

Multivariate Density Factorisation for Independent Component Analysis : An Unsupervised Artificial 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

giro0ci@paisley.ac.uk fyfe0ci@paisley.ac.uk

Abstract

We propose a novel homogenous nonlinear self-organising network which employs solely computationally simple hebbian and anti-hebbian learning, in approximating a linear independent component analysis (ICA). The learning algorithms diagonalise the transformed data covariance matrix and approximate an orthogonal rotation which maximises the sum of fourth order cumulants. This provides factorisation of the input multivariate density into the individual independent latent marginal densities. We apply this network to linear mixtures of data, which are inherently non-gaussian and have both Laplacian and bi-modal probability densities. We show that the proposed network is capable of factorising multivariate densities which are linear mixtures of independent latent playkurtic, leptokurtic and uniform distributions.

1. INTRODUCTION

Various forms of linear invertible transformations of data sets have been proposed for data analysis, one such transformation is the Karhunen-Loeve transform [1] sometimes referred to in the statistics and neural network literature as a Principal Component Analysis (PCA). As a multivariate data analysis tool PCA will identify projections with maximal variance which in some cases can be useful for cluster analysis [2]. PCA also allows optimal, in the mean square sense, dimensionality reduction and as such is a useful signal processing tool [2]. There are a number of self organising neural network architectures and learning algorithms which perform a PCA on a data set [3]. A multivariate data set which has been transformed via PCA will have marginal distributions which are uncorrelated ie independent to second order, it does not follow that they will be fully independent. Analysis of densities which are fully described by moments higher than second order will be limited by PCA, as only second order statistics are employed in the transformation. Gaussian densities are fully described by the first and second moments, that is the mean and variance. Non-Gaussian densities are described by higher order moments such as the third and fourth, that is skew and kurtosis. Inspection of the form of a Gram-Charlier or Edgeworth expansion of a non-normal distribution shows the dependence of the density shape on the cumulants of the data [4]. If we wish to transform a multivariate density such that the marginal distributions are independent then moments of order higher than second are required.

Independent component analysis (ICA) has been proposed as a transform which will factorise a multivariate density. It can be considered as a higher order PCA, as factorable marginal densities are independent to second order (PCA) and all higher orders. A simple illustrative example is given, two sets of five thousand data points are independently drawn from a uniform distribution. Taking these data points and plotting them as co-ordinate points on a plane we can see that they are indeed independent. No information can be inferred about the position of a point on the plane from information given by the value of any one of the co-ordinates. The joint density of both of these distributions is then factorable, and so we have for this simple two dimensional case

$$p_s(s_1, s_2) = p_{s_1}(s_1)p_{s_2}(s_2) \quad (1)$$

If we now transform these independent variables using an arbitrary non-diagonal rotation A such that $\mathbf{x} = A\mathbf{s}$ the components of \mathbf{s} will no longer be independent and

$$P_{\mathbf{s}}(s_1, s_2) \neq P_{s_1}(s_1)P_{s_2}(s_2)$$

Performing a PCA on the data vector \mathbf{x} , while providing independence up to second order (decorrelation), will leave the additional higher order moments which describe the distributions unaffected. Figure 1(a) shows the original data \mathbf{s} , the independence of the two components is apparent. After transformation using an arbitrary rotation

$$\mathbf{x} = A\mathbf{s} \quad \text{where for example} \quad A = \begin{bmatrix} 0.3497 & 0.2149 \\ 0.3424 & 0.6207 \end{bmatrix}$$

we can see from Figure 1(b) the lack of independence of the transformed components, the position of a point can now be inferred from knowledge given by the value of any one of the co-ordinates. We can consider this as a rise in the mutual information between the components. Performing a standard PCA on the vector \mathbf{x} such that $\mathbf{u} = W_{pca}\mathbf{x}$, where the columns of the transforming matrix will be the data covariance eigenvectors, will yield a decorrelated output. However, Figure 1(c) shows that there is still a good deal of information which can be inferred about the position of a point from knowledge of only one co-ordinate. Performing an ICA on the vector \mathbf{x} such that $\mathbf{u} = W_{ica}\mathbf{x}$ Figure 1(d) clearly shows the independence of the transformed data, and as such the factorisation of the joint density. The use of second order statistical measures, such as correlation, do not have the discriminating power to identify true, or higher order independence of a data set.

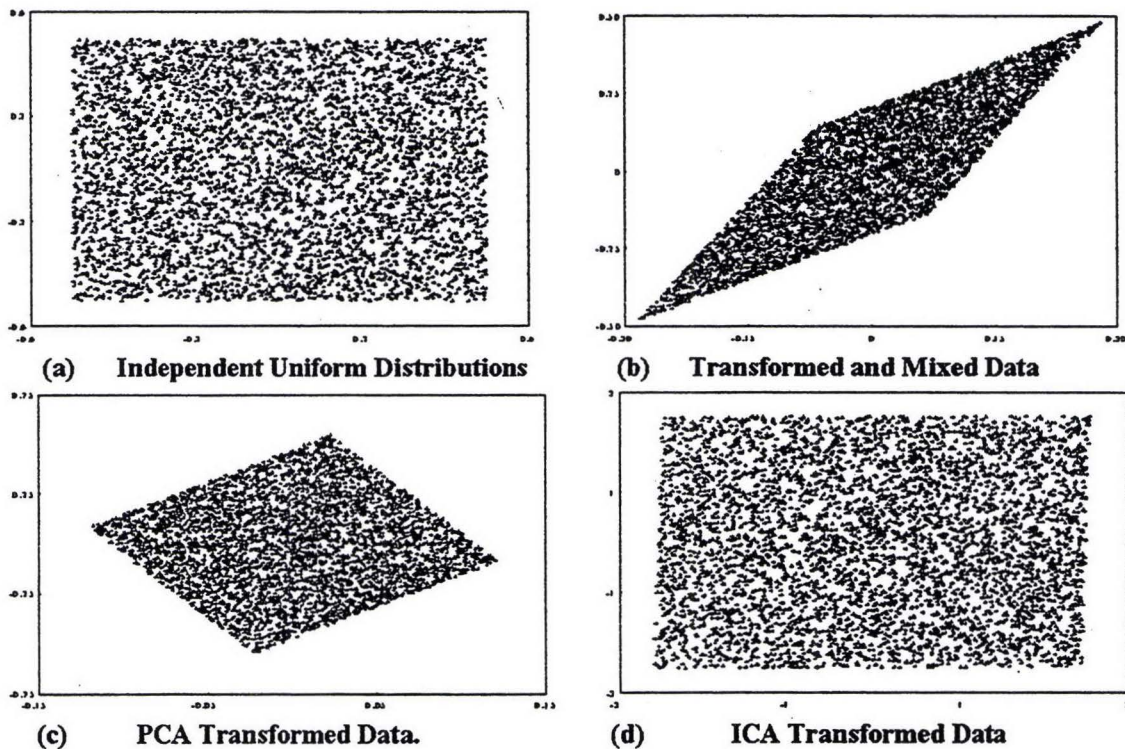


Figure 1: Examples of PCA and ICA Transformations.

If we consider ICA in the general case, then if a multivariate distribution can be modelled as a mixture of independent underlying latent variables, then performing an ICA transformation will factorise the density into the latent distributions.

2. INDEPENDENT COMPONENT ANALYSIS

ICA has become the subject of a good deal of research interest in recent years, prior to the formalism of the ICA transform by Comon [5] the signal processing community had investigated the problem of Blind Source Separation (BSS). BSS can be considered as the problem of extracting or identifying the original signals from a received mixture where both the mixing and the original signals are unknown [5]. ICA and BSS are essentially the same problem viewed from differing perspectives. Blind separation of sources is an underdetermined problem and as such traditional adaptive techniques are unsuitable as the source signal statistics, as well as the mixing and transfer channels, are unknown. Techniques have been developed based on information theoretic criteria and higher order statistics (HOS); 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 [5] develops a series of contrast functions based on an Edgeworth expansion of the marginal densities and batch methods are used in their maximisation. Cardoso [6] utilises the invariant properties of cumulants under orthogonal transformations; he develops series updating algorithms based on the maximisation of the sum of the square of fourth order cumulants.

Jutten and Herrault [7] 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, Cichocki *et al* [8]. Bell and Sejnowski [9] 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 matrix inversion required is a computational bottleneck and unrealistic from a hardware implementation viewpoint. Recently, Amari *et al* [10] have used the natural gradient descent algorithm which removes the matrix inversion requirement in Bell & Sejnowski's algorithm

Let \mathbf{x} be a variable in \mathfrak{R}^N with a probability density function (pdf) $p_x(\mathbf{u})$. If the vector \mathbf{x} has mutually independent components, and as such a factorable joint density, we can then write

$$\delta \left(p_x(\mathbf{u}), \prod_{i=1}^N p_{xi}(u_i) \right) = 0$$

The Kullback-Leibler divergence gives a measure of the mutual information between the components of \mathbf{x} .

$$I(p_x) = \int p_x(\mathbf{u}) \log \frac{p_x(\mathbf{u})}{\prod_{i=1}^N p_{xi}(u_i)} d\mathbf{u} \quad (2)$$

Approximating the marginal densities using an Edgeworth expansion (up to cumulant order 4) yields a measure of the mutual information or contrast between the components [5].

$$I(p_z) \cong J(p_y) - \frac{1}{48} \sum_{i=1}^N \left\{ 4K_{iii}^2 + K_{iiii}^2 + 7K_{iii}^2 - 6K_{iii}^2 K_{iii} \right\} \quad (3)$$

The term on the left hand side is the mutual information of the components of the vector \mathbf{z} , where $\mathbf{y} = \mathbf{Mz}$, \mathbf{M} being an orthogonal rotation. The first term on the right hand side is the negentropy of \mathbf{y} , and is fully defined in [5]. The second term is the sum of squares of third and fourth order cumulants of \mathbf{y} , it is clear

that maximisation of this term will minimise the mutual information between the vector components and as such can be used as a contrast. Further simplifying assumptions, [5], based on the pdf symmetry and the multilinearity of cumulants reduces the contrast to the sum of squares of fourth order marginal cumulants.

$$\Phi_{Max} = \sum_{i=1}^N \left\| K_4^{(i)} \right\|^2 \quad (4)$$

It is noted that the sum of squares of fourth order cumulants is invariant under linear orthogonal rotation and so for a whitened two dimensional vector we can write

$$\sum_{i=1}^{N=2} K_{iiii}^2 = (K_{1111}^2 + K_{2222}^2) + (K_{1222}^2 + K_{1112}^2) \Rightarrow \sum_{i=1}^{N=2} K_{iiii}^2 = \Phi + (K_{1222}^2 + K_{1112}^2)$$

By then maximising (4), under orthogonal constraints, we can see that this will minimise the cross cumulant terms and so yield an approximation to independent components.

3. NEURAL ICA BASED ON NONLINEAR PCA

Karhunen and Joutsensalo [11] develop a number of nonlinear variants of neural principal component analysis (PCA) learning and show their utility in sinusoidal frequency estimation. Karhunen [12] shows the separating properties of both the nonlinear and robust pca algorithms on negatively kurtotic simple periodic signals. An important point must be made here: the input mixture is pre-whitened to remove all second order statistics. This technique was originally exploited by Fyfe and Baddeley [13] for neural exploratory projection pursuit (EPP); effective removal of second order statistics allows the network learning algorithm (utilising an odd logistic activation function) to pick up higher order statistics.

The nonlinear pca algorithms and structures have been shown to exhibit separating properties, with the full nonlinear algorithm outperforming the robust algorithm [11]. Oja [14] has given a mathematically rigorous analysis to show that a separating matrix is the asymptotically stable stationary point of the averaged differential counterpart of the nonlinear PCA difference equation. The nonlinear algorithm is an approximative stochastic gradient algorithm for minimising the mean square representation error, Girolami and Fyfe [15] show that the algorithm can be considered as minimising an information theoretic contrast function enumerating the mutual information of the outputs. Wang et al [16] develop the bi-gradient algorithm to perform both principal and minor component analysis (PCA / MCA) with suitably chosen nonlinearities they show that the algorithm can separate mixtures of either sub-gaussian or super-gaussian signals. Sub-gaussian signals are characterised by having negative kurtosis and distributions which are flat or bimodal, whereas super-gaussian signals have positive kurtosis and have sharply peaked distributions.

Two of the current shortcomings with Neural structures for ICA is the requirement for all of the latent variables to come from one particular type of density, and the lack of robustness of separation when vectors of medium sized dimension are considered. The proposed neural network extends the linear feedback network of Fyfe [17] by utilising nonlinear lateral output connections and an activation function which will respond to higher order statistics from either leptokurtic or platykurtic distributions. The lateral connections have extended the robustness of the ICA performed on data with dimensions up to ten [18], which outperforms the entropy maximisation network of Bell and Sejnowski [9].

4. LATERALLY CONNECTED FEEDBACK NETWORK

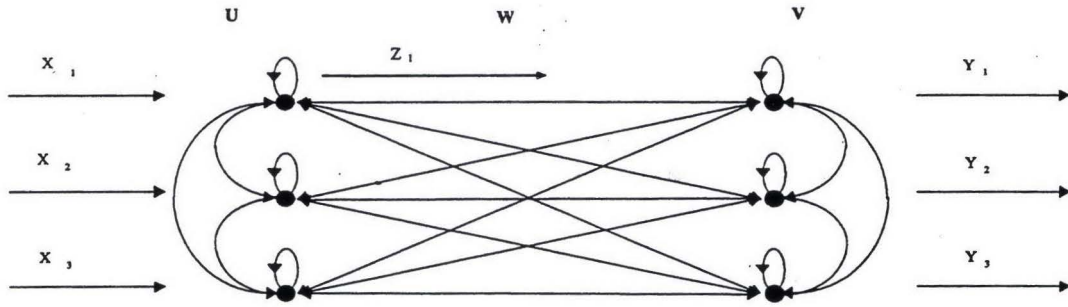


Figure 2 : Laterally Connected Feedback Network

The network is a two layer network with bi-directional connections internal to each layer and between layers. The network has fully connected lateral and feedforward nodes; the input nodes are linear, whereas the output nodes have nonlinear activation's. Consider zero mean source signals s with i.i.d components, mixed by the unknown linear matrix A , the received signals at the network input are then, in matrix format, $\mathbf{x}=\mathbf{A}s$. The output of the first layer of neurons is given as \mathbf{z} and so with the linear lateral connections at the input for an N dimensional input vector, the symmetric local activation's are

$$z_i = x_i + \sum_{k=1}^N u_{ik} x_k$$

$$\mathbf{z} = [\mathbf{I} + \mathbf{U}]\mathbf{x} \equiv \mathbf{U}_1 \mathbf{x} \quad (5)$$

The linear summation of the feedforward weights is defined as

$$r_i = \sum_{j=1}^N w_{ij} \sum_{k=1}^N u_{jk}^I x_k = \sum_{j=1}^N w_{ij} z_j$$

In matrix format

$$\mathbf{r} = \mathbf{W}\mathbf{U}_1 \mathbf{x} = \mathbf{W}\mathbf{z} \quad (6)$$

The lateral interconnections at the output are similar in nature to those at the input, however the neuron activation's are now nonlinear and so we have

$$y_i = f(r_i) + \sum_{j=1}^N v_{ij} f(r_j)$$

$$y_i = f\left(\sum_{j=1}^N w_{ij} \sum_{k=1}^N u_{jk}^I x_k\right) + \sum_{j=1}^N v_{ij} f\left(\sum_{k=1}^N w_{jk} \sum_{l=1}^N u_{kl}^I x_l\right) \quad (7)$$

We define the matrix $\mathbf{V}_1 \equiv [\mathbf{I} + \mathbf{V}]$ similar to (5) for clarity of representation, this then gives the full network output in matrix format as

$$\mathbf{y} = \mathbf{V}_1 \mathbf{f}(\mathbf{r}) \quad (8)$$

where the activation function of the output neurons is given as, Girolami and Fyfe [15]

$$\mathbf{f}(\mathbf{r}) = \mathbf{r} \pm \varphi(\mathbf{r}) \text{ and } \varphi(\mathbf{r}) = \tanh(\mathbf{r})$$

$$\mathbf{y} = \mathbf{V}_I \mathbf{W} \mathbf{U}_I \mathbf{x} - \mathbf{V}_I \varphi(\mathbf{W} \mathbf{U}_I \mathbf{x}) \quad (9)$$

$$\mathbf{y} = \mathbf{V}_I \mathbf{W} \mathbf{z} - \mathbf{V}_I \varphi(\mathbf{W} \mathbf{z}) \quad (10)$$

The following stochastic hebbian learning algorithms are employed for the network weights.

$$\Delta \mathbf{U} = \mu \left(\mathbf{I} - \mathbf{z} \mathbf{z}^T \right) \text{ and so } \left\langle \mathbf{U}_{(t+1)} \right\rangle = \left\langle \mathbf{U}_{(t)} \right\rangle + \left\langle \mu \left(\mathbf{I} - \mathbf{C}_{\mathbf{z} \mathbf{z}} \right) \right\rangle \quad (11)$$

(11) will remove all second order correlation's in the data and diagonalises the covariance matrix.

An identical anti-hebbian rule is used at the nonlinear output

$$\Delta \mathbf{V} \equiv \mu \left(\mathbf{I} - \mathbf{y} \mathbf{y}^T \right) \quad (12)$$

The feedforward section utilises hebbian learning of the nonlinear output and the residual of the input and weighted linear feedback. This provides an orthogonal rotation which stochastically maximises the objective function $\Phi(\mathbf{r})$ which in this case will be the individual terms of (4) that is the kurtosis [19].

$$\Delta \mathbf{W} = \eta_t \Phi'(\mathbf{r}) \left\{ \mathbf{z} - \mathbf{W} \mathbf{W}^T \mathbf{z} \right\} \quad (13)$$

We will now show that (11, 12, 13) work together to approximate a global linear independent component analysis.

The term $\Phi'(\mathbf{r})$ is the derivative of the objective function to be maximised which in this case will be the value of the fourth order marginal cumulant of the network output. Girolami and Fyfe [20] use an EPP network for ICA, however the stochastic maximisation (13) does not generate sufficient higher order statistics to ensure mutual information minimisation at the outputs for vectors of dimension greater than three. The addition of anti-hebbian lateral connections at the outputs of the network will yield the following using the learning of (12)

$$\Delta \mathbf{V} \equiv \mu \left(\mathbf{I} - \mathbf{y} \mathbf{y}^T \right) = \mu \left(\mathbf{I} - \left[\mathbf{V}_I \left[\mathbf{C}_{\mathbf{r} \mathbf{r}} + \varphi(\mathbf{r}) \varphi(\mathbf{r})^T \right] \mathbf{V}_I^T - \mathbf{V}_I \left[\varphi(\mathbf{r}) \mathbf{r}^T + \mathbf{r} \varphi(\mathbf{r})^T \right] \mathbf{V}_I^T \right] \right)$$

As $\langle \Delta \mathbf{V} \rangle \rightarrow \mathbf{0}$ and $\mathbf{V}_I \equiv \mathbf{I}$ then

$$\left\langle \left(\mathbf{I} - \left[\mathbf{C}_{\mathbf{r} \mathbf{r}} + \varphi(\mathbf{r}) \varphi(\mathbf{r})^T \right] - \left[\varphi(\mathbf{r}) \mathbf{r}^T + \mathbf{r} \varphi(\mathbf{r})^T \right] \right) \right\rangle \equiv \mathbf{0}$$

$$\text{as } \mathbf{C}_{\mathbf{r} \mathbf{r}} \equiv \mathbf{I} \text{ due to (11) and } \mathbf{W} \mathbf{W}^T = \mathbf{I} \Rightarrow \left\langle \varphi(\mathbf{r}) \mathbf{r}^T + \mathbf{r} \varphi(\mathbf{r})^T - \varphi(\mathbf{r}) \varphi(\mathbf{r})^T \right\rangle \equiv \mathbf{0}$$

Tanh is an odd function and so taking a Taylor expansion

$$\sum_k \varphi_{2k+1} \left\langle r_i^{2k+1} r_j \right\rangle + \sum_k \varphi_{2k+1} \left\langle r_i r_j^{2k+1} \right\rangle - \sum_{k,m} \varphi_{2k+1} \varphi_{2m+1} \left\langle r_i^{2k+1} r_j^{2m+1} \right\rangle = 0 \quad (14)$$

as $\mathbf{C}_{\mathbf{r} \mathbf{r}} \equiv \mathbf{I} \Rightarrow \langle r_i r_j \rangle = 0 \quad \forall i \neq j \Leftrightarrow \langle r_i^{2k+1} r_j^{2m+1} \rangle = 0 \quad \forall i \neq j$ and so (9) can be simply

considered as $\sum_k \varphi_{2k+1} \left\langle r_i^{2k+1} r_j \right\rangle + \sum_k \varphi_{2k+1} \left\langle r_i r_j^{2k+1} \right\rangle = 0 \Rightarrow \langle r_i^3 r_j \rangle + \langle r_i r_j^3 \rangle \rightarrow 0 \quad \forall i \neq j$

Which is simply the minimisation of all cross cumulants of order four. The stochastic marginal cumulant maximisation of (13) under the cross cumulant minimisation constraint yields an orthogonal ICA transformation of the input data. With the form of nonlinearity proposed (8) we can now perform ICA on multivariate data whose latent marginal components may exhibit either uni-modal or bi-modal distributions. This removes the requirement that latent variables come from one form of distribution. The simultaneous cross-cumulant minimisation (12), (14) provides a more robust ICA when dealing with data of high dimensionality thus improving on the performance of the solely feedforward architectures.

6. SIMULATION

We report on a simulation which is motivated by a signal processing problem, blind extraction of original signals from an unknown mixture. We have three original signals, two seconds of human speech which typically has a density which resembles a Laplacian distribution and can be described by its positive kurtosis. A similar sample of a pure sinusoidal tone, which has a distinctly bi-modal density, and two seconds of uniformly distributed noise. Figure (3) shows the original densities, this data is then mixed using an arbitrary full rank matrix, due to central limit effects the corresponding mixed densities are more gaussian than the originals. The actual signals are significantly degraded and as such are now an incoherent babble, indeed approaching Gaussian noise. With no *a priori* knowledge of the original data or mixing we present the corrupted data to the network, the weights converge after seven passes through the data. The normalised data output is shown along with the corresponding distributions. In terms of BSS we have blindly extracted the original signals; in terms of ICA, we have transformed the given data into a factored form.

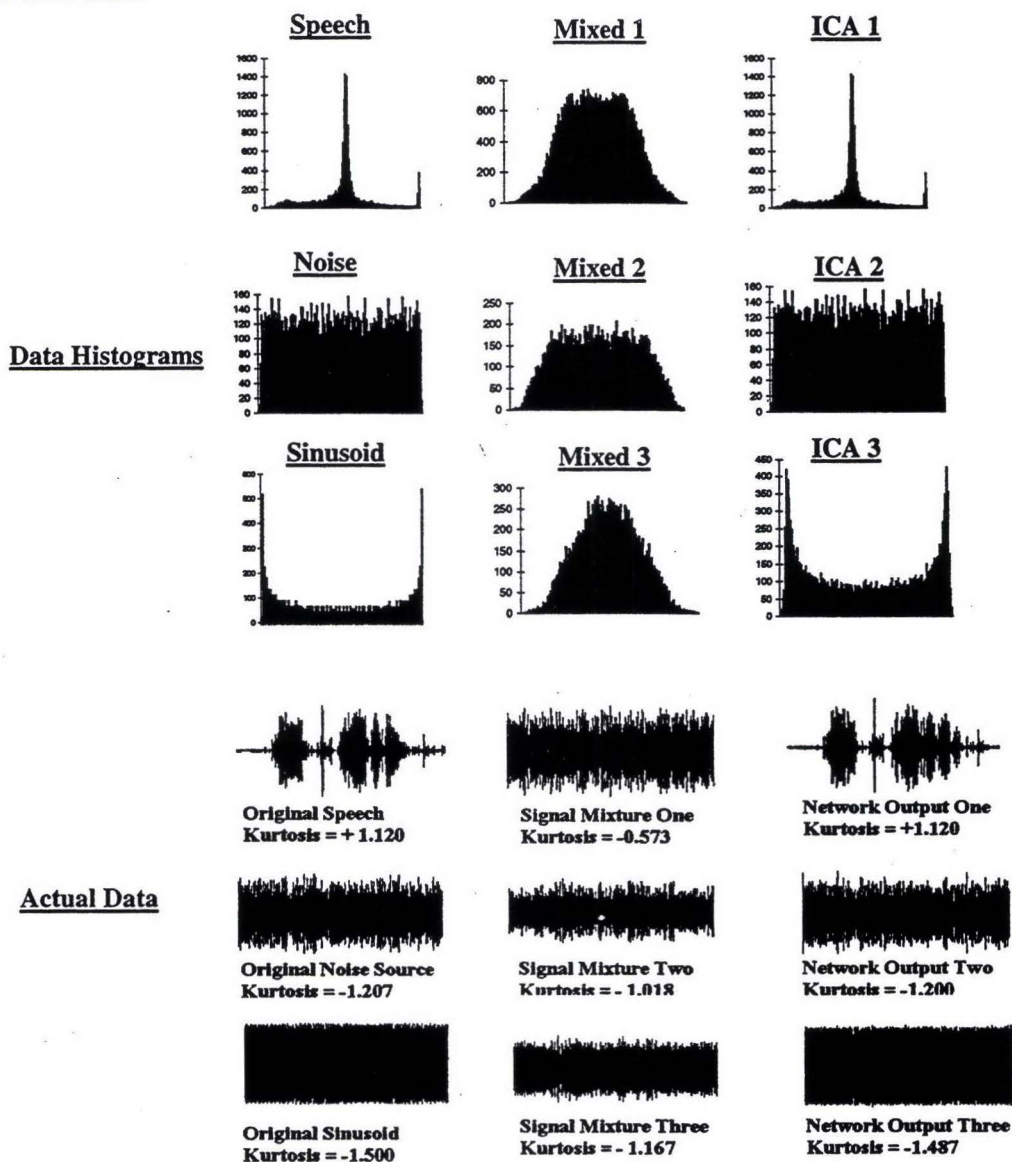


Figure 3 : Simulation Results.

7. CONCLUSIONS

We have developed a simple self organising neural network structure which is capable of performing a general ICA on a given data set. The form and learning of the network improves on existing neural implementations of ICA in that the lateral connections allow a more robust ICA to be performed on high dimensional data [21], and the proposed nonlinearity removes the requirement for solely platykurtic or leptokurtic marginal latent pdf's. This form of network and learning will find application in cluster analysis, exploratory projection pursuit and of course statistical signal processing.

8. REFERENCES

- [1] Parsons, T. Voice and speech processing. McGraw Hill, ISBN 0-07-048541-0, 1987.
- [2] Haykin, S. Neural Networks, A comprehensive foundation. Macmillan, ISBN 0-02-352761-7, 1995.
- [3] Sanger, T. Optimal unsupervised learning in a single-layer linear feedforward network. *Neural networks*, vol 2, pp. 459-473, 1989.
- [4] Masters, T. 'Advanced Algorithms For Neural Networks', John Wiley & Sons, 1995.
- [5] Comon, P. Independent Component Analysis, A New Concept ?. *Signal Processing*, 36, 287 - 314. 1994.
- [6] Cardoso, J.F. Adaptive source separation with uniform performance. *EUSIPCO-94*, Edinburgh, 1994.
- [7] Jutten, C Herault, J. Blind Separation of Sources, Part 1: An Adaptive Algorithm Based On Neuromimetic Architecture. *Signal Processing* 24 1- 10, 1991.
- [8] Cichocki, A Amari, S Yang, H. Recurrent Neural Networks for Blind Separation of Sources. *International Symposium on Nonlinear Theory and Applications Vol 1*. 37 - 42, 1995.
- [9] Bell, A and Sejnowski, T. An Information Maximisation Approach to Blind Separation and Blind Deconvolution. *Neural Computation* 7, 1129 - 1159, 1995.
- [10] Amari, S, Cichocki, A, and Yang, H. A new learning algorithm for blind signal separation. *Neural Information Processing*, Vol 8, M.I.T Press 1995.
- [11] Karhunen, J., Joutensalo, J. Representation and separation of signals using nonlinear PCA type learning. *Neural Networks* 7(1), pp.113-127, 1994.
- [12] Karhunen, J. Neural approaches to independent component analysis and source separation. *Proc. ESANN'96, (4'th European Symposium on Artificial Neural Networks)*, Bruges, Belgium, April 24-26 1996.
- [13] Fyfe, C and Baddeley, R. Non-linear data structure extraction using simple hebbian networks. *Biological Cybernetics*, 72(6):533-541, 1995.
- [14] Oja, E. The nonlinear PCA learning rule and signal separation-mathematical analysis. *Research Report A26*, Helsinki University of Technology, ISBN 951-22-2706-1, 1995.
- [15] Girolami, M and Fyfe, C. Stochastic ICA contrast maximisation using Oja's nonlinear PCA algorithm. *International journal of neural systems*, submitted, 1996.
- [16] Wang, L., Karhunen, J., Oja, E. A bigradient optimisation approach for robust PCA, MCA, and source separation. *Proc IEEE Int. Conf. on neural networks and signal processing*, Perth, Australia, 1995.
- [17] Fyfe, C. A fully parallel pca network. *IEEE/IEE Workshop on natural algorithms in signal processing*, 1993.
- [18] Girolami, M and Fyfe, C. Extraction of Independent Signal Sources using a Deflationary Exploratory Projection Pursuit Network with Lateral Inhibition. Submitted to *I.E.E Proceedings on Vision, Image and Signal Processing Journal*. Sept 1996.
- [19] Girolami, M and Fyfe, C. An Extended Exploratory Projection Pursuit Network with Linear and Nonlinear Anti-Hebbian Connections Applied to the Cocktail Party Problem. Submitted to *Neural Networks Journal*. May 1996.
- [20] 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.
- [21] Girolami, M and Fyfe, C. Negentropy and Kurtosis as Projection Pursuit Indices Provide Generalised ICA Algorithms', Invited Contribution, NIPS'96 Blind Signal Separation Workshop, (Org A. Cichocki & A. Back), Aspen Colorado, 7 Dec, 1996.