Kurtosis Extrema and Identification of Independent Components : A Neural Network Approach

Mark Girolami and Colin Fyfe

Department of Computing and Information Systems, University of Paisley, High Street, Paisley, Scotland, PA1 2BE
Telephone (+44) 141 848 3301, Fax (+44)141 848 3542
giro_ci0@paisley.ac.uk fyfe_ci0@paisley.ac.uk

ABSTRACT

We propose a nonlinear self-organising network which solely employs computationally simple hebbian and anti-hebbian learning in approximating a linear independent component analysis (ICA). Current neural architectures and algorithms which perform parallel ICA are either restricted to positively kurtotic data distributions [1] or data which exhibits one sign of kurtosis [2, 3, 12]. We show that the proposed network is capable of separating mixtures of speech, noise and signals with both platykurtic (positive kurtosis) and leptokurtic (negative kurtosis) distributions in a blind manner. A simulation is reported which successfully separates a mixture of twenty sources of music, speech, noise and fundamental frequencies.

1. INTRODUCTION

The blind signal separation problem has received considerable attention from both the signal processing and neural network communities in recent years. The motivation behind this activity is to solve the practical problem of identifying and separating individual original source signals from a received mixture within a multi-source and multi-sensor environment. The problem is further compounded as no *a priori* knowledge of the direct and cross coupling transfer paths can be assumed.

Algorithms have been developed based on information theoretic criteria and higher order statistics; if a signal has independent components then the product of the marginal probability densities is equal to the signal probability density. Using the Kullback-Leibler divergence as a measure of independence, Comon [4] develops a series of contrast functions based on an Edgeworth expansion of the marginal densities and batch methods are used in their maximisation.

The First Author is Supported by a Grant from NCR(Ltd) Technology Development Division.

Jutten and Herrault [5] were the first to develop a neural architecture and learning algorithm for blind separation; since then a number of variants on this architecture have appeared in the literature, Amari *et al* [6]. Bell and Sejnowski [1] developed a feedforward network and learning rule which minimises the mutual information at the output nodes; this yields excellent results for platykurtic signals such as speech, however, the algorithm is developed from a noise free model. The matrix inversion required is a computational bottleneck and unrealistic from a DSP implementation viewpoint. The simple multiply and accumulate operations of hebbian and anti-hebbian learning are attractive for DSP hardware implementations. Karhunen *et al* [2, 3] develop a number of nonlinear variants of principal component analysis (PCA) learning and demonstrate ICA performed on data such as natural images. The recently developed Bigradient algorithm [9] can deal with both leptokurtic and platykurtic data, the restriction of source signals which have kurtosis of the same sign is still required. Hyvarinen and Oja [7] have recently developed single neuron models for source extraction. Karhunen and Pajunen [8] have reported on the use of hierarchical [14] and transitional learning (from nonlinear to robust PCA learning) in separating mixtures of sources of varying kurtosis sign.

2. NETWORK ARCHITECTURE AND LEARNING

We consider the same signal model as in Adaptive Noise Cancellation (ANC) where the noise source is considered as a signal component. This assumes an *a priori* knowledge of the number of noise sources; for this work we shall adopt this particular assumption. We shall only consider instantaneous mixing here, however, we have reported on separation of convolved mixtures of signals in [10] for strictly causal minimum phase systems.

Figure 1 shows the network topology, the network is based around an exploratory projection pursuit network, Fyfe and Baddeley [11] utilise a nonlinear network with negative

feedback of activation to perform exploratory projection pursuit (EPP). Girolami and Fyfe [12], explored the use of an EPP network for separation of speech sources it was found that the choice of nonlinearity was crucial for the quality of separation, and that utilising symmetric local learning did not guarantee uniqueness of the outputs.

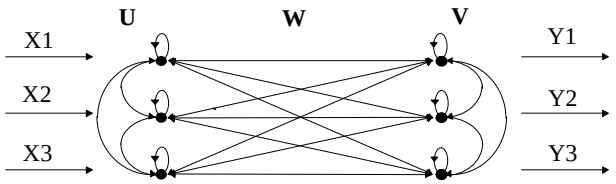


Figure 1 : The EEPP Network with lateral connections.

Consider *M* zero mean source signals **s** mixed by the unknown linear matrix **A** the received signals are then, in matrix format, **x=As**. The output of the first layer of neurons is given as **z,** and so with the linear lateral connections at the input

$$\mathbf{z} = \left[\mathbf{I} + \mathbf{U}\right]\mathbf{x} \equiv \mathbf{U}_{\mathbf{I}}\mathbf{x} \tag{1}$$

The first layer of neurons are used to decorrelate the incoming data, and so give an identity covariance matrix for the output of the layer. This decorrelation is sometimes referred to as sphereing or whitening [2, 3, 11] and is an essential pre-requisite for the learning of the output neurons. The learning rule (2) will provide whitened output data

$$\Delta\mathbf{U} = \alpha\left(\mathbf{I} - \mathbf{z}\mathbf{z}^{\mathrm{T}}\right) \tag{2}$$

If $\langle\Delta\mathbf{U}\rangle = \alpha\left(\mathbf{I} - \mathbf{C}_{\mathbf{zz}}\right) \to 0 \Rightarrow \mathbf{C}_{\mathbf{zz}} \to \mathbf{I}$ where $\mathbf{C}_{\mathbf{zz}}$ is the covariance matrix of **z**. The **z** values are fed forward through the **W** weights to the output neurons where there is a second layer of lateral weights. However before the activation is passed through this layer it is passed back to the originating **z** values in a hierarchical manner [8,14] as inhibition and then a nonlinear function of the inputs is calculated. The linear weighted sum value at the output neuron is given as $\mathbf{act} = \mathbf{W}^{T}\mathbf{U}_{\mathbf{I}}\mathbf{A}\mathbf{s}$ (3) and the associated residuals are given, in vector format, as

$$\mathbf{r}_i = \mathbf{z} - \sum_{j=1}^{i}\mathbf{act}_j\mathbf{w}_j \quad i = 1 \ldots M \tag{4}$$

The output neurons are nonlinear and a nonlinear functional is applied to the weighted sum. The outputs also have lateral connections to the nonlinear output neurons and so the output is defined in matrix format as

$$\mathbf{y} = \mathbf{V}_{\mathbf{I}}\,\mathrm{f}\left(\mathbf{act}\right) \tag{5}$$

Simple hebbian learning is used to update the feedforward weight values we can write this in vector notation as

$$\Delta\mathbf{w}_i = \beta\mathbf{r}_i\, y_i \tag{6}$$

and so $\Delta\mathbf{W} = \beta\left(\mathbf{z}\mathbf{y}^{\mathrm{T}} - \mathbf{W}\times\ \mathrm{upper}\left[\mathbf{act}\times\mathbf{y}^{T}\right]\right)$ (7)

Where the upper[.] operator sets the matrix argument upper triangular.

Similarly, anti-hebbian learning is applied at the output as at the input $\Delta\mathbf{V} = \gamma\left(\mathbf{I} - \mathbf{y}\mathbf{y}^{T}\right)$ (8)

It has been shown [11, 14] that (7) is an approximative stochastic algorithm which will maximise the integral of the neuron activation function that is $\int \mathrm{f}\left(\mathbf{act}\right)$, under orthonormal constraints. With the addition of (8) we have a further constraint to the feedforward learning that is

$$\text{If } \langle\Delta\mathbf{V}\rangle = \alpha\left(\mathbf{I} - \mathbf{C}_{\mathbf{yy}}\right) \to 0 \Rightarrow \mathbf{C}_{\mathbf{yy}} \to \mathbf{I}$$

As the neuron output is nonlinear, we will have higher order moments generated and as such (8) will provide a higher order decorrelation constraint. Consider the simple case of $\int \mathbf{V}_{\mathbf{I}}\,\mathrm{f}\left(\mathbf{act}\right) = \mathbf{act}^{4} - 3$ which in the expectation will be

$\left\langle\mathbf{act}^{4}\right\rangle - 3 = \kappa_4$ is the fourth order sample cumulant for zero mean and unit variance data. Also $\langle\Delta\mathbf{V}\rangle \to 0 \Rightarrow \mathbf{C}_{\mathbf{yy}} \to \mathbf{I} \Rightarrow \left\langle act_i^{3} act_j\right\rangle + \left\langle act_i\, act_j^{3}\right\rangle \to 0$ so we are effectively maximising the sum of fourth order cumulants under the constraints that the rotation is orthogonal and the sum of the fourth order cross cumulants is minimised. If all the sources have strictly positive or strictly negative kurtosis, which is the normalised fourth order cumulant, we then have an equivalence for

$$\sum\kappa_{iiii}^{2} = \left(\kappa_{1111}^{2} + \kappa_{2222}^{2}\right) + \left(\kappa_{1222}^{2} + \kappa_{1112}^{2}\right)$$

$$\Rightarrow \sum\kappa_{iiii}^{2} = \Phi + \left(\kappa_{1222}^{2} + \kappa_{1112}^{2}\right) \text{ where } \Phi = \sum_{i=1}^{N}\left\|\boldsymbol{\kappa}_4^{(i)}\right\|^{2}$$

This is the contrast for ICA developed by Comon; as Comon uses the square of the cumulant terms the sign of the individual source kurtosis is unimportant. It is noted that the sum of squares of fourth order cumulants is invariant under linear orthogonal rotation and so $\underset{\mathbf{W}^{\mathrm{T}}\mathbf{W}=\mathbf{I}}{\mathrm{Max}}\,\Phi = \sum_{i=1}^{N}\left\|\boldsymbol{\kappa}_4^{(i)}\right\|^{2}$ will by implication minimise the sum of squares of cross cumulants. This is maximised if the kurtosis of each individual output is extremised. So if we consider pairwise outputs the joint density of both distributions is then approximately factorable, $p_{\mathbf{s}}\left(\mathbf{u}_1, \mathbf{u}_2\right) = p_{s_1}\left(\mathbf{u}_1\right)p_{s_2}\left(\mathbf{u}_2\right)$. To ensure that kurtosis extrema occurs at each output we have proposed an adaptive nonlinearity which will respond to the kurtosis of the neuron input, this is given as

$$f\left(act_i\right) = \beta\left(act_i\right) - sign\left(\kappa_s^{(4)}\right)\tanh\left(act_i\right) \tag{9}$$

A detailed analysis of this form of nonlinearity is given in [13] consider a positively kurtotic, zero mean and unit variance neuron input *u*, then , using a truncated Taylor expansion $E\left\{f\left(u\right)\right\} = E\left\{u - \tanh\left(u\right)\right\}$

$= E\left\{u\right\} - E\left\{u - u^{3}/3 + 2u^{5}/15\right\} \cong E\left\{u^{3}/3\right\}$ which with (7) and (8) will maximise the individual kurtosis as

$E\{u^2\} \cong 1$. The same argument is applied to negatively kurtotic data so $E\{f(u)\} \cong -E\{u^3/3\}$ and the orthogonal rotation will minimise the data kurtosis. In keeping with the 'blind' form of the proposed separation we adopt the online kurtosis estimation, proposed by Hyvarinen and Oja [7] that is

$$\hat{m}_p[u_i(t+1)] = [1-\eta(t)]\hat{m}_p[u_i(t)] + \eta(t)u_i^p(t) \quad (10)$$

$$\hat{k}_4[u_i(t+1)] = \frac{\hat{m}_4[u_i(t)]}{\hat{m}_2^2[u_i(t)]} - 3 \quad (11)$$

(10) estimates online the p^{th} order moments of the data, $\eta(t)$ is a small learning constant. (11) is an estimate of the kurtosis for zero mean data. We can then use the adaptive nonlinearity

$$f[u_i(t)] = u_i(t) - \frac{\hat{k}_4[u_i(t)]}{\left\|\hat{k}_4[u_i(t)]\right\|}\tanh[u_i(t)] \quad (12)$$

The linear hebbian learning at the input diagonalises the input covariance matrix. The nonlinear hebbian and anti hebbian learning of the forward and output section rotate the output to extremise the individual output kurtosis, and minimise the output cross moments thus performing an approximation to linear independent component analysis. The lateral connections and asymmetric feedback, along with the adaptive nonlinearity (12) removes the restriction of uniformity of kurtosis and increases the scalability of the network. It should be noted that the single neuron models developed by Hyvarinen and Oja [7] allow sequential extraction of sources with arbitrary non-gaussian pdf.

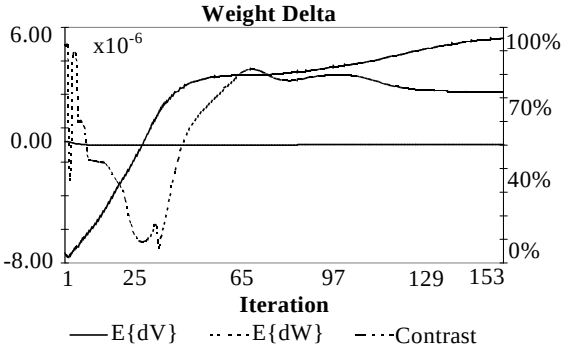
3. SIMULATIONS

Twenty sources were mixed using a 20 x 20 mixing matrix whose elements were randomly drawn from a uniform distribution in the range -1 ... +1. The twenty sources[1] included five second samples of music, speech, noise and fundamental tones. Figure 2.b shows the initial sources and their kurtosis values, the kurtosis values of the mixtures is given in the adjacent column. We note that these mixtures now have almost gaussian histograms, and small kurtosis. The final column shows the extracted sources and the output neurons (identified as a - t) at which they appeared. The value of the extracted signals kurtosis is listed, along with a numeric value of the signal to noise ratio (SNR). This SNR is calculated as follows

$$\text{SNR} = -10\log\left(\frac{\text{MSE}_{ij}}{\text{Var}_j}\right)$$

where MSE_{ij} is the mean square error of the normalised output *i* and the corresponding source *j*, with variance Var_j.

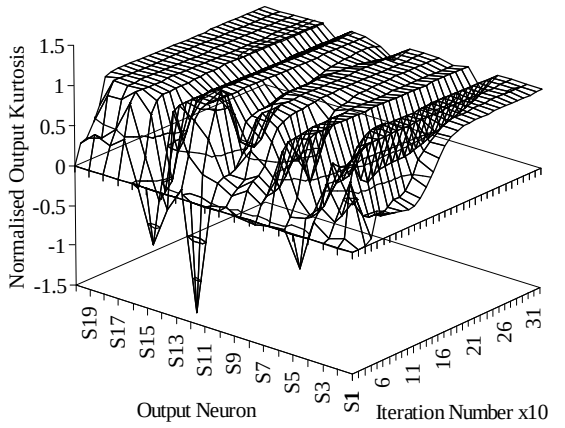
[1]The authors are grateful to Prof. B.Pearlmutter for making his source data available at http://phlegm.ucsd.edu/~bap/demos.html



(a)

Source	K_4	Mixture	K_4	Output	K_4	SNR (dB)
Music_01	2.48	1	-0.16	n	2.38	36.08
Music_02	0.96	2	-0.38	p	0.70	16.48
Music_03	2.53	3	-0.48	t	2.48	46.95
Music_04	1.47	4	-0.31	j	1.17	21.32
Music_05	2.74	5	-0.59	e	2.75	45.19
Music_06	1.16	6	-0.20	k	1.15	45.57
Music_07	0.73	7	-0.42	s	0.72	47.95
Noise_08	-1.17	8	-0.29	c	-1.16	42.12
Music_09	1.06	9	-0.30	m	1.06	39.54
Music_10	2.46	10	-0.32	r	2.45	59.99
Music_11	4.04	11	-0.25	o	4.02	52.94
Music_12	0.50	12	-0.47	f	0.47	30.04
Music_13	1.00	13	-0.31	h	1.00	50.03
Sine__ 14	-1.49	14	-0.56	b	-1.48	43.33
Sine__ 15	-1.49	15	-0.53	d	-1.48	40.82
Music_16	1.71	16	-0.27	g	1.71	58.91
Music_17	0.94	17	-0.44	l	0.94	46.23
Noise_18	-1.19	18	-0.19	a	-1.18	37.65
Music_19	0.45	19	-0.52	q	0.45	25.27
Music_20	0.56	20	-0.21	I	0.46	18.51

(b)



(c)

Figure 2 : (a) Weight and Contrast development, (b) Table of values, (c) Kurtosis Surface Development.

Figure 2(a) shows the contrast development $\Phi = \sum_{i=1}^{N} \left\| \kappa_4^{(i)} \right\|^2$ along with the feedforward (**W**) and output lateral weights (**V**), the contrast develops with the evolution of the feedforward weights. However it is clear that the lateral weights do not significantly change once the data has been sphered, due to the orthonormality of the **W** weights. A final contrast of 98% original is attained after 150 learning epochs, the final SNR's are very high, with only two sitting below 20dB. Figure 2(c) shows the normalised kurtosis development as a surface, this is plotted against each output and learning epoch, for perfect kurtosis extrema and hence separation we seek a horizontal plane, this is almost achieved after 150 epochs. It is also interesting to note the phased attainment of maximal kurtosis, due to the hierarchic feedback of activation. Figure 3, shows the original trace of MUSIC_10, mixture 10 and output r normalised, with almost perfect separation indicated by the 59.9dB SNR.

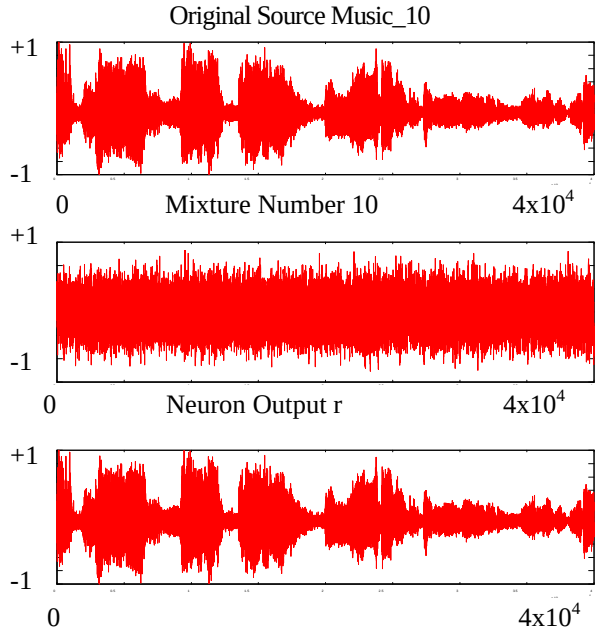


Figure 3: Original, Mixture and Extracted Source Trace.

4. CONCLUSIONS

We have developed a self-organising neural network which is capable of extracting a known number of independent sources from a mixture with no *a priori* knowledge of the mixing or the sources. Novel use of a nonlinearity which responds to the developing kurtosis of the output neurons allows mixtures of both sub and super gaussian sources to be separated. The use of lateral connections at the output and hierarchic feedback provides a more robust and scalable extraction when considering large mixtures of sources.

5. REFERENCES

[1] Bell, A and Sejnowski, T. An Information Maximisation Approach to Blind Separation and Blind Deconvolution. *Neural Computation* 7, 1129 - 1159, 1995.

[2] Karhunen, J., Wang, L. and Joutsensalo, J. Neural Estimation of Basis Vectors in Independent Component Analysis. *International Conference on Artificial Neural Networks*, Vol 1 317 - 322. 1995.

[3] Wang, L. Karhunen, J. Oja, E. and Vigario, R. Blind Separation of Sources using Nonlinear PCA Type Learning Algorithms. *International Conference on Neural Networks and Signal Processing, Nanjing*, China, December 1995.

[4] Comon, P. Independent Component Analysis, A New Concept ?. *Signal Processing*, 36, 287 - 314. 1994.

[5] Jutten, C. Herault, J. Blind Separation of Sources, Part 1: An Adaptive Algorithm Based On Neuromimetic Architecture in Signal Processing 24. 1991.

[6] Amari, S., Cichocki, A., Yang, H. Recurrent Neural Networks for Blind Separation of Sources. *International Symposium on Nonlinear Theory and Applications Vol 1*. 37 - 42. 1995.

[7] Hyvarinen, A and Oja, E. Simple neuron models for independent component analysis. technical report, Helsinki university of technology, laboratory of computer and information science, 1996.

[8] Karhunen, J and Pajunen, P. Hierarchic nonlinear PCA algorithms for neural blind source separation. Norsig-96, *I.E.E.E. Nordic signal processing symposium*, Espoo, Finland, September 24 - 27, 1996.

[9] Wang, L. Karhunen, J. Oja, E. A Bigradient Optimisation Approach for Robust PCA, MCA, and Source Separation. *ICNN'95*, Perth, Australia. December 1995.

[10] Girolami, M and Fyfe, C. A Temporal Model of Linear Anti-Hebbian Learning. *Neural Processing Letters* In Press, Vol 4, Issue 3, Jan 1997.

[11] Fyfe, C and Baddeley, R. Non-Linear Data Structure Extraction Using Simple Hebbian Networks. *Biological Cybernetics*, 72(6):533-541, 1995.

[12] Girolami, M and Fyfe, C. Blind Separation Of Sources Using Exploratory Projection Pursuit Networks. *International Conference on the Engineering Applications of Neural Networks*, (Ed A Bulsari) ISBN 952-90-7517-0, 249 - 252, 1996.

[13] Girolami, M., and Fyfe, C. Stochastic ICA contrast maximisation using Oja's nonlinear PCA algorithm. Accepted for Publication, *International Journal of Neural Systems*.

[14] Karhunen, J. and Joutsensalo, J. Generalisations of principal component analysis, optimisation problems, and neural networks. *Neural Networks*, vol. 8, no 4, pp 549 - 562, 1995.