ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 297-302

Independence is Far From Normal

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

Exploratory Projection Pursuit (EPP) is a statistical data analysis tool for identifying structure in high dimensional data. In this paper we consider two Neural implementations of EPP. The first; performed under orthonormal constraints; utilises criteria based on fourth order moments, this is shown to be a dual for Independent Component Analysis (ICA). The second is based on the Kullback divergence from normality (negentropy), and is seen to perform ICA on data which is a linear mixture of independent latent variables. Simulations are reported which show the exceptional convergence speed of the negentropy based algorithm when limited *a priori* knowledge of the source distributions and a simple momentum based acceleration scheme are employed.

1. Introduction

Exploratory Projection Pursuit is a statistical tool which allows structure in high dimensional data to be identified. This is achieved by projecting the data onto a low dimensional subspace and searching for structure in the projection. By defining indices which give a measure of how 'interesting' a given projection is, projection of the data onto a subspace which maximises the given index will then provide a maximally 'interesting' direction. Departures from a Gaussian distribution are viewed as 'interesting', as skewed or multi-modal distributions present certain structures within the data. If we then use an index which is a function of the direction of projection, index maximisation will then provide a direction furthest from gaussian.

Intrator [1] constructs a neural model for EPP derived from the Bienenstock, Cooper, Monro (BCM) neuron which is a model of cortical plasticity. Fyfe and Baddeley propose an alternative neural model of EPP based on the negative feedback network, a comparative study of both these EPP models can be found in [2]. ICA was first introduced by Jutten and Herrault [3] within the context of blind separation of sources (BSS); performing neural EPP driven by fourth order moments has been found to be equivalent to ICA for data with independent latent variables [7,16]. As EPP is concerned with driving the network output maximally from Gaussian, a criterion based on the Kullback divergence of a density and its normal equivalent is also considered.

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 297-302

2. Moment Based Neural EPP and ICA

An observation of an N dimensional random vector $\mathbf{x}(t) = (x_1(t), x_2(t), \ldots, x_N(t))^T$ is made at time point t such that the vector $\mathbf{x}(t)$ consists of N zero mean latent variables $\mathbf{s}(t) = (s_1(t), s_2(t), \ldots, s_N(t))^T$ projected onto a set of unknown vectors, ie. $\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$, where $\mathbf{A}$ is an $N \times N$ full rank matrix. To search for structure in the received data, $\mathbf{x}(t)$, a whitening process is first required so that the data covariance is an identity $\mathbf{C}_{\mathbf{xx}} = \mathbf{I}$ this can be achieved by simple decorrelation learning or PCA learning [4,7]. The reasons for this whitening are set forth in Fyfe [2]: if normalisation of all second order statistics takes place then the subsequent neural learning will only respond to the higher order statistics of the data. It is precisely these higher order statistics which characterise the non-gaussianity of a distribution and which guide the pursuit to discover maximally non-normal projections. Consider a criterion based on fourth order moments $\Phi(\mathbf{w}) = E\{u_i^4\}$ where $\mathbf{w}$ is the column weight vector of a feedforward structure which will project the input $\mathbf{x}(t)$ onto the output $u_i(t) = \mathbf{w}^T\mathbf{x}(t)$. [5] shows that the stochastic weight update $\Delta w_{ij} = \eta_t \frac{d\Phi}{du_i}\left\{x_j(t) - \sum_{k=1}^{M} w_{kj} \sum_{l=1}^{N} w_{kl} x_l(t)\right\}$ (1) is an approximative stochastic gradient algorithm to maximise the specific criterion function - in this case $\Phi(\mathbf{w}) = E\{u_i^4\}$ - under orthonormal constraints $\mathbf{w}^T\mathbf{w} = 1$. Now as $\mathbf{C}_{\mathbf{xx}} = \mathbf{I} \Rightarrow E\{u_i^2\} = \mathbf{w}^T\mathbf{C}_{\mathbf{xx}}\mathbf{w} = 1$, then maximisation of the output fourth moment is equivalent to maximisation of the fourth order cumulant or its normalised version which is kurtosis $\Phi(\mathbf{w}) \equiv \kappa_4 = E\{u_i^4\} - 3E^2\{u_i^2\} = E\{u_i^4\} - 3$. So the EPP learning searches for a maximally kurtotic subspace which will identify kurtotic structure in the data. Let us extend (1) to a full square weight matrix and use a network with N inputs and N outputs, now $\Phi(\mathbf{W}) = E\{\mathbf{u}^4\}$ and φ is the element - wise derivative of Φ then we have

$$\Delta\mathbf{W} = \eta_t\left[\mathbf{x}(t) - \mathbf{W}\mathbf{W}^T\mathbf{x}(t)\right]\varphi(\mathbf{x}^T\mathbf{W}) \quad (2)$$

This will provide a volume preserving linear rotation, as $\mathbf{W}\mathbf{W}^T = \mathbf{I}$ and $\det(\mathbf{W}) = \det(\mathbf{W}^T)$ then $\det(\mathbf{W}\mathbf{W}^T) = \det(\mathbf{I}) = 1 \Leftrightarrow \det(\mathbf{W}) = 1$. This has important implications in linking moment based neural EPP learning to ICA.

The Kullback divergence of a multivariate pdf from normality is defined as

$$J(p_{\mathbf{u}}(\mathbf{u})) = \int p_{\mathbf{u}}(\mathbf{u}) \log(p_{\mathbf{u}}(\mathbf{u}) / p_{\mathbf{G}}(\mathbf{u}))d\mathbf{u} \quad (3)$$

this is termed the negentropy. It is shown [6] that negentropy can be written as $J(p_{\mathbf{u}}(\mathbf{u})) = H(p_{\mathbf{G}}(\mathbf{u})) - H(p_{\mathbf{u}}(\mathbf{u}))$ where $H(p_{\mathbf{u}}(\mathbf{u}))$ is the entropy of the data density $\mathbf{u}$, and $H(p_{\mathbf{G}}(\mathbf{u}))$ is the equivalent entropy of a Gaussian density which has equal mean and covariance as p_u. Comon shows [6] that the mutual information of the vector $\mathbf{u}$

which for independent components is zero) can be written as,

$$I(p_{\mathbf{u}}(\mathbf{u})) = J(p_{\mathbf{u}}(\mathbf{u})) - \sum_{i=1}^{N} J(p_{ui}) + \tfrac{1}{2}\log\left(\prod \mathbf{C}_{uu(ii)} / \det \mathbf{C}_{uu}\right) \quad (4)$$

The first term is the multivariate negentropy as defined above, negentropy is invariant under volume preserving mappings, which is precisely the transformation that the network weights provide, and so $J(p_{\mathbf{u}}(\mathbf{u})) = J(p_{\mathbf{x}}(\mathbf{x}))$ which will not affect the minimisation of (4). It is also clear that as $\mathbf{C}_{uu} = E\{\mathbf{u}\mathbf{u}^T\} = \mathbf{W}E\{\mathbf{x}\mathbf{x}^T\}\mathbf{W}^T = \mathbf{W}\mathbf{C}_{xx}\mathbf{W}^T = \mathbf{I}$ then $\log(\det \mathbf{C}_{uu}) = \log\left(\prod \mathbf{C}_{uu(ii)}\right) = 0$ so the data pre-whitening eliminates the rightmost term in (4). Using Edgeworth PDF expansions and assuming the PDF's are symmetric, Comon derives for the mutual information after whitening

$$I(p_{\mathbf{u}}(\mathbf{u})) = J(p_{\mathbf{x}}(\mathbf{x})) - \sum_{i=1}^{N} J(p_{ui}) \cong J(p_{\mathbf{x}}(\mathbf{x})) - \sum_{i=1}^{N} \kappa_{iii}^2 \quad (5)$$

and so maximisation of the sum of squares of fourth order marginal cumulants $\sum_{i=1}^{N} \kappa_i^2$ will minimise (5). It is clear that (2) will approximate stochastic maximisation of the sum of squares of fourth order cumulants if all components of **s** have uniform sign of kurtosis. In practice this only holds for vectors which are mixtures of no more than three sources due to the manner in which the orthonormal constraint is embedded within the algorithm, simulations are reported in [16]. An extension of the simple feedforward architecture has been developed [7,17] and this has been shown to perform a more robust ICA with much faster convergence. We now consider an information theoretic indice for EPP / ICA.

3. Negentropy Based Neural EPP and ICA

As mixing of components introduces central limit effects [9], mixtures of data PDF's will have a tendency to be Gaussian (illustrative simulations in [10]). So we seek a data driven transformation which will maximally drive the densities furthest from normal. From (3) we have the multivariate negentropy as $J(p_{\mathbf{u}}(\mathbf{u})) = H(p_{\mathbf{G}}(\mathbf{u})) - H(p_{\mathbf{u}}(\mathbf{u}))$ which evaluated gives

$$J(\mathbf{W}) = \tfrac{1}{2}\log\left[(2\pi e)^N \det(\mathbf{C}_{uu})\right] + E\{\log[p_u(\mathbf{u})]\} \quad (7)$$

Taking instantaneous values everywhere, then the gradient terms for (7) are

$$\nabla\left\{\tfrac{1}{2}\log\left[(2\pi e)^N \det(\mathbf{C}_{uu})\right]\right\} = \nabla\left[\tfrac{1}{2}\log\left((2\pi e)^N \det(\mathbf{C}_{xx})\right) + \log[\det(\mathbf{W})]\right]$$

finally giving $\nabla\left\{\tfrac{1}{2}\log\left[(2\pi e)^N \det(\mathbf{C}_{uu})\right]\right\} = \mathrm{adj}(\mathbf{W})^T / \det(\mathbf{W}) = \left[\mathbf{W}^T\right]^{-1}$

We now have $\nabla J(\mathbf{W}) = \left[\mathbf{W}^T\right]^{-1} + \nabla\{\log[p_u(\mathbf{u})]\}$ As we are seeking to perform an ICA then we model the output density as the product of univariate independent densities.

$$H(p_u(\mathbf{u})) = \sum H(p_u(u_i)) - I(p_u(\mathbf{u})) \; ; \; J(p_u(\mathbf{u})) = H(p_G(\mathbf{u})) - \sum H(p_u(u_i)) + I(p_u(\mathbf{u}))$$

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 297-302

For $I(p_u(\mathbf{u})) \to 0$ then $J(p_u(\mathbf{u})) = H(p_G(\mathbf{u})) - \sum_{i=1}^{N} H(p_u(u_i))$. Maximally driving the network outputs from gaussianity under the parametric constraint that $p_u(\mathbf{u}) = \prod_{i=1}^{N} p(u_i)$ will yield an ICA if the marginal PDF's are suitably parameterised to match the independent components of the vector **s**. For the case of a single weight $\partial/\partial w_{ij} \log(p_u(\mathbf{u})) = p'(u_i)x_j \prod_{k\neq i}^{N} p(u_k) / \prod_{k=1}^{N} p(u_k) = p'(u_i)x_j / p(u_i)$ so $\nabla(\log(p_u(\mathbf{u}))) = \left[p'(u_i)\prod_{k\neq i}^{N} p(u_k)\right]^T \mathbf{x}^T / p(\mathbf{u}) = \nabla_u p_u(\mathbf{u})\mathbf{x}^T / p_u(\mathbf{u})$. Amari et al introduces the natural gradient in [12], as do Cardoso et al [13] though termed the relative gradient. Implementing the natural gradient we then have

$$\Delta\mathbf{W} \propto \tilde{\nabla}J = \nabla J\mathbf{W}^T\mathbf{W} = \left[\mathbf{I} + \left(\nabla_u p_u(\mathbf{u}) / p_u(\mathbf{u})\right)\mathbf{u}^T\right]\mathbf{W} \tag{8}$$

This is similar to the form of the maximum likliehood based algorithm derived for cICA by Pearlmutter and Parra [14]. With suitable parameterisation of the nonlinear term in (8) we find similarity with this form and the information theoretic based algorithms of Amari et al [12] and Bell & Sejnowski [11]. A generic form of nonlinearity is derived in [10] which will allow simultaneous separation of mixtures which contain both sub and super-gaussian sources. Cardoso [19] utilises information geometric arguments in building a framework of entropic contrasts for ICA and shows the equivalence of info-max [11], maximum likelihood [14] and maximum negentropy [10] approaches to ICA.

To improve on the convergence speed of the simple gradient based algorithm of (8) we can consider implementing second order information or techniques such as conjugate gradients. As a pre-cursor to these more sophisticated optimisation techniques we shall consider a simple momentum based acceleration scheme. Applying a momentum term to the gradient ascent algorithm (8) the parameter updates are finally given as

$$\Delta\mathbf{W}_{t+1} = \eta_t \tilde{\nabla}J + \mu_t \Delta\mathbf{W}_t \tag{9}$$

4. Simulation

We consider a linear mixture of five sources, each source is a five second sample of natural speech (each sampled at 8000 samples / sec). This data set has been used in [7,15] to test the extension of EPP learning and the non-linear PCA algorithm utilising Fahlman type activation functions. Typically these algorithms required seven epochs of learning for full separation, that is 280,000 iterations. It should also be noted that the data must be spatially pre-whitened, giving additional iterations. As the pdf of natural speech can be modelled as a Gamma distribution, and less accurately as a Laplace distribution, we shall parameterise the nonlinear term in (8) with the simple form of the Laplacian distribution for this simulation.

The left hand column of traces shows the input mixtures, the adjacent column shows the network output as the weights are adapting, Figure 1. The performance measure used is the distance of the composite matrix **WA** from a permutation, this is shown directly above the output traces. What we see is that after 12,000 iterations (1.5 sec's) we have complete separation of all five sources. It should be stressed that the

input data has not been pre-whitened. This corresponds to an increase in convergence speed of greater than 20 times over moment based EPP and non-linear PCA algorithms.

A comparison of (9) with the natural gradient form of the info-max algorithm [11] was made using a mixture of ten sources of speech and music, the momentum term gave an increase in convergence speed of 1.2~1.5x .

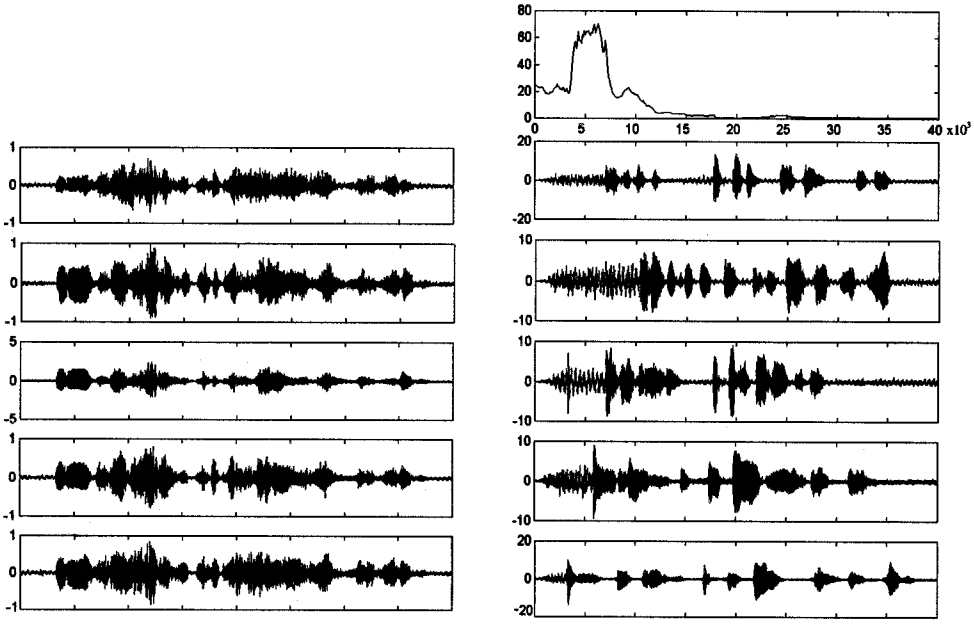

Figure 1: Input signals and output of network with performance measure.

5. Discussion

We have considered EPP and shown that indices based on fourth order moments and negentropy will also yield ICA algorithms for linear independent latent variable models. The convergence speed of the two algorithms has been compared using a mixture of speech sources, with the negentropy based algorithm exhibiting superior performance. A simple acceleration scheme has been applied to the gradient based algorithm and an increase in the convergence speed has been noted. Further work will include empirical comparisons between the standard stochastic gradient (8), conjugate gradient and Newton based algorithms, we will also consider a temporal form of the negentropy based algorithm and apply this to signals which have been convolved [18].

6. References

[1] Intrator, N. A neural network for feature extraction. *NIPS2*, San Mateo, CA, Morgan Kaufman, pp 719-726.

[2] C, Fyfe. A comparative study of two neural methods of exploratory projection pursuit. *Neural Networks*, Vol.9, No.6, pp1-6, 1996.
[3] Jutten, C Herault, J. Blind Separation of Sources, Part 1: An Adaptive Algorithm Based On Neuromimetic Architecture. *Signal Processing* 24 1- 10, 1991.
[4] 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.
[5] Karhunen, J., Joutensalo, J. Representation and separation of signals using nonlinear PCA type learning. *Neural Networks* **7**(1), pp.113-127, 1994.
[6] Comon, P. Independent Component Analysis, A New Concept ?. *Signal Processing*, 36, 287 - 314. 1994.
[7] 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.
[8] Karhunen, J., Oja, E., Wang, L., Vigario, R., Joutsensalo J. A class of neural networks for independent component analysis. *Research Report A28*, Lab of Computer Science, Helsink University of Technology, 1996.
[9] Loyd, E. Probability Vol II. Wiley, ISBN 0 471 27821 1, 1980.
[10] Girolami, M and Fyfe, C. Negentropy and Kurtosis as Projection Pursuit Indices Provide Generalised ICA Algorithms, *NIPS96, Blind Signal Separation Workshop*, (Eds Prof. A. Cichocki & A. Back), Aspen Colorado, 7 Dec, 1996.
[11] Bell, A and Sejnowski, T. An Information Maximisation Approach to Blind Separation and Blind Deconvolution. *Neural Computation* 7, 1129 - 1159, 1995.
[12] 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.
[13] Cardoso, J,F. Belouchrani, A. and Laheld, B. A new composite criterion for adaptive and iterative blind source separation. Proc of *ICASSP-94,* vol4, 273-276.
[14] Pearlmutter, B and Parra, L. A Context Sensitive Generalisation of ICA. *International Conference on Neural Information Processing*, Hong Kong, Sept. 24-27 1996. Springer.
[15] Girolami, M., and Fyfe, C. Stochastic ICA Contrast Maximisation Using Oja's Nonlinear PCA Algorithm. Submitted *International Journal of Neural Systems*, August 1996.
[16] Girolami, M and Fyfe, C. Blind Separation Of Sources Using Exploratory Projection Pursuit Networks. In Proc. *Speech and Signal Processing, International Conference on the Engineering Applications of Neural Networks*, ISBN 952-90-7517-0, 249 - 252, 1996.
[17] Girolami, M and Fyfe, C. Kurtosis Extrema and Identification of Independent Components : A Neural Network Approach. In Proc *ICASSP-97, I.E.E.E Conference on Acoustics, Speech and Signal Processing*. April 1997.
[18] Girolami, M and Fyfe, C. A Temporal Model of Linear Anti-Hebbian Learning. *Neural Processing Letters* Journal, Vol 4, Issue 3, pp 1-10, Jan 1997.
[19] Cardoso, J,F. Entropic Contrasts for Soure Separation, *NIPS96, Blind Signal Separation Workshop*, (Eds Prof. A. Cichocki & A. Back), Aspen Colorado, 7 Dec, 1996.