Independent Component Analysis for Mixed Sub-Gaussian and Super-Gaussian Sources

Te-Won Lee
Technische Universität Berlin AND
Computational Neurobiology Lab
The Salk Institute
10010 N. Torrey Pines Road
La Jolla, California 92037, USA
tewon@salk.edu

Terrence J. Sejnowski
Howard Hughes Medical Institute
Computational Neurobiology Lab
The Salk Institute
10010 N. Torrey Pines Road
La Jolla, California 92037, USA
terry@salk.edu

Abstract

An extension of the infomax algorithm of Bell & Sejnowski (1995) is presented that is able to separate the mixed sub- and super-Gaussian source distributions. The same learning rule has been derived by Girolami & Fyfe (1997) from the negentropy perspective for projection pursuit. Using a Laplacian prior we also propose a learning rule that is especially convenient to realize in hardware. The natural gradient extension is presented from different perspectives and the use of preprocessing steps is proposed to further speed up the convergence. Simulation results show that the algorithm is able to separate 20 source with a variety of source distributions. On real data, Jung et al. (1997) and McKeown et al. (1997) demonstrate the successful use of the extended ICA algorithm to analyze EEG and fMRI recordings.

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 [11, 12], telecommunications and medical signal processing [8, 14]. The goal of ICA is to recover independent sources given sensor outputs in which the sources have been linearly mixed. In contrast to correlation based solutions 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.

The blind source separation problem has been studied by researchers in the field of neural networks [1, 2, 5, 7, 9, 15, 16] and statistical signal processing [3, 6, 11]. Bell & Sejnowski [2] have developed an unsupervised learning algorithm based on entropy maximization in a feedforward neural network. The algorithm uses a sigmoidal activation function that is especially suited to separate sources that

have higher kurtosis than the Gaussian distribution (super-Gaussian). We use a related information-theoretic algorithm that preserves the simple architecture in Bell & Sejnowski and allows an extension to the separation of mixtures of super-Gaussian and sub-Gaussian sources. Girolami & Fyfe [7] have derived this learning rule from the negentropy viewpoint and use it for extended exploratory pursuit. We show here, that this algorithm can successfully separate 20 mixtures of the following sources: 10 sound tracks obtained from Pearlmutter, 6 speech/sound signals used in Bell & Sejnowski (1995), 3 uniformly distributed sub-Gaussian noise signals and one noise source with a Gaussian distribution.

Recently, Jung et al. [8] have successfully applied the extended ICA algorithm to remove artifacts in electroencephalographic (EEG) recordings. To this end, the raw data is blindly decomposed into independent components such as line noise, eye movements and muscle movements. After eliminating these artifactual components, the 'corrected' EEG data are free of these artifacts. Furthermore, McKeown et al. [14], show how the extended ICA algorithm can be used to find transient time-locked signals in fMRI data.

2 Algorithm

Bell and Sejnowski [2] have proposed a simple neural network algorithm that blindly separates mixtures $\mathbf{x}$ of independent sources $\mathbf{s}$ using infomax. They show that maximizing the joint entropy $H(\mathbf{y})$ of the output of a neural processor minimizes the mutual information among the output components $y_i = g(u_i)$ where $g(u_i)$ is an invertible bounded nonlinearity and $\mathbf{u} = \mathbf{W}\mathbf{x}$. Pearlmutter and Parra [15] derive the same learning rule from a Maximum-Likelihood (ML) density estimation approach using the Kullback-Leibler distance measure:

$$D(p,\hat{p}) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{\hat{p}(\mathbf{x};\mathbf{w})} d\mathbf{x} = H(p(\mathbf{x})) - \int p(\mathbf{x}) \log \hat{p}(\mathbf{x};\mathbf{w}) \tag{1}$$

$p(\mathbf{x})$ is the probability density function (pdf) of the observation $\mathbf{x}$ and $\hat{p}(\mathbf{x};\mathbf{w})$ is a parametric estimate of the distribution of the independent sources. In [4] Cardoso shows that infomax and ML are equivalent because the relation between the KL-distance and the ML differs by the constant Entropy $H(\mathbf{x})$ which is not dependent on $\mathbf{W}$.

$$L = -\int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{\hat{p}(\mathbf{x};\mathbf{w})} d\mathbf{x} - H(p_{\mathbf{x}}) \tag{2}$$

Girolami & Fyfe [7] start from the negentropy point of view and use a kurtosis measure for projection pursuit:

$$N(p_{\mathbf{u}}) = H(p_G) - H(p_{\mathbf{u}}) \tag{3}$$

where $H(p_G)$ is the entropy of the Gaussian distribution and $H(p_u)$ is the entropy of the estimated sources. Negentropy $N(p_{\mathbf{u}})$ from the ML perspective as a measure of the KL-distance of a transformed vector $\mathbf{u}$ to normality. Since the observation $\mathbf{x}$ is close to the Gaussian distribution for a linear mixing of independent variables due to the central limit theorem, the difference between maximizing the distance to the observation or to a Gaussian distribution does not matter in practice. In all three approaches the output entropy $H(\mathbf{y})$ of a neural processor is maximized which implies approximating the output density in the sense of minimum KL-distance, by a uniform density. The algorithm shapes the signal $\mathbf{u}$ according to the derivative of the activation function $p(u) = \partial g(u)/\partial u$ and makes u_i as independent as possible. Independence is achieved through the nonlinear squashing function for example a sigmoid function, which provides a combination of higher-order statistics through its Taylor series expansion. We can relate $p_{\mathbf{x}}(\mathbf{x})$ to $p_{\mathbf{y}}(\mathbf{y})$ by the determinant of the Jacobian:

$$p_{\mathbf{y}}(\mathbf{y}) = \frac{p_{\mathbf{x}}(\mathbf{x})}{|\det J(\mathbf{x})|} \tag{4}$$

Evaluating the expected value of the logarithmic representation for eq.4 gives the output entropy which can be maximized with respect to $\mathbf{W}$ [2] which is equivalent to maximizing the volume of the

Jacobian of the transfer function.

$$\frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} = \mathbf{W}^{-T} + \left(\frac{\frac{\partial p(\mathbf{u})}{\partial \mathbf{u}}}{p(\mathbf{u})}\right)\mathbf{x}^T \tag{5}$$

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

$$\Delta \mathbf{W} \propto \frac{\partial H(\mathbf{y})}{\partial \mathbf{W}}\mathbf{W}^T\mathbf{W} = \left[\mathbf{I} + \left(\frac{\frac{\partial p(\mathbf{u})}{\partial \mathbf{u}}}{p(\mathbf{u})}\right)\mathbf{u}^T\right]\mathbf{W} \tag{6}$$

As reported by Amari et al. [1], here $\mathbf{W}^T\mathbf{W}$ is an optimal rescaling of the entropy gradient, simplifying the learning rule in [2] and speeding convergence considerably.

Theoretically, the form of the nonlinearity $g(u)$ plays an essential role in the success of the algorithm. The ideal form for $g(u)$ is the cumulative density function (cdf) of the distributions of the independent sources. In practice however, if we choose $g(u)$ to be a sigmoid function the learning rule reduces to that proposed in [2]. The algorithm is then limited to separating sources with super-Gaussian distributions. Therefore, the purpose of an extended ICA algorithm is to provide a learning rule that can separate a variety of distributions. A more general, but computationally burdensome solution is to use contextual ICA [15] where the pdf is modeled in a parametric form by taking into account the temporal information. Pearlmutter and Parra choose to make p_i a weighted sum of logistic density functions with variable means and scales, and make these means linear functions of the recent history of source i.

$$p_i(u_i(t)|u(t-1),...;\mathbf{w}_i) = \sum_{k=1}^{K}\frac{m_{ik}}{\sigma_{ik}}\frac{\partial g(u_i)}{\partial u_i}\left(\frac{u_i(t)-\bar{u}_{ik}}{\sigma_{ik}}\right) \tag{7}$$

where m_{ik} are the weighting parameters and σ_{ik} are the scaling parameters. The component means $\bar{u}_{ik}$ are linear functions of the recent time samples of the source.

Another way of generalizing the learning rule to sources with either sub- or super-Gaussian distributions is to approximate the estimated pdf with an Edgeworth expansion or Gram-Charlier expansion [6]. The n^{th}-order Edgeworth expansion of the estimated sources $\mathbf{u}$ is given as the sum of the pdf of Gaussian approximations. Girolami & Fyfe [7] use a 4^{th}-order Edgeworth expansion and make approximations for two cases. For sub-Gaussians, the following approximation is possible:

$$\frac{\frac{\partial p(u_i)}{\partial u_i}}{p(u_i)} = +\tanh(u_i) - u_i \tag{8}$$

whereas for super-Gaussians, the approximation can be made as follows:

$$\frac{\frac{\partial p(u_i)}{\partial u_i}}{p(u_i)} = -\tanh(u_i) - u_i \tag{9}$$

The sign flip can be substituted by the normalized kurtosis k_4 which can be computed adaptively from the estimated sources $\mathbf{u}$:

$$k_4(u_i) = \frac{\text{cumulant}_4(u_i)}{(E\{u_i^2\})^2} = \frac{E\{u_i^4\} - 3(E\{u_i^2\})^2}{(E\{u_i^2\})^2} = \frac{E\{u_i^4\}}{(E\{u_i^2\})^2} - 3 \tag{10}$$

The learning rule extracted from eq.8, eq.9 and eq.6 is then:

$$\Delta \mathbf{W} \propto \frac{\partial H(\mathbf{y})}{\partial \mathbf{W}}\mathbf{W}^T\mathbf{W} = \left[\mathbf{I} - \text{sign}(k_4)\tanh(\mathbf{u})\mathbf{u}^T - \mathbf{u}\mathbf{u}^T\right]\mathbf{W} \tag{11}$$

Intuitively, for super-Gaussians the term $(-\tanh(\mathbf{u})\mathbf{u}^T)$ corresponds to an anti-Hebbian rule that tends to minimize the variance $\mathbf{u}$, whereas for sub-Gaussians the corresponding term is a Hebbian

rule that tends to maximize the variance $\mathbf{u}$. If we choose a sigmoidal activation function $y = 1/(1+\exp(-u))$ with the additional term $\mathbf{u}$ the learning rule changes to:

$$\Delta \mathbf{W} \propto \left[\mathbf{I} - \text{sign}(k_4)(1-2\mathbf{y})\mathbf{u}^T - \mathbf{u}\mathbf{u}^T\right]\mathbf{W} \tag{12}$$

Eq.11 and eq.12 do not differ in their performance since the tanh(.) and the sigmoidal activation function are proportional to each other. However, if we assume that the source distribution is Laplacian (e.g. speech) the activation function can be modeled as: $\int_{-\infty}^{u_i} \exp(-|v|)dv$ and the nonlinearity reduces to using the sign-function.

$$\Delta \mathbf{W} \propto \left[\mathbf{I} - \text{sign}(k_4)\text{sign}(\mathbf{u})\mathbf{u}^T - \mathbf{u}\mathbf{u}^T\right]\mathbf{W} \tag{13}$$

The nonlinearity in Eq.13 can be realized in an hardware implementation by a simple 1-bit quantizer. Note that in the case of separating only natural signals (mostly super-Gaussian) the learning rule is simply

$$\Delta \mathbf{W} \propto \left[\mathbf{I} - \text{sign}(\mathbf{u})\mathbf{u}^T\right]\mathbf{W} \tag{14}$$

2.1 Speeding up convergence

A significant improvement of the convergence is given by the 'natural' gradient [1]. Amari derives the optimized learning rule from an information geometry approach where in an Riemannian manifold the Fisher information matrix provides an optimal rescaling of the simple gradient.

$$g_{ij}(\mathbf{W}) = E\{\frac{\partial l(\mathbf{x},\mathbf{W})}{\partial w_i}\frac{\partial l(\mathbf{x},\mathbf{W})}{\partial w_j}\} \tag{15}$$

where $l(\mathbf{x},\mathbf{W}) = log(p(\mathbf{x}|\mathbf{W})$ is defined as the loss function and $g_{ij}(\mathbf{W})$ is the Fisher information matrix. Applied to the source separation problem, the Fisher information matrix reduces to the identity when the independent components are orthogonal to each other. Otherwise, the extension $\mathbf{W}^T\mathbf{W}$ provides a rescaling of the gradient for the non-orthogonal metric and the natural gradient is given by:

$$\nabla_n l(\mathbf{x},\mathbf{W}) = G^{-1}\,\nabla_e\, l(\mathbf{x},\mathbf{W}) \tag{16}$$

where ∇_e is the Euclidean (normal) gradient and G^{-1} is the transformation matrix.

Cardoso calls the metric the relative gradient and uses it in the equivariant learning rule [3]. MacKay [13] derives the natural gradient rule from the ML perspective and finds a metric from the curvature of the objective function given by the second derivative of the ML-function. The resulting learning rule is covariant which means that the steepest ascent in $\mathbf{W}$ is optimal with respect to the curvature of the objective function.

We use preprocessing methods to speed up the convergence. A common statistical tool before applying the learning rule in eq.11 is to remove the 2^{nd}-order correlation by prewhitening the observation vector $\mathbf{x}$. This preprocessing method speeds up convergence. The overall separation matrix $\mathbf{W}_{all}$ consists then of a second order sphering matrix $\mathbf{W}_S$ and the unmixing matrix found by the infomax algorithm: $\mathbf{W}_{all} = \mathbf{W}\cdot\mathbf{W}_\mathbf{S}$. The whitening matrix $\mathbf{W}_S$ can be computed by $\mathbf{W}_S = (E\{\mathbf{x}\mathbf{x}^T\})^{-\frac{1}{2}}$ Furthermore, the preprocessing method can be extended to 4^{th}-order correlation cancelation by simply adding another step that cancels out 4^{th}-order correlation.

$$\mathbf{x}_S = \mathbf{W}_S\mathbf{x} \quad \text{and} \quad \mathbf{W}_4 = (E\{\|\mathbf{x}_S\|^2\mathbf{x}_S\mathbf{x}_S^T\})^{-\frac{1}{2}} \tag{17}$$

$$\mathbf{W}_{all} = \mathbf{W}\cdot\mathbf{W}_4\cdot\mathbf{W}_\mathbf{S} \tag{18}$$

In [10] we can improve the performance of infomax by repetitively forcing 2^{nd}-order and 4^{th}-order correlations to zero which speeds up the convergence and improves separating sources with fewer data points.

During the learning process, we observe that a momentum helps to stabilize the convergence of the algorithm.

$$\Delta\mathbf{W}(n+1) = (1-\alpha)\Delta\mathbf{W}(n) + \alpha\mathbf{W}(n) \tag{19}$$

where α takes into account the history of $\mathbf{W}$ and increases with increasing number of iterations.

3 Simulation and Experimental Results

An obvious question given the different learning rules in eq.11, eq.12, eq.13 is to what extent the nonlinear activation function $g(\mathbf{u})$ has to approximate the cdf of s to separate the sources in practice. To answer the question, we perform two simulation experiments.

1. Separation of 10 sound sources obtained from Pearlmutter (for comparison purpose to cICA in [15]).
2. Separation of the following 20 sources: 10 sound tracks obtained from Pearlmutter, 6 speech/sound signals used in Bell & Sejnowski, 3 uniformly distributed sub-Gaussian noise signals and one noise source with a Gaussian distribution.

For the first simulation experiment, we use eq.13, eq.19 and the preprocessing steps in eq.18. The 10 mixed sources can be separated in one or two pass through the data. For 55000 data points and a blocksize of 10, one pass is equivalent to 275 iterations. The learning rate was fixed at 0.0001. As a measure of performace, we use the measure proposed by Amari et al. in [1] which can be related to the SNR measure [10].

$$E = \sum_{i=1}^{N}\left(\sum_{j=1}^{N}\frac{|p_{ij}|}{\max_k |p_{ik}|}-1\right)+\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\frac{|p_{ij}|}{\max_k |p_{kj}|}-1\right) \tag{20}$$

where $P = \mathbf{W}\cdot\mathbf{A}$ and P is close to the identity matrix (except of permutation and scaling) when the sources are separated. Figure 1 shows the performance index during the learning process (two iterations) for eq.13 and eq.12. Both learning rules converge to the correct solution. However, eq.13 converges faster than eq.12 although the pdf of the sound sources are closer to the derivative of the sigmoidal activation function than the Laplacian prior.

$$\begin{bmatrix}
-0.0869 & -0.3835 & 0.1439 & -0.0987 & -0.0555 & 0.9308 & -0.3567 & -0.5379 & 0.1690 & 14.7862 \\
-11.1760 & -0.0126 & 0.1423 & 0.0500 & -0.0799 & 0.0177 & 0.0715 & 0.2052 & -0.1206 & -0.6785 \\
0.1503 & 0.0780 & -0.0792 & -0.0227 & 10.1875 & -0.0204 & 0.1454 & 0.0556 & 0.0730 & 0.1695 \\
0.3949 & 0.6057 & -0.6988 & -0.0672 & 0.1378 & 0.3241 & -0.0833 & 0.8545 & 7.6403 & -0.1570 \\
0.0436 & 0.7586 & 14.8921 & 0.0325 & 0.0260 & -0.1665 & 0.1828 & -0.3120 & -0.1940 & 0.0386 \\
0.1077 & 12.8930 & -0.5396 & -0.2330 & -0.4262 & -0.2100 & -0.1163 & 0.0474 & 0.0786 & 0.1806 \\
0.4542 & 0.1663 & -0.0242 & 6.5322 & 0.2396 & 0.9798 & -0.3873 & -0.929 & 0.0643 & -0.0857 \\
0.3137 & 0.1424 & 0.2285 & 0.0305 & -0.1384 & -17.2518 & -0.3871 & -0.2518 & 0.1943 & 0.3918 \\
-0.5391 & -0.8079 & 0.6236 & 0.8411 & -0.1786 & 0.4674 & -0.0374 & 10.4839 & -0.9210 & 0.1258 \\
-0.0778 & -0.2582 & 0.1458 & -0.1020 & 0.4879 & 0.0091 & -10.2541 & 0.5898 & 0.3325 & -0.9416
\end{bmatrix}$$

The above matrix shows the performance matrix P after 550 iterations and is close to the identity matrix after rescaling and reordering. Therefore, compared to cICA the original infomax algorithm shows the same performance without having to learn the nonlinear transfer function.

For the second experiment, we performed the same preprocessing steps and used eq.11. The blocksize was 100 and 50 passes through the data (1375 iterations) were neccessary for convergence. Figure 2 shows the performance after the rows were manually reordered and normalized to unity. A listening test shows a clear separation of all sources from their mixture. In this case, when we used eq.13, the noise source with Gaussian distribution cannot be separated completely from the mixtures. Hence, the Laplacian prior is not suitable to separate a source with an approximately Gaussian distribution. In other words, the sigmoidal activation function seems to provide a good compromise between the Gaussian and the spiky Laplacian distribution that provides an approximation for a wide range of source distributions.

3.1 Experimental Results on EEG and fMRI Data

We have applied the learning rules in eq.11 - eq.13 and the preprocessing steps to analyze EEG recordings and fMRI data.

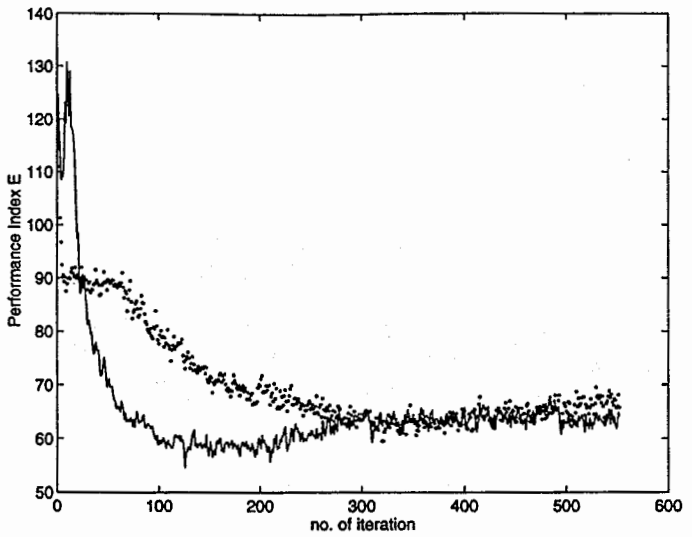

Figure 1: Performance index for the separation of 10 sound sources. The doted line is the performance for eq.12 and the continuous line is the performance for eq.13

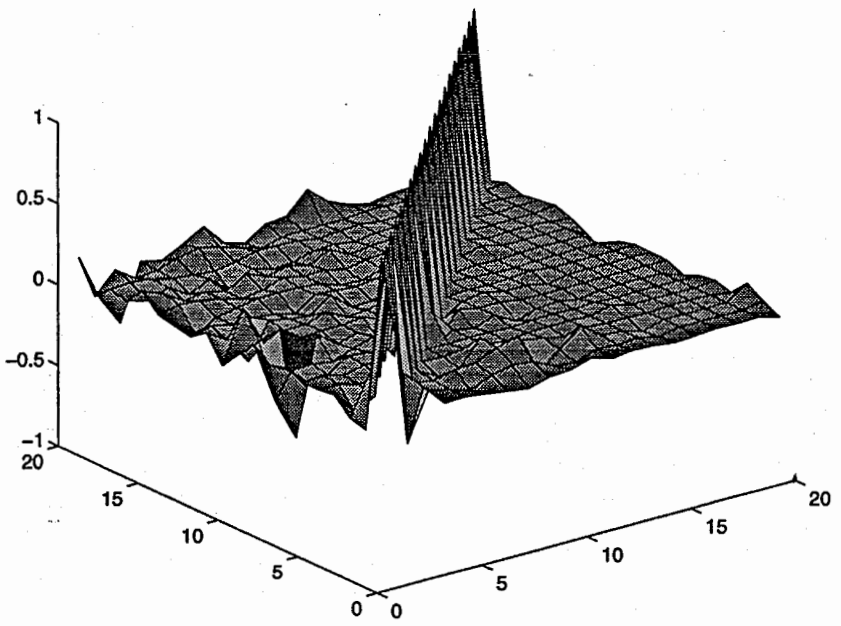

Figure 2: Results from the separation of 20 sources. This figure shows the separation performance matrix P after normalizing and reordering. For perfect separation P is an identity matrix.

In Jung et al. [8], we showed how the proposed algorithm can be used to remove artifacts from EEG recordings. The recorded EEG signals are often contaminated with artifacts which have to be filtered out. By using the extended ICA algorithm eq.12, we can separate eye movement artifacts, periodic muscle spiking, line noise and cardiac contamination (EKG noise). The sources have sub and super Gaussian distributions. After eliminating these five artifactual components, the 'corrected' EEG data are free of these artifacts.

In McKeown et al. [14], we show how the algorithm can be used to find time courses in voxels of fMRI data which correspond to the time course of the experiments.

4 Conclusions

The extended ICA algorithm presented here is a promising generalization of ICA for mixed sub-Gaussian and super-Gaussian sources. The algorithm is robust and efficient and has been used successfully on several large data sets derived from electrical and blood flow measurements of functional activity in the brain.

Acknowledgments

T.W. Lee is supported by the German Academic Exchange Program. We are grateful to O. Coenen, T-P. Jung and M. McKeown for discussions and comments.

References

[1] S. Amari, A. Cichocki, and H. Yang. A New Learning Algorithm for Blind Signal Separation. In *Advances in Neural Information Processing Systems 8*, 1996.

[2] A. Bell and T. Sejnowski. An Information Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7:1129–1159, July 1995.

[3] J-F. Cardoso and B. Laheld. Equivariant adaptive source separation. *IEEE Trans. on Signal Processing*, 45,2:434-444, Dec. 1996.

[4] J-F. Cardoso. Infomax and maximum likelihood for blind source separation. to appear in *IEEE Signal Processing Letters.*

[5] A. Cichocki, R. Unbehauen, and E. Rummert. Robust learning algorithm for blind separation of signals. *Electronics Letters*, 30, 17, 1386-1387, 1994

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

[7] M. Girolami and C. Fyfe. Extraction of Independent Signal Sources using a Deflationary Exploratory Projection Pursuit Network with Lateral Inhibition. submitted *I.E.E Proceedings on Vision, Image and Signal Processing Journal.*

[8] T.P Jung, C. Humphries, T.W. Lee, S. Makeig, M. McKeown, V. Iragui, T. Sejnowski. Extended ICA Removes Artifacts from Electroencephalographic Recordings. submitted to *Advances in Neural Information Processing Systems*, May 1997.

[9] 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, vol8:487–504, May. 1997.

[10] B.U. Koehler, T. Lee and R. Orglmeister. Improving the performance of infomax using statistical signal processing. submitted to *ICANN 1997*, Lausanne.

[11] R. Lambert. Multichannel blind deconvolution: Fir matrix algebra and separation of multipath mixtures. Thesis, University of Southern California, Department of Electrical Engineering, May 1996.

[12] T-W. Lee, A.J. Bell and R. Orglmeister. Blind source separation of real-world signals. *Proc. ICNN*, Houston, USA, 1997.

[13] D. MacKay. Maximum likelihood and covariant algorithms for independent component analysis. ***submitted to a journal***, Dec. 1996.

[14] M. McKeown, T.P. Jung, S. Makeig, G. Brown, S. Kindermann, T.W. Lee, T. Sejnowski. Transiently Time-locked fMRI Activations revealed by independent component analysis submitted to *Proceedings of the National Academy of Sciences*, May 1997

[15] B. Pearlmutter and L. Parra. A context-sensitive generalization of ICA. In ICONIP'96 . In press.

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