Independent Component Analysis using an Extended Infomax Algorithm for Mixed Sub-Gaussian and Super-Gaussian Sources

Te-Won Lee[1,4], **Mark Girolami**[2] **and Terrence J. Sejnowski**[1,3]
tewon@salk.edu, giro0ci@paisley.ac.uk, terry@salk.edu

[1]Howard Hughes Medical Institute
Computational Neurobiology Laboratory
The Salk Institute
La Jolla, California 92037, USA

[2]Department of Computing
and Information Systems
University of Paisley
PA1 2BE, Scotland

[3]Department of Biology
University of California, San Diego
La Jolla, California 92093, USA

[4] Institut für Elektronik
Technische Universität Berlin
Einsteinufer 17 10587 Berlin, Germany

Neural Computation, Vol:11(2), 409-433, 1999

Number of text pages: 24
Number of figures: 12
Number of tables: 2

Abstract

An extension of the infomax algorithm of Bell and Sejnowski (1995) is presented that is able to blindly separate mixed signals with sub- and super-Gaussian source distributions. This was achieved by using a simple type of learning rule first derived by Girolami (1997) by choosing negentropy as a projection pursuit index. Parameterized probability distributions that have sub- and super-Gaussian regimes were used to derive a general learning rule that preserves the simple architecture proposed by Bell and Sejnowski (1995), is optimized using the natural gradient by Amari (1998), and uses the stability analysis of Cardoso and Laheld (1996) to switch between sub- and super-Gaussian regimes. We demonstrate that the extended infomax algorithm is able to easily separate 20 sources with a variety of source distributions. Applied to high-dimensional data from electroencephalographic (EEG) recordings, it is effective at separating artifacts such as eye blinks and line noise from weaker electrical signals that arise from sources in the brain.

1 Introduction

Recently, blind source separation by Independent Component Analysis (ICA) has received attention because of its potential signal processing applications such as speech enhancement 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 reduces the statistical dependencies of the signals, attempting to make the signals as independent as possible.

The blind source separation problem has been studied by many researchers in neural networks and statistical signal processing (Jutten and Herault, 1991; Comon, 1994; Cichocki et al., 1994; Bell and Sejnowski, 1995; Cardoso and Laheld, 1996; Amari et al., 1996; Pearlmutter and Parra, 1996; Deco and Obradovic, 1996; Oja, 1997; Karhunen et al., 1997; Girolami and Fyfe, 1997a). See the introduction of Nadal and Parga (1997) for a historical review of ICA, and Karhunen (1996) for a review of different neural based blind source separation algorithms. More general ICA reviews are in Cardoso (1998a) and Lee et al. (1998a).

Bell and Sejnowski (1995) have developed an unsupervised learning algorithm based on entropy maximization in a single-layer feedforward neural network. The algorithm is effective in separating sources that have super-Gaussian distributions: sharply peaked probability density functions (p.d.f.s) with heavy tails. As illustrated in section 4 of Bell and Sejnowski (1995) the algorithm fails to separate sources that have negative kurtosis (e.g. uniform distribution). Pearlmutter and Parra (1996) have developed a contextual ICA algorithm within the maximum likelihood estimation (MLE) framework that is able to separate a more general range of source distributions. Motivated by computational simplicity, we use an information-theoretic algorithm that preserves the simple architecture in Bell and Sejnowski (1995) and allows an extension to the separation of mixtures of super-Gaussian and sub-Gaussian sources. Girolami (1997) derived this type of learning rule from the viewpoint of negentropy maximization[1] for exploratory projection pursuit (EPP) and ICA. These algorithms can be used on-line as well as off-line. Off-line algorithms that can also separate mixtures of super-Gaussian and sub-Gaussian sources were proposed by Cardoso and Soloumiac (1993), Comon (1994) and Pham and Garrat (1997).

The extended infomax algorithm preserves the simple architecture in Bell and Sejnowski (1995) and the learning rule converges rapidly with the 'natural' gradient proposed by Amari et al. (1996); Amari (1998) or 'relative' gradient proposed by Cardoso and Laheld (1996). In computer simulations, we show that this algorithm can successfully separate 20 mixtures of the following sources: 10 sound tracks [2], 6 speech and sound signals used in Bell and Sejnowski (1995), 3 uniformly distributed sub-Gaussian noise signals and one noise source with a Gaussian distribution. To test the extended infomax algorithm on more challenging real world data, we performed experiments with EEG recordings and show that it can clearly separate electrical artifacts from brain activity. This

[1] Relative entropy is the general term for negentropy. Negentropy maximization refers to maximizing the sum of marginal negentropies.

[2] obtained from Pearlmutter in `http://sweat.cs.unm.edu/~bap/demos.html`

technique shows great promise for analyzing EEG recordings (Makeig et al., 1997; Jung et al., 1998) and functional magnetic resonance imaging (fMRI) data (McKeown et al., 1998).

The paper is organized as follows: In section 2, the problem is stated and a simple but general learning rule that can separate sub- and super-Gaussian sources is presented. This rule is applied to simulations and real data in section 3. In section 4, there is a brief discussion of other algorithms and architectures, potential applications to real world problems, limitations and further research problems.

2 The Extended Infomax Algorithm

Assume that there is an M-dimensional zero-mean vector $\mathbf{s}(t) = [s_1(t), \cdots, s_M(t)]^T$, such that the components $s_i(t)$ are mutually independent. The vector $\mathbf{s}(t)$ 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.

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

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

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

where A is a full rank $N \times M$ scalar matrix. As the components of the observed vectors are no longer independent, the multivariate p.d.f. will not satisfy the p.d.f. product equality. In this paper, we shall consider the case where, the number of sources is equal to the number of sensors $N = M$. If the components of $\mathbf{s}(t)$ are such that at most one source is normally distributed then it is possible to extract the sources $\mathbf{s}(t)$ from the received mixtures $\mathbf{x}(t)$ (Comon, 1994). The mutual information of the observed vector is given by the Kullback-Leibler (KL) divergence of the multivariate density from the product of the marginal (univariate) densities:

$$I(x_1, x_2, \cdots, x_N) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} p(x_1, x_2, \cdots, x_N) \log \frac{p(x_1, x_2, \cdots, x_N)}{\prod_{i=1}^{N} p_i(x_i)} dx_1 dx_2 \cdots dx_N. \tag{3}$$

For simplicity, we write:

$$I(\mathbf{x}) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{\prod_{i=1}^{N} p_i(x_i)} d\mathbf{x}. \tag{4}$$

The mutual information will always be positive and will only equal zero when the components are independent (Cover and Thomas, 1991).

The goal of ICA is to find a linear mapping $\mathbf{W}$ such that the unmixed signals $\mathbf{u}$

$$\mathbf{u}(t) = \mathbf{W}\mathbf{x}(t) = \mathbf{W}\mathbf{A}\mathbf{s}(t) \tag{5}$$

are statistically independent. The sources are recovered up to scaling and permutation. There are many ways for learning $\mathbf{W}$. (Comon, 1994) minimizes the degree of dependence among outputs using contrast functions approximated by the Edgeworth expansion of the KL divergence. The higher-order statistics are approximated by cumulants up to 4th order. Other methods related to minimizing mutual information can be derived from the infomax approach. Nadal and Parga (1994) showed that in the low-noise case, the maximum of the mutual information between the input and output of a neural processor implied that the output distribution was factorial. Roth and Baram (1996) and Bell and Sejnowski (1995) independently derived stochastic gradient learning rules for this maximization and applied them, respectively to forecasting, time series analysis, and the blind separation of sources. A similar adaptive method for source separation has been proposed by Cardoso and Laheld (1996).

2.1 A simple but general learning rule

The learning algorithm can be derived using the maximum likelihood formulation. The MLE approach to blind source separation was first proposed by Gaeta and Lacoume (1990), Pham and Garrat (1997) and was pursued more recently by Pearlmutter and Parra (1996) and Cardoso (1997). The probability density function of the observations $\mathbf{x}$ can be expressed as (Amari and Cardoso, 1997):

$$p(\mathbf{x}) = |\det(\mathbf{W})| p(\mathbf{u}) \tag{6}$$

where $p(\mathbf{u}) = \prod_{i=1}^{N} p_i(u_i)$ is the hypothesized distribution of $p(\mathbf{s})$. The log-likelihood of equation 6 is

$$L(\mathbf{u}, \mathbf{W}) = \log|\det(\mathbf{W})| + \sum_{i=1}^{N} \log p_i(u_i). \tag{7}$$

Maximizing the log-likelihood with respect to $\mathbf{W}$ gives a learning algorithm for $\mathbf{W}$ (Bell and Sejnowski, 1995):

$$\Delta\mathbf{W} \propto \left[(\mathbf{W}^T)^{-1} - \varphi(\mathbf{u})\mathbf{x}^T\right] \tag{8}$$

where

$$\varphi(\mathbf{u}) = -\frac{\frac{\partial p(\mathbf{u})}{\partial \mathbf{u}}}{p(\mathbf{u})} = \left[-\frac{\frac{\partial p(u_1)}{\partial u_1}}{p(u_1)}, \cdots, -\frac{\frac{\partial p(u_N)}{\partial u_N}}{p(u_N)}\right]^T. \tag{9}$$

An efficient way to maximize the log-likelihood is to follow the 'natural' gradient (Amari, 1998):

$$\Delta\mathbf{W} \propto \frac{\partial L(\mathbf{u}, \mathbf{W})}{\partial \mathbf{W}} \mathbf{W}^T\mathbf{W} = \left[\mathbf{I} - \varphi(\mathbf{u})\mathbf{u}^T\right]\mathbf{W} \tag{10}$$

as proposed by Amari et al. (1996) or relative gradient, Cardoso and Laheld (1996). Here $\mathbf{W}^T\mathbf{W}$ rescales the gradient, simplifies the learning rule in equation 8 and speeds convergence considerably. It has been shown that the general learning algorithm in equation 10 can be derived from several theoretical viewpoints such as MLE (Pearlmutter and Parra, 1996), infomax (Bell and Sejnowski, 1995) and negentropy maximization (Girolami and Fyfe, 1997b). Lee et al. (1998a) review these techniques and show their relation to each other.

The parametric density estimate $p_i(u_i)$ plays an essential role in the success of the learning rule in equation 10. Local convergence is assured if $p_i(u_i)$ is the derivative of the log-densities of the sources (Pham and Garrat, 1997). If we choose $g_i(u)$ to be a logistic function ($g_i(u_i) = \tanh(u_i)$) so that $\varphi(\mathbf{u}) = 2\tanh(\mathbf{u})$ the learning rule reduces to that in Bell and Sejnowski (1995) with the natural gradient:

$$\Delta\mathbf{W} \propto \left[\mathbf{I} - 2\tanh(\mathbf{u})\mathbf{u}^T\right]\mathbf{W}. \tag{11}$$

Theoretical considerations as well as empirical observations [3] have shown that this algorithm is limited to separating sources with super-Gaussian distributions. The sigmoid function used in Bell and Sejnowski (1995) provides a priori knowledge about the source distribution, i.e. the super-Gaussian shape of the sources. However, they also discuss a 'flexible' sigmoid function (a sigmoid function with parameters p, r so that $g(u_i) = \int g(u_i)^p(1 - g(u_i))^r$) can be used to match the source distribution. The idea of modeling a parametric nonlinearity has been further investigated and generalized by Pearlmutter and Parra (1996) in their contextual ICA (cICA) algorithm. They model the p.d.f. in a parametric form by taking into account the temporal information and by choosing $p_i(u_i)$ as a weighted sum of several logistic density functions with variable means and scales. Moulines et al. (1997) and Xu et al. (1997) model the underlying p.d.f. with mixtures

[3] as detailed in section 4 of Bell and Sejnowski (1995)

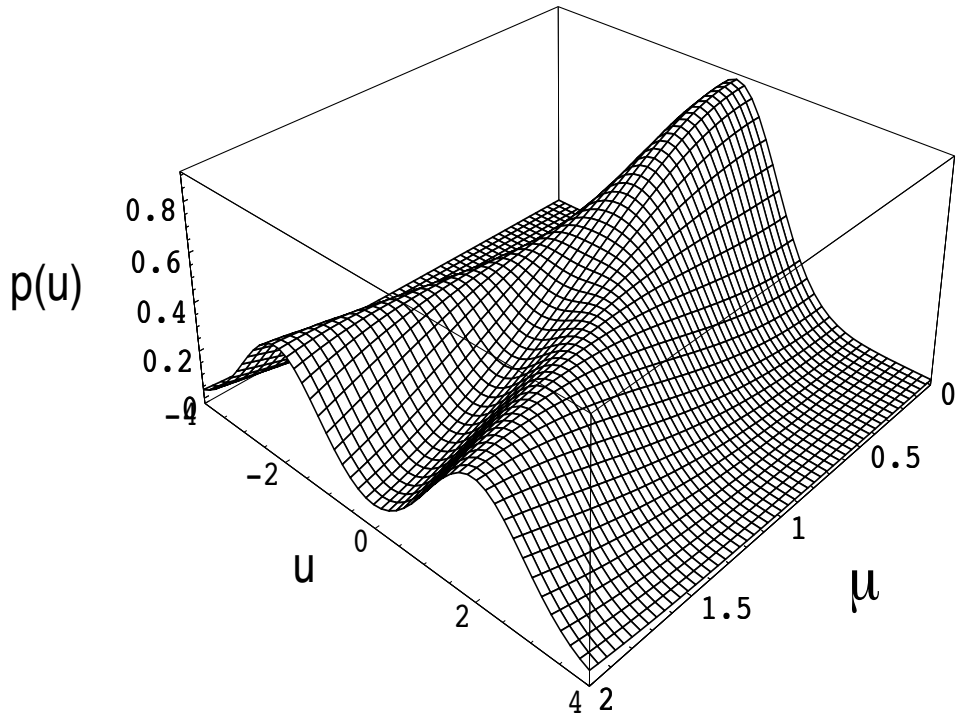

Figure 1: Estimated sub-Gaussian density models for the extended infomax learning rule with $\sigma^2 = 1$ and $\mu_i = \{0 \cdots 2\}$. The density becomes clearly bimodal when $\mu_i > 1$.

of Gaussians and show that they can separate sub and super-Gaussian sources. These parametric modeling approaches are in general computationally expensive. In addition, our empirical results on EEG and event related potentials (ERP) using contextual ICA indicate that cICA can fail to find independent components. We conjecture that this is due to the limited number of recorded time points (e.g. 600 data points for ERPs) from which a reliable density estimate is difficult.

2.2 Deriving a learning rule to separate sub- and super-Gaussian sources

The purpose of the extended infomax algorithm is to provide a simple learning rule with a fixed nonlinearity that can separate sources with a variety of distributions. One way of generalizing the learning rule to sources with either sub- or super-Gaussian distributions is to approximate the estimated p.d.f. with an Edgeworth expansion or Gram-Charlier expansion (Stuart and Ord, 1987) as proposed by Girolami and Fyfe (1997b). In Girolami (1997) a parametric density estimate was used to derive the same learning rule without making any approximations as we show below.

A symmetric strictly sub-Gaussian density can be modeled using a symmetrical form of the Pearson mixture model (Pearson, 1894) as follows (Girolami, 1998, 1997).

$$p(u) = \frac{1}{2}\left(N(\mu, \sigma^2) + N(-\mu, \sigma^2)\right) \tag{12}$$

where $N(\mu, \sigma^2)$ is the normal density with mean μ and variance σ^2. Figure 1 shows the form of the density $p(u)$ for $\sigma^2 = 1$ with varying $\mu = [0 \cdots 2]$. For $\mu = 0$ $p(u)$ is a Gaussian model whereas for e.g. $\mu_i = 1.5$ the $p(u)$ is clearly bimodal. The kurtosis k_4 (normalized 4^{th}-order cumulant) of $p(u)$ is

$$\kappa = \frac{c_4}{c_2^2} = \frac{-2\mu^4}{(\mu^2 + \sigma^2)^2} \tag{13}$$

where c_i is the i^{th}-order cumulant .(Girolami, 1997) Depending on the values of μ and σ^2 the kurtosis lies between -2 and 0. So equation 12 defines a strictly sub-Gaussian symmetric density when $\mu > 0$.

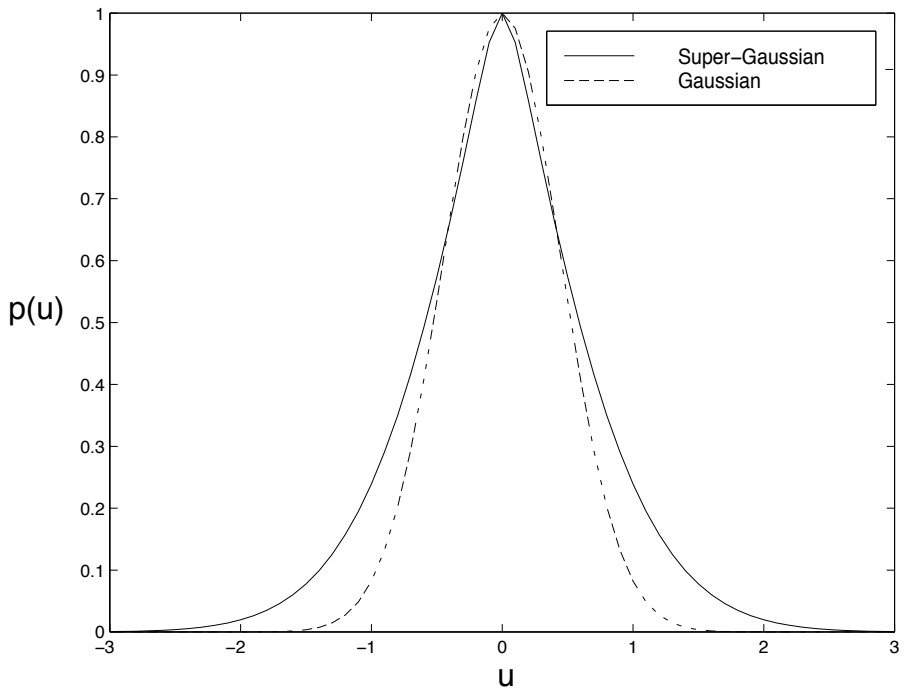

Figure 2: Density model for the super-Gaussian distribution. The super-Gaussian model has a heavier tail than the normal density.

Defining $a = \frac{\mu}{\sigma^2}$ and applying equation 12 we may write for $\varphi(u)$

$$\varphi(u) = -\frac{\frac{\partial p(u)}{\partial u}}{p(u)} = \frac{u}{\sigma^2} - a\left(\frac{\exp(au) - \exp(-au)}{\exp(au) + \exp(-au)}\right). \tag{14}$$

Using the definition of the hyperbolic tangent we can write

$$\varphi(u) = \frac{u}{\sigma^2} - \frac{\mu}{\sigma^2}\tanh\left(\frac{\mu}{\sigma^2}u\right). \tag{15}$$

Setting $\mu = 1$ and $\sigma^2 = 1$ equation 15 reduces to

$$\varphi(u) = u - \tanh(u). \tag{16}$$

The learning rule for strictly sub-Gaussian sources is now (equation 10 and equation 16)

$$\Delta\mathbf{W} \propto \left[\mathbf{I} + \tanh(\mathbf{u})\mathbf{u}^T - \mathbf{u}\mathbf{u}^T\right]\mathbf{W}. \tag{17}$$

In the case of unimodal super-Gaussian sources we adopt the following density model

$$p(u) \propto p_G(u)\mathrm{sech}^2(u) \tag{18}$$

where $p_G(u) = N(0,1)$ is a zero-mean Gaussian density with unit variance. Figure 2 shows the density model for $p(u)$. The nonlinearity $\varphi(u)$ is now

$$\varphi(u) = -\frac{\frac{\partial p(u)}{\partial u}}{p(u)} = u + \tanh(u). \tag{19}$$

The learning rule for super-Gaussian sources is (equation 10 and equation 19)

$$\Delta \mathbf{W} \propto \left[\mathbf{I} - \tanh(\mathbf{u})\mathbf{u}^T - \mathbf{u}\mathbf{u}^T\right]\mathbf{W}. \tag{20}$$

The difference between the super-Gaussian learning rule in equation 20 and the sub-Gaussian learning rule equation 17 is the sign before the tanh-function.

$$\Delta \mathbf{W} \propto \begin{cases} \left[\mathbf{I} - \tanh(\mathbf{u})\mathbf{u}^T - \mathbf{u}\mathbf{u}^T\right]\mathbf{W} & : \quad \text{super} - \text{Gaussian} \\ \left[\mathbf{I} + \tanh(\mathbf{u})\mathbf{u}^T - \mathbf{u}\mathbf{u}^T\right]\mathbf{W} & : \quad \text{sub} - \text{Gaussian} \end{cases} \tag{21}$$

The learning rules differ in the sign before the tanh-function and can be determined using a switching criterion. Girolami (1997) employs the sign of kurtosis of the unmixed sources as a switching criterion. However, as there is no general definition for sub- and super-Gaussian sources we chose a switching criterion based on stability criteria presented in the next subsection.

2.3 Switching between nonlinearities

The switching between the sub- and super-Gaussian learning rule is

$$\Delta \mathbf{W} \propto \left[\mathbf{I} - \mathbf{K}\tanh(\mathbf{u})\mathbf{u}^T - \mathbf{u}\mathbf{u}^T\right]\mathbf{W} \begin{cases} k_i = 1 & : \quad \text{super} - \text{Gaussian} \\ k_i = -1 & : \quad \text{sub} - \text{Gaussian} \end{cases} \tag{22}$$

where k_i are elements of the N-dimensional diagonal matrix $\mathbf{K}$. The switching parameter k_i can be derived from the generic stability analysis of separating solutions as employed by Cardoso and Laheld (1996) [4], Pham and Garrat (1997) and Amari et al. (1997). In the stability analysis the mean field is 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 are on average pulled back to the stationary point

A sufficient condition guaranteeing asymptotic stability can be derived (Cardoso, 1998a, 1998b) so that

$$\kappa_i > 0 \qquad 1 \leq i \leq N \tag{23}$$

where κ_i is

$$\kappa_i = E\{\varphi_i'(u_i)\}E\{u_i^2\} - E\{\varphi_i(u_i)u_i\} \tag{24}$$

and

$$\varphi_i(u_i) = u_i + k_i \tanh(u_i). \tag{25}$$

Substituting equation 25 in equation 24 gives

$$\begin{aligned} \kappa_i &= E\{k_i \text{sech}^2(u_i) + 1\}E\{u_i^2\} - E\{[k_i \tanh(u_i) + u_i]u_i\} & (26) \\ &= k_i \left(E\{\text{sech}^2(u_i)\}E\{u_i^2\} - E\{[\tanh(u_i)]u_i\}\right). & (27) \end{aligned}$$

To ensure $\kappa_i > 0$ the sign of k_i must be the same as the sign of $E\{\text{sech}^2(u_i)\}E\{u_i^2\} - E\{[\tanh(u_i)]u_i\}$. Therefore we can use the learning rule in equation 22 where the k_i's are

$$k_i = \text{sign}\left(E\{\text{sech}^2(u_i)\}E\{u_i^2\} - E\{[\tanh(u_i)]u_i\}\right). \tag{28}$$

2.4 The hyperbolic-Cauchy density model

We present another parametric density model that may be used for the separation of sub- and super-Gaussian sources. We define the parametric mixture density as

$$p(u) \propto \text{sech}^2(u + b) + \text{sech}^2(u - b). \tag{29}$$

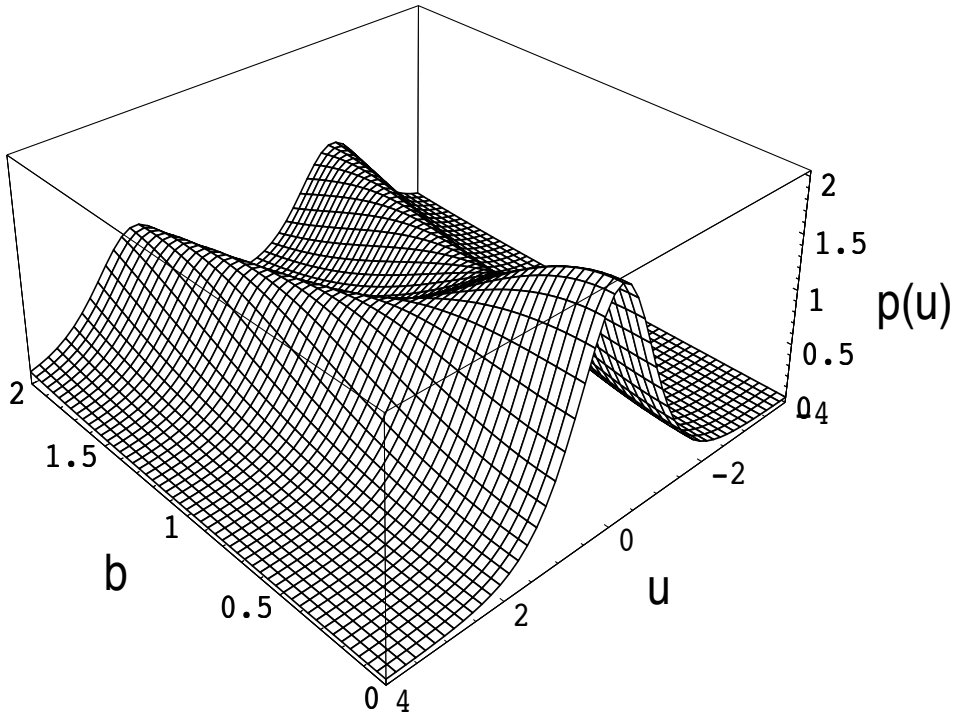

Figure 3: $p(u)$ as a function of b. For $b = 0$ the density estimate is suited to separate super-Gaussian sources. If for example $b = 2$ the density estimate is bimodal and therefore suited to separate sub-Gaussian sources.

Figure 3 shows the parametric density as a function of b. For $b = 0$ the parametric density is proportional to the hyperbolic-Cauchy distribution and is therefore suited for separating super-Gaussian sources. For $b = 2$ the parametric density estimator has a bimodal [5] distribution with negative kurtosis and is therefore suitable for separating sub-Gaussian sources:

$$\varphi(u) = -\frac{\partial}{\partial u} \log p(u) = -2\tanh(u) + 2\tanh(u+b) + 2\tanh(u-b). \tag{30}$$

The learning algorithm for sub- and super-Gaussian sources is now (equation 30 and equation 10)

$$\Delta \mathbf{W} \propto \left[\mathbf{I} + 2\tanh(\mathbf{u})\mathbf{u}^T - 2\tanh(\mathbf{u}+\mathbf{b})\mathbf{u}^T - 2\tanh(\mathbf{u}-\mathbf{b})\mathbf{u}^T\right]\mathbf{W}. \tag{31}$$

When $\mathbf{b} = \mathbf{0}$ (where $\mathbf{0}$ is a N-dim. vector with elements 0) then the learning rule reduces to

$$\Delta \mathbf{W} \propto \left[\mathbf{I} - 2\tanh(\mathbf{u})\mathbf{u}^T\right]\mathbf{W}. \tag{32}$$

which is exactly the learning rule in Bell and Sejnowski (1995) with the natural gradient extension. For $\mathbf{b} > \mathbf{1}$, the parametric density is bimodal (as shown in figure 3) and the learning rule is suitable for separating signals with sub-Gaussian distributions. Here again we may use the sign of the general stability criteria in equation 23 and κ_i in equation 24 to determine b_i so that we can switch between $b_i = 0$ and for example $b_i = 2$. In figure 4 we compare the range of kurtosis values of the parametric mixture density models in equation 12 and equation 29. The kurtosis value is shown as a function of the shaping parameter μ for the symmetric Pearson density model and b for the hyperbolic-Cauchy mixture density model. The kurtosis for the Pearson model is strictly negative except for $\mu = 0$ when the kurtosis is zero. Because the kurtosis for the hyperbolic-Cauchy model ranges from positive to negative, it may be used to separate signals with both sub- and super-Gaussian densities.

[4] see eqs. 40 and 41 in their paper.

[5] Symmetric bimodal densities considered in this paper are sub-Gaussian, however this is not always the case.

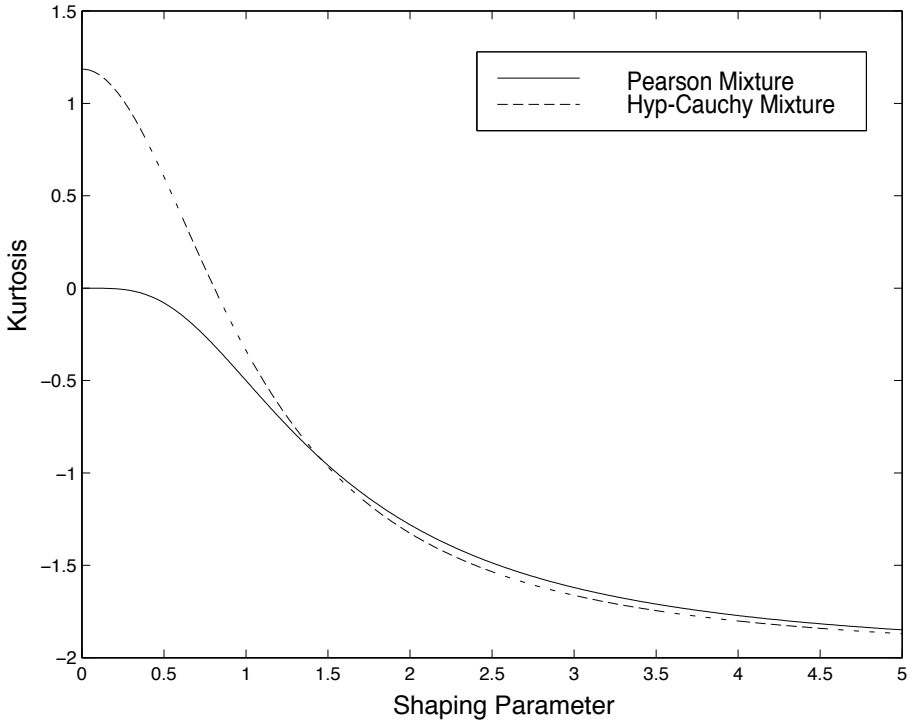

Figure 4: The kurtosis value is shown as a function of the shaping parameter μ and b (μ for the Pearson density model and b for the hyperbolic-Cauchy density model). Both models approach $k_4 = -2$ as the shaping parameter increases. The kurtosis for the Pearson model is strictly negative except for $\mu = 0$. The kurtosis for the hyperbolic-Cauchy model ranges from positive to negative so that we may use this single parametric model to separate signals with sub- and super-Gaussian densities.

3 Simulations and Experimental Results

Extensive simulations and experiments were performed on recorded data to verify the performance of the extended infomax algorithm equation 21. First, we show that the algorithm is able to separate a large number of sources with a wide variety of sub- and super-Gaussian distributions. We compared the performance of the extended infomax learning rule in equation 10 to the original infomax learning rule equation 11. Second, we performed a set of experiments on EEG data, which are high dimensional and include various noise sources.

3.1 10 Mixed Sound Sources

We obtained 10 mixed sound sources which were separated by contextual ICA as described in Pearlmutter and Parra (1996). No prewhitening is required since the transformation $\mathbf{W}$ is not restricted to a rotation in contrast to nonlinear PCA (Karhunen et al., 1997). All 55000 data points were passed 20 times through the learning rule using a block size (batch) of 300. This corresponds to 3666 iterations (weight updates). The learning rate was fixed at 0.0005. Figure 5 shows the error measure during learning. Both learning rules converged. The small variations of the extended infomax algorithm (upper curve) were due to the adaptation process of $\mathbf{K}$. The matrix $\mathbf{K}$ was initialized to the identity matrix and during the learning process the elements of $\mathbf{K}$ converge to -1 or 1 to extract sub- or super-Gaussian sources respectively. In this simulation example, sources 7,8 and 9 are close to Gaussian and slight variations of their density estimation change the sign. Annealing of the learning rate reduced the variation. All the music signals had super-Gaussians distribution and therefore were separable by the original infomax algorithm. The sources are already well separated after one pass through the data (about 10 sec on a Sparc 10 workstation using MATLAB) as shown in table 1:

−0.09	−0.38	0.14	−0.10	−0.06	0.93	−0.36	−0.54	0.17	**14.79**
-11.18	−0.01	0.14	0.05	−0.08	0.02	0.07	0.21	−0.12	−0.68
0.15	0.078	−0.08	−0.02	**10.19**	−0.02	0.15	0.05	0.07	0.17
0.39	0.61	−0.70	−0.07	0.14	0.32	−0.08	0.85	**7.64**	−0.16
0.04	0.76	**14.89**	0.03	0.03	−0.17	0.18	−0.31	−0.19	0.04
0.11	**12.89**	−0.54	−0.23	−0.43	−0.21	−0.12	0.05	0.07	0.18
0.45	0.16	−0.02	**6.53**	0.24	0.98	−0.39	−0.97	0.06	−0.08
0.31	0.14	0.23	0.03	−0.14	**-17.25**	−0.39	−0.25	0.19	0.39
−0.54	−0.81	0.62	0.84	−0.18	0.47	−0.04	**10.48**	−0.92	0.12
−0.08	−0.26	0.15	−0.10	0.49	0.01	**-10.25**	0.59	0.33	−0.94

Table 1: The performance matrix $\mathbf{P}$ (equation 34) for 10 mixed sound sources after one pass through the data. After one pass through the data $\mathbf{P}$ is already close to the identity matrix after rescaling and reordering.

For all experiments and simulations, a momentum term helped to accelerate the convergence of the algorithm:

$$\Delta\mathbf{W}(n+1) = (1-\alpha)\Delta\mathbf{W}(n) + \alpha\mathbf{W}(n) \tag{33}$$

where α takes into account the history of $\mathbf{W}$ and α can be increased with increasing number of weight updates (as $n \to \infty$, $\alpha \to 1$).

The performance during the learning process we monitored by the error measure that was proposed by Amari et al. (1996):

$$E = \sum_{i=1}^{N}\left(\sum_{j=1}^{N}\frac{|p_{ij}|}{\max_k|p_{ik}|} - 1\right) + \sum_{j=1}^{N}\left(\sum_{i=1}^{N}\frac{|p_{ij}|}{\max_k|p_{kj}|} - 1\right) \tag{34}$$

where p_{ij} are elements of the performance matrix $\mathbf{P} = \mathbf{WA}$. $\mathbf{P}$ is close to a permutation of the scaled identity matrix when the sources are separated. Figure 5 shows the error measure during the learning process.

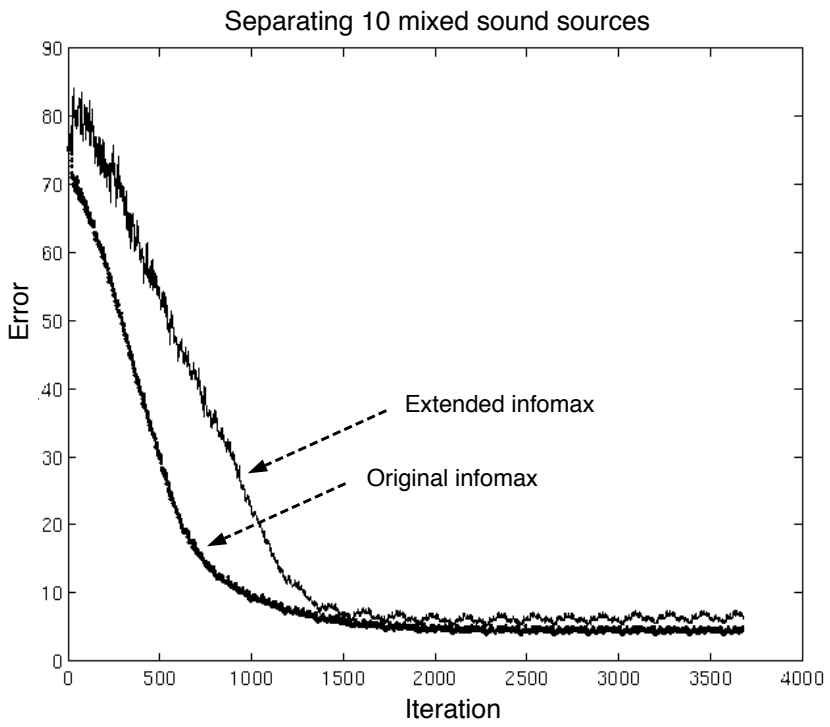

Figure 5: Error measure E in equation 34 for the separation of 10 sound sources. The upper curve is the performance for extended infomax and the lower curve shows the performance for the original infomax. The separation quality is shown in table 1.

To compare the speed of the extended infomax algorithm with another closely related ones, we separated the 10 mixed sound sources using the extended exploratory projection pursuit network with inhibitory lateral connections Girolami and Fyfe (1997a). The single feedforward neural network converged several times faster than this architecture using the same learning rate and a block size of 1. Larger block sizes can be used in the feedforward network but not the feedback networks, which increases the convergence speed considerably due to a more reliable estimate of the switching matrix $\mathbf{K}$.

3.2 20 Mixed Sound Sources

We separated the following 20 sources: 10 sound tracks obtained from Pearlmutter, 6 speech & sound signals used in Bell and Sejnowski (1995), 3 uniformly distributed sub-Gaussian noise signals and one noise source with a Gaussian distribution. The densities of the mixtures were close to the Gaussian distributions. The following parameters were used: learning rate fixed at 0.0005, block size of 100 data points, 150 passes through the data (41250 iterations).

Figure 6 shows the performance of the matrix $\mathbf{P}$ after the rows were manually reordered and normalized to unity. $\mathbf{P}$ is close to the identity matrix and its off diagonal elements indicate the amount of error. In this simulation we employ k_4 as a measure of the recovery of the sources. The original infomax algorithm separated most of the positive kurtotic sources. However, it failed to extract several sources including two super-Gaussian sources (music 7 & 8) with low kurtosis (0.78 and 0.46 respectively). In contrast, figure 7 shows that the performance matrix $\mathbf{P}$ for the extended infomax algorithm is close to the identity matrix. In a listening test, there was a clear separation of all sources from their mixtures. Note that although the sources ranged from Laplacian distribu-

Source number	Source type	Original kurtosis	Recovered kurtosis (infomax)	Recovered kurtosis (ext. infomax)	SNR (ext. infomax)
1	Music 1	2.4733	2.4754	2.4759	43.4
2	Music 2	1.5135	1.5129	1.5052	55.2
3	Music 3	2.4176	2.4206	2.4044	44.1
4	Music 4	1.076	1.0720	1.0840	31.7
5	Music 5	1.0317	1.0347	1.0488	43.6
6	Music 6	1.8626	1.8653	1.8467	48.1
7	Music 7	0.7867	0.8029	0.7871	32.7
8	Music 8	0.4639	0.2753	0.4591	29.4
9	Music 9	0.5714	0.5874	0.5733	36.4
10	Music 10	2.6358	2.6327	2.6343	46.4
11	Speech 1	6.6645	6.6652	6.6663	54.3
12	Speech 2	3.3355	3.3389	3.3324	50.5
13	Music 11	1.1082	1.1072	1.1053	48.1
14	Speech 3	7.2846	7.2828	7.2875	50.5
15	Music 12	2.8308	2.8198	2.8217	52.6
16	Speech 4	10.8838	10.8738	10.8128	57.1
17	Uni. Noise 1	-1.1959	-0.2172	-1.1955	61.4
18	Uni. Noise 2	-1.2031	-0.2080	-1.2013	67.7
19	Uni. Noise 3	-1.1966	-0.2016	-1.1955	63.6
20	Gauss. Noise	-0.0148	-0.0964	-0.0399	24.9

Table 2: Kurtosis of the 20 original signal sources and the kurtosis of the recovered signals from original infomax and extended infomax. The source signals range from highly kurtotic speech signals, Gaussian noise (kurtosis is zero) to noise sources with uniform distribution (negative kurtosis). Boxes are placed around sources that failed to clearly separate. In addition, the SNR is computed for extended infomax.

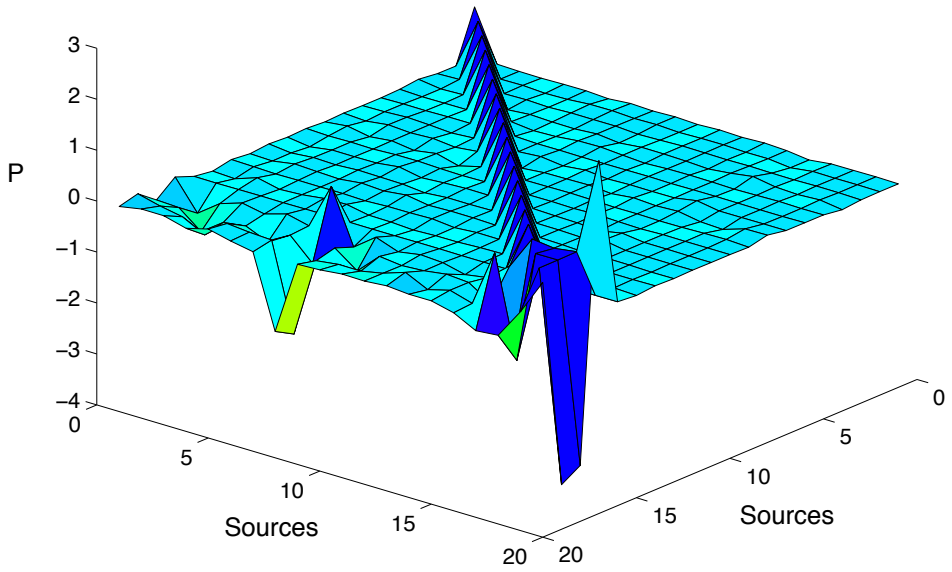

Figure 6: Performance matrix $\mathbf{P}$ for the separation of 20 sources using the original infomax algorithm after normalizing and reordering. Most super-Gaussian sources were recovered. However, the three sub-Gaussian sources (17,18,19), the Gaussian source (20) and two super-Gaussian sources (7, 8) remain mixed and aliased in other sources. In total, 14 sources were extracted and 6 channels remained mixed (see Table 2).

tions ($p(s) \propto \exp(-|s|)$, e.g. speech), Gaussian noise, to uniformly distributed noise, they were all separated using one nonlinearity.

The simulation results suggest that the super-Gaussian and sub-Gaussian density estimates in equation 12 and equation 18 are sufficient to separate the true sources. The learning algorithms in equation 21 and equation 31 performed almost identically.

3.3 EEG Recordings

In electroencephalographic (EEG) recordings of brain electrical activity from the human scalp, artifacts such as line noise, eye movements, blinks and cardiac signals (EKG) pose serious problems in analyzing and interpreting the recordings. Regression methods have been used to partially remove eye movement from the EEG data (Berg and Scherg, 1991); other artifacts such as electrode noise, cardiac signals and muscle noise are even more difficult to remove. Recently, Makeig et al. (1996) have applied ICA to the analysis of EEG data using the original infomax algorithm. They showed that some artifactual components can be isolated from overlapping EEG signals including alpha and theta bursts.

We analyzed EEG data that were collected to develop a method of objectively monitoring the alertness of operators listening for auditory signals (Makeig and Inlow, 1993). During a half-hour session, the subject was asked to push a button whenever they detected an auditory target stimulus. EEG was collected from 14 electrodes located at sites of the International 10-20 System (Makeig et al., 1997) at a sampling rate of 312.5 Hz. The extended infomax algorithm was applied to the

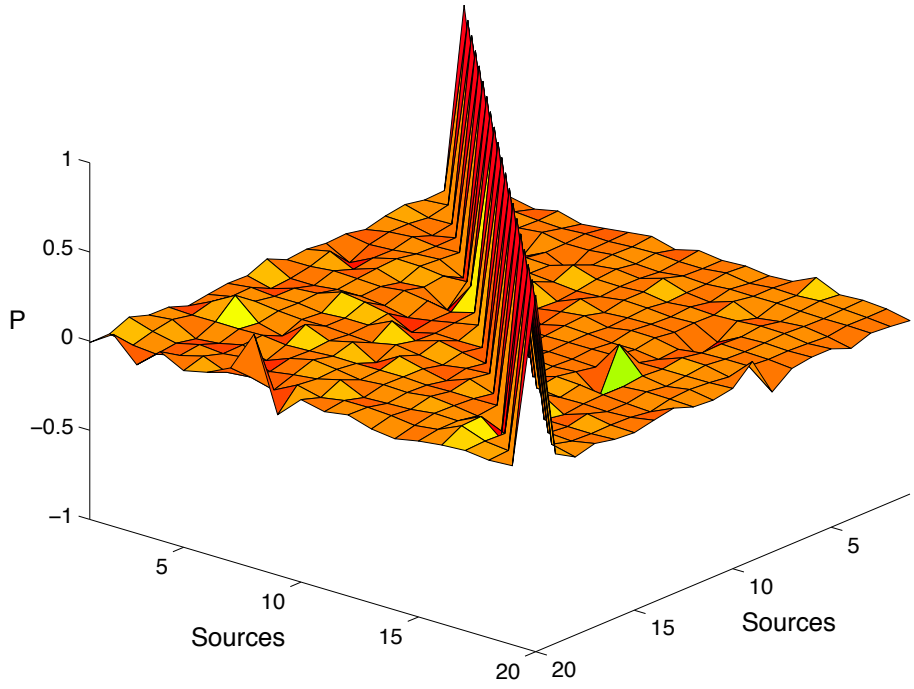

Figure 7: Performance matrix **P** for the separation of 20 sources using the extended infomax algorithm after normalizing and reordering. **P** is approximately the identity matrix which indicates nearly perfect separation.

14 channels of 10 seconds of data with the following parameters: learning rate fixed at 0.0005, 100 passes with block size of 100 (3125 weight updates). The power spectrum was computed for each channel and the power in a band around 60 Hz was used to compute the relative power for each channel and each separated component.

Figure 8 shows the time course of 14 channels of EEG and figure 9 shows the independent components found by the extended infomax algorithm. Several observations on the ICA components in figure 9 and its power spectrum are of interest:

- Alpha bursts (about 11 Hz) were detected in components 1 and 5. Alpha band activity (8-12 Hz) occurs most often when eyes are closed and the subject is relaxed. Most subjects have more than one alpha rhythm, with somewhat different frequencies and scalp patterns.
- Theta bursts (about 7 Hz) were detected in components 4, 6 and 9. Theta-band rhythms (4-8 Hz) may occur during drowsiness and transient losses of awareness or microsleeps (Makeig and Inlow, 1993), but frontal theta bursts may occur during intense concentration.
- An eye blink was isolated in component 2 at 8 sec.
- 60 Hz line noise was concentrated in component 3 (see bottom of figure 10).

Figure 10 (top)shows power near 60 Hz distributed in all EEG channels but predominantly in components 4, 13 and 14. Figure 10 (middle) shows that the original infomax cannot concentrate the line noise into one component. In contrast, extended infomax (figure 10, bottom panel) concentrates it mainly in one sub-Gaussian component, channel 3.

Figure 11 shows another EEG data set with 23 channels including 2 EOG (electrooculogram) channels. The eye blinks near 5 sec and 7 sec contaminated all of the channels. Figure 12 shows the ICA components without normalizing the components with respect to their contribution to the raw data. ICA component 1 in figure 12 contained the pure eye blink signal. Small periodic muscle spiking at the temporal sites (T3 and T4) was extracted into ICA component 14.

Experiments with several different EEG data sets confirmed that the separation of artifactual signals was highly reliable. In particular, severe line noise signals could always be decomposed into one or two components with sub-Gaussian distributions. Jung et al. (1998) show further that eye movement also can be extracted.

4 Discussion

4.1 Applications to real world problems

The results reported here for the separation of eye-movement artifacts from EEG recordings have immediate application to medical and research data. Independently, Vigario et al. (1996) reported similar findings for EEG recordings using a fixed-point algorithm for ICA (Hyvaerinen and Oja, 1997). It would be useful to compare this and other ICA algorithms on the same data sets to assess their merits. Compared to traditional techniques in EEG analysis extended infomax requires less supervision and is easy to apply (see Makeig et al. (1997); Jung et al. (1998)). In addition to the very encouraging results on EEG data given here, McKeown et al. (1998) have demonstrated another successful use of the extended infomax algorithm on fMRI recordings. They investigated task-related human brain activity in fMRI data. In this application, they considered both spatial and temporal ICA and found that the extended infomax algorithm extracted sub-Gaussian temporal components that could not be extracted with the original infomax algorithm.

4.2 Limitations and future research

The extended infomax learning algorithm makes several assumptions that limit its effectiveness.

First, the algorithm requires the number of sensors to be the same or greater than the number of sources ($N \geq M$). The case when there are more sources than sensors, $N < M$, is of theoretical and practical interest. Given only one or two sensors that observe more than two sources can

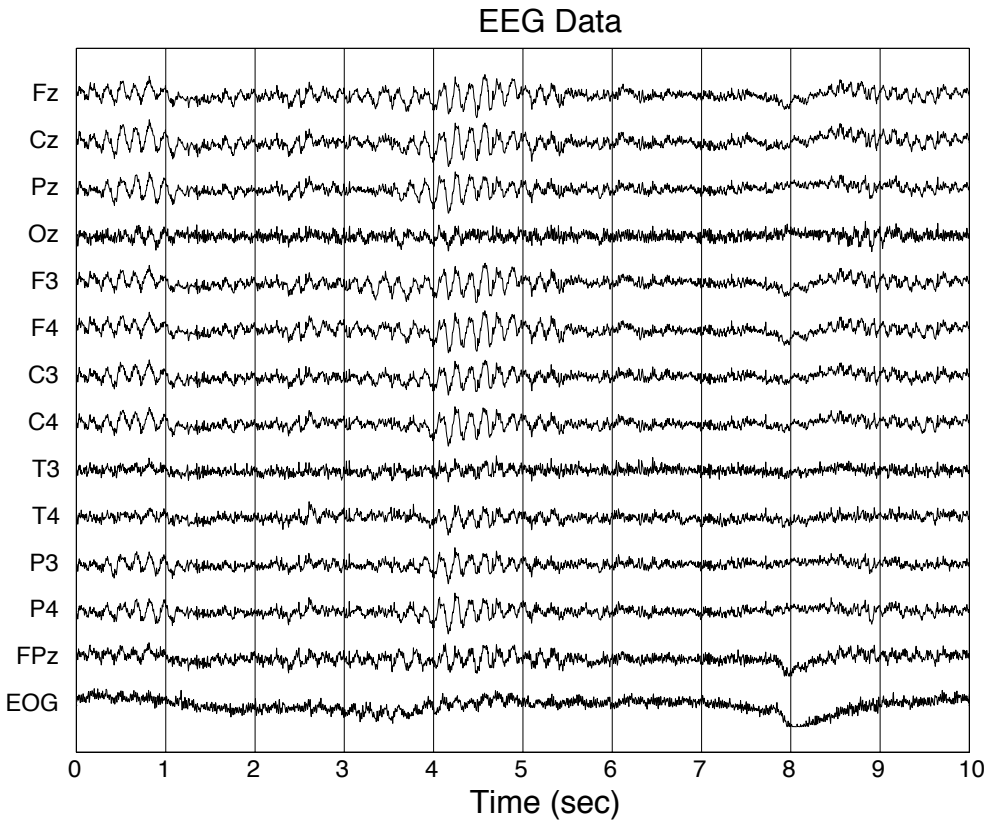

Figure 8: A 10-sec portion of the EEG time series with prominent alpha rhythms (8-21 Hz). The location of the recording electrode from the scalp is indicated on the left of each trace. The electrooculogram (EOG) recording is taken from the temples.

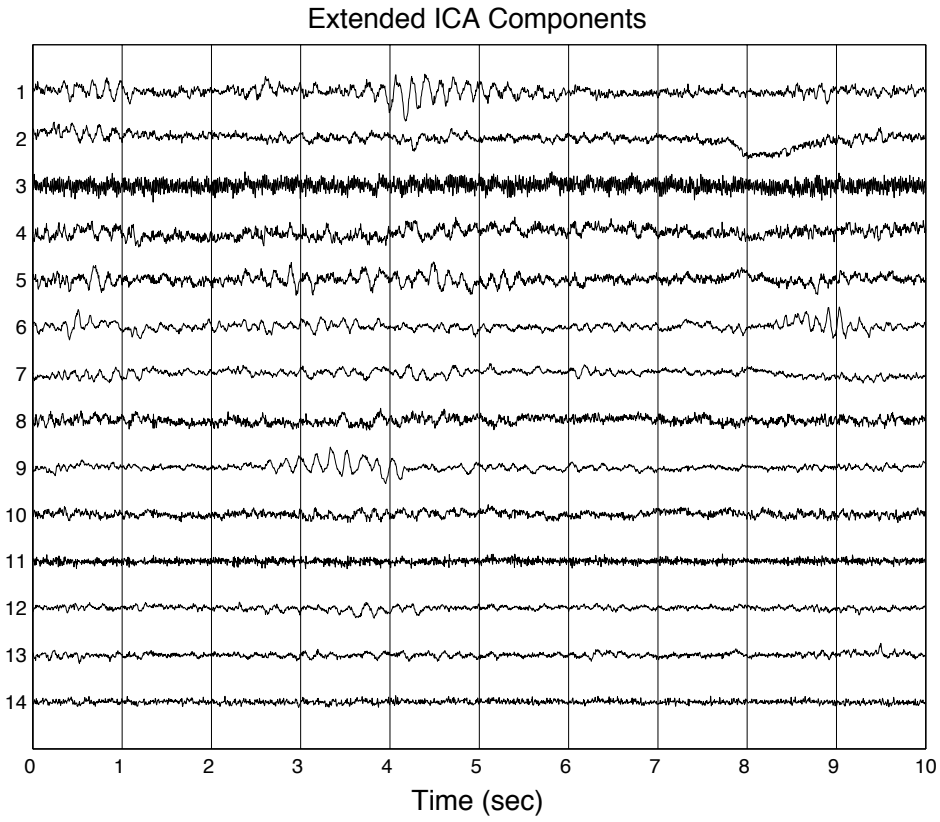

Figure 9: The 14 ICA components extracted from the EEG data in figure 8. The components 3, 4, 7, 8 and 10 have sub-Gaussian distributions and the others have super-Gaussian distributions. There is an eye movement artifact at 8 seconds. Line noise is concentrated in component 3. The prominent rhythms in components 1,4,5,6 and 9 have different time courses and scalp distributions.

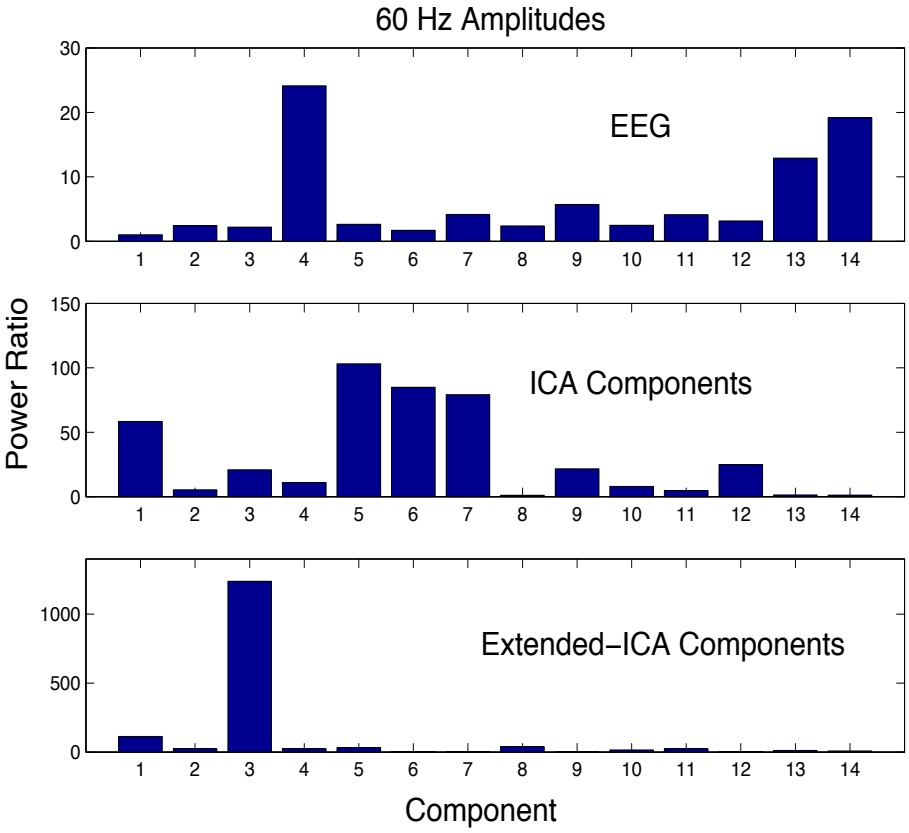

Figure 10: Top: Ratio of power near 60 Hz over 14 components for EEG data in figure 8. Middle: Ratio of power near 60 Hz for the 14 infomax ICA components. Bottom: Ratio of power near 60 Hz for the 14 extended infomax ICA components in figure 9. Note the difference in scale by a factor of 10 between the original infomax and the extended infomax.

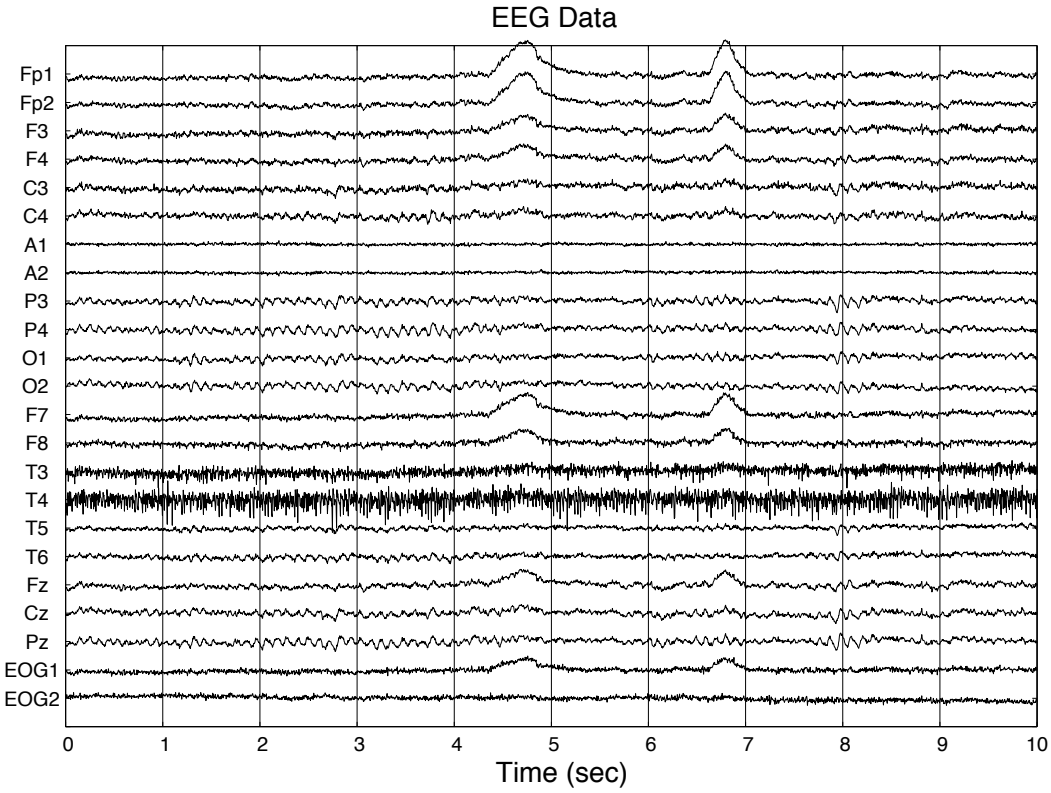

Figure 11: EEG data set with 23 channels including 2 EOG channels. At around 4-5 sec and 6-7 sec artifacts from severe eye blinks contaminate the data set.

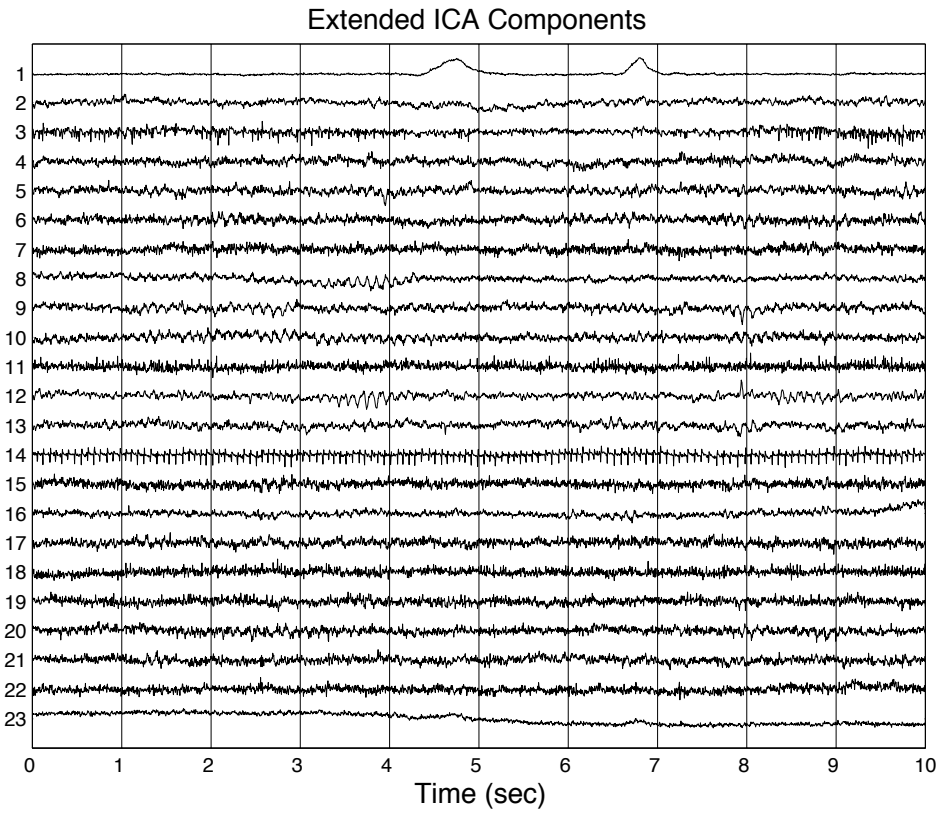

Figure 12: Extended infomax ICA components derived from the EEG recordings in figure 11. The eye blinks are clearly concentrated in component 1. Component 14 contains the steady state signal.

we still recover all sources? Preliminary results by Lewicki and Sejnowski (1998) suggest that an overcomplete representation of the data can to some extent extract the independent components using a priori knowledge of the source distribution. This has been applied by Lee et al. (1998b) to separate three sources from two sensors.

Second, researchers have recently tackled the problem of nonlinear mixing phenomena. Yang et al. (1997), Taleb and Jutten (1997) and Lee et al. (1997) propose extensions when linear mixing is combined with certain nonlinear mixing models. Other approaches use self-organizing feature maps to identify nonlinear features in the data (Lin and Cowan, 1997; Pajunen and Karhunen, 1997). More recently, Hochreiter and Schmidhuber (1998) have proposed low complexity coding and decoding approaches for nonlinear ICA.

Third, sources may not be stationary, i.e. sources may appear and disappear and move (speaker moving in a room). In these cases, the weight matrix $\mathbf{W}$ may change completely from one time point to the next. This is a challenging problem for all existing ICA algorithms. A method to model the context switching (non-stationary mixing matrix) in an unsupervised way is proposed in Lee et al. (1998c).

Fourth, sensor noise may influence separation and should be included in the model (Nadal and Parga, 1994; Moulines et al., 1997; Attias, 1998). Much more work needs to be done to determine the effect of noise on performance.

In addition to these limitations, there are other issues that deserve further research. In particular, it remains an open question to what extent the learning rule is robust to parametric mismatch given a limited number of data points.

Despite these limitations, the extended infomax ICA algorithm presented here should have many applications where both sub-Gaussian and super-Gaussian sources need to be separated without additional prior knowledge of their statistical properties.

4.3 Conclusions

The extended infomax ICA algorithm proposed here is a promising generalization that satisfies a general stability criterion for mixed sub-Gaussian and super-Gaussian sources (Cardoso and Laheld, 1996). Based on the learning algorithm first derived by Girolami (1997) and the natural gradient, the extended infomax algorithm has shown excellent performance on several large real data sets derived from electrical and blood flow measurements of functional activity in the brain. Compared to the originally proposed infomax algorithm (Bell and Sejnowski, 1995), the extended infomax algorithm separates a wider range of source signals whilst maintaining its simplicity.

Acknowledgments

T.W. Lee is supported by the German Academic Exchange Program. M. Girolami is supported by a grant from NCR Financial Systems (Ltd), Knowledge Laboratory, Advanced Technology Development Division, Dundee, Scotland. T.J. Sejnowski is supported by the Howard Hughes Medical Institute. We are much indebted to Jean-François Cardoso for insights and helpful comments on the stability criteria and Tony Bell for general comments and discussions. We are grateful to Tzyy-Ping Jung and Scott Makeig for EEG data as well as useful discussions and comments, and to Olivier Coenen for helpful comments. We thank the reviewers for fruitful comments.

References

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, in press.

Amari, S. and Cardoso, J.-F. (1997). Blind source separation — semiparametric statistical approach. *IEEE Trans. on Signal Processing*, 45(11):2692–2700.

Amari, S., Chen, T.-P., and Cichocki, A. (1997). Stability analysis of adaptive blind source separation. *Neural Networks*, 10(8):1345–1352.

Amari, S., Cichocki, A., and Yang, H. (1996). A New Learning Algorithm for Blind Signal Separation. In *Advances in Neural Information Processing Systems 8*, pages 757–763.

Attias, H. (1998). Blind separation of noisy mixtures: An em algorithm for independent factor analysis. *Neural Computation*, submitted.

Bell, A. J. and Sejnowski, T. J. (1995). An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7:1129–1159.

Berg, P. and Scherg, M. (1991). Dipole models of eye movements and blinks. *Electroencephalog. clin. Neurophysiolog.*, pages 36–44.

Cardoso, J. and Soloumiac, A. (1993). Blind beamforming for non-Gaussian signals. *IEE Proceedings-F*, 140(46):362–370.

Cardoso, J.-F. (1997). Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Letters*, 4(4):112–114.

Cardoso, J.-F. (1998a). Blind signal processing: a review. *Proceedings of IEEE*. to appear.

Cardoso, J.-F. (1998b). *Unsupervised adaptive filtering*, chapter Entropic contrasts for source separation. S. Haykin (editor) Prentice Hall. to appear.

Cardoso, J.-F. and Laheld, B. (1996). Equivariant adaptive source separation. *IEEE Trans. on S.P.*, 45(2):434–444.

Cichocki, A., Unbehauen, R., and Rummert, E. (1994). Robust learning algorithm for blind separation of signals. *Electronics Letters*, 30(17):1386–1387.

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

Cover, T. and Thomas, J., editors (1991). *Elements of Information Theory*, volume 1. John Wiley and Sons, New York.

Deco, G. and Obradovic, D. (1996). *An Information-Theoretic Approach to Neural Computing*. Springer Verlag, ISBN 0-387-94666-7.

Gaeta, M. and Lacoume, J.-L. (1990). Source separation without prior knowledge: the maximum likelihood solution. *Proc. EUSIPO*, pages 621–624.

Girolami, M. (1997). Self-organizing artificial neural networks for signal separation. Ph.d. thesis, Department of Computing and Information Systems, Paisley University, Scotland.

Girolami, M. (1998). An alternative perspective on adaptive independent component analysis algorithms. *Neural Computation*, to appear.

Girolami, M. and Fyfe, C. (1997a). 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.

Girolami, M. and Fyfe, C. (1997b). Generalised independent component analysis through unsupervised learning with emergent bussgang properties. In *Proc. ICNN*, pages 1788–1891, Houston, USA.

Hochreiter, S. and Schmidhuber, J. (1998). Feature extraction through lococode. *Neural Computation*, to appear.

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

Jung, T.-P., Humphries, C., Lee, T.-W., Makeig, S., McKeown, M., Iragui, V., and Sejnowski, T. J. (1998). Extended ica removes artifacts from electroencephalographic recordings. *Advances in Neural Information Processing Systems 10*, pages 894–900.

Jutten, C. and Herault, J. (1991). Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10.

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

Karhunen, J., Oja, E., Wang, L., Vigario, R., and Joutsensalo, J. (1997). A class of neural networks for independent component analysis. *IEEE Trans. on Neural Networks*, 8:487–504.

Lee, T.-W., Girolami, M., Bell, A. J., and Sejnowski, T. J. (1998a). A unifying framework for independent component analysis. *International Journal on Mathematical and Computer Models*, in press.

Lee, T.-W., Koehler, B., and Orglmeister, R. (1997). Blind separation of nonlinear mixing models. In *IEEE NNSP*, pages 406–415, Florida, USA.

Lee, T.-W., Lewicki, M. S., Girolami, M., and Sejnowski, T. J. (1998b). Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, submitted.

Lee, T.-W., Lewicki, M. S., and Sejnowski, T. J. (submitted 1998c). Unsupervised classification with non-gaussian mixture models using ica. In *Advances in Neural Information Processing Systems 11*. MIT Press.

Lewicki, M. and Sejnowski, T. J. (1998). Learning nonlinear overcomplete represenations for efficient coding. In *Advances in Neural Information Processing Systems 10*, pages 815–821.

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

Makeig, S., Bell, A. J., Jung, T., and Sejnowski, T. J. (1996). Independent Component Analysis of Electroencephalographic Data. *Advances in Neural Information Processing Systems 8*, pages 145–151.

Makeig, S. and Inlow, M. (1993). Changes in the eeg spectrum predict fluctuations in error rate in an auditory vigilance task. In *Society for Psychophysiology*, volume 28:S39.

Makeig, S., Jung, T., Bell, A. J., Ghahremani, D., and Sejnowski, T. J. (1997). Blind Separation of Event-related Brain Response into spatial Independent Components. *Proc. of the National Academy of Sciences*, 94:10979–10984.

McKeown, M., Makeig, S., Brown, G., Jung, T.-P., Kindermann, S., Lee, T.-W., and Sejnowski, T. J. (1998). Spatially independent activity patterns in functional magnetic resonance imaging data during the stroop color-naming task. *Proceedings of the National Academy of Sciences*, 95:803–810.

Moulines, E., Cardoso, J.-F., and Cassiat, E. (1997). Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proc. ICASSP'97*, volume 5, pages 3617–3620, Munich.

Nadal, J.-P. and Parga, N. (1994). Non linear neurons in the low noise limit : a factorial code maximizes information transfer. *Network*, 5:565–581.

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

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

Pajunen, P. and Karhunen, J. (1997). A maximum likelihood approach to nonlinear blind source separation. In *ICANN*, pages 541–546, Lausanne.

Pearlmutter, B. and Parra, L. (1996). A context-sensitive generalization of ICA. In *International Conference on Neural Information Processing*, pages 151–157.

Pearson, K. (1894). Contributions to the mathematical study of evolution. *Phil. Trans. Roy. Soc. A*, 185(71).

Pham, D.-T. and Garrat, P. (1997). Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Trans. on Signal Proc.*, 45(7):1712–1725.

Roth, Z. and Baram, Y. (1996). Multidimensional density shaping by sigmoids. *IEEE Trans. on Neural Networks*, 7(5):1291–1298.

Stuart, A. and Ord, J. (1987). *Kendall's Advanced Theory of Statistic, 1, Distribution Theory*. John Wiley, New York.

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

Vigario, R., Hyvaerinen, A., and Oja, E. (1996). Ica fixed-point algorithm in extraction of artifacts from eeg. In *IEEE Nordic Signal Processing Symposium*, pages 383–386, Espoo, Finland.

Xu, L., Cheung, C., Yang, H., and Amari, S. (1997). Maximum equalization by entropy maximization and mixture of cumulative distribution functions. In *Proc. of ICNN'97*, pages 1821–1826, Houston.

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