PERGAMON

Computers and Mathematics with Applications 39 (2000) 1–21

An International Journal
computers & mathematics
with applications

www.elsevier.nl/locate/camwa

A Unifying Information-Theoretic Framework for Independent Component Analysis

TE-WON LEE
Howard Hughes Medical Institute, Computational Neurobiology Laboratory
The Salk Institute, La Jolla, CA 92037, U.S.A.
and
Institut für Elektronik, Technische Universität Berlin
Einsteinufer 17, 10587 Berlin, Germany
tewon@salk.edu

M. GIROLAMI
Department of Computing and Information Systems
University of Paisley, Paisley PA1 2BE, Scotland, U.K.
giro0ci@paisley.ac.uk

A. J. BELL
Howard Hughes Medical Institute
Computational Neurobiology Laboratory
The Salk Institute, La Jolla, CA 92037, U.S.A.

T. J. SEJNOWSKI
Howard Hughes Medical Institute, Computational Neurobiology Laboratory
The Salk Institute, La Jolla, CA 92037, U.S.A.
and
Department of Biology, University of California, San Diego, La Jolla, CA 92093, U.S.A.
terry@salk.edu

Abstract—We show that different theories recently proposed for independent component analysis (ICA) lead to the same iterative learning algorithm for blind separation of mixed independent sources. We review those theories and suggest that information theory can be used to unify several lines of research. Pearlmutter and Parra [1] and Cardoso [2] showed that the infomax approach of Bell and Sejnowski [3] and the maximum likelihood estimation approach are equivalent. We show that negentropy maximization also has equivalent properties, and therefore, all three approaches yield the same learning rule for a fixed nonlinearity. Girolami and Fyfe [4] have shown that the nonlinear principal component analysis (PCA) algorithm of Karhunen and Joutsensalo [5] and Oja [6] can also be viewed from information-theoretic principles since it minimizes the sum of squares of the fourth-order marginal cumulants, and therefore, approximately minimizes the mutual information [7]. Lambert [8] has proposed different Bussgang cost functions for multichannel blind deconvolution. We show how the Bussgang property relates to the infomax principle. Finally, we discuss convergence and stability as well as future research issues in blind source separation. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords—Blind source separation, ICA, Entropy, Information maximization, Maximum likelihood estimation.

Lee was supported by the Office of Naval Research. Girolami was supported by a grant from NCR Financial Systems (Ltd), Knowledge Laboratory, Advanced Technology Development Division, Dundee, Scotland. Bell and Sejnowski were supported by the Howard Hughes Medical Institute and the Office of Naval Research.

0898-1221/00/$ - see front matter © 2000 Elsevier Science Ltd. All rights reserved. Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX
PII: S0898-1221(00)00101-2

1. INTRODUCTION

Recently, blind source separation by independent component analysis (ICA) has received attention because of its potential applications in signal processing such as in speech recognition systems, telecommunications, and medical signal processing. The goal of ICA is to recover independent sources given only sensor observations that are unknown linear mixtures of the unobserved independent source signals. In contrast to correlation-based transformations such as principal component analysis (PCA), ICA not only decorrelates the signals (2^{nd}-order statistics) but also reduces higher-order statistical dependencies, attempting to make the signals as independent as possible.

Two different research communities have considered the analysis of independent components. On one hand, the study of separating mixed sources observed in an array of sensors has been a classical and difficult signal processing problem. The seminal work on blind source separation was by Herault and Jutten [9] where they introduced an adaptive algorithm in a simple feedback architecture that was able to separate several unknown independent sources. Their approach has been further developed by Jutten and Herault [10], Karhunen and Joutsensalo [5], and Cichocki *et al.* [11]. Comon [7] elaborated the concept of independent component analysis and proposed cost functions related to the approximate minimization of mutual information between the sensors.

In parallel to blind source separation studies, unsupervised learning rules based on information theory were proposed by Linsker [12]. The goal was to maximize the mutual information between the inputs and outputs of a neural network. This approach is related to the principle of redundancy reduction suggested by Barlow [13] as a coding strategy in neurons. Each neuron should encode features that are as statistically independent as possible from other neurons over a natural ensemble of inputs; decorrelation as a strategy for visual processing was explored by Atick [14]. Nadal and Parga [15] showed that in the low-noise case, the maximum of the mutual information between the input and output of a neural network implied that the output distribution was factorial; that is, the multivariate probability density function (p.d.f.) can be factorized as a product of marginal p.d.f.s. Roth and Baram [16] and Bell and Sejnowski [3] independently derived stochastic gradient learning rules for this maximization and applied them, respectively, to forecasting, time series analysis, and the blind separation of sources. Bell and Sejnowski [3] put the blind source separation problem into an information-theoretic framework and demonstrated the separation and deconvolution of mixed sources. Their adaptive methods are more plausible from a neural processing perspective than the cumulant-based cost functions proposed by Comon [7]. A similar adaptive method for source separation was proposed by Cardoso and Laheld [17].

Other algorithms for performing ICA have been proposed from different viewpoints. Maximum likelihood estimation (MLE) approaches to ICA were first proposed by Gaeta and Lacoume [18] and elaborated by Pearlmutter and Parra [1]. Girolami and Fyfe [19,20], motivated by information-theoretic indices for exploratory projection pursuit (EPP), used marginal negentropy[1] as a projection index and showed that kurtosis-seeking projection pursuit will extract one of the underlying sources from a linear mixture. A multiple output EPP network was developed to allow full separation of all the underlying sources [20]. Nonlinear PCA algorithms for ICA which have been developed by Karhunen and Joutsensalo [5], Xu [22] and Oja [6] can also be viewed from the infomax principle since they approximately minimize the sum of squares of the fourth-order marginal cumulants [7] and therefore, approximately minimize the mutual information of the network outputs [4]. Bell and Sejnowski [3] have pointed out a similarity between their infomax algorithm and the Bussgang algorithm in signal processing and Lambert [8] elucidated the connection between three different Bussgang cost functions. We show here how the Bussgang property relates to the infomax principle and how all of these seemingly different approaches can

[1] A general term for negentropy is relative entropy [21].

be put into a unifying framework for the source separation problem based on an information theoretic approach.

The original infomax learning rule for blind separation by Bell and Sejnowski [3] was suitable for super-Gaussian sources. Girolami and Fyfe [19] derive, by choosing negentropy as a projection pursuit index, a learning rule that is able to blindly separate mixed sub- and super-Gaussian source distributions. Lee, Girolami and Sejnowski [23] show that the learning rule is an extension of the infomax principle satisfying a general stability criterion and preserving the simple architecture of Bell and Sejnowski [3]. When optimized using the natural gradient [24], or equivalently, the relative gradient [17], the learning rule gives superior convergence. Simulations and results on real-world physiological data show the power of the proposed methods [23].

This paper is organized as follows. In Section 2, we formulate the problem and the assumptions usually made in ICA. Section 3 reviews the infomax approach by Bell and Sejnowski [3]. Sections 4–8 describe, respectively, the relation between infomax, MLE, negentropy maximization, nonlinear PCA, higher-order statistics, and the Bussgang property. In Section 9, we discuss convergence properties and stability of the proposed algorithms. Potential applications and further research issues are discussed in Section 10.

2. PROBLEM STATEMENT AND ASSUMPTIONS

Assume that there is an M-dimensional zero mean vector $\mathrm{s}(t) = [s_1(t), \ldots, s_M(t)]^\top$, whose components are mutually independent. The vector $\mathrm{s}(t)$ corresponds to M independent scalar valued source signals $s_i(t)$. We can write the multivariate p.d.f. of the vector as the product of marginal independent distributions.

$$p(\mathrm{s}) = \prod_{i=1}^{M} p_i\left(s_i\right). \tag{1}$$

A data vector $\mathbf{x}(t) = [x_1(t), \ldots, x_N(t)]^\top$ is observed at each time point t, such that

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \tag{2}$$

where $\mathbf{A}$ is an $N \times M$ scalar matrix. The mixing is assumed to be instantaneous so there is no time-delay between the source i mixing into channel j. Generalizations to time-delayed and convolved sources are considered in the discussion. Instantaneous mixtures occur when the difference in time of arrival between the sensors can be neglected. As the components of the observed vectors are no longer independent, the multivariate p.d.f. will not satisfy the product equality in equation (1). The mutual information $I(\mathbf{x})$ of the observed vector is given by the Kullback-Leibler (KL) divergence $D(. \parallel .)$ of the multivariate density from the density written in product form

$$I(\mathbf{x}) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{\prod_{i=1}^{N} p_i(x_i)} \, d\mathbf{x} = D\left(p(\mathbf{x}) \,\Big\|\, \prod_{i=1}^{N} p_i\left(x_i\right)\right). \tag{3}$$

The mutual information is positive and is equal to zero only when the components x_i are independent [21].

The goal of ICA is to find a linear transformation $\mathbf{W}$ of the dependent sensor signals $\mathbf{x}$ that makes the outputs as independent as possible:

$$\mathbf{u}(t) = \mathbf{W}\mathbf{x}(t) = \mathbf{W}\mathbf{A}\mathbf{s}(t), \tag{4}$$

where $\mathbf{u}$ is an estimate of the sources. The sources are exactly recovered when $\mathbf{W}$ is the inverse of $\mathbf{A}$ up to a permutation and scale change.

$$\mathbf{P} = \mathbf{R}\mathbf{S} = \mathbf{W}\mathbf{A}, \tag{5}$$

where $\mathbf{R}$ is a permutation matrix and $\mathbf{S}$ is the scaling matrix. The two matrices define the performance matrix $\mathbf{P}$ so that if $\mathbf{P}$ is normalized and reordered, a perfect separation leads to the identity matrix. For the linear mixing and unmixing model, we adopt the following assumptions [7,17].

(1) The number of sensors is greater than or equal to the number of sources $N \geq M$.
(2) The sources $\mathbf{s}(t)$ are at each time instant mutually independent.
(3) At most one source is normally distributed.
(4) No sensor noise or only low additive noise signals are permitted.

Assumption 1 is needed to make $\mathbf{A}$ a full rank matrix. Assumption 2 is the basis of ICA and can be expressed as follows:

$$p(\mathbf{s}(t)) = \prod_{i=1}^{M} p(s_i(t)) . \tag{6}$$

For Assumption 3, the unmixing of two Gaussian sources is ill posed when the sources are white random processes. Nonwhite Gaussian processes may be recovered with time-decorrelation methods if they have different spectra [25]. However, pure Gaussian processes are rare in real data. Assumption 4 is necessary to satisfy the infomax condition, in which the mutual information between outputs is only minimized for the low noise case [12,15]. However, one can imagine that noise is an independent source itself and if as many sensor outputs are available as the number of sources, the noise signal can be segregated from the mixtures.

3. INFORMATION MAXIMIZATION

Nadal and Parga [15] showed that in the low-noise case, the maximum of the mutual information between the inputs $\mathbf{x}$ and outputs $\mathbf{y}$ of a neural processor implied that the output distributions were factorial. In other words, maximizing the information transfer in a nonlinear neural network minimizes the mutual information among the outputs (factorial code) when optimization is performed over both the synaptic weights $\mathbf{W}$ and the nonlinear transfer function $g(\mathbf{u})$. Roth and Baram [16] and Bell and Sejnowski [3] independently derived stochastic gradient learning rules for this maximization and applied them, respectively, to forecasting, time series analysis, and the blind separation of sources. Bell and Sejnowski [3] proposed a simple learning algorithm for a feedforward neural network that blindly separates linear mixtures $\mathbf{x}$ of independent sources $\mathbf{s}$ using information maximization. They show that maximizing the joint entropy $H(\mathbf{y})$ of the output of a neural processor can approximately minimize the mutual information among the output components $y_i = g(u_i)$ where $g(u_i)$ is an invertible monotonic nonlinearity and $\mathbf{u} = \mathbf{Wx}$.

The joint entropy at the outputs of a neural network is

$$H(y_1, \ldots, y_N) = H(y_1) + \cdots + H(y_N) - I(y_1, \ldots, y_N), \tag{7}$$

where $H(y_i)$ are the marginal entropies of the outputs and $I(y_1, \ldots, y_N)$ is their mutual information. Maximizing $H(y_1, \ldots, y_N)$ consists of maximizing the marginal entropies and minimizing the mutual information. The outputs $\mathbf{y}$ are amplitude-bounded random variables, and therefore, the marginal entropies are maximum for a uniform distribution of y_i. Maximizing the joint entropy will also decrease $I(y_1, \ldots, y_N)$ since the mutual information is always positive. For $I(y_1, \ldots, y_N) = 0$, the joint entropy is the sum of marginal entropies

$$H(y_1, \ldots, y_N) = H(y_1) + \cdots + H(y_N). \tag{8}$$

The maximal value for $H(y_1, \ldots, y_N)$ is achieved when the mutual information among the bounded random variables $y_1, \ldots, y_N$ is zero and their marginal distribution is uniform. As we will show below, this implies that the nonlinearity $g(u_i)$ has the form of the cumulative density function (c.d.f.) of the true source distribution s_i. Bell and Sejnowski [3] chose the nonlinearity to be a fixed logistic function. This is equivalent to assuming a prior distribution of the sources:

a super-Gaussian distribution with heavy tails and a peak centered at the mean. The weights $\mathbf{W}$ are determined by maximizing the joint entropy with respect to $\mathbf{W}$. We can rewrite the derivative of equation (7) with respect to $\mathbf{W}$ can be written in terms of the KL divergence between the multivariate uniform distribution denoted as $p_1(\mathbf{y})$ and multivariate uniform estimate $p(\mathbf{y})$.

$$\frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} = \frac{\partial}{\partial \mathbf{W}} \left(-D\left(p_1(\mathbf{y}) \parallel p(\mathbf{y})\right)\right). \tag{9}$$

In the limit when the transfer function $g(u_i)$ and $\mathbf{W}$ are optimized, the joint entropy $H(\mathbf{y})$ is maximum and $p(\mathbf{y}) = p_1(\mathbf{y})$ so that $I(\mathbf{y}) = 0$. If $g(u_i)$ is an invertible mapping from u_i to y_i, the KL divergence in equation (9) is equal to the KL divergence between the estimate of the source distribution $p(\mathbf{u})$ and the sources $p(\mathbf{s})$,

$$D\left(p_1(\mathbf{y}) \parallel p(\mathbf{y})\right) = D\left(p(\mathbf{s}) \parallel p(\mathbf{u})\right), \tag{10}$$

since the KL divergence is invariant under an invertible transformation. If the mutual information between the outputs is zero $I(y_1, \ldots, y_N) = 0$, the mutual information before the nonlinearity $I(u_1, \ldots, u_N)$ must also be zero since the nonlinearity does not introduce any dependencies. The relation between y_i and u_i is [26]

$$p\left(y_i\right) = \frac{p\left(u_i\right)}{\left|\frac{\partial g(u_i)}{\partial u_i}\right|}. \tag{11}$$

For a uniform distribution of y_i, it follows that

$$p\left(u_i\right) = \left|\frac{\partial g\left(u_i\right)}{\partial u_i}\right|. \tag{12}$$

This means that u_i is an independent variable with a distribution that is approximately the form of the derivative of the nonlinearity. In the case of the logistic function, the appropriate p.d.f. is shown in Figure 1 (bottom). The distributions of music and speech signals can be approximated by this p.d.f.

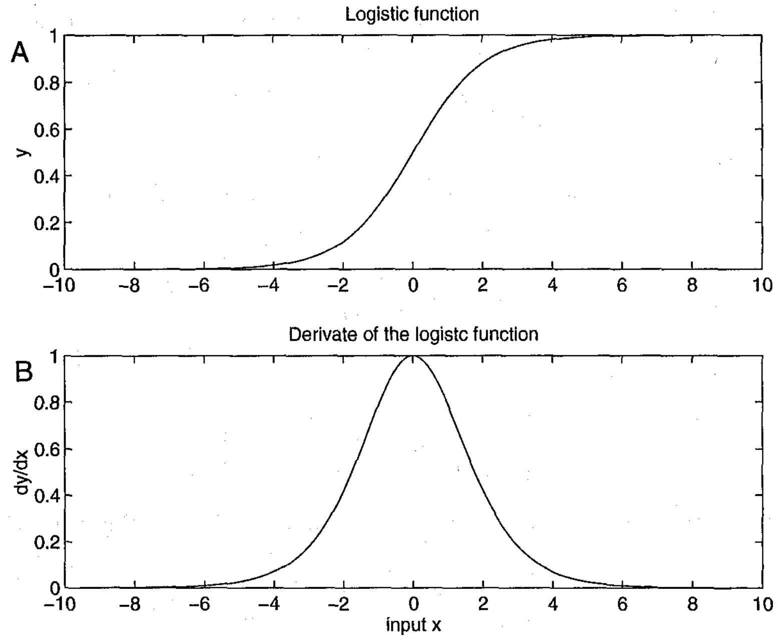

Figure 1. (a) logistic function ($y = 1/(1 + \exp(-x))$), and (b) its derivative ($\frac{\partial y}{\partial x} = y(1-y)$).

Bell and Sejnowski [3] separated mixtures of several music and speech signals using infomax with a logistic activation function. Will infomax always minimize the mutual information? Bell and Sejnowski [3] answer this question in a thought experiment where they illustrate that when there is a mismatch between the source p.d.f. and the slope of the nonlinearity, a maximal joint entropy value can be achieved with $I(\mathbf{y}) > 0$ that is higher than the joint entropy with $I(\mathbf{y}) = 0$ (due to lower marginal entropies). In those cases, infomax will not minimize the mutual information. This occurs when there is an excessive mismatch between the nonlinearity and cumulative density function (c.d.f.) of the true source distribution.

A simple architecture that can realize the mapping from $\mathbf{x}$ to $\mathbf{y}$ is a single-layer feedforward neural network with a nonlinear output activation function. The nonlinearity $g_i(u)$ is essential for minimizing the mutual information to perform ICA. Another motivation for the choice of $g_i(u)$, e.g., a sigmoid function, is that it provides a combination of higher-order statistics through its Taylor series expansion that is essential to minimize higher-order correlations. The learning rule can be derived by maximizing the output entropy $H(\mathbf{y})$ of a neural processor, as proposed by Bell and Sejnowski [3]. We can relate $p(\mathbf{x})$ to $p(\mathbf{y})$ by the determinant of the Jacobian matrix $\mathbf{J}(\mathbf{x})$ (see [26])

$$p(\mathbf{y}) = \frac{p(\mathbf{x})}{|\det \mathbf{J}(\mathbf{x})|}. \tag{13}$$

See Figure 1 in [3] for a visual interpretation of the optimal information flow. Evaluating the expected value of the logarithmic representation for equation (13) gives the output entropy $H(\mathbf{y})$:

$$H(\mathbf{y}) = -E\{\log p(\mathbf{y})\} = E\{\log|\det \mathbf{J}(\mathbf{x})|\} - E\{\log p(\mathbf{x})\}. \tag{14}$$

This can be maximized with respect to $\mathbf{W}$ and it is equivalent to maximizing the absolute value of the Jacobian determinant of the transfer function

$$\frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} = \frac{\partial}{\partial \mathbf{W}} \log|\det(\mathbf{W})| + \frac{\partial}{\partial \mathbf{W}} \log \prod_{i=1}^{N} \left|\frac{\partial y_i}{\partial u_i}\right|. \tag{15}$$

For the first term in equation (15): $\frac{\partial}{\partial \mathbf{W}} \log|\det(\mathbf{W})| = \mathbf{W}^{-\mathsf{T}}$. In the second term, the product splits up into sums of log-terms, in which only one is dependent on a particular W_{ij}, and hence,

$$\frac{\partial}{\partial \mathbf{W}} \log \prod_{i=1}^{N} \left|\frac{\partial y_i}{\partial u_i}\right| = -\varphi(\mathbf{u})\mathbf{x}^{\mathsf{T}}, \tag{16}$$

where $\varphi(\mathbf{u})$ is the gradient vector of the log likelihood called the score function [27]

$$\varphi(\mathbf{u}) = -\frac{\frac{\partial p(\mathbf{u})}{\partial \mathbf{u}}}{p(\mathbf{u})} = \left[-\frac{\frac{\partial p(u_1)}{\partial u_1}}{p(u_1)}, \dots, -\frac{\frac{\partial p(u_N)}{\partial u_N}}{p(u_N)}\right]^{\mathsf{T}}. \tag{17}$$

The general learning rule is now

$$\frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} = \left[\left(\mathbf{W}^{\mathsf{T}}\right)^{-1} - \varphi(\mathbf{u})\mathbf{x}^{\mathsf{T}}\right]. \tag{18}$$

An efficient way to maximize the joint entropy is to follow the 'natural' gradient

$$\Delta \mathbf{W} \propto \frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} \mathbf{W}^{\mathsf{T}}\mathbf{W} = \left[\mathbf{I} - \varphi(\mathbf{u})\mathbf{u}^{\mathsf{T}}\right] \mathbf{W}, \tag{19}$$

as proposed by Amari *et al.* [28], or equivalently the relative gradient [17]. Here $\mathbf{W}^{\mathsf{T}}\mathbf{W}$ rescales the gradient, simplifies the learning rule in equation (18) and speeds convergence considerably.

As we show here, the general learning algorithm in equation (19) can be derived from several theoretical viewpoints such as MLE [1], infomax [3], and negentropy maximization [4].

An elegant way of parameterizing the learning rule in equation (19) to separate mixed sub- and super-Gaussians has been proposed by Girolami [29] and Girolami and Fyfe [19] by choosing negentropy as a projection pursuit index. In [29], a parametric density model is employed for sub- and super-Gaussian sources resulting in a simple form for $\varphi_i(u_i)$:

$$\varphi_i(u_i) = \begin{cases} u_i + \tanh(u_i), & \text{super-Gaussian,} \\ u_i - \tanh(u_i), & \text{sub-Gaussian,} \end{cases} \tag{20}$$

giving

$$\Delta \mathbf{W} \propto \left[\mathbf{I} - \mathbf{K}\tanh(\mathbf{u})\mathbf{u}^T - \mathbf{u}\mathbf{u}^T\right]\mathbf{W}, \tag{21}$$

where $\mathbf{K}$ is a diagonal matrix with elements $\operatorname{sign}(k_4(u_i))$ and $k_4(u_i)$ is the kurtosis of the source estimate u_i. An extended infomax algorithm also yields the learning rule in equation (21) where $\mathbf{K}$ is a function of the nonlinearity used in $\varphi(\mathbf{u})$ satisfying a general stability criterion [30,31], presented in Section 9.1.

4. NEGENTROPY MAXIMIZATION

Another approach related to minimizing the mutual information between the u_is is maximizing negentropy [32]. Girolami and Fyfe [19,20], motivated by information-theoretic indices for exploratory projection pursuit (EPP), used marginal negentropy as a projection index. EPP is a statistical method that allows structure in high-dimensional data to be identified [33]. This is achieved by projecting the data onto a low-dimensional subspace and searching for structure in the projection. Projections that identify non-Gaussian structure such as multiple modes are interesting from the point of view of identifying potential higher-order structure within high-dimensional data. Projections that are maximally non-Gaussian are highly desirable in pursuing informative views of the data [33]. Girolami [32] showed that if the observed data fits a latent variable model [34], which conforms to the deterministic ICA mixing model, then a kurtosis-seeking projection pursuit will extract one of the underlying sources. A multiple output EPP network was also developed to allow full separation of all the underlying sources [32]. Marriot in [35] noted that approximately symmetrical and almost Gaussian (low kurtosis) clustered projections can sometimes be difficult to identify with indices based on third- and fourth-order moments and suggested the use of indices based on information theoretic criteria. Girolami [32] developed single and multiple output algorithms for EPP based on negentropy maximization. He showed that a negentropy maximizing pursuit will perform a general ICA on sources which may be either sub- or super-Gaussian. The negentropy of the output neurons can be stochastically maximized by driving their distributions maximally away from Gaussian distributions. Girolami [32] showed that maximizing the output data negentropy is identical to minimizing the mutual information of the output data which has been shown to be equivalent to ICA for observed data that can be modeled as a sum of independent latent variables. A brief derivation follows.

Negentropy is defined as the KL divergence between $p(\mathbf{u})$ and the Gaussian distribution $p_G(\mathbf{u})$ with the same mean and covariance as $p(\mathbf{u})$ (see [21])

$$J(\mathbf{u}) = D\left(p(\mathbf{u}) \parallel p_G(\mathbf{u})\right) = \int p(\mathbf{u}) \log \frac{p(\mathbf{u})}{p_G(\mathbf{u})}\, d\mathbf{u}, \tag{22}$$

where $\mathbf{u}$ is the vector of estimated sources given the parameters $\mathbf{W}$ ($\mathbf{u} = \mathbf{W}\mathbf{x}$). The parametric form of the output is factorable $\prod_{i=1}^N p(u_i)$ with the equality $p(\mathbf{u}) = \prod_{i=1}^N p(u_i)$ holding only when all u_is are independent, i.e., the mutual information is zero ($I(\mathbf{u}) = 0$). Assume that $\mathbf{u}$ is decorrelated and that u_is are factorable but not factorized ($J(\mathbf{u}) \neq \sum_{i=1}^{N} J(u_i)$).

$$\sum_{i=1}^{N} J(u_i) = \sum_{i=1}^{N} D\left(p(u_i) \,\|\, p_G(u_i)\right) \tag{23}$$

$$= \int p(u_1) \log \frac{p(u_1)}{p_G(u_1)}\, du_1 + \cdots + \int p(u_N) \log \frac{p(u_N)}{p_G(u_N)}\, du_N \tag{24}$$

$$= \int p(\mathbf{u}) \log \frac{\prod_{i=1}^{N} p(u_i)}{\prod_{i=1}^{N} p_G(u_i)}\, d\mathbf{u} \tag{25}$$

$$= \int p(\mathbf{u}) \log \frac{\prod_{i=1}^{N} p(u_i)}{p_G(\mathbf{u})}\, d\mathbf{u} \tag{26}$$

$$= \int p(\mathbf{u}) \log \frac{\prod_{i=1}^{N} p(u_i)}{p(\mathbf{u})}\, d\mathbf{u} + \int p(\mathbf{u}) \log \frac{p(\mathbf{u})}{p_G(\mathbf{u})}\, d\mathbf{u} \tag{27}$$

$$= D\left(\prod_{i=1}^{N} p(u_i) \,\|\, p(\mathbf{u})\right) + J(\mathbf{u}) \tag{28}$$

$$[illegible]) + J(\mathbf{u}). \tag{29}$$

[illegible] be written as a sum of KL divergences. The substitution $p_G(\mathbf{u}) =$ [illegible] follows from the assumption that $\mathbf{u}$ is decorrelated. The first term [illegible] mutual information ($-I(\mathbf{u})$). The second term can be [illegible]

$$[illegible] \int p(\mathbf{u}) \log p_G(\mathbf{u})\, d\mathbf{u} \tag{30}$$

$$[illegible] \mathbf{V})|) - \frac{1}{2} \log\left((2\pi e)^N \det\left(\langle \mathbf{u}\mathbf{u}^{\mathsf{T}} \rangle\right)\right). \tag{31}$$

[illegible] the equality of equations (30) and (31). First, [illegible] $\det(\mathbf{W})|)$ because of the p.d.f. transformation [illegible] $\int p(\mathbf{u}) \log p_G(\mathbf{u})\, d\mathbf{u}$ is the entropy [illegible] n $p(\mathbf{u})$ and $p_G(\mathbf{u})$ yield the same [illegible] correlated, its covariance matrix is [illegible] follows that

$$[illegible] \tag{32}$$

[illegible]

$$[illegible] \tag{33}$$

[illegible] matrix are not [illegible] entropies with

$$[illegible] \tag{34}$$

This leads to exactly the same learning rule as in Section 3.3 using infomax. Maximizing $\sum_{i=1}^{N} J(u_i)$ with respect to $\mathbf{W}$ in equation (31) gives

$$\begin{aligned}\frac{\partial}{\partial \mathbf{W}} \sum_{i=1}^{N} J(u_i) &= \frac{\partial}{\partial \mathbf{W}} \left[\int p(\mathbf{u}) \log \left(\prod_{i=1}^{N} p(u_i) \right) d\mathbf{u} + \frac{1}{2} \log \left((2\pi e)^N \right) \right] \\ &= \frac{\partial}{\partial \mathbf{W}} \left[E\left\{ \log \left(\prod_{i=1}^{N} p(u_i) \right) \right\} + \log(\det(\mathbf{W})) + \frac{1}{2} \log \left((2\pi e)^N \right) \right].\end{aligned} \tag{35}$$

Note that as in equation (19), only the first and second terms in equation (36) depend on $\mathbf{W}$:

$$\frac{\partial}{\partial \mathbf{W}} \sum_{i=1}^{N} J(u_i) = \mathbf{W}^{-\mathrm{T}} + \left(\frac{\frac{\partial p(\mathbf{u})}{\partial \mathbf{u}}}{p(\mathbf{u})} \right) \mathbf{x}^{\mathrm{T}}. \tag{36}$$

Although the derivation of the learning rule in equation (36) depends on the assumption that $\mathbf{u}$ is decorrelated, Girolami [32] showed that a slightly different objective function related to maximizing the marginal negentropies leads to the same learning algorithm in equation (36) without making the assumption that $\mathbf{u}$ is decorrelated.

5. MAXIMUM LIKELIHOOD ESTIMATION

The goal of MLE is to model the observation $\mathbf{x}$ as being generated from latent variables $\mathbf{s}$ via a linear mapping $\mathbf{A}$. In the noiseless case, we can use a parametric density estimator $\hat{p}(\mathbf{x}; \mathbf{a})$ to find the parameter vector $\mathbf{a}$ that minimizes the difference between the generative model $\hat{p}(\mathbf{x}; \mathbf{a})$ and the observed distribution $p(\mathbf{x})$. Note that $\mathbf{a}$ can be considered the basis vectors of $\mathbf{A}$ so that $\hat{p}(\mathbf{x}; \mathbf{a})$ is an estimate of the observed vector $p(\mathbf{x})$. The difference between the estimate and the observation can be measured using the KL divergence

$$D(p(\mathbf{x}), \hat{p}(\mathbf{x}; \mathbf{a})) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{\hat{p}(\mathbf{x}; \mathbf{a})} d\mathbf{x} = H(\mathbf{x}) - \int p(\mathbf{x}) \log \hat{p}(\mathbf{x}; \mathbf{a}), \tag{37}$$

where $p(\mathbf{x})$ is the p.d.f. of the observation $\mathbf{x}$ and $\hat{p}(\mathbf{x}; \mathbf{a})$ is a parametric estimate of the distribution $p(\mathbf{x})$. The divergence $D(p(\mathbf{x}) \parallel \hat{p}(\mathbf{x}; \mathbf{a}))$ is zero only if our estimate $\hat{p}(\mathbf{x}; \mathbf{a})$ matches the observation $p(\mathbf{x})$. Pearlmutter and Parra [1] and Cardoso [2] showed that infomax and MLE are equivalent for ICA, as briefly described here. The normalized log-likelihood of $\hat{p}(\mathbf{x}; \mathbf{a})$ is

$$L(\mathbf{a}) = \frac{1}{N} \sum_{i=1}^{N} \log \hat{p}(\mathbf{x}_i; \mathbf{a}), \tag{38}$$

where N is the number of independent realizations of $\mathbf{x}$. The log-likelihood converges in probability, by the law of large numbers, to its expectation

$$L(\mathbf{a}) = \int p(\mathbf{x}) \log \hat{p}(\mathbf{x}; \mathbf{a})\, d\mathbf{x}. \tag{39}$$

Note that this can be rewritten

$$\begin{aligned}L(\mathbf{a}) &= -\int p(\mathbf{x}) \log p(\mathbf{x})\, d\mathbf{x} - \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{\hat{p}(\mathbf{x}; \mathbf{a})} d\mathbf{x} \\ &= H(\mathbf{x}) - D(p(\mathbf{x}) \parallel \hat{p}(\mathbf{x}; \mathbf{a})).\end{aligned} \tag{40}$$

Since $H(\mathbf{x})$ is not dependent on $\mathbf{W}$, maximizing the log-likelihood minimizes the KL divergence between the observed density $p(\mathbf{x})$ and the estimated density $\hat{p}(\mathbf{x}; \mathbf{a})$,

$$\frac{\partial L(\mathbf{a})}{\partial \mathbf{W}} = -\frac{\partial}{\partial \mathbf{W}} D(p(\mathbf{x}) \parallel \hat{p}(\mathbf{x}; \mathbf{a})). \tag{41}$$

Since $\mathbf{A}$ is an invertible matrix and the KL divergence is invariant under an invertible transformation, minimizing the KL divergence in equation (41) minimizes the KL divergence between the estimate of the sources $p(\mathbf{u})$ and the true source distribution $p(\mathbf{s})$.

$$\frac{\partial L(\mathbf{a})}{\partial \mathbf{W}} = -\frac{\partial}{\partial \mathbf{W}} D(p(\mathbf{s}) \parallel \hat{p}(\mathbf{u})). \tag{42}$$

Therefore, equation (42) and equation (9) are equivalent for ICA.

6. HIGHER-ORDER MOMENTS AND CUMULANTS

In the previous sections, the nonlinearity of the output approximated the c.d.f. of the true source density. Here we examine cumulants to study the higher-order correlations between the sources.

If the observed vector has a covariance matrix $\langle \mathbf{x}\mathbf{x}^{\mathsf{T}} \rangle = E\{\mathbf{x}\mathbf{x}^{\mathsf{T}}\}$, then the mutual information in equation (3) can be expressed as [7]

$$I(\mathbf{x}) = J(\mathbf{x}) - \sum_{i=1}^{N} J(x_i) + \frac{1}{2} \log \frac{\left(\prod_{i=1}^{N} \langle x_i^2 \rangle \right)}{\det(\langle \mathbf{x}\mathbf{x}^{\mathsf{T}} \rangle)}, \tag{43}$$

where $\langle x_i^2 \rangle$ in equation (43) are the diagonal elements of the covariance matrix. $J(\mathbf{x})$ is the multivariate negentropy as in equation (22) and $J(x_i)$ are the marginal negentropies

$$J(x_i) = \int p(x_i) \log \frac{p(x_i)}{p_G(x_i)}\, dx_i. \tag{44}$$

If a spatial whitening transformation (diagonalization of the covariance matrix) is used to remove the second-order redundancy in the data $\tilde{\mathbf{x}} = \mathbf{V}\mathbf{x}$, where $\mathbf{V}$ denotes the whitening transformation matrix and $\langle \tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathsf{T}} \rangle = \mathbf{I}$, then $\det(\langle \tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathsf{T}} \rangle) = 1$ and the mutual information of the spatially white data reduces to

$$I(\tilde{\mathbf{x}}) = J(\tilde{\mathbf{x}}) - \sum_{i=1}^{N} J_i(\tilde{x}_i). \tag{45}$$

A further transformation $\mathbf{u} = \mathbf{W}\tilde{\mathbf{x}}$ using higher-order correlations is required to reduce the remaining redundancy within the vector for non-Gaussian sources. This transformation seeks an orthogonal matrix that accounts for the correct rotation of the data. Comon [7] minimized the degree of dependence among outputs using contrast functions approximated by the Edgeworth expansion of the KL divergence. He determined the orthogonal matrix from the higher-order cumulants. Note that cumulants are used to describe characteristics of non-Gaussian processes. The truncated Edgeworth expansion [36] of $p(u_i)$ written in terms of its n^{th}-order cumulants and Hermite polynomials, denoted as k_n and h_n, respectively, is

$$\begin{aligned} p(u_i) = p_G(u_i) \Bigg[1 &+ \frac{1}{3!} k_3 h_3(u_i) + \frac{1}{4!} k_4 h_4(u_i) + \frac{10}{6!} k_4^2 h_6(u_i) \\ &+ \frac{1}{5!} k_5 h_5(u_i) + \frac{35}{7!} k_3 k_4 h_7(u_i) + \frac{280}{9!} k_3^3 h_9(u_i) \\ &+ \frac{1}{5!} k_5 h_5(u_i) + \frac{56}{8!} k_3 k_5 h_8(u_i) + \frac{35}{8!} k_4^2 h_8(u_i) \\ &+ \frac{2100}{10!} k_3^2 k_4 h_{10}(u_i) + \frac{15400}{12!} k_3^4 h_{12}(u_i) \Bigg], \end{aligned} \tag{46}$$

where $p_G(u_i)$ denotes the Gaussian density. The cumulants k_n are coefficients related to the form of the p.d.f. of u_i, and they can be expressed in terms of moments. The terms $h_k(u_i)$ are the orthogonal Hermite polynomials defined as [36]

$$(-1)^k \frac{\partial^k p_G(u_i)}{\partial u^k} = h_k(u_i)\, p_G(u_i), \tag{47}$$

which can be computed recursively

$$\begin{aligned} h_0(u_i) &= 1, \\ h_k(u_i) &= u_i h_{k-1} - (k-1) h_{k-2}. \end{aligned} \tag{48}$$

The validity of the truncated series expansion approximation in equation (46) is discussed in [36]. Expansion terms higher than fourth order can lead to excessive fluctuations at the tails of the distribution leading potentially to negative values. Therefore, the expansion in equation (46) is truncated at fourth order. After substituting the expression for marginal negentropies $J(u_i)$ in equation (44) into equation (46) [7], $J(u_i)$ becomes

$$J(u_i) \cong \frac{1}{12}k_3^2(i) + \frac{1}{48}k_4^2(i) + \frac{7}{48}k_3^4(i) + \frac{1}{8}k_3^2(i)k_4(i). \tag{49}$$

If we make the assumption that the p.d.f. of the signals under consideration are approximately symmetric, then the third-order cumulants will have a negligible contribution in equation (49). The mutual information in equation (43) of the transformed data $\mathbf{u}$ is now approximated by

$$I(\mathbf{u}) \cong J(\mathbf{u}) - \frac{1}{48}\sum_{i=1}^{N} k_4^2(i). \tag{50}$$

$J(\mathbf{u})$ is invariant under an orthogonal transformation

$$\begin{aligned}
J(\mathbf{u}) &= \int p(\mathbf{u}) \log \frac{p(\mathbf{u})}{p_G(\mathbf{u})}\, d\mathbf{u} \\
&= H(\mathbf{u}) - \frac{1}{2}\log\left((2\pi e)^N \det\left(\langle \mathbf{u}\mathbf{u}^\top \rangle\right)\right) \\
&= H(\tilde{\mathbf{x}}) + \log(\det(\mathbf{W})) - \frac{1}{2}\log\left((2\pi e)^N \det\left(\mathbf{W}\langle \tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top \rangle \mathbf{W}^\top\right)\right) \\
&= H(\tilde{\mathbf{x}}) - \frac{1}{2}\log\left((2\pi e)^N \det\left(\langle \tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top \rangle\right)\right) \\
&= H(\tilde{\mathbf{x}}) - H_G(\tilde{\mathbf{x}}) = J(\tilde{\mathbf{x}}),
\end{aligned} \tag{51}$$

where $H_G(\tilde{\mathbf{x}})$ is the entropy of a normal density, the following matrix determinant equalities have been employed:

$$\det\left(\mathbf{W}\langle \tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top \rangle \mathbf{W}^\top\right) = \det(\mathbf{W})\det\left(\langle \tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top \rangle\right)\det\left(\mathbf{W}^\top\right), \tag{52}$$

$$\det\left(\mathbf{W}^\top\right) = \det(\mathbf{W}). \tag{53}$$

Since $\mathbf{u}$ is a result of a rotation of $\tilde{\mathbf{x}}$, the negentropy $J(\mathbf{u})$ is equal to $J(\tilde{\mathbf{x}})$ and the approximation for mutual information can be rewritten as

$$I(\mathbf{u}) \cong J(\tilde{\mathbf{x}}) - \frac{1}{48}\sum_{i=1}^{N} k_4^2(i). \tag{54}$$

Thus, under an orthogonal transformation, the mutual information of the data can be approximately minimized by maximizing the sum of squares of the fourth-order marginal cumulants. Maximizing the contrast function is approximately equivalent to maximizing the sum of marginal negentropies. This corroborates the claim that maximizing the marginal negentropies with respect to $\mathbf{W}$ minimizes mutual information.

$$\frac{\partial}{\partial \mathbf{W}} I(\mathbf{u}) \cong \frac{\partial}{\partial \mathbf{W}}\left(-\sum_{i=1}^{N} k_4^2(i)\right). \tag{55}$$

Therefore, Comon [7] proposed the following contrast function:

$$\Phi_{\max} = \sum_{i=1}^{N} k_4^2(i). \tag{56}$$

Here, the higher-order statistics are approximated by cumulants up to 4th-order and their maximization requires intensive computation using a batch-based method.

Simulation results in [37,38] and results on experimental data in [39] indicate that 4th-order cumulants are not sufficient to completely separate mixtures of several sources (e.g., $N > 5$). This suggests that a correct rotation of the whitened data requires more than 4th-order statistics with increasing number of sources. This is consistent with the assumption that the Taylor expansion of the nonlinear function used in infomax, negentropy maximization, and MLE provides statistics higher than fourth order necessary to make the data as independent as possible.

7. NONLINEAR PCA

The nonlinear extension of Oja's principle component analysis (PCA) subspace network [40], originally developed by Karhunen and Joutsensalo [5] and Xu [22], has no apparent connection to the infomax principle, but has been shown to separate whitened linear mixtures of sources [41–43]. A major shortcoming of the algorithm is that is has been restricted to the separation of sub-Gaussian sources, because of stability requirements. Another property is that the data have to be prewhitened. Those two characteristics have led Girolami and Fyfe [4] to relate the nonlinear PCA algorithm to the infomax principle showing that it is an approximate online adaptive equivalent of the batch algorithm proposed by Comon [7].

In this section, we summarize the results in [4] and their generalization to cope with sub- and super-Gaussian source distributions. It is an alternative form of the nonlinear PCA rule which satisfies the dynamic and asymptotic stability criteria for the algorithm [4].

In nonlinear PCA, the input signals $\mathbf{x}$ are first prewhitened giving $\tilde{\mathbf{x}}$, where $\langle \tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top \rangle = \mathbf{I}$. The learning rule is an approximate stochastic gradient descent algorithm that minimizes the mean-squared error incurred in representing a vector by a nonlinear projection $f(\mathbf{W}\tilde{\mathbf{x}})$ onto a basis of reduced dimensionality

$$\tilde{\mathbf{x}} = \tilde{\mathbf{x}}' + e = \mathbf{W} f\left(\mathbf{W}\tilde{\mathbf{x}}\right) + e, \tag{57}$$

where $\tilde{\mathbf{x}}'$ is a nonlinear estimate of $\tilde{\mathbf{x}}$ and e denotes the estimation error. Next we minimize a cost function $C(\mathbf{W})$ to find a linear transformation $\mathbf{W}$ giving $\mathbf{u} = \mathbf{W}\tilde{\mathbf{x}}$, where $\mathbf{u}$ are the estimated sources and $\mathbf{W}$ is constrained to be orthonormal $\mathbf{W}^\top\mathbf{W} = \mathbf{I}$.

$$C(\mathbf{W}) = \mathbf{1}^\top E\left\{\left(\tilde{\mathbf{x}} - \mathbf{W} f\left(\mathbf{W}\tilde{\mathbf{x}}\right)\right)^2\right\}, \tag{58}$$

where $C(\mathbf{W})$ is a scalar resulting from an inner product and $\mathbf{1}^\top$ is a row vector of length N with ones as its elements. Rewriting equation (58) in its transpose form gives

$$C(\mathbf{W}) = E\left\{\left(\tilde{\mathbf{x}}^\top\tilde{\mathbf{x}} - f^\top\left(\mathbf{W}\tilde{\mathbf{x}}\right) - \tilde{\mathbf{x}}^\top\mathbf{W} f\left(\mathbf{W}\tilde{\mathbf{x}}\right) + f^\top\left(\mathbf{W}\tilde{\mathbf{x}}\right)\mathbf{W}^\top\mathbf{W} f\left(\mathbf{W}\tilde{\mathbf{x}}\right)\right)\right\}. \tag{59}$$

Since the observed data is spatially white, it follows that $E\{\mathbf{W}^\top\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top\mathbf{W}\} = \mathbf{I}$. We assume unit variance for the independent components u_i. We can rewrite the cost function as

$$C(\mathbf{W}) = N + E\left\{\left(-f^\top(\mathbf{u})\mathbf{u} - \mathbf{u}^\top f(\mathbf{u}) + f^\top(\mathbf{u}) f(\mathbf{u})\right)\right\}. \tag{60}$$

For a polynomial such as $(f(\mathbf{u}) = \mathbf{u}^3/3)$ or $(f(\mathbf{u}) = -\mathbf{u}^3/3)$ or a hyperbolic nonlinear function which has a cubic as the dominating element, we can write for $f(\mathbf{u}) = \mathbf{u}^3/3$,

$$C(\mathbf{W}) \cong N + E\left\{-\left(\frac{\mathbf{u}^3}{3}\right)^\top \mathbf{u} - \mathbf{u}^\top\frac{\mathbf{u}^3}{3} + \left(\frac{\mathbf{u}^3}{3}\right)^\top \frac{\mathbf{u}^3}{3}\right\}. \tag{61}$$

Now the rightmost term $(\mathbf{u}^3)^\top\mathbf{u}^3/9$ can be neglected as $(2/3)\mathbf{u}^\top\mathbf{u}^3 \gg (\mathbf{u}^3)^\top\mathbf{u}^3/9$ is satisfied for white standardized data [29]. The cost function can be rewritten as

$$C(\mathbf{W}) \cong N - \frac{2}{3}E\left\{\sum_{i=1}^{N} u_i^4\right\} = (N-2) - \frac{2}{3}\left(E\left\{\sum_{i=1}^{N} u_i^4\right\} - 3\right), \tag{62}$$

where N is the number of sources and the term $(E\{u_i^4\} - 3)$ is the expression for the fourth-order marginal cumulant (unnormalized kurtosis). Hence, for spatially white standardized data, the cost function can be considered as the negative sum of the marginal fourth-order cumulants of the linearly transformed data.

$$C(\mathbf{W}) \cong -\sum_{i=1}^{N} k_4(i). \tag{63}$$

Minimizing the cost function in equation (63) is equivalent to maximizing the sum of fourth-order cumulants when the kurtosis of the estimated sources is positive (super-Gaussian). Optimization of equation (63) with respect to $\mathbf{W}$ is equivalent to maximization of the sum of squares of the marginal fourth-order cumulants, for mixtures of strictly super-Gaussian sources. The function

$$\Phi_{\max} = \sum_{i=1}^{N} k_4^2(i) \tag{64}$$

is equivalent to Comon's contrast function in equation (56). Comon [7] has shown that maximizing this contrast function approximately minimizes the mutual information.

Let us consider the case when the activation function is $f(\mathbf{u}) = -\mathbf{u}^3/3$. Applying the same reduction as above, the cost function has the following form:

$$C(\mathbf{W}) \cong \sum_{i=1}^{M} k_4(i). \tag{65}$$

Minimizing the cost function in equation (65) is equivalent to maximizing the sum of fourth-order cumulants when the kurtosis of the estimated sources is negative (sub-Gaussian). Hence, the contrast function is the same as in equation (64), but for a different nonlinear term. The negatively cubic term can be interpreted as accounting for a different prior on the source distribution. We can summarize the differences in the learning rules in equations (63) and (65) and formulate a general cost function [4].

$$C(\mathbf{W}) \equiv -\mathrm{sign}\,(f(\mathbf{u})) \sum_{i=1}^{M} k_4(i), \tag{66}$$

where $\mathrm{sign}(f(\mathbf{u}))$ is the sign function of the nonlinearity used at the output neurons and $f(\mathbf{u}) = \pm\mathbf{u}^3/3$. Note that this new form of minimization of the signal representation error criterion is valid for observed data which is zero-mean and spatially white.

The MSE (mean-squared-error) of the cost function in equation (66) relates to the mutual information as shown in Section 6 under the further assumption that probability densities are more or less symmetric so that the third-order cumulant terms within expansion can be removed from the fourth-order approximation of the Edgeworth expansion. The mutual information can then be approximated as follows (see Section 6):

$$I(\mathbf{u}) \cong J(\tilde{\mathbf{x}}) - \frac{1}{48}\sum_{i=1}^{N} k_4^2(i). \tag{67}$$

As in Section 6, maximizing the marginal negentropies with respect to $\mathbf{W}$ minimizes mutual information giving

$$\frac{\partial}{\partial \mathbf{W}} I(\mathbf{u}) \cong \frac{\partial}{\partial \mathbf{W}}\left(-\sum_{i=1}^{N} k_4^2(i)\right), \tag{68}$$

which corroborates that maximizing the sum of marginal cumulants or minimizing the MSE of the cost function derived for nonlinear PCA can be interpreted as an approximate information-theoretic contrast for ICA.

8. BUSSGANG ALGORITHMS

Bussgang algorithms have been introduced by Bellini [44] to perform blind deconvolution. Lambert [8] proposes three different multichannel blind deconvolution (separation and deconvolution) algorithms based on three classes of Bussgang cost functions. These algorithms are similar to the information-theoretic learning algorithm [3], but the relationship to the infomax algorithm is not obvious. Intuitive explanations have been proposed by Bell and Sejnowski [3], Lambert and Bell [45], Girolami and Fyfe [20], and Lee *et al.* [46]. Here, we show how the Bussgang algorithm can be interpreted as an information-theoretic cost function.

A white zero-mean stochastic process u_t has the Bussgang property if it satisfies [47]

$$E\left\{u_t u_{t+k}\right\} = E\left\{f\left(u_t\right) u_{t+k}\right\}, \tag{69}$$

where the subscript denotes time-points t and its time-shifted version $t+k$ and the Bussgang nonlinearity $f(.)$ is some monotone nonlinear function. The Bussgang property in equation (69) states that the autocorrelation function of u_t is equal to the cross-correlation function between the process u_t and the output of a nonlinearity $f(u_t)$ where both correlation functions are measured for the same lag.

We may rewrite the Bussgang property in equation (69) for spatial processes as follows:

$$E\left\{u_i u_j\right\} = E\left\{f\left(u_i\right) u_j\right\}, \tag{70}$$

where the subscripts i, j denote independent (white) stochastic processes. In fact, equation (69) differs from equation (70) only insofar as the subscripts refer to spatial rather than temporal samples, which allows us to relate the Bussgang property to the spatial ICA formulation. Now, the left side of equation (70) describes the second-order cross-correlation between two estimated sources and the right side of equation (70) accounts for higher-order cross-correlation between these estimates due to the nonlinearity $f(.)$ that can be thought of as a combination of higher-order terms in a Taylor series expansion.

A common way to derive a learning rule in blind deconvolution is to estimate the mean-squared error between the estimate u_i and the true source s_i. However, since the true source is not available, another estimator is needed. A valid estimator would be a nonlinear estimate $f(u_i)$ where the form of the function $f(.)$ has to reflect some information about the true signals s_i. Define a cost function C that minimizes the MSE between the source estimate u_i and a Bussgang nonlinear estimate $f(u_i)$. For simplicity, we consider one source estimate u_i:

$$C = E\left\{\left(u_i - f\left(u_i\right)\right)^2\right\}. \tag{71}$$

The form of the Bussgang nonlinearity can be derived from the maximum *a posteriori* (MAP) model by forming a conditional log-likelihood model given the observed data as follows.

For an independent source s_i, the estimated source u_i can be modeled as the source s_i plus an independent noise source n such that $u_i = s_i + n$. Let us define an error variable z_i as the difference between the true source signal and the estimated source signal

$$z_i = s_i - u_i. \tag{72}$$

We assumed that u_i can be estimated by the nonlinear function $f(u_i)$ giving

$$z_i = s_i - f\left(u_i\right). \tag{73}$$

The conditional density of the source given the variable z_i can be described by the MAP model

$$p\left(s_i \mid z_i\right) = p\left(z_i \mid s_i\right) p\left(s_i\right). \tag{74}$$

Assume that $p(z_i \mid s_i)$ can be modeled as a white zero-mean Gaussian process giving

$$p\left(z_i \mid s_i\right) = K \exp\left(-\frac{\left(z_i - s_i\right)^2}{2\sigma_u^2}\right), \tag{75}$$

where K is a constant and $\sigma_u^2 = \sigma_s^2 + \sigma_n^2$ is the variance of u_i. The justification of a Gaussian process for the conditional estimator $p(z_i \mid s_i)$ is that the sum of N ($N \gg 1$) zero-mean independent sources s_i sum up to a Gaussian observation due to the central limit theorem. Substituting equation (75) in equation (74) and taking the logarithm of the conditional estimate in equation (74), it follows that

$$\log\left(p\left(s_i \mid z_i\right)\right) = \log(K) - \frac{\left(z_i - s_i\right)^2}{2\sigma_u^2} + \log\left(p\left(s_i\right)\right). \tag{76}$$

The derivative of equation (76) with respect to s_i gives

$$\frac{\partial \log\left(p\left(s_i \mid z_i\right)\right)}{\partial s_i} = \frac{\left(z_i - s_i\right)}{\sigma_u^2} + \frac{\partial \log\left(p\left(s_i\right)\right)}{\partial s_i}. \tag{77}$$

When the estimation error is minimized, equation (77) is zero and solving for z_i gives the following expression:

$$z_i = s_i - \sigma_u^2 \frac{\partial \log\left(p\left(s_i\right)\right)}{\partial s_i}. \tag{78}$$

Now comparing equation (78) with equation (73) and assuming unit variance for u_i ($\sigma_u^2 = 1$), the form for the Bussgang nonlinear estimator must satisfy

$$f\left(s_i\right) = \frac{\frac{\partial p(s_i)}{\partial s_i}}{p\left(s_i\right)}, \tag{79}$$

which is proportional to the derivative of the log-density of the true source distribution.

We can apply equation (79) to the initial Bussgang property by rewriting equation (70) in matrix form

$$\begin{aligned} E\left\{\mathbf{u}\mathbf{u}^\top\right\} &= E\left\{f(\mathbf{u})\mathbf{u}^\top\right\}, \\ E\left\{\mathbf{u}\mathbf{u}^\top\right\} - E\left\{f(\mathbf{u})\mathbf{u}^\top\right\} &= \mathbf{0}, \\ E\left\{\mathbf{W}\mathbf{A}\,\mathbf{s}\mathbf{s}^\top\mathbf{A}^\top\mathbf{W}^\top\right\} - E\left\{f(\mathbf{u})\mathbf{u}^\top\right\} &= \mathbf{0}. \end{aligned} \tag{80}$$

The left side of equation (80) is the identity matrix when we assume that $\mathbf{W} = \mathbf{A}^{-1}$. Multiplying equation (80) with $\mathbf{W}$ gives

$$\left[\mathbf{I} - E\left\{f(\mathbf{u})\mathbf{u}^\top\right\}\right]\mathbf{W} = \mathbf{0}. \tag{81}$$

The optimal Bussgang nonlinearity $f(\mathbf{u})$ when applied to ICA must be equivalent to

$$f(\mathbf{u}) = -\frac{\frac{\partial p(\mathbf{u})}{\partial \mathbf{u}}}{p(\mathbf{u})}, \tag{82}$$

which is precisely the score function $\varphi(\mathbf{u})$ in equation (17). Therefore, we have

$$\left[\mathbf{I} - E\left\{\varphi(\mathbf{u})\mathbf{u}^\top\right\}\right]\mathbf{W} = \mathbf{0}, \tag{83}$$

which is exactly the convergence criterion for the infomax learning rule in equation (19). The justification of the Bussgang nonlinearity in equation (79) also corroborates the infomax principle and its application to blind source separation and blind deconvolution.

9. PROPERTIES OF BLIND SOURCE SEPARATION ALGORITHMS

We review and discuss important properties of ICA algorithms used in the source separation problem.

9.1. Convergence

The insights in Sections 3–8 suggest that the estimation of the true source densities should be crucial to extract the sources. Many researchers have therefore tried to find the separating matrix $\mathbf{W}$ as well as a parametric estimate of the nonlinearity associated with the source density [1,48,49]. Pearlmutter and Parra [1] proposed a contextual ICA (cICA) that assumed a weighted sum of parametric logistic functions to model the source density. Moulines, Cardoso and Gassiat [48] and Xu *et al.* [49] model the underlying p.d.f. with mixtures of Gaussians showing that they can also separate sub- and super-Gaussian sources. These parametric modeling approaches in general are computationally expensive because they learn complex density parameters. Empirical results by Lee *et al.* [23] and Makeig [50] on electroencephalographic (EEG) data and data from event related potentials (ERP) using cICA indicate that it can fail to find independent components when the number of time-samples is too small to give a reliable density estimate (e.g., 600 data points for ERPs).

However, simulation results performed by many researchers show that ICA algorithms with a fixed nonlinearity converge to a separating solution although the nonlinearity is only a crude approximation of the underlying sources. Bell and Sejnowski [3] reported that the infomax algorithm would separate 10 super-Gaussian sources such as music and speech using only one logistic function that imposes a super-Gaussian prior. Lee *et al.* [23] report that the 10 sound sources used by Pearlmutter and Parra [1] can be separated easily and with faster convergence than cICA using the logistic function instead of a parametric density estimator. These empirical results suggest that simple density estimators may be sufficient to separate mixed sources. However, a more detailed analysis needs to be done to determine the conditions on the source densities under which the algorithm converges to the correct separation [30].

The units of the learning algorithm on the left and right side of equation (18) do not match, and hence, the rate of convergence depends on the scales of the axes. The natural gradient [24] or relative gradient [17] greatly improves convergence of ICA by making the gradient invariant to the scale on the axes. The normal entropy gradient (Euclidean gradient) assumes that the space of $\mathbf{W}$ is orthonormal; that is, each $\mathbf{w}_{ij}$ is of unit length and points in an orthogonal direction to the others. In this case, the metric tensor is the identity matrix $\mathbf{w}_{ij}.\mathbf{w}_{kl} = \delta_{(ij)(kl)}$. However, Amari [24] has shown how $\mathbf{W}$ is a Riemannian space with a nonorthonormal metric tensor. Fortunately, in the case of ICA, convergence can proceed as fast as if the space of $\mathbf{W}$ was orthonormal, if we only rescale the Euclidean gradient to the natural gradient as follows:

$$\tilde{\nabla} f(\mathbf{W}) = \nabla f(\mathbf{W})\mathbf{W}^{\mathsf{T}}\mathbf{W}, \tag{84}$$

where $\tilde{\nabla} f(\mathbf{W})$ denotes the natural gradient, and $\nabla f(\mathbf{W})$ denotes the Euclidean gradient. The units now match and convergence is optimal. A detailed derivation of this intuitive explanation is presented by Amari [24] and Yang and Amari [51].

9.2. Stability

A generic stability analysis of separating solutions was examined by Cardoso and Laheld [17], Pham and Garrat [52] and Amari *et al.* [53], and more recently by Cardoso [30]. In the analysis, the mean field updates were approximated by a first-order perturbation in the parameters of the separating matrix. The linear approximation near the stationary point is the gradient of the mean field at the stationary point. The real part of the eigenvalues of the derivative of the mean field must be negative so that the parameters on average return to the stationary point.

A sufficient condition guaranteeing asymptotic stability can be derived [30] so that

$$\kappa_i > 0, \qquad 1 \le i \le N, \tag{85}$$

where κ_i is

$$\kappa_i = E\left\{\varphi_i'(u_i)\right\} E\left\{u_i^2\right\} - E\left\{\varphi_i(u_i)\, u_i\right\}, \tag{86}$$

and

$$\varphi_i(u_i) = u_i + k_i \tanh(u_i). \tag{87}$$

Substituting equation (87) in equation (86) gives

$$\begin{aligned} \kappa_i &= E\left\{k_i \mathrm{sech}^2(u_i) + 1\right\} E\left\{u_i^2\right\} - E\left\{\left[k_i \tanh(u_i) + u_i\right] u_i\right\} && (88) \\ &= k_i \left(E\left\{\mathrm{sech}^2(u_i)\right\} E\left\{u_i^2\right\} - E\left\{\left[\tanh(u_i)\right] u_i\right\}\right). && (89) \end{aligned}$$

To ensure $\kappa_i > 0$, the sign of k_i must be the same as the sign of $E\{\mathrm{sech}^2(u_i)\}E\{u_i^2\} - E\{[\tanh(u_i)]u_i\}$. Therefore, we can use the learning rule in equation (20) where the k_is are

$$k_i = \mathrm{sign}\left(E\left\{\mathrm{sech}^2(u_i)\right\} E\left\{u_i^2\right\} - E\left\{\left[\tanh(u_i)\right] u_i\right\}\right). \tag{90}$$

10. DISCUSSION

10.1. Conclusions

We have unified several lines of research on ICA within an information-theoretic framework and conjecture that this framework is well suited to further investigate ICA from many different theoretical viewpoints.

10.2. Applications to Real World Problems

The extended infomax algorithm has recently been applied to real world problems such as analyzing electroencephalographic (EEG) data [54,55] and functional magnetic resonance imaging (fMRI) data [39,56]. Makeig *et al.* [56] showed that the Bell and Sejnowski [3] algorithm was able to linearly decompose EEG activity and artifacts. Jung *et al.* [55] show that the extended infomax algorithm could additionally extract sub-Gaussian artifacts such as line noise and eye movements. McKeown *et al.* [39] have used the extended infomax algorithm to investigate task-related human brain activity in fMRI data. They made the assumption that the sources underlying the fMRI recordings were spatially independent rather than temporally independent as in the case of EEG, and found both consistently and transiently task-related brain activations.

Another difficult real-world problem is the separation of convolved and time-delayed sources. Recently, blind separation experiments of real recorded audio data have been addressed by several researchers. Two voices recorded in a room with two microphones (the cocktail party problem) can be separated taking into account time delayed and convolved sources [8,46,57–61]. An application to underwater communication has been considered by Li and Sejnowski [62].

10.3. Limitations and Future Research

Nonlinear mixing problem

ICA starts with a linear model for mixing. Researchers have recently tackled the problem of nonlinear mixing. Burel [63], Yang *et al.* [64], Taleb and Jutten [65], Lee *et al.* [66], and Pajunen and Karhunen [67] proposed extensions when linear mixing is combined with certain nonlinear mixing models. Other approaches include self-organizing feature maps to identify nonlinear features in the data [68]. However, these methods are not generally applicable. It may be necessary to restrict the model sufficiently to allow a solution.

Overcomplete ICA

Overcomplete ICA is applicable when there are more sources than sensors $N < M$. With only one or two sensors, can more than two sources be recovered? An interesting discussion is presented by Jutten and Cardoso [69]. More recently, preliminary results by Lewicki and Sejnowski [70] suggest that an overcomplete representation of the data can to some extent extract the independent components using *a priori* knowledge of the source distribution.

Noisy ICA

Only a few papers have studied ICA in the presence of noise [15], and much more work needs to be done to determine the effects of noise on performance. Noise can be treated in a Bayesian framework using overcomplete basis functions [70] or generative models [71].

Nonstationarity problem

If the sources are stationary, i.e., sources may appear, disappear or move (speaker moving in a room), the weight matrix **W** could change completely from one time point to the next. Unsupervised methods are needed to handle the abrupt changes that occur in real environments. Nadal and Parga [72] have proposed some analytical methods for time-dependent mixtures. Murata *et al.* [73] suggest an adaptation of the learning rate to cope with changing environments.

There are many other potential applications of ICA and a few are mentioned here.

ICA on recordings from the olfactory system

Hopfield [74] has suggested that the olfactory system may use a factorial code representation. We are currently applying ICA to data from the olfactory system [75] to test this hypothesis.

ICA in communications

Complex-valued signal mixing occurs in radio channels. This problem occurs in current mobile communication applications such as CDMA (code division multiple access) systems. Torkkola [76] has incorporated prior knowledge about the source distributions into the nonlinear transfer function to adaptively determine time-varying mixing matrices. In simulations, he showed that infomax can successfully be applied to unmix radio signals in fading channels.

ICA for data mining

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential for helping companies focus on the most important information in their data warehouses. For example, Lizhong and Moody [77] have explored ICA for financial data modeling. More recently, Girolami *et al.* [78] suggested ICA based projection pursuit networks for data clustering and data mining.

Biological evidence of ICA?

Learning rules that only require local information are more biologically plausible. The learning rules in equations (18) and (19) for a single feedforward architecture are nonlocal; i.e., the neurons must have information about the synaptic weights of neighboring neurons without being connected to them. However, there are a number of local learning rules for ICA such as that for the recurrent Herault-Jutten architecture and Linsker's [79] network for Bell and Sejnowski's ICA learning rule. The extended exploratory projection pursuit network with inhibitory lateral connections [80] also has a local learning rule. It is therefore possible that some form of ICA learning rule is used in the brain. Field [81] has suggested that factorial codes are an efficient coding strategy for visual sensory processing, and ICA applied to natural images yields localized and oriented filters similar to those found in visual cortex [82]. ICA has been used to extract local features for face recognition systems [83]. Factorial coding principles might be found in other brain areas such as the cerebellum that might use efficient coding schemes for motor control and prediction.

ICA as a conditional density estimator for classification problems

Roth and Baram [16] have used ICA as a conditional density estimator for classification and time-series prediction. Their results indicate that ICA can be useful as a conditional density estimator in classical pattern recognition issues.

Hardware implementation of ICA

An analog VLSI chip implementation of the Herault-Jutten algorithm was fabricated by Cohen and Andreou [84]. The implementation of the extended infomax algorithm in VLSI will be a challenging goal. An extension to time-delays and convolved mixtures could have significance for many practical applications that are computationally demanding.

REFERENCES

1. B. Pearlmutter and L. Parra, A context-sensitive generalization of ICA, In *ICONIP'96*, pp. 151–157, (1996).
2. J.-F. Cardoso, Infomax and maximum likelihood for blind source separation, *IEEE Signal Processing Letters* **4**, 109–111 (1997).
3. A.J. Bell and T.J. Sejnowski, An information maximization approach to blind separation and blind deconvolution, *Neural Computation* **7**, 1129–1159 (1995).
4. M. Girolami and C. Fyfe, Stochastic ICA contrast maximization using Oja's nonlinear PCA algorithm, *International Journal of Neural Systems* (to appear).
5. J. Karhunen and J. Joutsensalo, Representation and separation of signals using nonlinear PCA type learning, *Neural Networks* **7**, 113–127 (1994).
6. E. Oja, The nonlinear PCA learning rule in independent component analysis, *Neurocomputing* **17**, 25–45 (1997).
7. P. Comon, Independent component analysis—A new concept?, *Signal Processing* **36** (3), 287–314 (1994).
8. R. Lambert, Multichannel blind deconvolution: FIR matrix algebra and separation of multipath mixtures, Thesis, University of Southern California, Department of Electrical Engineering, (1996).
9. J. Herault and J. Jutten, Space or time adaptive signal processing by neural network models, In *Neural Networks for Computing: AIP Conference Proceedings 151*, (Edited by J.S. Denker), American Institute for Physics, New York, (1986).
10. C. Jutten and J. Herault, Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture, *Signal Processing* **24**, 1–10 (1991).
11. A. Cichocki, R. Unbehauen and E. Rummert, Robust learning algorithm for blind separation of signals, *Electronics Letters* **30** (17), 1386–1387 (1994).
12. R. Linsker, Local synaptic learning rules suffice to maximize mutual information in a linear network, *Neural Computation* **4**, 691–702 (1992).
13. H.B. Barlow, Possible principles underlying the transformation of sensory messages, In *Sensory Communication*, (Edited by W.A. Rosenblith), MIT Press, (1961).
14. J.J. Atick, Could information theory provide an ecological theory of sensory processing?, *Network* **3**, 213–251 (1992).
15. J.P. Nadal and N. Parga, Non-linear neurons in the low-noise limit: A factorial code maximizes information transfer, *Network* **4**, 295–312 (1994).
16. Z. Roth and Y. Baram, Multidimensional density shaping by sigmoids, *IEEE Trans. on Neural Networks* **7** (5), 1291–1298 (1996).
17. J.-F. Cardoso and B. Laheld, Equivariant adaptive source separation, *IEEE Trans. on Signal Processing* **45** (2), 434–444 (1996).
18. M. Gaeta and J. Lacoume, Sources separation without *a priori* knowledge: The maximum likelihood solution, *Eusipco 90*, Barcelona, pp. 621–624 (1990).
19. M. Girolami and C. Fyfe, Generalised independent component analysis through unsupervised learning with emergent Bussgang properties, In *Proc. International Conference on Neural Networks*, Houston, TX, pp. 1788–1891, (1997).
20. M. Girolami and C. Fyfe, Extraction of independent signal sources using a deflationary exploratory projection pursuit network with lateral inhibition, *I.E.E. Proceedings on Vision, Image and Signal Processing Journal* **14** (5), 299–306 (1997).
21. T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley and Sons, New York, (1991).
22. L. Xu, Least MSE reconstruction: A principle for self organizing nets, *Neural Networks* **6**, 627–648 (1993).
23. T.-W. Lee, M. Girolami. and T. Sejnowski, Independent component analysis using an extended infomax algorithm for mixed sub-Gaussian and super-Gaussian sources, *Neural Computation* (submitted).
24. S. Amari, Natural gradient works efficiently in learning, *Neural Computation* (to appear).
25. L. Molgedey and H. Schuster, Separation of independent signals using time-delayed correlations, *Physical Review Letters* **72** (23), 3634–3637 (1994).
26. A. Papoulis, *Probability, Random Variables and Stochastic Processes*, 2$^{\text{nd}}$ edition, McGraw-Hill, New York, (1984).
27. S. Amari and J.-F. Cardoso, Blind source separation—Semiparametric statistical approach, *IEEE Trans. on Signal Processing* **45**, 2692–2700 (1997).
28. S. Amari, A. Cichocki and H. Yang, A new learning algorithm for blind signal separation, In *Advances in Neural Information Processing Systems 8*, pp. 757–763, (1996).

29. M. Girolami, An alternative perspective on adaptive independent component analysis algorithms, Technical Report, Computing and Information Systems, Paisley University, Scotland, ISSN 1461-6122, (1997).
30. J.-F. Cardoso, Blind signal processing: Statistical principles, In *Proceedings of the IEEE, 86*, pp. 2009–2025, (1998).
31. J.-F. Cardoso, *Unsupervised Adaptive Filtering*, Chapter "Entropic contrasts for source separation", (Edited by S. Haykin), First presented at the NIPS*96 on 'Blind Signal Processing' organized by A. Cichocki, (1998).
32. M. Girolami, Self-organizing artificial neural networks for signal separation, Ph.D. Thesis, Department of Computing and Information Systems, Paisley University, Scotland, (1997).
33. J.H. Friedman, Exploratory projection pursuit, *Journal of the American Statistical Association* **82** (397), 249–266 (1987).
34. B.S. Everitt, *An Introduction to Latent Variable*, Chapman and Hall, London, (1984).
35. M.C. Jones and R. Sibson, What is projection pursuit, *The Royal Statistical Society* **A150**, 1–36 (1987).
36. A. Stuart and J.K. Ord, *Kendall's Advanced Theory of Statistic, 1, Distribution Theory*, John Wiley, New York, (1987).
37. T.-W. Lee and T. Sejnowski, Independent component analysis for sub-Gaussian and super-Gaussian mixtures, In *4^{th} Joint Symposium on Neural Computation*, Vol. 7, pp. 132–140, Institute for Neural Computation, (1997).
38. T.-W. Lee, *Independent Component Analysis: Theory and Applications*, Kluwer Academic, Boston, (1998).
39. M. McKeown, T.-P. Jung, S. Makeig, G. Brown, S. Kindermann, T.-W. Lee and T. Sejnowski, Transiently time-locked fMRI activations revealed by independent component analysis, *Proceedings of the National Academy of Sciences* **95**, 803–810 (1997).
40. E. Oja, The nonlinear PCA learning rule in independent component analysis, *Neurocomputing* **17**, 25–45 (1997).
41. J. Karhunen, E. Oja, L. Wang, R. Vigario and J. Joutsensalo, A class of neural networks for independent component analysis, *IEEE Trans. on Neural Networks* **8**, 487–504 (1997).
42. E. Oja and J. Karhunen, Signal separation by nonlinear Hebbian learning, In *Proceedings IEEE ICNN 95*, pp. 83–87, (1995).
43. J. Karhunen, L. Wang and R. Vigario, Nonlinear PCA type approach for source separation and independent component analysis, In *Proc. ICNN*, Perth, Australia, pp. 995–1000, (1995).
44. S. Bellini, Blind deconvolution, In *Bussgang Techniques for Blind Deconvolution and Equalization*, (Edited by Haykin), Prentice Hall, New York, (1994).
45. R. Lambert and A. Bell, Blind separation of multiple speakers in a multipath environment, In *ICASSP*, Munich, Germany, April 21–26, pp. 423–426, (1997).
46. T.-W. Lee, A.J. Bell and R. Orglmeister, Blind source separation of real-world signals, *IEEE Proc. ICNN*, Houston, TX, pp. 2129–2135 (1997).
47. S. Bellini, Bussgang techniques for blind deconvolution and equalization, In *Blind Deconvolution*, (Edited by S. Haykin), Prentice Hall, (1994).
48. E. Moulines, J.-F. Cardoso and E. Gassiat, Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models, *Proc. ICASSP'97*, Munich, (5), pp. 3617–3620, (1997).
49. L. Xu, C. Cheung, H. Yang and S. Amari, Maximum equalization by entropy maximization and mixture of cumulative distribution functions, *IEEE Proc. ICNN*, Houston, TX, pp. 1821–1826, (1997).
50. S. Makeig, T. Jung, A.J. Bell, D. Ghahremani and T.J. Sejnowski, Blind separation of event-related brain response into spatial independent components, *Proc. of the National Academy of Sciences* **94**, 10979–10984 (1997).
51. H. Yang and S. Amari, Adaptive on-line learning algorithms for blind separation—Maximum entropy and minimum mutual information, *Neural Computation* **9**, 1457–1482 (1997).
52. D.T. Pham and P. Garrat, Blind separation of mixture of independent sources through a quasi-maximum likelihood approach, *IEEE Trans. on Signal Proc.* **45**, 1712–1725 (1997).
53. S. Amari, T.-P. Chen and A. Cichocki, Stability analysis of adaptive blind source separation, *Neural Networks* **10** (8), 1345–1352 (1997).
54. S. Makeig, T.-P. Jung, A. Bell, D. Ghahramani and T. Sejnowski, Independent component analysis of electroencephalographic data, *Proceedings of the National Academy of Sciences* **94**, 10979–10984 (1997).
55. T.-P. Jung, C. Humphries, T.-W. Lee, S. Makeig, M. McKeown, V. Iragui and T. Sejnowski, Extended ICA removes artifacts from electroencephalographic recordings, *Advances in Neural Information Processing Systems* (to appear).
56. S. Makeig, A. Bell, T.-P. Jung and T. Sejnowski, Independent component analysis of electroencephalographic data, In *Advances in Neural Information Processing Systems 8*, pp. 145–151, MIT Press, Cambridge, MA, (1996).
57. K. Torkkola, Blind separation of convolved sources based on information maximization, In *IEEE Workshop on Neural Networks for Signal Processing*, Kyoto, Japan, pp. 423–432, (1996).
58. D. Yellin and E. Weinstein, Multichannel signal separation: Methods and analysis, *IEEE Transactions on Signal Processing* **44** (1), 106–118 (1996).
59. M. Girolami and C. Fyfe, Extraction of independent signal sources using a deflationary exploratory projection pursuit network with lateral inhibition, *I.E.E. Proceedings on Vision, Image and Signal Processing Journal* **14** (5), 299–306 (1997).

60. A. Cichocki, S. Amari and J. Cao, Neural network models for blind separation of time delayed and convolved signals, *IECE Trans. Fundamentals* **E82-A** (1997).

61. H.-L. Nguyen Thi and C. Jutten, Blind source separation for convolutive mixtures, *Signal Processing* **45** (2) (1995).

62. S. Li and T.J. Sejnowski, Adaptive separation of mixed broadband sound sources with delays by a beamforming Herault-Jutten network, *IEEE Journal of Oceanic Engineering* **20** (1), 73–79 (1994).

63. G. Burel, A non-linear neural algorithm, *Neural Networks* **5**, 937–947 (1992).

64. H. Yang, S. Amari and A. Cichocki, Information back-propagation for blind separation of sources from non-Mor-linear mixtures, *Proc. of ICNN'97*, Houston, TX, pp. 2141–2146, (1997).

65. A. Taleb and C. Jutten, Nonlinear source separation: The post-nonlinear mixtures, *ESANN'97*, pp. 279–284, (1997).

66. T.-W. Lee, B. Köhler and R. Orglmeister, Blind source separation of nonlinear mixing models, *IEEE Proc. NNSP*, Florida, pp. 406–415, (1997).

67. P. Pajunen and J. Karhunen, A maximum likelihood approach to nonlinear blind source separation, In *Proceedings of the 1997 Int. Conf. on Artificial Neural Networks (ICANN'97)*, Lausanne, pp. 541–546, (1997).

68. J. Lin and J. Cowan, Faithful representation of separable input distributions, *Neural Computation* **9** (6), 1305–1320 (1997).

69. C. Jutten and J.-F. Cardoso, Source separation: Really blind?, In *Proc. NOLTA*, pp. 79–84, (1995).

70. M. Lewicki and T. Sejnowski, Learning nonlinear overcomplete representations for efficient coding, *Advances in Neural Information Processing Systems* (to appear).

71. G. Hinton and Z. Ghahramani, Generative models for discovering sparse distributed representations, *Philosophical Transactions Royal Society B* **352**, 1177–1190 (1997).

72. J.P. Nadal and N. Parga, Redundancy reduction and independent component analysis: Conditions on cumulants and adaptive approaches, *Neural Computation* **9**, 1421–1456 (1997).

73. N. Murata, K.-R. Mueller, A. Ziehe and S. Amari, Adaptive on-line learning in changing environments, In *Advances in Neural Information Processing Systems 9*, pp. 599–605, MIT Press, (1997).

74. J. Hopfield, Olfactory computation and object perception, *Proc. Natl. Acad. Sci. USA* **88**, 6462–6466 (1991).

75. Kauer and White, personal communication.

76. K. Torkkola, Blind separation of radio signals in fading channels, In *Advances in Neural Information Processing Systems 10*, MIT Press (to appear).

77. W. Lizhong and J. Moody, Multi-effect decompositions for financial data modeling, In *Advances in Neural Information Processing Systems 9*, pp. 995–1001, MIT Press, (1997).

78. M. Girolami, A. Cichocki and S. Amari, A common neural network model for exploratory data analysis and independent component analysis, Technical Report, bip-97-001, Brain Information Processing Group, RIKEN, Japan, (1997).

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

80. M. Girolami and C. Fyfe, A temporal model of linear anti-Hebbian learning, *Neural Processing Letters Journal* **4** (3), 1–10 (1997).

81. F. Field, What is the goal of sensory coding?, *Neural Computation* **6**, 559–601 (1994).

82. A.J. Bell and T.J. Sejnowski, The 'independent components' of natural scenes are edge filters, *Vision Research* **37** (23), 3327–3338 (1997).

83. M. Bartlett and T.J. Sejnowski, Viewpoint invariant face recognition using independent component analysis and attractor networks, In *Advances in Neural Information Processing Systems 9*, pp. 817–823, MIT Press, (1997).

84. M.H. Cohen and A.G. Andreou, Current-mode subthreshold MOS implementation of the Herault-Jutten autoadaptive network, *IEEE J. Solid-State Circuits* **27** (5), 714–727 (1992).

85. S. Haykin, *Adaptive Filter Theory*, 2$^{\text{nd}}$ edition, Prentice-Hall, (1991).

86. A. Hyvärinen and E. Oja, A fast fixed-point algorithm for independent component analysis, *Neural Computation* **9**, 1483–1492 (1997).

87. J. Karhunen, Neural approaches to independent component analysis and source separation, In *Proc. of ESANN'96, 4th European Symposium on Artificial Neural Networks*, Bruges, Belgium, pp. 249–266, (1996).

88. P. Pajunen, Blind source separation of binary sources with less sensors than sources, *Proc. of ICNN'97*, Houston, TX, pp. 1994–1997, (1997).

89. D.T. Pham, P. Garrat and C. Jutten, Separation of a mixture of independent sources through a maximum likelihood approach, *Proc. EUSIPCO*, pp. 771–774, (1992).

90. D.T. Pham, Blind separation of instantaneous mixture of sources via an independent component analysis, *IEEE Trans. on Signal Proc.* **44** (11), 2768–2779 (1996).

91. R. Vigario, A. Hyvärinen and E. Oja, ICA fixed-point algorithm in extraction of artifacts from EEG, In *Proc. 1996 IEEE Nordic Signal Processing Symposium NORSIG'96*, Espoo, Finland, pp. 383–386, (1996).