equiv (s_1, \ldots, s_N)^T \sim p(s) \equiv \prod_i p(s_i)$ as hidden sources; $x \equiv (x_1, \ldots, x_M)^T \sim p(x)$ as sensory inputs; $u \equiv (u_1, \ldots, u_N)^T \sim p(u)$ as neural outputs; $z \equiv (z_1, \ldots, z_M)^T \sim p(z)$ as background noises; $\epsilon \equiv (\epsilon_1, \ldots, \epsilon_M)^T \sim p(\epsilon)$ as prediction errors; and $f \in \mathbb{R}^M$, $g \in \mathbb{R}^N$, and $h \in \mathbb{R}^M$ as nonlinear functions (see also Table 1). The generative process of the external world (or the environment) is described by a stochastic equation as:

$$\text{Generative process}: \ x = f(s) + z. \tag{1}$$

Recognition and generative models of the neural network are defined as follows:

$$\text{Recognition model}: \ u = g(x), \tag{2}$$

$$\text{Generative model}: \ x = h(u) + \epsilon. \tag{3}$$

Figure 1 illustrates the structure of the system under consideration. For the generative model, the prior distribution of u is defined as $p_u(u) = \prod_i p_u(u_i)$ and the likelihood function as $p_\epsilon(\epsilon) = p^*(x|h(u)) = \mathcal{N}[\epsilon; 0, \Sigma_\epsilon]$, where p^* indicates a statistical model and $\mathcal{N}$ is a Gaussian distribution. Moreover, suppose $\theta \sim p(\theta)$, $W(\in \mathbb{R}^{N\times M}) \sim p(W)$, and $V(\in \mathbb{R}^{M\times N}) \sim p(V)$ as parameter sets for f, g, and h, respectively, $\lambda \sim p(\lambda)$ as a hyper-parameter set for $p(s)$ and $p(z)$, and $\gamma \sim p(\gamma)$ as a hyper-parameter set for $p_u(u)$ and $p_\epsilon(\epsilon)$. Here, hyper-parameters are defined as parameters that determine the shape of distributions (e.g., the covariance matrix of $p_\epsilon(\epsilon)$). Note that W and V

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2018 doi:10.20944/preprints201711.0020.v2

Table 1. Glossary of expressions.

Expression	Description
Generative process	A set of stochastic equations that generate the external world dynamics
Recognition model	A model in the neural network that imitates the inverse of the generative process
Generative model	A model in the neural network that imitates the generative process
$s \in \mathbb{R}^N$	Hidden sources
$x \in \mathbb{R}^M$	Sensory inputs
θ	A set of parameters
λ	A set of hyper-parameters
$\vartheta \equiv \{s, \theta, \lambda\}$	A set of hidden states of the external world
$u \in \mathbb{R}^N$	Neural outputs
$W \in \mathbb{R}^{N\times M}, V \in \mathbb{R}^{M\times N}$	Synaptic strength matrices
γ	State of neuromodulators
$\varphi \equiv \{u, W, V, \gamma\}$	A set of the brain internal states
$z \in \mathbb{R}^M$	Background noises
$\epsilon \in \mathbb{R}^M$	Prediction errors
$p(x)$	The actual probability density of x
$p(\varphi\|x), p(x, \varphi), p(\varphi)$	Actual probability densities (posterior densities)
$p_u(u), p_\epsilon(\epsilon), p_\varphi(\varphi)$	Prior densities
$p^*(x), p^*(\varphi\|x), p^*(x, \varphi)$	Statistical models
$dx \equiv \prod_i dx_i$	Finite spatial resolution of x
$\langle\bullet\rangle_{p(x)} \equiv \int \bullet p(x)dx$	Expectation of $\bullet$ over $p(x)$
$H[p(x)] \equiv \langle -\log(p(x)dx)\rangle_{p(x)}$	Shannon entropy of $p(x)dx$
$\langle -\log(p^*(x)dx)\rangle_{p(x)}$	Cross entropy of $p^*(x)dx$ over $p(x)$
$\mathcal{D}_{KL}[p(\bullet)\|\|p^*(\bullet)] \equiv \left\langle \log \frac{p(\bullet)}{p^*(\bullet)} \right\rangle_{p(\bullet)}$	KLD between $p(\bullet)$ and $p^*(\bullet)$
$I[x;\varphi] \equiv \mathcal{D}_{KL}[p(x,\varphi)\|\|p(x)p(\varphi)]$	Mutual information between x and φ
$S(x) \equiv \log \frac{p(x)}{p^*(x)}$	Surprise
$\overline{S} \equiv \langle S(x)\rangle_{p(x)}$	Surprise expectation
$F(x) \equiv S(x) + \mathcal{D}_{KL}[p(\varphi\|x)\|\|p^*(\varphi\|x)]$	Free energy
$\overline{F} \equiv \langle F(x)\rangle_{p(x)}$	Free energy expectation
$X[x;\varphi] \equiv \left\langle \log \frac{p^*(x,\varphi)}{p(x)p(\varphi)} \right\rangle_{p(x,\varphi)}$	Utilizable information between x and φ

are assumed as synaptic strength matrices for feedforward and backward paths, respectively, while γ is assumed as a state of neuromodulators similarly to [13–15]. Eqs. (1)-(3) are transformed into probabilistic representations

$$\begin{aligned} \text{Generative process}: \; p(s,x|\theta,\lambda) &= p(x|s,\theta,\lambda)p(s|\lambda) \\ &= \int \delta(x - f(s;\theta) - z)p(z|\lambda)p(s|\lambda)dz \\ &= p(z = x - f|\lambda)p(s|\lambda), \end{aligned} \tag{4}$$

$$\begin{aligned} \text{Recognition model}: \; p(x,u|W) &= p(x|u,W)p(u|W) \\ &= p(u|x,W)p(x) \\ &= \delta(u - g(x;W))p(x), \end{aligned} \tag{5}$$

$$\begin{aligned} \text{Generative model}: \; p^*(x,u|V,\gamma) &= p^*(x|u,V,\gamma)p_u(u|\gamma) \\ &= \int \delta(x - h(u;V) - \epsilon)p_\epsilon(\epsilon|\gamma)p_u(u|\gamma)d\epsilon \\ &= p_\epsilon(\epsilon = x - h|\gamma)p_u(u|\gamma). \end{aligned} \tag{6}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2018 doi:10.20944/preprints201711.0020.v2

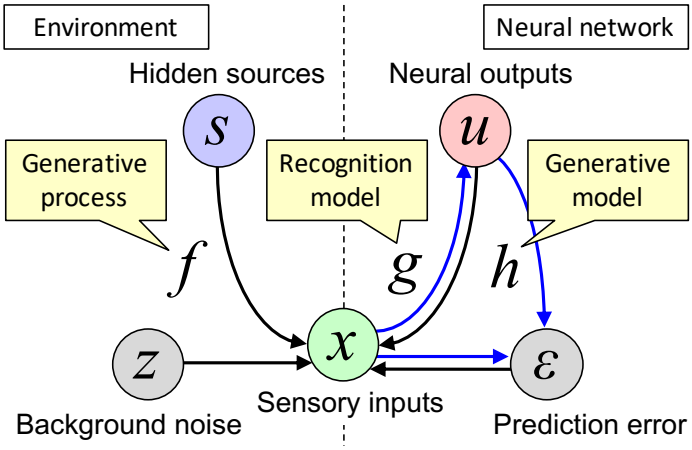

Figure 1. Schematic images of a generative process of the environment (left) and recognition and generative models of the neural network (right). Note that the neural network can access only the states in the right side of the dashed line, including x (see text in Section 3). Black arrows are causal relationships, while blue arrows are information flows of the neural network. See main text and Table 1 for meanings of variables and functions.

Note that $\delta(\bullet)$ is Dirac's delta function and $p^*(x|u,V,\gamma) \equiv p(x|u,V,\gamma,m)$ is a statistical model given a model structure m. For simplification, let $\vartheta \equiv \{s,\theta,\lambda\}$ be a set of hidden states of the external world and $\varphi \equiv \{u,W,V,\gamma\}$ be a set of internal states of the neural network. By multiplying $p(\theta,\lambda)$ to Eq. (4) and $p(W,V,\gamma)$ to Eqs. (5)(6), Eqs. (4)-(6) become

$$\text{Generative process}: \ p(x,\vartheta) = p(x|\vartheta)p(\vartheta) = p(z = x - f)p(\vartheta), \tag{7}$$

$$\text{Recognition model}: \ p(x,\varphi) = p(x|\varphi)p(\varphi) = p(\epsilon = x - h)p(\varphi), \tag{8}$$

$$\text{Generative model}: \ p^*(x,\varphi) = p^*(x|\varphi)p_\varphi(\varphi) = p_\epsilon(\epsilon = x - h)p_\varphi(\varphi), \tag{9}$$

where p_φ is the prior distribution for φ and $p^*(x,\varphi) \equiv p(x,\varphi|m)$ is a statistical model given a model structure m, which is determined by the shapes of p_φ and p_ϵ. The expression of $p^*(x,\varphi)$ is used instead of $p(x,\varphi|m)$ to emphasize the difference between $p(x,\varphi)$ and $p^*(x,\varphi)$. While $p(x,\varphi)$ is the actual joint probability of (x,φ) (which corresponds to the posterior distribution), $p^*(x,\varphi)$, i.e., the product of the likelihood function and the prior distribution, represents the generative model that the neural network expects (x,φ) to follow. As shown later, the inference and learning are achieved by minimizing the difference between $p(x,\varphi)$ and $p^*(x,\varphi)$.

2.2. Information stored in the neural network

Information is defined as the negative log of probability [36]. When $\mathrm{Prob}(x)$ is the probability of given sensory inputs x, its information is given by $-\log \mathrm{Prob}(x)$ [nat], where 1 nat = 1.4427 bits. When x takes continuous values, by coarse graining, $-\log \mathrm{Prob}(x)$ is replaced with $-\log(p(x)dx)$, where $p(x)$ is the probability density of x and $dx \equiv \prod_i dx_i$ is the product of the finite spatial resolutions of x's elements. The expectation of $-\log(p(x)dx)$ over $p(x)$ gives the Shannon entropy (or average information), which is defined by

$$H[p(x)] \equiv \langle -\log(p(x)dx)\rangle_{p(x)} \ [\text{nat}], \tag{10}$$

where $\langle\bullet\rangle_{p(x)} \equiv \int \bullet p(x)dx$ represents the expectation of $\bullet$ over $p(x)$. Note that the use of $-\log(p(x)dx)$ instead of $-\log p(x)$ is useful because this $H[p(x)]$ is non-negative (because $d\mathrm{Prob}(x) = p(x)dx$ takes a value between 0 and 1), while the addition of constant $-\log dx$ has no

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2018 doi:10.20944/preprints201711.0020.v2

effect except for sliding the offset value. If and only if $p(x)$ is Dirac's delta function, $H[p(x)] = 0$ is realized. For the system under consideration (Eqs. (7)–(9)), the information shared between the external world states (x, ϑ) and the internal states of the neural network φ is defined by mutual information [37]

$$I[(x,\vartheta);\varphi] \equiv \left\langle \log \frac{p(x,\vartheta,\varphi)}{p(x,\vartheta)p(\varphi)} \right\rangle_{p(x,\vartheta,\varphi)} \text{[nat]}. \tag{11}$$

Note that $p(x, \vartheta, \varphi)$ is the joint probability of (x, ϑ) and φ. Moreover $p(x, \vartheta)$ and $p(\varphi)$ are their marginal distributions, respectively. This mutual information takes a non-negative value and quantifies how much (x, ϑ) and φ are related with each other. High mutual information indicates the internal states are informative to explain the external world states, while zero mutual information means they are independent of each other.

However, the only information that the neural network can directly access is the sensory input. This is the case because the system under consideration can be described as Bayesian network, see [38, 39] for the detail on Markov blanket. Hence, entropy of the external world states under a fixed sensory input gives the information that the neural network cannot infer. Moreover, there is no feedback control from the neural network to the external world in this setup. Thus, under a fixed x, ϑ and φ are conditionally independent of each other. From $p(\vartheta, \varphi|x) = p(\vartheta|x)p(\varphi|x)$, we can obtain

$$I[(x,\vartheta);\varphi] = \left\langle \log \frac{p(\vartheta|x)p(\varphi|x)p(x)}{p(\vartheta|x)p(x)p(\varphi)} \right\rangle_{p(\vartheta|x)p(\varphi|x)p(x)} = \left\langle \log \frac{p(\varphi|x)}{p(\varphi)} \right\rangle_{p(\varphi,x)} = I[x;\varphi]. \tag{12}$$

Using Shannon entropy, $I[x; \varphi]$ becomes

$$I[x;\varphi] = H[p(x)] - H[x|\varphi] \quad \text{[nat]}, \tag{13}$$

where

$$H[x|\varphi] \equiv \Big\langle -\log\Big(p(x|\varphi)dx\Big)\Big\rangle_{p(x,\varphi)} \equiv \Big\langle -\log\Big(p(\epsilon)dx\Big)\Big\rangle_{p(\epsilon)p(\varphi)} \equiv \langle H[p(\epsilon)]\rangle_{p(\varphi)} \tag{14}$$

is the conditional entropy of x given φ. Thus, maximization of $I[(x, \vartheta); \varphi]$ is the same as maximization of $I[x; \varphi]$ for this system. Because $I[x; \varphi]$, $H[p(x)]$, and $H[x|\varphi]$ are non-negative, $I[x; \varphi]$ has the range $0 \leq I[x; \varphi] \leq H[p(x)]$. Zero mutual information occurs if and only if x and φ are independent, while $I[x; \varphi] = H[p(x)]$ occurs if and only if x is fully explained by φ. In this manner, $I[x; \varphi]$ describes the information about the external world stored in the neural network. Note that this $I[(x, \vartheta); \varphi]$ can be expressed using the Kullback–Leibler divergence (KLD) [40] as $I[x; \varphi] \equiv \mathcal{D}_{KL}\Big[p(x, \varphi)||p(x)p(\varphi)\Big]$. KLD takes a non-negative value and indicates the divergence between two distributions.

The infomax principle states that "the network connections develop in such a way as to maximize the amount of information that is preserved when signals are transformed at each processing stage, subject to certain constraints" [28], see also [29–31]. According to the infomax principle, the neural network is hypothesized to maximize $I[x; \varphi]$ to perceive the external world. However, $I[x; \varphi]$ does not fully explain the inference capability of a neural network. For example, if neural outputs just express the sensory input itself ($u = x$), $I[x; \varphi] = H[p(x)]$ is easily achieved, but this does not mean that the neural network can predict input statistics. This is considered in the next section.

2.3. Free-energy principle

If one has a statistical model determined by model structure m, the information calculated based on m is given by the negative log likelihood $-\log p(x|m)$, which is termed as the (marginal) surprise of the sensory input and expresses the unpredictability of the sensory input for the individual. The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2018 doi:10.20944/preprints201711.0020.v2

neural network is considered to minimize the surprise in the sensory input using the knowledge about the external world, to perceive the external world [13]. To infer if an event is likely to happen based on the past observation, a statistical (i.e., generative) model is necessary; otherwise it is difficult to generalize sensory inputs [41]. Note that the surprise is the marginal over the generative model; hence, the neural network can reduce the surprise by optimizing its internal states, while Shannon entropy of the input is determined by the environment. When the actual probability density and a generative model are given by $p(x)$ and $p^*(x) \equiv p(x|m)$, respectively, the cross entropy $\langle -\log(p^*(x)dx)\rangle_{p(x)}$ is always larger than or equal to Shannon entropy $H[p(x)]$ because of the non-negativity of KLD. Hence, in this study, the input surprise is defined by

$$S(x) \equiv -\log p^*(x) + \log p(x) \quad [\text{nat}] \tag{15}$$

and its expectation over $p(x)$ by

$$\overline{S} \equiv \langle S(x)\rangle_{p(x)} = \mathcal{D}_{KL}[p(x)||p^*(x)] = \langle -\log(p^*(x)dx)\rangle_{p(x)} - H[p(x)] \quad [\text{nat}]. \tag{16}$$

This definition of $S(x)$ is to ensure $\overline{S}$ is non-negative and $\overline{S} = 0$ if and only if $p^*(x) = p(x)$. Since $H[p(x)]$ is determined by the environment and constant for the neural network, minimization of this $\overline{S}$ is the same meaning as minimization of $\langle -\log(p^*(x)dx)\rangle_{p(x)}$.

Because the sensory input is generated by the external world generative process, consideration of the structure and dynamics placed in the background of the sensory input can provide accurate inference. According to the internal model hypothesis, animals develop the internal model in their brain to increase the accuracy and efficiency of inference [12–15,17–19]; thus, internal states of the neural network φ are hypothesized to imitate the hidden states of the external world ϑ. A problem is that $-\log p^*(x) = -\log(\int p^*(x,\varphi)d\varphi)$ is intractable for the neural network, because the integral of $p^*(x,\varphi)$ placed in the logarithm function. The FEP hypothesizes that the neural network calculates an upper bound of $-\log p^*(x)$ instead of the exact value, which is more tractable [13]. This upper bound is termed as variational free energy:

$$F(x) \equiv S(x) + \mathcal{D}_{KL}[p(\varphi|x)||p^*(\varphi|x)] = \langle -\log p^*(x,\varphi) + \log p(x,\varphi)\rangle_{p(\varphi|x)} \,[\text{nat}]. \tag{17}$$

Note that $p(\varphi|x)$ expresses the belief about hidden states of the external world encoded by internal states of the neural network, termed as the recognition density. Due to the non-negativity of KLD, $F(x)$ is guaranteed to be an upper bound of $S(x)$ and $F(x) = S(x)$ holds if and only if $p^*(\varphi|x) = p(\varphi|x)$. Furthermore, the expectation of $F(x)$ over $p(x)$ is defined by

$$\overline{F} \equiv \langle F(x)\rangle_{p(x)} = \mathcal{D}_{KL}[p(x,\varphi)||p^*(x,\varphi)] = \langle U(x,\varphi)\rangle_{p(x,\varphi)} - H[p(x,\varphi)] \quad [\text{nat}], \tag{18}$$

where $U(x,\varphi) \equiv -\log(p^*(x,\varphi)dxd\varphi)$ is termed as the internal energy and $H[p(x,\varphi)] \equiv \langle -\log(p(x,\varphi)dxd\varphi)\rangle_{p(x,\varphi)}$ is the joint entropy of x and φ. $\overline{F}$ indicates the difference between the actual probability $p(x,\varphi)$ and the generative model $p^*(x,\varphi)$. Because of the non-negativity of KLD, $\overline{F}$ is always larger than or equal to $\overline{S}(\geq 0)$ and $\overline{F} = \overline{S} = 0$ holds if and only if $p^*(x,\varphi) = p(x,\varphi)$.

Internal energy $U(x,\varphi)$ quantifies the amplitude of the prediction error at a given moment [13]. Minimization of $\langle U(x,\varphi)\rangle_{p(x,\varphi)}$ is the so-called maximum a posteriori (MAP) estimation (or the maximum likelihood estimation if the priors are uniform distributions) [10] and provides a solution that (at least locally) minimizes the prediction error. Whereas, maximization of $H[p(x,\varphi)]$ increases the independency between internal states, which helps neurons to establish an efficient representation as pointed out by Jaynes' max entropy principle [42,43]. This is essential for BSS [32–35] because the optimal parameters that minimize $\langle U(x,\varphi)\rangle_{p(x,\varphi)}$ are not always determined identically. Due to this, the MAP estimation alone does not always identify the generative process behind the sensory inputs. As $\overline{F}$ is the sum of costs for the MAP estimation and BSS, free-energy minimization is the rule to simultaneously minimize the prediction error and maximize the independency of the internal states.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2018 doi:10.20944/preprints201711.0020.v2

It is recognized that animals perform BSS [2–7]. Interestingly, even *in vitro* neural networks perform BSS which is accompanied by significant reduction of free energy in accordance with the FEP and Jaynes' max entropy principle [44].

2.4. Information available for inference

We now consider how free energy expectation $\overline{F}$ relates to mutual information $I[x;\varphi]$. According to unconscious inference and the internal model hypothesis, the aim of a neural network is to predict x, and for this purpose, it infers hidden states of the external world. While the neural network is conventionally hypothesized to express sufficient statistics of the hidden states of the external world [14], here it is hypothesized that internal states of the neural network are random variables and the probability distribution of them imitates the probability distribution of the hidden states of the external world. Thereby, the aim of the neural network is to match the probability distribution of the internal states with that of the hidden states of the external world. To do so, the neural network shifts the actual probability of internal states $p(x,\varphi) = p(\epsilon)p(\varphi)$ closer to those of the generative model $p^*(x,\varphi) = p_\epsilon(\epsilon)p_\varphi(\varphi)$ that the neural network expects (x,φ) to follow. From this viewpoint, the difference between these two distributions is associated with the loss of information.

The amount of information available for inference can be calculated using the following three values related to information loss: (i) Because $H[p(x)]$ is information of the sensory input and $I[x;\varphi]$ is information stored in the neural network, $H[p(x)] - I[x;\varphi] = \langle H[p(\epsilon)]\rangle_{p(\varphi)}$ indicates the information loss in the recognition model (Fig. 2). (ii) The difference between actual and desired (prior) distributions of internal states $\mathcal{D}_{KL}[p(\varphi)||p_\varphi(\varphi)]$ quantifies the information loss for inferring internal states (i.e., blind state separation). (iii) The difference between distributions of the actual reconstruction error and the prediction error under the given model $\langle \mathcal{D}_{KL}[p(x|\varphi)||p^*(x|\varphi)]\rangle_{p(\varphi)} = \langle \mathcal{D}_{KL}[p(\epsilon)||p_\epsilon(\epsilon)]\rangle_{p(\varphi)}$ quantifies the information loss for representing inputs using internal states. Therefore, by subtracting these three values from $H[p(x)]$, a mutual-information-like measure representing the inference capability is obtained:

$$\begin{aligned} X[x;\varphi] &\equiv H[p(x)] - \langle H[p(\epsilon)]\rangle_{p(\varphi)} - \mathcal{D}_{KL}[p(\varphi)||p_\varphi(\varphi)] - \langle \mathcal{D}_{KL}[p(\epsilon)||p_\epsilon(\epsilon)]\rangle_{p(\varphi)} \\ &= \left\langle \log \frac{p^*(x,\varphi)}{p(x)p(\varphi)} \right\rangle_{p(x,\varphi)} \quad [\text{nat}], \end{aligned} \tag{19}$$

which is called utilizable information in this study. This utilizable information $X[x;\varphi]$ is defined by replacing $p(x,\varphi)$ in $I[x;\varphi]$ with $p^*(x,\varphi)$, immediately yielding

$$\overline{F} = I[x;\varphi] - X[x;\varphi] \quad [\text{nat}]. \tag{20}$$

Hence, $\overline{F}$ represents the gap between the amount of information stored in the neural network and the amount that is available for inference, which is equivalent to the information loss in the generative model. Note that the sum of losses in the recognition and generative models $H[p(x)] - X[x;\varphi] = \overline{F} + \langle H[p(\epsilon)]\rangle_{p(\varphi)}$ is an upper bound of $\overline{F}$ because of the non-negativity of $\langle H[p(\epsilon)]\rangle_{p(\varphi)}$ (Fig. 2). Because $\langle H[p(\epsilon)]\rangle_{p(\varphi)}$ is generally nonzero, $F(x) + \langle H[p(\epsilon)]\rangle_{p(\varphi)}$ does not usually reach zero, even when $p(x,\varphi) = p^*(x,\varphi)$.

Furthermore, $X[x;\varphi]$ is transformed into

$$X[x;\varphi] = H[p(x)] - L_X - L_A, \tag{21}$$

where

$$L_X \equiv \langle -\log(p_\epsilon(\epsilon)d\epsilon)\rangle_{p(\epsilon)p(\varphi)} \tag{22}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2018 doi:10.20944/preprints201711.0020.v2

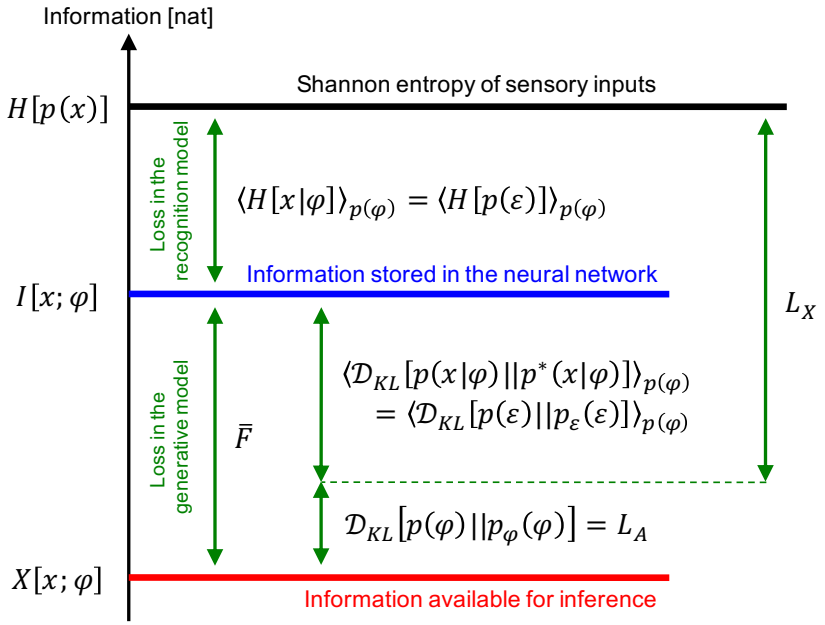

Figure 2. Relationship between information measures. The mutual information between the inputs and internal states of the neural network ($I[x;\varphi]$) is less than or equal to the Shannon entropy of the inputs ($H[p(x)]$) because of the information loss in the recognition model. The utilizable information ($X[x;\varphi]$) is less than or equal to the mutual information and the gap between them gives the expectation of the variational free energy ($\bar{F}$), which quantifies the loss in the generative model. The sum of PCA and ICA costs ($L_X + L_A$) is equal to the gap between the Shannon entropy and the utilizable information, expressing the sum of losses in the recognition and generative models.

is the so-called reconstruction error, which is similar to the reconstruction error for principal component analysis (PCA) [45], while

$$L_A \equiv \mathcal{D}_{KL}[p(\varphi)||p_\varphi(\varphi)] \tag{23}$$

is a generalization of Amari's cost function for independent component analysis (ICA) [46].

PCA is one of the most popular dimensionality reduction methods. It is used to remove background noise and extract important features from sensory inputs [45,47]. In contrast, ICA is a BSS method used to decompose a mixture set of sensory inputs into independent hidden sources [33,35,46,48,49]. Theoreticians hypothesize that the PCA- and ICA-like learning underlies BSS in the brain [3]. This kind of extractions of the hidden representation is also an important problem in machine learning [50,51]. Equation (21) indicates that $X[x;\varphi]$ consists of PCA- and ICA-like parts, i.e., maximization of $X[x;\varphi]$ can perform both dimensionality reduction and BSS (Fig. 2). Their relationship is discussed in the next section.

3. Comparison between the free-energy principle and related theories

In this section, the FEP is compared with other theories. As described in the Methods, the aim of the infomax principle is to maximize mutual information $I[x;\varphi]$ (Eq. (13)), while the aim of the FEP is to minimize free energy expectation $\bar{F}$ (Eq. (18)), and maximization of utilizable information $X[x;\varphi]$ (Eq. (19)) means to do both of them simultaneously.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2018 doi:10.20944/preprints201711.0020.v2

3.1. Infomax principle

The generative process and recognition- and generative models defined in Eqs. (1)-(3) are assumed. For simplification, suppose W, V and γ follow Dirac's delta functions; then, the goal of the infomax principle is simplified as maximization of mutual information between x and u:

$$I[x;u] = \left\langle \log \frac{p(x,u)}{p(x)p(u)} \right\rangle_{p(x,u)} = H[p(x)] - H[x|u] = H[p(u)] - H[u|x]. \tag{24}$$

If $\dim(x) \geq \dim(u)$ and a linear recognition model $u = g(x) = Wx$ with full-rank matrix W is supposed, because $H[u|x] = 0$ and u has an infinite range, $I[x;u] = H[p(u)]$ monotonically increases as the variance of u increases. Thus, maximization of $I[x;u]$ cannot perform either PCA or ICA. To perform PCA and ICA based on the infomax principle, one needs to consider mutual information between sensory inputs and nonlinearly transformed neural outputs $\psi(u) = (\psi(u_1), \ldots, \psi(u_N))^T$ with an injective nonlinear function $\psi(\bullet)$. This mutual information is given by

$$I[x;\psi(u)] = \left\langle \log \frac{p(x,\psi(u))}{p(x)p(\psi(u))} \right\rangle_{p(x,\psi(u))} = H[p(\psi(u))] - H[\psi(u)|x]. \tag{25}$$

When nonlinear neural outputs have a finite range (e.g., between 0 and 1), the variance of u should be maintained in the appropriate range. The infomax based ICA [48,49] is formulated based on this constraint. From $p(\psi(u)) = |\partial u/\partial\psi(u)|p(u) = (\prod_i \psi'(u_i))^{-1}p(u)$, $H[p(\psi(u))]$ becomes $H[p(\psi(u))] = \langle -\log\{(\prod_i \psi'(u_i))^{-1}p(u)du\}\rangle_{p(u)} = H[p(u)] + \langle \sum_i \log\psi'(u_i)\rangle_{p(u)}$. Since $H[\psi(u)|x] = 0$ hold, Eq. (25) becomes

$$I[x;\psi(u)] = H[p(u)] + \left\langle \sum_i \log\psi'(u_i) \right\rangle_{p(u)}. \tag{26}$$

In what follows, it is described that maximization of Eq. (26) as well as the FEP performs PCA and ICA.

3.2. Principal component analysis

Both the infomax principle and the FEP give a cost function of PCA. Suppose $\dim(x) > \dim(u)$, $V = W^T$, and $-\log\psi'(u_i) = u_i^2/2 + \text{const}$. From Eq. (24), $H[p(u)] = H[p(x)] - \langle H[p(\epsilon)]\rangle_{p(\varphi)}$ holds. Since the prediction error is given by $\epsilon = x - W^T u = (I - W^T W)x$, we obtain $\langle H[p(\epsilon)]\rangle_{p(\varphi)} = \langle -\log\{p(x)|\partial x/\partial\epsilon|dx\}\rangle_{p(x,\varphi)} = H[p(x)] + \langle \log|I - W^T W|\rangle_{p(\varphi)}$. Thus, Eq. (26) becomes

$$I[x;\psi(u)] = -\left\langle \log|I - W^T W| \right\rangle_{p(\varphi)} - \frac{1}{2}\left\langle |u|^2 \right\rangle_{p(u)}. \tag{27}$$

The first term of Eq. (27) is maximized if $WW^T = I$ holds (i.e., W is an orthogonal matrix). To maximize the second term, outputs u need to be involved in a subspace spanned by the first to the N-th major principal components of x. Therefore, maximization of Eq. (27) performs PCA.

PCA is also derived by minimization of L_X (Eq. (22)) under the assumption that $p_\epsilon(\epsilon)$ is a Gaussian distribution $p_\epsilon(\epsilon) = \mathcal{N}[\epsilon; 0, \gamma^{-1}I]$ with a scalar hyper-parameter $\gamma > 0$. This is given by

$$L_X = \left\langle \frac{\gamma}{2}\epsilon^T\epsilon - \frac{1}{2}\log|\gamma| \right\rangle_{p(\varphi)} + \text{const}. \tag{28}$$

The derivative of Eq. (28) gives the update rule for the least square error PCA [45], which is similar to the well-known Oja's subspace rule for PCA [47]. This L_X is also the same form as the cost function for auto-encoder [50]. Moreover, when the priors of u, W, V, and γ are flat, free energy expectation (Eq. (18)) becomes $\overline{F} = L_X - \langle H[p(\epsilon)]\rangle_{p(\varphi)} - H[p(u)] = L_X + \text{const}$.; thus, under this condition $\overline{F}$ is equivalent to the PCA cost function.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2018 doi:10.20944/preprints201711.0020.v2

3.3. Independent component analysis

Both the infomax principle and the FEP give a cost function of ICA. Suppose that sources $s_1, \ldots, s_N$ independently follow an identical distribution $p_0(s_i)$. The infomax based ICA is derived from Eq. (26) [48,49]. If $\psi(u_i)$ is defined to satisfy $\psi'(u_i) = p_0(u_i)$, negative mutual information $-I[x;\psi(u)]$ becomes KLD between actual and prior distributions up to constant term,

$$-I[x;\psi(u)] - \log du = \Big\langle \log p(u) - \log p_0(u) \Big\rangle_{p(u)} = \mathcal{D}_{KL}[p(u)||p_0(u)] \equiv L_A. \tag{29}$$

This L_A is known as Amari's ICA cost function [46]. While both $-I[x;\psi(u)]$ and L_A provide the same gradient descent rule, formulating $I[x;\psi(u)]$ requires the nonlinearly transformed neural outputs $\psi(u)$. By contrast, L_A straightforwardly represents that minimization of KLD between $p(u)$ and $p_0(u)$ performs ICA. Indeed, if $\dim(u) = \dim(x) = N$, the background noise is small, and the priors of W, V, and γ are flat, we obtain $\overline{F} = \mathcal{D}_{KL}[p(u)||p_0(u)] = L_A$. Therefore, ICA is a subset of the inference problem considered in the FEP, and the derivation from the FEP is simpler while both the infomax principle and the FEP give the same ICA algorithm.

Furthermore, when $\dim(x) > \dim(u)$, minimization of $\overline{F}$ can perform both dimensionality reduction and BSS. When the priors of W, V, and γ are flat, free energy expectation (Eq. (18)) approximately becomes $\overline{F} \approx L_X + L_A + \text{const.} = -X[x;u] + \text{const.}$ The ratio of PCA to ICA is controlled by γ. Unlike the case with scalar γ described above, if $\Sigma_\epsilon(\gamma)$ is fine tuned by high-dimensional γ to minimize $\overline{F}$, $\Sigma_\epsilon = \langle \epsilon\epsilon^T \rangle_{p(\epsilon)}$ is obtained. Under this condition, L_X is equal to $H[x|u]$ up to constant term and thereby $\overline{F} = L_A + \text{const.}$ is obtained. This indicates that $\overline{F}$ consists only of the ICA part. These comparisons suggest that low-dimensional γ is better to perform noise reduction.

4. Simulation and results

The difference between the infomax principle and the FEP is illustrated by a simple simulation using a linear generative process and a linear neural network (Fig. 3). For simplification, it is assumed that the dynamics of u quickly converge to the optimum that minimizes $F(x)$ compared to the change of s (adiabatic approximation).

For the results shown in Fig. 3, s denotes two-dimensional hidden sources following an identical Laplace distribution with zero mean and unit variance; x denotes four-dimensional sensory inputs; u denotes two-dimensional neural outputs; z denotes four-dimensional background Gaussian noises following $\mathcal{N}[z;0,\Sigma_z]$; θ denotes a 4×2-dimensional mixing matrix; W is a 2×4-dimensional synaptic strength matrix for the bottom-up path; V is a 4×2-dimensional synaptic strength matrix for the top-down path; and the priors of W, V, and γ are flat priors. Sensory inputs are determined by $x = \theta s + z$, while neural outputs are determined by $u = Wx$. The prediction error is given by $\epsilon = x - Vu$ and used to calculate $H[p(\epsilon)]$ and L_A. Horizontal and vertical axes in the figure are conditional entropy $H[x|\varphi]$ (Eq. (14)) and free energy expectation $\overline{F}$ (Eq. (18)), respectively. Simulations were conducted 100 times with randomly selected θ and Σ_z for each condition. For each simulation, 10^8 random sample points were generated and probability distributions were calculated using the histogram method.

First, when W is randomly chosen and V is defined by $V = W^T$, both $H[x|\varphi]$ and $\overline{F}$ are scattered (black circles in Fig. 3) because neural outputs represent random mixtures of sources and noises. Next, when W is optimized according to either Eq. (27) or (28) under the constraint of $V = W^T$, the neural outputs express the major principal components of the inputs; i.e., the network performs PCA (blue circles in Fig. 3). This is the case when $H[x|\varphi]$ is minimized. In contrast, when W, V, and $\Sigma_\epsilon(\gamma)$ are optimized according to the FEP (see Eq. (**??**)), the neural outputs represent the independent components that match the prior source distribution; i.e., the network performs BSS or ICA while minimizing the prediction error (red circles in Fig. 3). For linear generative processes, the minimization of $\overline{F}$ can reliably and accurately perform both dimensionality reduction and BSS

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2018 doi:10.20944/preprints201711.0020.v2

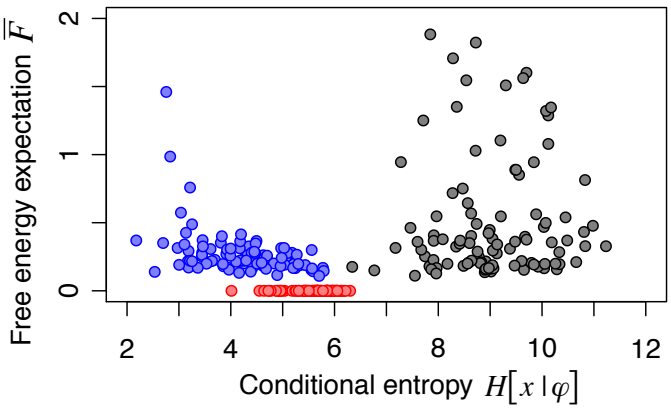

Figure 3. Difference between the infomax principle and FEP when sources follow a non-Gaussian distribution. Black, blue, and red circles indicate the results when W is a random matrix, optimized for the infomax principle (i.e., PCA), and optimized for the FEP, respectively.

because the outputs become independent of each other and match the prior belief if and only if the outputs represent true sources up to permutation and sign-flip. Because the utilizable information consists of PCA and ICA cost functions (see Eq. (21)), the maximization of $X[x;\varphi]$ leads to a solution that is a compromise between the solutions for the infomax principle and the FEP. Interestingly, the infomax optimization (i.e., PCA) provides a W that makes $\overline{F}$ closer to zero than random states, which indicates that the infomax optimization contributes to the free energy minimization. Note that, for nonlinear systems, there are many different transformations that make the outputs independent of each other [52]. Hence, there is no guarantee that minimization of $\overline{F}$ can identify the true sources of nonlinear generative models.

In sum, the aims of the FEP and infomax principle are similar to each other. In particular, when both the sources and noises follow Gaussian distributions, their aims become the same. Conversely, the optimal synaptic weights under the FEP are different from those under the infomax principle when sources follow non-Gaussian distributions. Under this condition, the maximization of the utilizable information leads to a compromise solution between those for the FEP and the infomax principle.

5. Discussion

In this study, the FEP is linked with the infomax principle, PCA, and ICA. It is more likely that the purpose of a neural network in a biological system is to minimize the surprise of sensory inputs to realize better inference rather than maximize the amount of stored information. For example, the visual input captured by a video camera contributes to the stored information, but this amount of information is not equal to the amount of information available for inference. The surprise expectation represents the difference between actual and inferred observations; the free energy expectation provides the difference between recognition and generative models. Utilizable information is introduced to quantify the inference and generalization capability of sensory inputs. Using this approach, the free energy expectation can be explained as the gap between the information stored in the neural network and that available for inference. Moreover, the derivation of ICA is simplified by the FEP. To perform ICA based on the infomax principle, one needs to tune the nonlinearity of the neural outputs to ensure the derivative of the nonlinear I/O function matches the prior distribution. Conversely, under the FEP, ICA is straightforwardly derived from the KLD between the actual probability distribution and the prior distribution of u. Especially, in the absence of background noise and prior knowledge of the parameters and hyper-parameters, the free energy

 doi:10.20944/preprints201711.0020.v2

expectation is equivalent to the surprise expectation as well as Amari's ICA cost function, which indicates that ICA is a subproblem of the FEP.

The FEP is a rigorous and promising theory from theoretical and engineering viewpoints because various learning rules are derived from the FEP [14,15]. However, to be a physiologically plausible theory of the brain, the FEP needs to satisfy certain physiological requirements. There are two major requirements: first, physiological evidence that shows the existence of learning or self-organizing processes under the FEP is required. The model structure under the FEP is consistent with the structure of cortical microcircuits [19]. Moreover, *in vitro* neural networks performing BSS reduce free energy [44]. It is known that the spontaneous prior activity of a visual area enables it to learn the properties of natural pictures [53]. These results suggest the physiological plausibility of the FEP. Nevertheless, further experiments and consideration of information-theoretical optimization under physiological constraints [54] are required to prove the existence of the FEP in the biological brain. Second, the update rule must be a biologically plausible local learning rule; i.e., synaptic strengths must be changed by signals from connected cells or widespread liquid factors. While the synaptic update rule for a discrete system is local [17], the current rule for a continuous system [14] is a non-local rule. Recently developed biologically-plausible three-factor learning models in which Hebbian learning is mediated by a third modulatory factor [55–58] may help reveal the neuronal mechanism underlying unconscious inference. Therefore, it is necessary to investigate how actual neural networks infer the dynamics placed in the background of the sensory input and if this is consistent with the FEP, see also [59] for the relationship between the FEP and spike-timing dependent plasticity [60,61]. This may help develop a biologically plausible learning algorithm through which an actual neural network might develop its internal model. Characterization of information from physical viewpoint may also help understand how the brain physically embodies the information [62,63]. In the subsequent work, we would like to see their relationship.

In summary, this study investigated the differences between two types of information—information stored in the neural network and information available for inference. It was demonstrated that free energy represents the gap between these two types of information. This result clarifies the difference between the FEP and related theories and can be utilized for understanding unconscious inference from a theoretical viewpoint.

Acknowledgments: This work was supported by RIKEN Center for Brain Science.

Conflicts of Interest: The author declares no competing financial interests. The founding sponsor had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. DiCarlo, J.J.; Zoccolan, D.; Rust, N.C. How does the brain solve visual object recognition? *Neuron* **2012**, *73*, 415-434.
2. Bronkhorst, A.W. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica* **2000**, *86*, 117-128.
3. Brown, G.D.; Yamada, S.; Sejnowski, T.J. Independent component analysis at the neural cocktail party. *Trends in neurosciences* **2001**, *24*, 54-63.
4. Haykin, S.; Chen, Z. The cocktail party problem. *Neural Comput* **2005**, *17*, 1875-1902.
5. Narayan, R.; Best, V.; Ozmeral, E.; McClaine, E.; Dent, M.; Shinn-Cunningham, B.; Sen, K. Cortical interference effects in the cocktail party problem. *Nat Neurosci* **2007**, *10*, 1601-1607.
6. Mesgarani, N.; Chang, E.F. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* **2012**, *485*, 233-236.
7. Golumbic, E.M.Z.; Ding, N.; Bickel, S.; Lakatos, P.; Schevon, C.A.; McKhann, G.M.; Schroeder, C.E. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron* **2013**, *77*, 980-991.

8. Dayan, P.; Abbott, L.F. *Theoretical neuroscience: computational and mathematical modeling of neural systems*; MIT Press, London, 2001.
9. Gerstner, W.; Kistler, W. *Spiking Neuron Models. Single Neurons, Populations, Plasticity*; Cambridge University Press, Cambridge, 2002.
10. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer Verlag, 2006.
11. von Helmholtz, H. *Treatise on physiological optics (Vol. 3)*; The Optical Society of America, 1925.
12. Dayan, P.; Hinton, G.E.; Neal, R.M.; Zemel, R.S. The helmholtz machine. *Neural Comput* **1995**, *7*, 889-904.
13. Friston, K.; Kilner, J.; Harrison, L. A free energy principle for the brain. *Journal of Physiology-Paris* **2006**, *100*, 70-87.
14. Friston, K.J. Hierarchical model in the brain. *PLoS Comput Biol* **2008**, *4*, e1000211.
15. Friston, K. The free-energy principle: a unified brain theory? *Nat Rev Neurosci* **2010**, *11*, 127-138.
16. Friston, K. A free energy principle for biological systems. *Entropy* **2012**, *14*, 2100-2121.
17. Friston, K.; FitzGerald, T.; Rigoli, F.; Schwartenbeck, P.; Pezzulo, G. Active inference: A process theory. *Neural Comput* **2017**, *29(1)*, 1-49.
18. George, D.; Hawkins, J. Towards a mathematical theory of cortical micro-circuits. *PLoS Comput Biol*, **2009**, *5*, e1000532.
19. Bastos, A.M.; Usrey, W.M.; Adams, R.A.; Mangun, G.R.; Fries, P.; Friston, K.J. Canonical microcircuits for predictive coding. *Neuron* **2012**, *76*, 695-711.
20. Rao, R.P.; Ballard, D.H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci* **1999**, *2*, 79-87.
21. Friston, K. A theory of cortical responses. *Philos Trans R Soc Lond B Biol Sci* **2005**, *360*, 815-836.
22. Kilner, J.M.; Friston, K.J.; Frith, C.D. Predictive coding: an account of the mirror neuron system. *Cognitive Processing* **2007**, *8*, 159-166.
23. Friston, K.; Mattout, J.; Kilner, J. Action understanding and active inference. *Biological Cybernetics*, **2011**, *104*, 137-160.
24. Friston, K.J.; Frith, C.D. Active inference, communication and hermeneutics. *Cortex* **2015**, *68*, 129-143.
25. Friston, K.; Frith, C. A duet for one. *Consciousness and Cognition* **2015**, *36*, 390-405.
26. Fletcher, P.C.; Frith, C.D. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat Rev Neurosci*, **2009**, *10*, 48-58.
27. Friston, K.J.; Stephan, K.E.; Montague, R.; Dolan, R.J. Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry*, **2014**, *1*, 148-158.
28. Linsker, R. Self-organization in a perceptual network. *Computer* **1988**, *21*, 105-117.
29. Linsker, R. Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Comput* **1992**, *4*, 691-702.
30. Lee, T.W.; Girolami, M.; Bell, A.J.; Sejnowski, T.J. A unifying information-theoretic framework for independent component analysis. *Comput Math Appl* **2000**, *39*, 1-21.
31. Simoncelli, E. P.; Olshausen, B. A. Natural image statistics and neural representation. *Ann Rev Neurosci* **2001**, *24*, 1193-1216.
32. Belouchrani, A.; Abed-Meraim, K.; Cardoso, J.F.; Moulines, E. A blind source separation technique using second-order statistics. *Signal Processing IEEE Trans on* **1997**, *45*, 434-444.
33. Choi, S.; Cichocki, A.; Park, H.M.; Lee, S.Y. Blind source separation and independent component analysis: A review. *Neural Information Processing-Letters and Reviews* **2005**, *6*, 1-57.
34. Cichocki, A.; Zdunek, R.; Phan, A.H.; Amari, S.I. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*; John Wiley & Sons, 2009.
35. Comon, P.; Jutten, C. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*; Academic Press, 2010.
36. Shannon, C.E.; Weaver, W. *The mathematical theory of communication*; University of Illinois Press, 1998.
37. Cover, T.M.; Thomas, J.A. *Elements of information theory* John Wiley & Sons, New York, NY, 1991.
38. Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*; San Fransisco, CA: Morgan Kaufmann, 1988.
39. Friston, K.J. Life as we know it. *J R Soc Interface*, **2013**, *10*, 20130475.
40. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann Math Stat*, **1951**, *22*, 79-86.
41. Arora, S.; Risteski, A. Provable benefits of representation learning. *arXiv*, arXiv:1706.04601.

42. Jaynes, E.T. Information theory and statistical mechanics. *Physical Review*, **1957**, *106*, 620-630.
43. Jaynes, E.T. Information theory and statistical mechanics. II. *Physical Review*, **1957**, *108*, 171-190.
44. Isomura, T.; Kotani, K.; Jimbo, Y. Cultured Cortical Neurons Can Perform Blind Source Separation According to the Free-Energy Principle. *PLoS Comput Biol* **2015**, *11*, e1004643.
45. Xu, L. Least mean square error reconstruction principle for self-organizing neural-nets. *Neural Netw* **1993**, *6*, 627-648.
46. Amari, S.I.; Cichocki, A.; Yang, H.H. A new learning algorithm for blind signal separation. *Adv Neural Inf Proc Sys* **1996**, *8*, 757-763.
47. Oja, E. Neural networks, principal components, and subspaces. *Int J Neural Syst* **1989**, *1*, 61-68.
48. Bell, A.J.; Sejnowski, T.J. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput* **1995**, *7*, 1129-1159.
49. Bell, A.J.; Sejnowski, T.J. The "independent components" of natural scenes are edge filters. *Vision Res* **1997**, *37*, 3327-3338.
50. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504-507.
51. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436.
52. Hyvärinen, A.; Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Netw* **1999**, *12*, 429-439.
53. Berkes, P.; Orbán, G.; Lengyel, M.; Fiser, J. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science* **2011**, *331*, 83-87.
54. Sengupta, B.; Stemmler, M.B.; Friston, K.J. Information and efficiency in the nervous system–a synthesis. *PLoS Comput Biol* **2013**, *9*, e1003157.
55. Frémaux, N.; Gerstner, W. Neuromodulated Spike-Timing-Dependent Plasticity, and Theory of Three-Factor Learning Rules. *Front Neural Circuits* **2016**, *9*, doi:10.3389/fncir.2015.00085.
56. Isomura, T.; Toyoizumi, T. A Local Learning Rule for Independent Component Analysis. *Sci Rep* **2016**, *6*, 28073.
57. Kuśmierz, Ł.; Isomura, T.; Toyoizumi T. Learning with three factors: modulating Hebbian plasticity with errors. *Curr Opin Neurobiol* **2017**, *46*, 170-177.
58. Isomura, T.; Toyoizumi, T. Error-gated Hebbian rule: a local learning rule for principal and independent component analysis. *Sci Rep* **2018**, *8*, 1835.
59. Isomura, T.; Sakai, K.; Kotani, K.; Jimbo, Y. Linking neuromodulated spike-timing dependent plasticity with the free-energy principle. *Neural Comput* **2016**, *28*, 1859-1888.
60. Markram, H.; Lübke, J.; Frotscher, M.; Sakmann, B. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* **1997**, *275*, 213-215.
61. Bi, G.Q.; Poo, M.M. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J Neurosci* **1998**, *18*, 10464-10472.
62. Karnani, M.; P a akk onen, K.; Annila, A. The physical character of information. *Proc R Soc Lond A Math Phys Eng Sci* **2009**, *465*, 2155-2175.
63. Annila, A. On the character of consciousness. *Front Sys Neurosci* **2016**, *10*, 27.